DESIGN AND PERFORMANCE OF HIGH-SPEED COMMUNICATION SYSTEMS OVER TIME-VARYING RADIO CHANNELS

by

Andrea Goldsmith

Memorandum No. UCB/ERL M94/75

15 September 1994

DESIGN AND PERFORMANCE OF HIGH-SPEED COMMUNICATION SYSTEMS OVER TIME-VARYING RADIO CHANNELS

by

Andrea Goldsmith

Memorandum No. UCB/ERL M94/75

15 September 1994

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

DESIGN AND PERFORMANCE OF HIGH-SPEED COMMUNICATION SYSTEMS OVER TIME-VARYING RADIO CHANNELS

by

Andrea Goldsmith

Memorandum No. UCB/ERL M94/75

15 September 1994

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Abstract

DESIGN AND PERFORMANCE OF HIGH-SPEED COMMUNICATION SYSTEMS OVER TIME-VARYING RADIO CHANNELS

by

Andrea Goldsmith

Doctor of Philosophy in Engineering
Electrical Engineering and Computer Sciences
University of California at Berkeley
Professor Pravin P. Varaiya, Chair

The next generation of wireless networks will require more efficient use of the underlying time-varying channel to accommodate the demand for voice, video, and data transmission. In this thesis, we investigate methods to increase the spectral efficiency of point-to-point and multiuser communication systems operating over time-varying radio channels. We begin by developing several models for the time-varying channel. Specifically, we model both deterministic and stochastic multipath channels, cellular channels, and state space channels. Next, we propose spectrally-efficient communication techniques for time-varying channels when the channel is estimated and this estimate fed back to the transmitter. We determine the maximum spectral efficiency under this assumption, and show that this maximum is achieved when three parameters are adapted to the channel variation: transmit power, data rate, and coding scheme.

When a feedback path is not available, the receiver can use a priori knowledge about the channel statistics to decode the input sequence. For a discrete-time channel with Markov variation, we propose a decision-feedback decoding algorithm that uses the channel's Markovian structure to determine the maximum-likelihood input sequence. We calculate the capacity and cutoff rate of this decoding scheme, and compare them to the inherent rate limits of the channel. Finally, we discuss some multiresolution coding techniques for the case when no a priori channel information is available. This type of coding allows some loss of nonessential data in order to achieve overall higher data rates.

We then consider performance of multiuser systems, where interference limits the

total number of users and their respective data rates. We determine the achievable rate regions of multiuser time-varying channels with channel estimation and transmitter feedback under CDMA, FDMA, and TDMA spectrum-sharing techniques. We also discuss several uses for power control beyond traditional interference balancing. In particular, we propose a hybrid power/rate control scheme which adapts to the system traffic load and the channel characteristics of each user. This policy maintains fairness in the network while taking advantage of favorable propagation conditions. Finally, we propose an architecture and protocol suite for interconnecting wireless subnets with different specifications and requirements.

Pravin Varaiya 29 August '94

Professor Pravin P. Varaiya, Chair Date

Contents

List of Figures

List of Symbols

$p(x), \pi(x), \pi_x$	Probability of x.
$p(x\|y)$	Probability of x conditioned on y.
$E[x], \overline{x}$	Expectation of x.
$\hat{x}$	Estimate of x.
$\lfloor x \rfloor$	Largest integer less than or equal to x.
$\|x\|$	Absolute value of x.
x^m	$(x_1, \ldots, x_m)$.
x_n^m	$(x_n, \ldots, x_m)$.
$\{x_n\}_{n=1}^m$	$x_1, \ldots, x_m$.
$1[x]$	Indicator of x ($1[x] = 1$ for x true, $1[x] = 0$ for x false).
$[x]^+$	$x1[x > 0]$.
$A_x(\tau)$	Autocorrelation of the wide-sense stationary process x.
$\|\mathcal{X}\|$	Number of elements in the set $\mathcal{X}$.
$\mathcal{P}(\mathcal{X})$	Set of all distributions over the set $\mathcal{X}$.
$\mathcal{X} \times \mathcal{Y}$, $\{(x,y); x \in \mathcal{X}, y \in \mathcal{Y}\}$	Set of all (x,y) with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
$\mathcal{X}^n$	Set of all $(x_1, \ldots, x_n)$ with $x_i \in \mathcal{X}$.
$f : \mathcal{X} \to \mathcal{Y}$	The function f has domain $\mathcal{X}$ and range $\mathcal{Y}$.
$A \Rightarrow B$	A implies B.
$A \triangleq B$	A is defined as B.
log	Natural logarithm.
$\log_n$	Logarithm base n.
$H(X)$	Entropy of X.
$H(Y\|X)$	Entropy of Y conditioned on X.
$I(X;Y)$	Mutual information between X and Y.

$C(P)$	Channel capacity with power P.
i.i.d.	Independent and identically distributed.
i.d.	Identically distributed.
iff	If and only if.
rms	Root mean squared.
WSS	Wide-sense stationary.
AWGN	Additive white Gaussian noise.
Q-AWN	Quantized additive white noise.
MSE	Mean-squared error.
BER	Bit error rate.
SNR	Signal-to-noise power ratio.
UEP	Unequal error protection.
LOS	Line-of-sight.
BSC	Binary symmetric channel.
DMC	Discrete memoryless channel.
FSMC	Finite-state Markov channel.

Acknowledgements

I am deeply grateful to my advisor, Pravin Varaiya, for his endless support, guidance, and encouragement. Working with Pravin and learning from him has been a true privilege, and his creativity, depth of knowledge, and unfaltering standards of excellence have inspired me throughout my tenure at Berkeley. It was Pravin's unlimited encouragement, patience, and availability to answer my questions and listen to my ideas that kept me focused and enthusiastic throughout the course of this work.

I am also very grateful to Larry Greenstein of AT&T Bell Laboratories for his strong support and encouragement. Larry provided me with a dynamic and stimulating research environment at Bell Labs. He also taught me a great deal about communications, mobile radio, and research in general. His mentoring and guidance are deeply appreciated.

In addition, I would like to thank Jean Walrand, Jim Pitman, and Jean-Paul Linnartz for serving on my qualifying exam or reading committee. Their comments and suggestions both clarified and strengthened many parts of this document.

This work has benefited from valuable discussions with friends and colleagues at Berkeley and Bell Labs, including John Barry, Gerhard Fettweis, Jerry Foschini, Joe Kahn, Tony Rustako, Ahmad Shank-Bahai, and Ravi Subramanian. I would like to thank the participants in the communications seminar group at Berkeley for many stimulating talks and discussions. I'm also grateful to the support staff at Berkeley for their friendly and skilled assistance.

Many friends at Berkeley contributed to making the years of graduate school enjoyable. I will mention a few of special importance. I am grateful for Sonia's friendship and sense of humor, which helped to keep my small catastrophes in perspective. Pradeep greatly enriched my years at Berkeley with his friendship; I treasured our midnight tea parties, philosophical discussions, and poetic email exchanges at the crack of dawn. My dearest friend Nina has shared and enhanced my life in every respect; her laughter, spirit, and friendship brightened every moment of our passage through graduate school.

Finally, my most heartfelt gratitude goes to the two most special people in my life. My father Werner supported me with endless love, caring, and pride. His big hugs and unique eccentricities always warmed my heart, and he also inspired me with a love of learning and academia. To my husband Arturo, I am deeply grateful for his love, support, and devotion. He has taught me the true meaning of companionship, and enriched both my life and my dreams. I thank them both for always being there and believing in me; this work is dedicated to them.

This research was supported in part by the PATH program, University of California, Berkeley, and in part by an IBM graduate fellowship.

Chapter 1

Introduction

The vision of wireless communication providing high-speed high-quality information exchange between two portable devices located anywhere in the world is the communications frontier of the next century. The great popularity of cordless telephones, cellular telephones, radio paging, and other emerging portable communication technologies demonstrates a great demand for such services. The network infrastructure must support all of these services, and will likely encompass wireline networks as well. Efficient interconnection of the subnetworks, each with different protocols and requirements, will require standardization of interfaces and internetworking protocols, as well as intelligent networking capabilities to exchange information across subnet boundaries.

What has emerged from worldwide research and development activity in this area is the need for the following technological advances to implement this wireless vision:

- Hardware for low-power handheld computer and communication terminals.
- Techniques to improve the quality and spectral efficiency of communication over wireless channels.
- Better means of sharing the limited spectrum to accommodate the different wireless applications.
- An architecture and protocol suite to integrate the various subnetworks and systems into an interconnected network.

We now give a more complete overview of the proposed integrated wireless communication network, and the technical issues involved in its implementation. Many of these

technical issues will be carefully examined in the thesis chapters, where we propose and evaluate new methods to approach them, and compare these methods with other techniques that are currently being implemented or suggested in the literature.

1.1 Global Wireless Networks

The current network architecture of cellular and cordless phone systems is shown in Figure 1.1. The local base of the cordless phone connects into the Public Switched Telephone Network (PSTN) in the same way as a wireline telephone, with communication between the base station and wireless handset via low power radio. The cellular system has a similar architecture, except that the base stations are controlled by an intermediate mobile telephone switching office (MTSO), which provides central control of all the base station and mobile units (call routing, transfer between base stations, etc.). Calls that are initiated and terminated within the same cellular system do not go through the local exchange; they are routed directly by the MTSO. All other calls go through the local exchange. The inefficiency, cost, and bottlenecks associated with this centralized control scheme has led to more decentralized proposals for future-generation architectures.

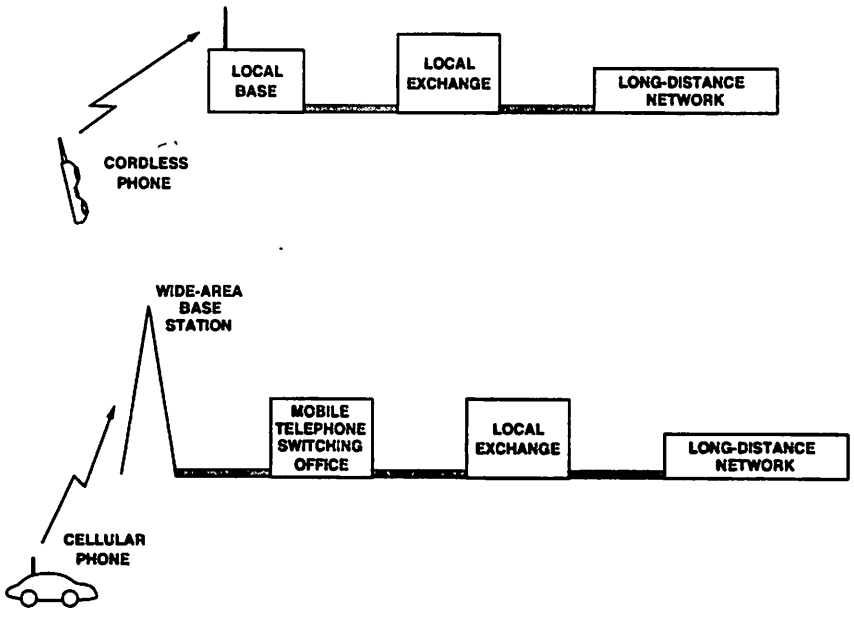

Figure 1.1: Current Network Architecture

The architecture for a global wireless infrastructure is still under development, as

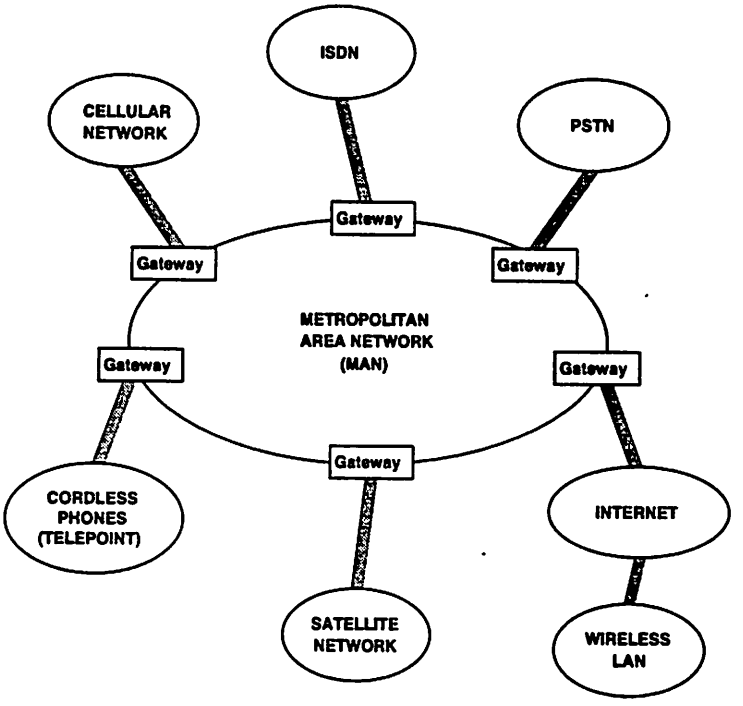

Figure 1.2: Interconnection of Wireless Communication Subsystems

we will discuss in more detail in Chapter 6. For the near future, however, it is likely that wireless communication subsystems will be connected to existing wide-area voice and data networks either directly (as the current systems are connected to the PSTN), or through a metropolitan area network, as shown in Figure 1.2. In either case, gateways will be necessary for routing and protocol conversion between the wireless subsystems and the networks that support their interconnection. This will provide backward compatibility with existing systems, however, it will impede the development of an architecture that is optimized to fully integrate the emerging personal communication technologies. The conflicting goals of maintaining backward compatibility while optimizing for an emerging system is not new; however, it presents a challenge for the design of a wireless infrastructure, given the diverging standards for the different wireless technologies and the desire to interconnect them through a single global network. Research and design of the global infrastructure have not yet received much attention in academic or industrial circles, probably because the wireless subsystems are still under development, and the economic potential of the infrastructure design is less immediate than that of the wireless subsystems. However, addressing the issue now will provide subsystem developers with standardization guidelines

that will significantly ease subsystem integration. It is unlikely that the goal of global wireless internetworking can be implemented if all the wireless subsystems are designed independently, as is currently the case.

1.2 Technical Issues

In addition to the internetworking difficulties, the physical limitations of the wireless communication link present a fundamental technical challenge for reliable high-speed communication equivalent to that currently available on wireline networks. The channel is susceptible to time-varying noise, interference, and multipath. Moreover, the radio spectrum is a limited resource, and even with the recent increase in spectrum allocation for wireless applications, this resource will be stretched to its capacity to accommodate the various wireless services. Techniques to increase spectral efficiency and effectively share the radio resource are the main focus of this thesis.

Limitations in the power and size of the communication and computing devices also present a major design consideration. Vehicular communication devices have few power or size limitations. However, most personal communication devices are meant to be carried in a briefcase, purse, or pocket. These devices must be small and lightweight, which translates to low power requirements, since small batteries must be used. However, many of the signal processing techniques required for efficient spectral utilization and networking demand much processing power, precluding the use of low power devices. Hardware advances for low power circuits with high processing ability will relieve some of this conflict; however, placing the processing burden on fixed location sites with large power resources has and will continue to dominate wireless system designs. The associated bottlenecks and single points-of-failure are clearly undesirable for the overall system.

1.3 Thesis Outline

The overall approach of each thesis chapter is to first present the theoretical capacity limits of the channel under consideration. We then use the capacity analysis as a foundation for deriving novel communication techniques that come close to this theoretical upper bound. Since optimal performance generally implies more hardware complexity and sensitivity, we also consider suboptimal techniques which are more robust and practical for

actual implementation.

The thesis outline is as follows. We begin in Chapter 2 with a detailed description of time-varying radio channels. We first develop both deterministic and statistical models for multipath channels. We then combine the multipath model with shadowing and interference to obtain two models for urban cellular radio channels: the macrocell model for large coverage areas and the microcell model for small coverage areas. We conclude the chapter with the general state space channel, which models almost any type of channel variation, including time-varying impulse response channels and channels which vary arbitrarily.

Chapter 3 describes techniques for spectrally-efficient communication over time-varying channels when the channel is estimated and this estimate fed back to the transmitter. This allows the transmitter to adapt to the changing channel. We first determine the performance limits, in terms of channel capacity, of such channels. We then propose an adaptive power control and coded-modulation technique for narrowband fading channels which comes close to achieving this performance limit. We conclude the chapter with a discussion of channel estimation. In particular, we compute the effect of estimation error on our adaptive coded-modulation technique. We also bound the capacity loss resulting from periodic channel estimation, where no data is transmitted during this estimation time. The consequent loss in data rate is more than just the fraction of time spent estimating the channel, since the periodic estimation sequence restricts the data encoding.

In many cases, a feedback path between the receiver and transmitter is not available. In Chapter 4 we develop receiver processing techniques to increase spectral efficiency in this case. We first derive the Shannon capacity of time-varying channels without feedback when the channel variation is Markov. We then propose a maximum-likelihood decision-feedback decoder for this channel. Our decoding scheme achieves a higher spectral efficiency than the interleaving and memoryless encoding method typically used on this channel without a significant increase in complexity. We conclude the chapter with a discussion of unequal error protection codes. These codes prioritize the source encoded bit stream to ensure that high-priority bits are received even under worst-case channel conditions.

We then turn to spectrum sharing for multiuser systems. In Chapter 5 we first discuss the performance limits of multiuser systems in the context of multiuser information theory. We then evaluate several spectrum-sharing techniques; in particular, we compare the two competing technologies for the North American digital cellular standard: CDMA and FDMA. Power control was originally proposed for CDMA systems to eliminate the near-

far problem[1]. However, this type of power control, which equalizes the received power of all users, tends to waste power to compensate for bad channels, and also increases interference to other receivers. We therefore propose a hybrid power control technique which equalizes the power of all users, then incrementally increases power and data rates of the users with the best channels. This technique alleviates the near-far problem while taking advantage of good propagation conditions to increase spectral efficiency.

Wireless networks will be examined in Chapter 6. We first propose an architecture using a hierarchical cellular structure, and show how it integrates with existing wireless and wireline networks. We then describe some of the protocols necessary for routing and mobility management within this heterogeneous network infrastructure. We also outline the other protocols required for network operation, including security, pricing, and network control. Conclusions and extensions to this thesis are discussed in the final chapter.

[1]The near-far problem arises in multiuser systems when two transmitters using the same frequency band, but with different channel characteristics, access the same receiver. The transmitter with the good channel will tend to overpower the other transmitter.

Chapter 2

Time-Varying Channels

The wireless radio channel poses a severe challenge as a medium for reliable high-speed communication. Not only is it susceptible to noise, interference, and multipath, but because the users are presumed to be moving, these channel impediments change over time in unpredictable ways. In this chapter, we will characterize channel variations for several different types of channels. We first define the additive noise channel. Since receiver hardware always introduces some noise, models for any time-varying channel should include an additive noise term, unless the noise is negligible relative to other channel impediments. We then consider multipath effects, which can cause two types of signal degradation: the amplitude of the received signal may vary over time, and the received signal may be distorted or spread in time, resulting in intersymbol interference.

Next, we discuss models for urban cellular channels, both macrocells (one to five mile coverage area) and microcells (one thousand foot coverage). For cellular channels, interference resulting from spatial reuse of the same frequency band adds to the other channel impediments. Moreover, the propagation characteristics of cellular systems change with cell size. We conclude the chapter with an abstract state space model, where the channel variation between states is governed by a stationary stochastic process. This model is applied to both discrete and continuous channels.

2.1 Additive Noise Channels

The additive noise channel models noise introduced by hardware components at the receiver front end. The channel model is illustrated in Figure 2.1, where the noise term

$n(t)$ is a stochastic process. Receiver noise is commonly modeled as a zero-mean Gaussian process. When the noise is white, the channel is referred to as an additive white Gaussian noise (AWGN) channel. Performance and communication techniques for the additive noise channel have been studied in depth since the late 1940s; we will consider additive noise effects only in conjunction with other channel impediments, such as multipath fading and interference.

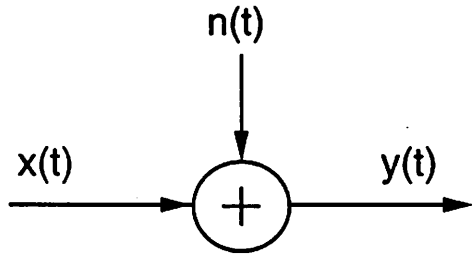

Figure 2.1: Additive Noise Channel.

2.2 Multipath Channels

In a typical urban environment, a radio signal transmitted from a fixed source to a mobile receiver experiences extreme variation in both amplitude and phase. This variation is due to multipath, which arises when the transmitted signal is reflected, diffracted, or scattered by an object. These additional copies of the transmitted signal can be attenuated in power, delayed in time, and shifted in phase and/or frequency from the line-of-sight (LOS) signal path[1]. Multipath affects the received signal in two ways: the constructive and destructive interference of the multiple paths causes the received signal amplitude to vary, and the time delay of each path causes intersymbol interference if the signal bandwidth is larger than the inverse of the delay spread. We will discuss both of these phenomena for several different multipath models in the subsections below.

We assume that the distances are small enough not to be affected by the earth's curvature [1]. If the transmitter, receiver, and reflectors are all immobile, then the constructive and destructive interference of the multiple paths, and their delays relative to the LOS

[1]The line-of-sight path is the straight line path between the transmitter and receiver. This path may also be blocked or attenuated.

path, are fixed. However, if the source or receiver are moving, then the characteristics of the multiple paths vary with time. These time variations are deterministic when the number, location, and characteristics of the reflectors are known, otherwise, statistical models must be used.

In §2.2.1, we describe propagation models which assume a finite number of reflectors with known location and dielectric properties. The details of the multipath propagation in this case can be solved using Maxwell's equations with appropriate boundary conditions. However, the computational complexity of this solution makes it impractical as a general modeling tool [2]. Ray-tracing techniques approximate the propagation of electromagnetic waves by representing the wavefronts as simple particles: the model determines the reflection and refraction effects on the wavefront but ignores the more complex scattering phenomenon predicted by Maxwell's coupled differential equations. The error of the ray tracing approximation is smallest when the receiver is many wavelengths from the nearest scatterer, and all the scatterers are large relative to a wavelength and fairly smooth, as with window reflections. Comparison of the ray tracing method with empirical data shows it to be a good model for signal propagation in rural areas, or along city streets where both the transmitter and receiver are close to the ground [3].

We conclude §2.2.1 with a general ray tracing model that has attenuated, diffracted, and scattered multipath components. This model uses all of the geometrical and dielectric properties of the buildings surrounding the transmitter and receiver, and therefore the model almost always requires on-site empirical measurements. Computer programs based on this model, which use a local building database for calculations, are currently available [4]; these programs are now widely used for system planning in both indoor and outdoor environments.

If the number of reflectors is large, or the reflector surfaces are not smooth, then we can use statistical approximations based on the law of large numbers. The fading model described in §2.2.2 yields the propagation statistics in this case, which vary depending on the signal bandwidth. Much work has been done on statistical modeling of radio propagation over a large urban area [1, 5, 6, 7]; we will derive and summarize the commonly used statistical models for both wideband and narrowband signals. Hybrid models, which combine ray tracing and statistical fading, can also be found in the literature [8, 9], however we will not describe them here.

2.2.1 Ray Tracing Models

This section describes several ray tracing models of increasing complexity. We start with a simple two-path model, which predicts signal variation resulting from a ground reflection interfering with the LOS path. This model characterizes signal propagation in isolated areas with few reflectors, such as rural roads or highways. We then present a ten-ray reflection model, which predicts the variation of a signal propagating along a straight, building-lined street with the transmit and receive antennas placed below the skyline. Finally, we describe a general model which predicts signal propagation for any building and transceiver configuration. The two-ray model only requires information about the antenna heights, while the ten-ray model requires antenna height and street width information, and the general model requires these parameters as well as detailed information about the geometry and dielectric properties of the surrounding buildings.

We assume that the transmitted signal is given by

$$s(t) = u(t)e^{j(2\pi ft+\phi_0)}, \tag{2.1}$$

where $u(t)$ is a complex baseband signal with bandwidth B_u and power P_u, f is the carrier frequency, and ϕ_0 is an arbitrary initial phase. Throughout this section, we will suppress the additive receiver noise as defined in §2.1, since it is added to the sum of multipath components. Similarly, we suppress the phase term $e^{j(2\pi ft+\phi_0)}$, since it is a constant multiplier of all the multipath components. In addition to the random phase, there is a doppler frequency shift of each multipath component equal to $v\cos\psi/\lambda$, where ψ is the arrival angle of the multipath ray, v is the receiver velocity, and $\lambda = c/f$ is the signal wavelength. Thus, $v_m \triangleq v\cos\psi$ is the relative velocity between the transmitter and receiver. We will ignore this doppler term in the ray tracing models of this section, since for typical urban vehicle speeds (60mph) and frequencies (900 MHz), it is less 70Hz [1, 5]. However, we will include doppler effects in the statistical models of §2.2.2.

Two-Path Model

The two-path model is used when a single ground reflection dominates the multipath effect, as illustrated in Figure 2.2. The received signal consists of two components: the direct or LOS component, which is just the transmitted signal propagating through free space, and a reflected component, which is the transmitted signal reflected off the ground.

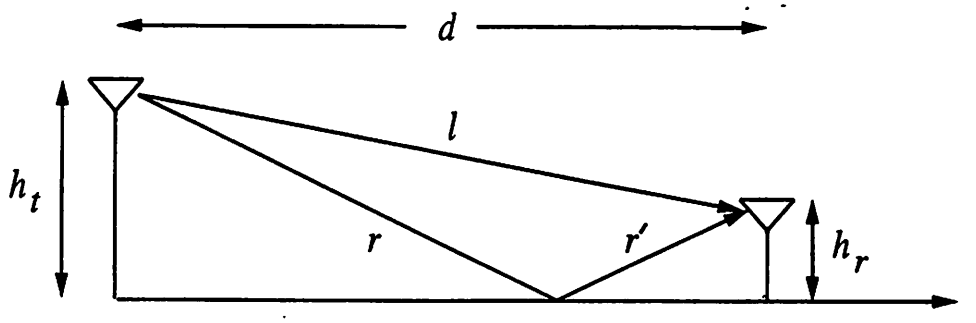

Figure 2.2: Two-Path Model.

The received LOS component is determined from the free-space propagation loss formula:

$$r_{LOS}(t) = u(t)\frac{\lambda G_l e^{j(2\pi l/\lambda)}}{4\pi l}, \tag{2.2}$$

where l is the length of the LOS path and G_l is the product of the transmit and receive antenna field radiation patterns in the LOS direction. The reflected ray is shown in Figure 2.2 by the segments r and r'. If we ignore the effect of surface wave attenuation[2], then by superposition, the received signal for the two-path model is

$$r_{2path}(t) = \frac{\lambda}{4\pi}\left[\frac{G_l u(t)e^{j(2\pi l/\lambda)}}{l} + \frac{RG_r u(t+\tau)e^{j2\pi(r+r')/\lambda}}{r+r'}\right], \tag{2.3}$$

where $\tau = (r+r'-l)/\lambda c$ is the time delay of the ground reflection, R is the ground reflection coefficient, and G_r is the product of the transmit and receive antenna field radiation patterns corresponding to rays r and r', respectively. If the transmitted signal is narrowband relative to the time delay ($\tau << B_u^{-1}$), then $u(t) \approx u(t+\tau)$. Thus, the received power of the two-path model for narrowband transmission is

$$P_r = P_u\left[\frac{\lambda}{4\pi}\right]^2\left|\frac{G_l}{l} + \frac{RG_r e^{j\Delta\phi}}{r+r'}\right|^2, \tag{2.4}$$

where $\Delta\phi$ is the phase difference between the two received signal components. If d denotes the horizontal separation of the antennas, h_t denotes the transmitter height, and h_r denotes the receiver height, then this phase difference is given by

$$\Delta\phi = \frac{2\pi(r'+r-l)}{\lambda} = \frac{2\pi}{\lambda}\left[\left[\left(\frac{h_t+h_r}{d}\right)^2+1\right]^{1/2} - \left[\left(\frac{h_t-h_r}{d}\right)^2+1\right]^{1/2}\right]. \tag{2.5}$$

[2]This is a valid approximation for antennas located more than a few wavelengths from the ground.

Equation (2.4) has been shown to agree very closely with empirical data [10]. The delay spread of the two-path model is just the excess delay of the ground reflection: $(r+r'-l)/c$. When $d > 5h_t h_r$, $r + r' - l \approx 2h_t h_r/d$, and thus $\Delta\phi \approx 4\pi h_t h_r/\lambda d$.

The ground reflection coefficient is given by [1, 11]

$$R = \frac{\sin\theta - Z}{\sin\theta + Z}, \tag{2.6}$$

where

$$Z = \begin{cases} \sqrt{\epsilon_r - \cos^2\theta}/\epsilon_r & \text{for vertical polarization} \\ \sqrt{\epsilon_r - \cos^2\theta} & \text{for horizontal polarization} \end{cases}, \tag{2.7}$$

and ϵ_r is the dielectric constant of the ground, which for earth or road surfaces is approximately that of a pure dielectric ($\epsilon_r = 15$).

From (2.5), if $d > 5h_t h_r$, then $r + r' - l \approx 2h_t h_r/d$, and thus

$$\Delta\phi \approx 4\pi h_t h_r/\lambda d. \tag{2.8}$$

For asymptotically large d, $r + r' \approx l \approx d$, $\theta \approx 0$, $G_l \approx G_r$, and $R \approx -1$. Substituting these approximations into (2.4), we see that in this asymptotic limit, the received signal power is approximately

$$P_r \approx \left[\frac{\lambda G_l}{4\pi d}\right]^2 \left[\frac{4\pi h_t h_r}{\lambda d}\right]^2. \tag{2.9}$$

Thus, in the asymptotic limit of large d, the received power falls off inversely with the fourth power of d. In [10], plots of (2.4) as a function of distance illustrate this asymptotic limit; up to a certain critical distance d_c, the wave experiences constructive and destructive interference of the two rays, resulting in a wave pattern with a sequence of maxima and minima. At distance d_c, the final maximum is reached, after which the signal power falls off proportionally to d^{-4}. An approximation for d_c can be obtained by setting $\Delta\phi$ to π in (2.8), obtaining $d_c = 4h_t h_r/\lambda$. The critical distance is used in the design of cellular systems to determine optimal cell size, as we will discuss in §2.3.

If we average out the local maxima and minima in (2.4), the resulting average power loss can be approximated by dividing the power loss curve into two regions. For $d < d_c$, the average power falloff with distance corresponds to free space loss. For $d > d_c$, the falloff with distance is approximated by the fourth-power law in (2.9). These approximations are captured with the following simplified model for average received power [12, 13], which assumes that $G_l \approx G_r$:

$$P_r = P_u G_l \left[\frac{\lambda}{4\pi}\right]^2 \frac{1}{L(d)}, \tag{2.10}$$

where

$$L(d) \triangleq \left[\frac{d}{d_0}\right]^2 \sqrt[q]{1+\frac{d}{d_c}^{(m-2)q}} \tag{2.11}$$

is a linear approximation for the power falloff. For this approximation, $m \triangleq 4$ is the exponent of the power falloff in the asymptotic limit of large d, d_0 is an empirical constant that reflects the constructive addition of the two paths before the transition region, and q is a parameter that determines the smoothness of the path loss at the transition region close to d_c.

Dielectric Canyon (Ten-Ray Model)

We now examine a model for urban area transmissions developed by Amitay [3]. This model assumes rectilinear streets[3] with buildings along both sides of the street and transmitter and receiver antenna heights that are well below the tops of the buildings. The building-lined streets act as a dielectric canyon to the propagating signal. Theoretically, an infinite number of rays can be reflected off the building fronts to arrive at the receiver; in addition, rays may also be back-reflected from buildings behind the transmitter or receiver. However, since some of the signal energy is dissipated with each reflection, signal paths corresponding to more than three reflections can generally be ignored. When the street layout is relatively straight, back reflections are usually negligible also. Experimental data shows that a model of ten reflection rays closely approximates signal propagation through the dielectric canyon [3]. The ten rays incorporate all paths with one, two, or three reflections: specifically, there is the LOS, the ground-reflected (GR), the single-wall reflected (SW), the double-wall reflected (DW), the triple-wall (TW) reflected, the wall-ground (WG) reflected and the ground-wall (GW) reflected paths. There are two of each type of wall-reflected path, one for each side of the street. An overhead view of the LOS, ground, single-wall, double-wall, and triple-wall reflected rays is shown in Figure 2.3. Rays reflected off vehicles are not included in this model.

For the ten-ray model, the received signal is given by

$$r_{10ray}(t) = \frac{\lambda}{4\pi}\left[\frac{G_l u(t) e^{j(2\pi l)/\lambda}}{l} + \sum_{i=1}^{9} \frac{R_i G_{ri} u(t+\tau_i) e^{j(2\pi r_i)/\lambda}}{r_i}\right], \tag{2.12}$$

[3] A rectilinear city is flat, with linear streets that intersect at 90° angles, as in midtown Manhattan.

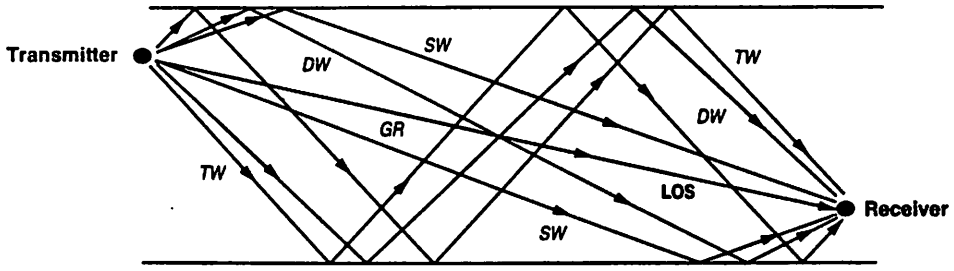

Figure 2.3: Eight Rays of the Ten-Ray Model.

where r_i denotes the path length of the ith reflected ray, $\tau_i = (r_i - l)/\lambda c$, and G_{ri} is the product of the transmit and receive antenna gains corresponding to the ith ray. For each reflection path, the coefficient R_i is either a single reflection coefficient given by (2.6) or, if the path corresponds to multiple reflections, the product of the reflection coefficients corresponding to each reflection. The dielectric constants used in (2.6) are approximately the same as the ground dielectric, so $\epsilon_r = 15$ is used for all the calculations of R_i. If we again assume that $u(t) \approx u(t + \tau_i)$ for all i, then the received power corresponding to (2.12) is

$$P_r = P_u \left[\frac{\lambda}{4\pi}\right]^2 \left|\frac{G_l}{r} + \frac{\sum_{i=1}^{9} R_i G_{ri} e^{j\Delta\phi_i}}{r_i}\right|^2, \tag{2.13}$$

where $\Delta\phi_i = 2\pi(r_i - l)/\lambda$. The delay spread for this model is $\max_i[(r_i - l)/c]$.

Power falloff with distance in both the ten-ray model (2.13) and urban empirical data [10, 14, 15] is proportional to d^{-2}, even at relatively large distances. Moreover, this falloff exponent is relatively insensitive to the transmitter height, as long as the transmitter is significantly below the building skyline. This falloff with distance squared is due to the dominance of the wall-reflected rays, which decay as d^{-2}, over the combination of the LOS and ground-reflected rays (the two-path model), which decays as d^{-4}. Other empirical studies [12, 16, 17] have obtained power falloff with distance proportional to $d^{-\gamma}$, where γ lies anywhere between two and four. The difference in falloff exponents among the various empirical studies indicates the difficulty in obtaining a single model to encompass all the vagaries of urban signal propagation. However, we can generalize (2.10) by using an expression for $L(d)$ with a more general falloff exponent than (2.11). Two different expressions for $L(d)$ have been used in the literature:

$$L(d) \triangleq d^2 \sqrt[q]{1 + \frac{d}{d_c}^{(m-2)q}} \tag{2.14}$$

is used in [12], where the falloff parameter m, the critical distance d_c, and the smoothing parameter q are derived empirically for different streets. In [18], the expression

$$L(d) \triangleq \frac{1}{A + Bd^2 + Cd^4} \tag{2.15}$$

is used, where the (A, B, C) coefficients are derived from empirical data. The two models are quite similar: (2.14) is more general since it can encorporate more values for the distance falloff, however (2.15) can be used over a variety of street layouts with different falloff characteristics [18]. For this reason we use (2.15) as our falloff model in later sections.

General Ray Tracing

General Ray Tracing (GRT) can be used to predict field strength and delay spread for any building configuration and antenna placement [4, 19, 20]. For this model, the building database (height, location, and dielectric properties) and the transmitter and receiver locations relative to the buildings must be specified exactly. Since this information is site-specific, the GRT model is not used to obtain general theories about system performance and layout; rather, it explains the basic mechanism of urban propagation, and can be used to obtain delay and signal strength information for a particular transmitter and receiver configuration.

The GRT method uses geometrical optics to trace the propagation of the LOS and reflected signal components, as well as signal components from building diffraction and diffuse scattering. There is no limit to the number of multipath components at a given receiver location: the strength of each component is derived explicitly based on the building locations and dielectric properties. In general, the LOS and reflected paths provide the dominant components of the received signal, since diffraction and scattering losses are high. However, in regions close to scattering or diffracting surfaces, which are typically blocked from the LOS and reflecting rays, these other multipath components may dominate.

The propagation model for direct and reflected paths was outlined in the previous section. Wedge diffraction provides an accurate model for the mechanism by which signals are diffracted around street corners [20, 21, 22], although the knife-edge diffraction model is sometimes preferred for its simplicity [1, 11]. The geometry of wedge diffraction is shown in Figure 2.4. The geometrical theory of diffraction (GTD) yields the following formula for

the received diffracted signal:

$$r(t) = u(t)\, D\, \frac{G_d e^{j(2\pi(d+d'))/\lambda}}{d'} \sqrt{\frac{d'}{d(d'+d)}}, \tag{2.16}$$

where G_d is the antenna gain, and D represents the diffraction coefficient, which depends on the signal polarization, the wedge angle, and the angles of incidence and diffraction (ϕ and ϕ'). Theoretical and heuristic expressions for D can be found in [22] and [21], respectively. The latter reference obtains numerical results for the diffraction coefficient, yielding losses that exceed 100dB for some incident angles. Calculation of the diffraction coefficient generally requires a computer, although simple approximations have also been derived [23].

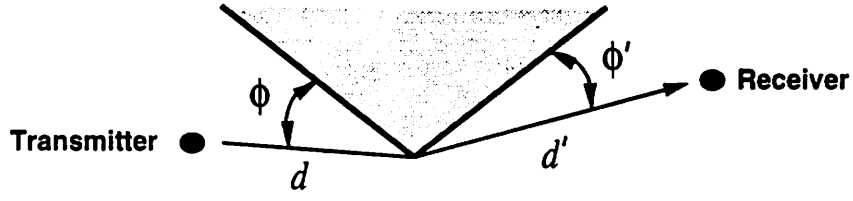

Figure 2.4: Wedge Diffraction.

In addition to the wedge-diffracted ray, there may also be multiply diffracted rays, or rays that are both reflected and diffracted. Models exist for including all possible permutations of reflection and diffraction [23, 24]; however, the attenuation of the corresponding signal components is generally so large that these components are negligible relative to the noise.

A scattered ray, shown in Figure 2.5 by the segments s' and s, has a path loss proportional to the product of s and s'. This multiplicative dependence is due to the additional spreading loss the ray experiences after scattering. The received signal due to a scattered ray is given by the bistatic radar equation [25]:

$$r(t) = u(t) \frac{\lambda G_s \sigma e^{j(2\pi(s+s'))\lambda}}{4\pi s s'}, \tag{2.17}$$

where σ is the radar cross section of the scattering object, and G_s is the antenna gain. The value of σ depends on the roughness, size, and shape of the scattering object. Empirical values of σ were determined in [26] for different buildings in several cities.

The total received electric field is determined from the superposition of all the components due to the multiple paths. Thus, if we have a LOS ray, N_r reflected waves, N_d

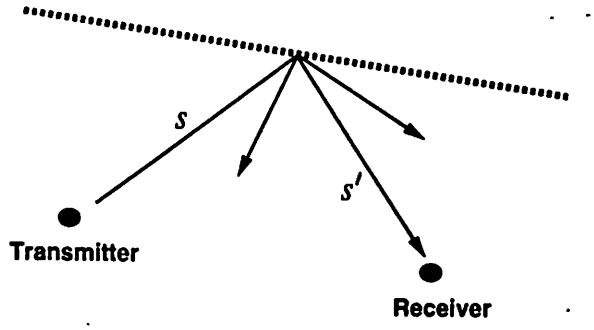

Figure 2.5: Scattering.

diffracted rays, and N_s diffusely scattered rays, the total received signal is

$$\begin{aligned} r_{total}(t) &= \left[\frac{\lambda}{4\pi}\right]\left[\frac{G_l u(t)e^{j(2\pi l)/\lambda}}{l} + \sum_{i=1}^{N_r}\frac{R_i G_{r_i} u(t-\tau_i)e^{j(2\pi r_i/\lambda)}}{r_i}\right. \\ &+ \sum_{i=1}^{N_d}\frac{D_i G_{d_i} u(t-\tau_i)e^{j(2\pi(d_i+d_i'))/\lambda}}{d_i'}\sqrt{\frac{d_i'}{d_i(d_i'+d_i)}} \\ &+ \left.\sum_{i=1}^{N_s}\frac{\sigma_i G_{s_i} u(t-\tau_i)e^{j(2\pi(s_i+s_i'))\lambda}}{s_i s_i'}\right], \end{aligned} \tag{2.18}$$

where τ_i is the time delay of the given multipath component. The corresponding received power is $P_{total} = E|r_{total}(t)|^2$.

Any of these multipath components may have an additional attenuation factor if its propagation path is blocked by buildings or other objects. In this case, the attenuation factor of the obstructing object multiplies the component's path loss term in (2.18). This attenuation loss will vary widely, depending on the material and depth of the object. An attenuation loss of 12dB is commonly used as an average of empirical measurements [27].

2.2.2 Statistical Fading Models

The models of the previous sections all require detailed information about the number and nature of the multipath components. In this section, we describe a statistical model for the received signal. There are generally two phenomena that cause fluctuations in the received signal as the receiver or transmitter moves. First, as discussed in the previous section, multiple signal reflections arrive at the receiver shifted in phase, which causes constructive and destructive interference. The resulting variations in the signal amplitude, called *signal fading*, vary over distances proportional to a signal wavelength; thus, this type

of fading is referred to as *fast* fading. When the number of multipath components is large, the law of large numbers can be used to approximate the fast fading effects with Gaussian statistics. We first describe this approximation, which results in Rayleigh statistics of the short-term signal envelope variation. As in the previous section, we exclude the noise term introduced at the receiver front end in the equations below.

In addition to interference effects, the LOS and reflected paths may also be attenuated by buildings or other objects. This type of fading, or *shadowing*, varies over distances that are proportional to the size of the buildings, and is thus referred to as *slow* fading. When the number of signal attenuators is large, a Gaussian approximation for the attenuation distance can be used for the slow fading statistics; this results in a log-normal distribution for the signal variation over large distances.

The fast and slow fading phenomena give rise to a multiplicative model for the received power:

$$p(t) = r(t)s(t), \tag{2.19}$$

where $r(t)$ is the value of the Rayleigh fading, $s(t)$ is the value of the log-normal shadowing, and the two processes are statistically independent. From this independence, $\overline{p} = \overline{r}\,\overline{s}$. If P, R, and S denote the dB values of p, r, and s, respectively, then the received power has the additive form $P(t) = R(t) + S(t)$.

The Rayleigh fading model applies to both satellite and terrestrial communication systems: multipath is generated in satellite systems from tropospheric scatter [28], and in terrestrial systems from building reflections. Slow log-normal shadowing is unique to terrestrial urban communication systems when the transmitter or receiver is placed above the building skyline [1, 11, 29]. For rural, semiurban, and urban propagation with both the transmitter and receiver below the skyline, the ray tracing techniques of the previous section better characterize both fast and slow fading of the received signal.

Short-Term Fluctuations

The statistical model for short-term multipath fluctuation of the received signal amplitude is based on a physical propagation environment consisting of a large number of isolated reflectors with unknown locations and reflection properties. Let the transmitted signal be given by (2.1). If we initially assume that the LOS component is obstructed, the

corresponding received signal is the sum of all multipath components:

$$r(t) = \sum_i \frac{\lambda}{4\pi r_i} R_i G_{ri} u(t + \tau_i) e^{j2\pi(f-\delta f_i)(t+\tau_i)}. \tag{2.20}$$

The unknowns in this expression are the multipath component delays ($\tau_i = r_i/c$), doppler shifts (δf_i), reflection coefficients (R_i), and antenna gains (G_{ri}). These parameters change with time, and we assume that their variation is stationary. In general, omnidirectional antennas are used, so the antenna gains are approximately equal. We also assume that the path length spread, defined by

$$S \triangleq \max_i r_i - \min_i r_i, \tag{2.21}$$

is small relative to the carrier wavelength: thus, the attenuation with distance, $\frac{\lambda}{4\pi r_i}$, will be approximately the same for each reflected path. However, since $r_i >> \lambda$, small differences in r_i can lead to extreme phase differences in the received components. This suggests that $\theta_i \triangleq 2\pi f \tau_i$ should be modeled as an i.i.d. random variable uniformly distributed on $[-\pi, \pi]$: this phase model has been confirmed by empirical measurements [7].

Under these assumptions, the received signal is approximated by

$$r(t) \approx A \sum_i R_i u(t - \tau_i) e^{j2\pi[(f-\delta f_i)t + \theta_i - \delta f_i \tau_i]}, \tag{2.22}$$

where A equals the product of the distance loss and antenna gain, which is the same for all i. If we assume that the R_is are also i.i.d. and independent of the θ_is, then the first and second moments of the received process are

$$\mathrm{E}[r(t)] = 0, \tag{2.23}$$

$$\mathrm{E}[r(t)r(s)] = \left[\sum_i \frac{A^2}{2} \overline{R_i^2} u(t - \tau_i) u^*(s - \tau_i) e^{j2\pi(f-\delta f_i)(t-s)}\right], \tag{2.24}$$

where $\overline{x}$ denotes the expectation of x, and x^* denotes its complex conjugate.

If the received process is Gaussian, then the first and second order statistics specify it completely. The process will be Gaussian if the R_is are Rayleigh distributed, or will approach the Gaussian distribution for any R_i distribution as the number of scatterers becomes large [7].

When the doppler spread is zero (i.e., the channel is static), we say that the channel is *time dispersive*. There are three types of signal distortion caused by the time dispersive channel: incoherent combining (fading), distortion, and time spreading.

If the multipath delay spread, defined by $L \triangleq S/c$, is small relative to the inverse signal bandwidth ($L << B_u^{-1}$), then $u(t+\tau_i) \approx u(t)$, and we can rewrite (2.22) as

$$r(t) \approx u(t) \left(\sum_i A R_i e^{j(2\pi r_i)/\lambda} \right). \tag{2.25}$$

Equation (2.25) differs from the original transmitted signal by the complex scale factor in parentheses. It can be shown that, under the assumptions stated above, this scale factor is Rayleigh distributed [7], so the variation of the received signal envelope is Rayleigh. This has also been confirmed experimentally [5, 6]. An approximation for the autocorrelation of $r(t)$ for narrowband transmission is [1]

$$A_r(\tau) \triangleq \mathrm{E}[(r(t)-\overline{r})(r(t+\tau)-\overline{r})] = \sigma^2 J_0^2(2\pi f_m \tau), \tag{2.26}$$

where σ^2 is the variance of $r(t)$, J_0 denotes a 0th order Bessel function, and $f_m = v_m/\lambda$. When the LOS component is not blocked, the envelope variation follows a Rician distribution [30]. Either Rayleigh or Rician amplitude variations can cause severe performance degradation of narrowband modulation techniques [31].

Another form of distortion occurs due to the multipath delay spread L. A short transmitted pulse will result in a received signal that is at least as long as the multipath delay spread. Thus, the duration of the received signal may be significantly increased. If we transmit short data pulses sequentially, this time spreading will result in intersymbol interference. Equalization techniques, which basically invert the channel impulse response, may be used to counter this effect [31].

As B_u increases so that $L \approx B_u^{-1}$, the approximation $u(t-\tau_i) \approx u(t)$ is no longer valid. Thus, the received signal is a sum of copies of the original signal, where each copy is delayed in time by τ_i and shifted in phase by θ_i. The signal copies will combine destructively when their phase terms differ significantly, and will distort the direct path signal when $u(t-\tau_i)$ differs from $u(t)$. However, wideband signal modulation techniques can be used to counter the distortion and fading effects of the time dispersive channel. Spread spectrum [32] is one such method. The basic idea is to multiply the narrowband information signal with a wideband modulating sequence such that the approximation $u(t-\tau_i) \approx u(t)$ is no longer valid. This modulation technique allows the receiver to separate out the delayed multipath components [6]. All but one path is eliminated by a matched filter, hence there is no multipath interference. The receiver may also use a bank of filters matched to $u(t-\tau_i)$

for all i. The matched filter outputs are coherently combined, resulting in a higher effective SNR than would be obtained with just one of the multiple paths.

Wideband signals can be approximated using Turin's model [6] if the incoming paths form subpath clusters. In this model, paths that are approximately the same length ($|\tau_i - \tau_j| << B_u^{-1}$) are not resolvable at the receiver. Thus, they are combined into a single subpath. A finite number of resolvable subpaths is assumed. The received signal is then

$$r(t) = \sum_{i=1}^{I} A_i u(t - \tau_i) e^{j\theta_i}, \tag{2.27}$$

where I is the number of resolvable subpaths, and A_i, τ_i, and θ_i are, respectively, the subpath amplitude, delay, and phase.

A discrete-time version of this model is obtained by dividing the time axis into equal intervals from zero to the maximum expected multipath delay spread [6]. The interval width is less than the receiver resolution, and each subpath is restricted to lie in one of these time interval "bins." We define the random variable ψ_i to be one if a subpath falls in the ith bin, and zero otherwise. The statistics of A_i and θ_i, conditioned on $\psi_i = 1$, can then be taken from the narrowband Rayleigh fading model, or derived from empirical measurements.

This completes the discrete-time approximation for a single channel impulse response. As the channel impulse response changes, a sequence of these models is required. Thus, the time-varying wideband channel model must include both the first order statistics of $(I, \tau_i, \rho_i, \theta_i)$ for each instantaneous channel, as well as the temporal and spatial correlations (assumed Markov) between them. More details on the model and the empirically-derived distributions for $(I, \tau_i, \rho_i, \theta_i)$ can be found in [33].

Long-Term Fluctuations

The signal fading described in the previous section results from out-of-phase combining of the multipath components. Since these phases rotate π degrees every half wavelength, the signal amplitude changes rapidly over short distances (approximately every foot for a 900 MHz signal). If these local variations are averaged out, the local mean will also vary with distance due to two effects: the propagation loss with distance described above for the ray tracing models, and the changing configuration of surrounding buildings and obstacles which attenuate both the LOS and the multipath components. Based on the two- and ten-ray models, it is generally assumed that the propagation loss with distance

is proportional to d^{-4} in a rural environment, and d^{-2} in an urban environment. The more complex models described above (e.g., (2.11), (2.14), and (2.15)) may also be used to determine propagation loss with distance.

Empirical data is commonly used to predict the expected power loss versus distance from building attenuation [1, 19, 34]. Although these loss measurements vary depending on the test location, the distribution of the mean received signal is approximately log-normal in most empirical studies. Thus, the dB value of the mean received signal is Gaussian. This statistical model can be justified by the following attenuation model [35].

The attenuation of a signal as it travels through a building of depth d is approximately equal to

$$s(d) = ce^{-\alpha d}, \tag{2.28}$$

where c is an adjustment constant and α is an attenuation constant that depends on the building materials and interior. If we assume that α is approximately equal for all buildings in a given region, then the attenuation of a signal as it propagates through this region is

$$s(d_t) = ce^{-\alpha d_t}, \tag{2.29}$$

where d_t is the sum of the building widths through which the signal travels. If there are many buildings between the transmitter and receiver, then we can approximate d_t by a Gaussian random variable. Thus, $\log s(d) = \log c - \alpha d_t$ will have a Gaussian distribution with mean μ and standard deviation σ. The value of σ will depend on the environment, and usually ranges between four and twelve dB [1, 18, 36].

The autocorrelation function for the fluctuation of the signal attenuation about this mean is not well documented in the literature. However, measurements in [36] support an autoregressive autocorrelation model of the form

$$A_S(\tau) = \sigma^2 e^{-v\tau/X_c}, \tag{2.30}$$

where $S = \log s$ is wide-sense stationary, σ is the standard deviation of the mean value, v is the vehicle velocity, and X_c is the decorrelation distance, which is a function of the surrounding building sizes and layout. Values of X_c for various measurement conditions are reported in [36].

2.3 Cellular Channels

In order to accommodate the demand for wireless communication, efficient use of the limited available frequency spectrum is essential. Cellular systems exploit the power falloff with distance of a transmitted signal to reuse the same frequency channel or time slot at another spatially separated location [37]. The coverage area is divided into *cells* where, in each cell, only one user is assigned to a particular channel or time slot. With frequency division, the total system bandwidth is divided into orthogonal channels centered around a frequency f_i, and each frequency channel is reused at a spatially separated cell, as illustrated in Figure 2.6. With time division, the signal occupies the entire frequency band, and is divided into time slots t_i which are reused in distant cells. Time division is depicted by Figure 2.6 if the f_is are replaced by t_is.

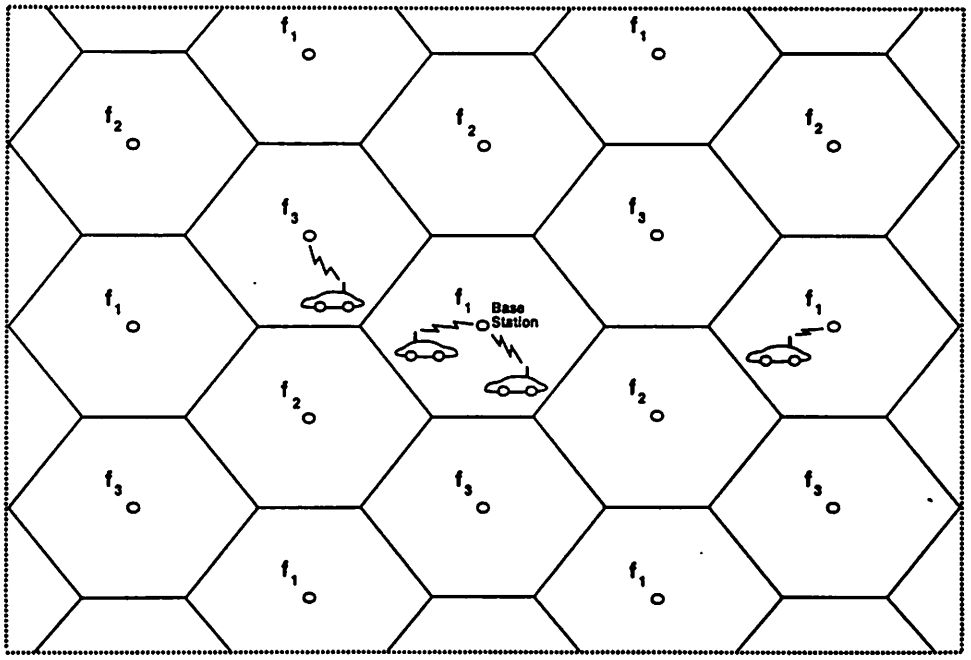

Figure 2.6: Cellular Systems.

Operation within a cell is controlled by a central base station, and the base stations connect to a high-bandwidth wide-area network such as the public telephone system. When a mobile user crosses the boundary between two cells, its communication channel is switched,

or *handed off*, to the base station in the new cell. The shape of the cell is determined by the power footprint of the transmitting base station, which is circular if the transmit and receive antennas are isotropic and propagation follows a free-space loss model. However, urban propagation does not follow the free-space model, so blockage and multipath fading cause significant distortion of this circular shape.

The spatial separation of cells that share the same frequency band or time slot should be as small as possible to cover the largest possible area with a single channel. However, as the spatial reuse distance shrinks, the interference from cells operating in the same frequency or time slot grows. To complicate matters further, both the transmitted and interfering signals experience the long- and short-term multipath fluctuations described in the previous section. To help determine the spatial reuse, data rates, and system layout, accurate models for cellular transmission are required.

Coverage areas can also be divided using spread spectrum code division techniques [38]. For this method, each user within a cell modulates the information signal with a wideband semi-orthogonal coding sequence. The base station can separate each of the received signals by separately decoding each spreading sequence. However, since the codes are semi-orthogonal, the users within a cell interfere with each other (intracell interference), and codes that are reused in other cells also cause interference (intercell interference). Both the intracell and intercell interference power is reduced by the spreading gain of the code. Moreover, interference in spread spectrum systems can be further reduced through multiuser detection and interference cancellation. We will compare code division with the other spectrum-sharing techniques in Chapter 5.

In this section, we consider models for two types of urban cellular systems, based on the size of the cell. Since propagation conditions in suburban and rural areas are more favorable than in cities, these urban models generally reflect worst-case propagation conditions. The first model is for urban macrocells. Macrocells correspond to cells where the base stations are placed on the tops of tall buildings, and transmit enough power to cover one to five miles. These cells are used in the current analog cellular telephone systems of the United States, Europe, and Japan.

If all parameters scale with distance then by shrinking the size of a cell by a factor of N we can accommodate N times more users in a given area, since each cell accommodates the same number of users in a smaller area. However, in order to shrink the size of the cells, the base stations transmit at a much lower power than in macrocells, and therefore must

be placed closer to the ground. From the previous section, we know that lower antenna placement fundamentally changes the mechanism of signal propagation. We therefore use a microcell model for the case when the transmitters are less than fifty feet high. Transmit power in microcells is generally sufficient to cover about a thousand feet; this cell diameter is chosen since it corresponds to the point at which the power falloff of a transmitted signal versus distance increases from d^{-2} to d^{-4}, thereby significantly reducing the power from distant interferers.

We will refer to the transmission link from the mobile to the base station as the *forward* link, and the link from the base station to the mobile as the *reverse* link. The forward links are separated in frequency from the reverse links, so the base stations interfere with each other, but not with the mobiles, and vice versa. Based on both empirical and analytical models, the interference is generally much greater than the receiver noise, so receiver noise will be neglected in our cellular models and analysis.

2.3.1 Macrocells

Macrocell models have been well documented in [1, 11], and the references therein; in this section we summarize these results. The macrocell model requires propagation characteristics of both the transmitted signal within the cell, and the interference from other cells. Since the building concentration in an urban environment is quite dense, both the transmitted signal and the interferers are blocked or reflected from numerous objects. Thus, the statistical propagation model of §2.2.2 applies. When isotropic antennas are used, the long-term received signal variation in both the forward and reverse links are closely approximated by a free-space propagation model with additional log-normal shadowing. Based on empirical measurements, the variance of the log-normal shadowing for typical urban environments ranges from three to eight dB. With this model, energy radiates out from each antenna in a uniform circular pattern. In order to cover a given area with nonoverlapping (tessellating) cells, a hexagonal cell shape is used as the closest tessalating shape to a circle, as shown in Figure 2.7.

For narrowband transmissions, the short-term fluctuation of the desired and interfering signal envelopes generally follows a Rayleigh distribution. If the transmitted signal has a LOS path to the receiver then the fluctuation of the desired signal is Rician[4]. The

[4]The macrocell model generally uses Rayleigh fading as a worst-case assumption.

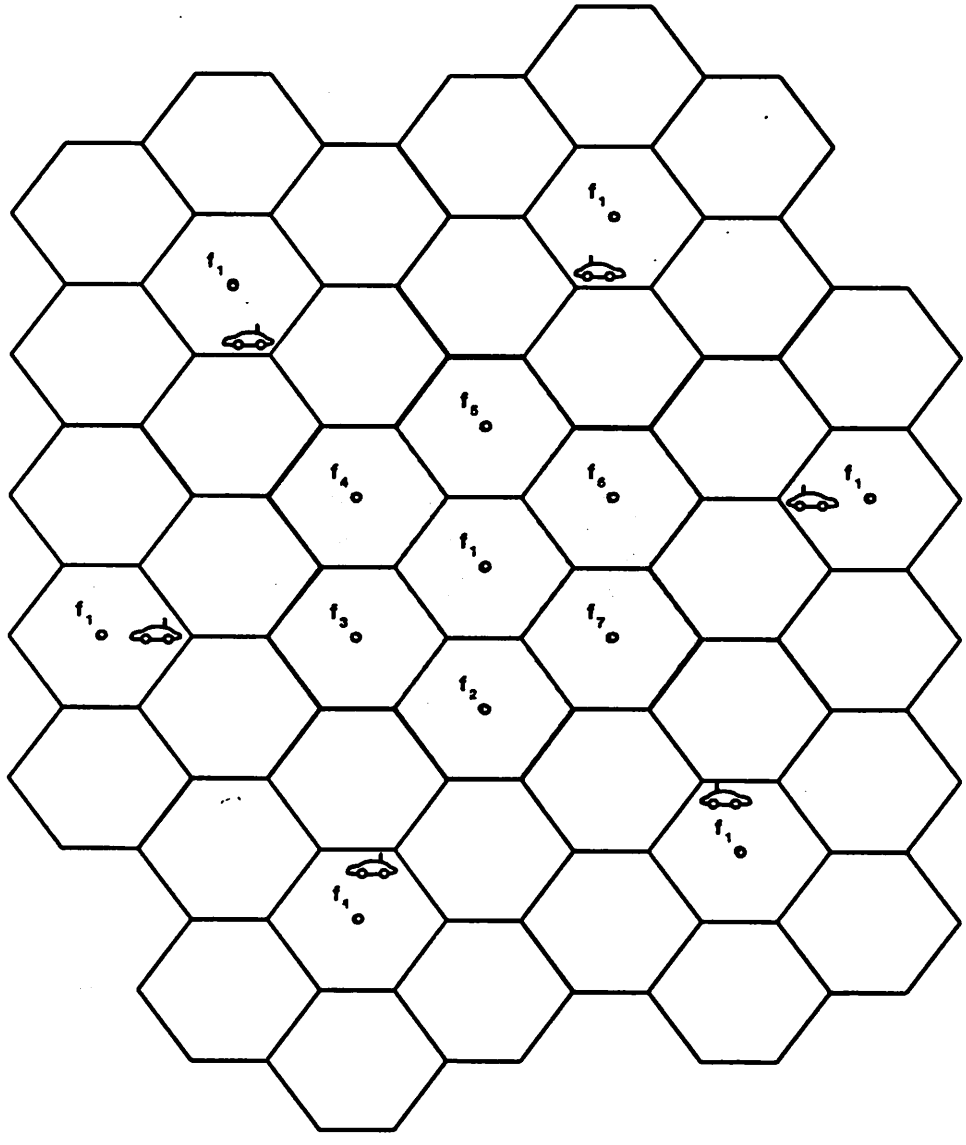

Figure 2.7: Hexagonal Cell Geometry.

number of interfering signals is random, but usually only the interferers in the closest ring of cells are taken into account; these interferers are shown in Figure 2.7 for a cluster size of seven, where the cluster size refers to the number of available reuse frequencies or time slots. Since code division has better intercell interference rejection than time or frequency division [37], a cluster size of one is generally used with this technique. For the reverse link, the distance between an interfering and the transmitting base station (and therefore the maximum interference power) is known. However, since the mobiles may be anywhere within a cell, the average interference power on the forward link is a random variable, with its maximum value determined by placing all of the interfering mobiles on the closest cell boundary, as in Figure 2.7. This figure depicts the interference for time or frequency division. With code division, there are many more interferers both within the same cell, and in adjacent cells, however their interference power is reduced by the spreading gain of the code.

The long-term variation of a wideband signal is characterized by the same log-normal shadowing as in the narrowband case. The short-term fluctuation of both transmitted and interfering wideband signals are characterized by Turin's subpath model (§2.2.2). However, if spread spectrum techniques are used, the number of intracell and intercell interferers is quite large, so we can apply a Central Limit Theorem approximation to the interference and model it as Gaussian noise.

2.3.2 Microcells

In microcells, there are two types of signal propagation: LOS propagation, which refers to propagation between base stations and mobiles with a direct path between them, and non-LOS propagation, which refers to the case where there is no LOS path. In the latter case, the signal must "bend" around one or more corners via diffraction, scattering, or reflection to reach the intended receiver, as shown in Figure 2.8.

The LOS propagation in microcells is accurately modeled with the ten-ray model described in §2.2.1 [3]. However, using ray tracing to model non-LOS propagation requires detailed information about the building and street layout, geometry, and dielectric properties. This information requires field measurements for the particular cell of interest, and the resulting model only applies for that particular site. A more general non-LOS model for cities with rectilinear street layouts is developed in [18]. This measurement-based model is

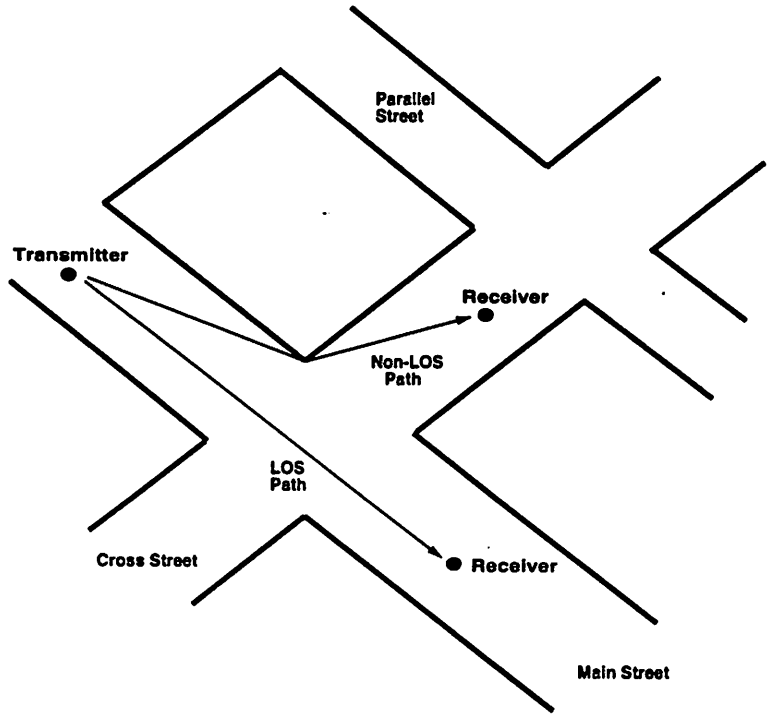

Figure 2.8: Microcell Propagation.

obtained from data collected in Manhattan at 900 MHz [10]. The model includes a prediction method for the mean average power, and a statistical model for both short-term and long-term variations about this mean.

For the microcell geometry of Figure 2.8, within a particular cell the street containing the cell transmitter is called the *main* street, streets perpendicular to the main street are called *cross* streets, and streets parallel to the main street are called *parallel* streets. We will use x to denote the distance variable along a main street, and y to denote the distance variable along a cross street. The model doesn't explicitly determine the power loss on parallel streets, since cross street data can be interpolated to obtain these values.

Constant Average Power

In [18], empirical contours of constant average power with both the long- and short-term fluctuations averaged out have the shape of concave diamonds which are elongated along the main street in both directions. The concave nature of these diamonds suggests that the mechanism by which the signal energy couples into cross streets is via scattering since,

if reflections were the dominant mechanism, the attenuation along any cross street point (x, y) would be a function of $x + y$, where x is the distance traveled along the main street before turning the corner, and y is the subsequent distance traveled along the cross street. This would result in constant power contours with straight sides. However, if the signal couples into the cross street via scattering, the attenuation at (x, y) would be proportional to the product of $f(x)$ and $g(y)$, where f and g are functions that characterize the power loss versus distance along the main and cross streets, respectively. This product form for the power at (x, y) leads to a concave shape for the constant power contours.

If the signal propagates along the main and cross streets according to the free-space loss formula, then the functions f and g would just be linear equations of x^{-2} and y^{-2}, respectively. The path loss model can be generalized to urban propagation using the fitting function of (2.15): for the main street, the path loss at a distance x from the transmitter is approximated by

$$f(x) = \frac{1}{A_m + B_m x^2 + C_m x^4}. \tag{2.31}$$

Similarly, for the cross street, the path loss at distance y from the intersection with the main street is approximated by

$$g(y) = \frac{1}{A_c + B_c y^2 + C_c y^4}. \tag{2.32}$$

The set of coefficients (A_m, B_m, C_m) are chosen to minimize the mean-squared error (MSE) between (2.31) and the empirical path loss data on the main street, and the set (A_c, B_c, C_c) minimizes the MSE between (2.32) and the path loss data for a given cross street [18]. These forms for f and g include power falloff with distance of both d^{-2} and d^{-4}, since both have been observed in urban empirical measurements.

Let $L(x, y)$ denote the path loss at a particular point within a cell, where $(0, 0)$ denotes the coordinates of the transmitter. Consider a particular cross street located x_0 feet from the transmitter. Assuming that f and g predict the path loss perfectly, then the dB path loss at the intersection of the main street and this cross street is given by

$$L(0, x_0) = -10 \log(A_m + B_m x_0^2 + C_m x_0^4). \tag{2.33}$$

Similarly, from (2.32), the dB path loss at any point y along the cross street is given by

$$L(y, x_0) = -10 \log(A_c + B_c y^2 + C_c y^4). \tag{2.34}$$

Setting $y = 0$ in (2.34) and equating it to (2.33) yields

$$-10\log A_c = -10\log(A_m + B_m x_0^2 + C_m x_0^4). \tag{2.35}$$

If we now factor out A_c in (2.34) and substitute the right side of (2.35), we get that the attenuation at the point (y, x_0) is

$$L(y, x_0) = -10\log\left[\left[A_m + B_m x_0^2 + C_m x_0^4\right]\left[1 + \frac{B_c}{A_c}y^2 + \frac{C_c}{A_c}y^4\right]\right]. \tag{2.36}$$

The attenuation equation (2.36) uses the coefficients $(A_c, B_c; C_c)$ derived for a particular cross street. However, in [18] it was found that for every cross street in the data set, the fourth power falloff with distance dominated the other terms, so

$$\frac{C_c}{A_c}y^4 >> 1 + \frac{B_c}{A_c}y^2. \tag{2.37}$$

Moreover, the ratio C_c/A_c was approximately constant over all cross street measurements. If we denote the mean of this ratio by C_A, then substituting this approximation and (2.37) into (2.36) yields an attenuation model for all cross streets in the cell:

$$L(y, x_0) = -10\log C_A y^4\left[A_m + B_m x_0^2 + C_m x_0^4\right]. \tag{2.38}$$

Assume now that $C_c/A_c \approx C_A$ for cross streets in any rectilinear city, where C_A is derived from the Manhattan data. Then the model (2.38) can be applied to any rectilinear city to predict the path loss on cross streets. However, the model still requires a method to obtain the main street propagation coefficients (A_m, B_m, C_m) for the cell site of interest. But from [3], the ten-ray model predicts path loss as a function of distance within a 2dB margin of error along the main street. If we use this model instead of empirical data to obtain the (A_m, B_m, C_m) coefficients, then we only require knowledge of the base and mobile antenna locations and the width of the main street to predict signal attenuation throughout the cell.

We saw that for macrocells, the tessalating shape which approximated the circular constant power contours was a hexagon. Since constant power contours for microcells form concave diamonds, it is possible to inscribe a square inside each of the diamonds to form tessalating cells covering the area of interest. Thus, square cells form the building blocks for microcellular geometries.

Short-Term Fluctuations

The short term fluctuations in microcells are caused by the same phenomenon as in macrocells: the constructive and destructive interference of the multiple paths. The main street usually has an unblocked LOS path, so its short-term fluctuation follows a Rician distribution. The cross streets have no LOS path, and the statistics of the short-term fluctuation on these streets was found in [18] to be approximately Rayleigh.

Long-Term Fluctuations

The long-term signal fluctuation in the microcell model reflects variations in the average path loss formulas of (2.31) and (2.32). For the Manhattan measurements of [18], the statistics of these variations were shown to be log-normal, with an rms value of three to five dB. Thus, the statistics of the long-term signal strength variation in microcells is the same as in macrocells (with a lower variance). However, the cause of this variation is quite different. In macrocells the long-term variation is caused by building blockage. Since microcell signals propagate around buildings, there is no such phenomenon. The long-term fluctuation in microcells has been shown, both empirically and using the ten-ray model, to be caused by multipath [10]. Thus, in microcells multipath gives rise to both the long-term and the short-term fluctuations.

2.4 State Space Channels

The state space channel model applies to general discrete and continuous time-varying channels, whose variation is governed by a stochastic process taking values over a state space of time-invariant channels. We first describe the discrete-time state space channel model, then extend it to continuous-time.

2.4.1 Discrete-Time Model

The variation of the discrete-time state space channel is determined by a discrete-time stochastic process $\{S_n, n \geq 0\}$ with state space $\mathcal{C}$. The state space is a set of discrete memoryless channels (DMCs) with common input and output alphabets, denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. We call S_n the channel state at time n. The input and output of the channel at time n are denoted by x_n and y_n, respectively, and we assume that the

channel inputs are independent of its states. We will use the notation $r^n \triangleq (r_1, \ldots, r_n)$ and $r_m^{n+m} \triangleq (r_m, \ldots, r_{n+m})$ for $r = x, y$, or S.

The discrete-time channel is defined by its conditional input/output probability at time n, which is determined by the channel state at time n,

$$p(y_n|x_n, S_n) = \sum_{c \in C} p_c(y_n|x_n) I[S_n = c], \tag{2.39}$$

where $p_c(y|x) = p(y|x, S = c)$, and I[·] denotes the indicator function. The memory of the state space channel is due to the correlation structure of the process $\{S_n\}$. We assume that the state at any point in time is independent of past input/output pairs, when conditioned on past states:

$$p(S_{n+1}|S^n, x^n, y^n) = p(S_{n+1}|S^n). \tag{2.40}$$

In addition, since the channels in C are memoryless,

$$p(y^N|x^N, S^N) = \prod_{n=1}^{N} p(y_n|x_n, S_n). \tag{2.41}$$

If the inputs are also independent, then

$$p(y^N, x^N|S^N) = \prod_{n=1}^{N} p(y_n, x_n|S_n), \tag{2.42}$$

and

$$p(y^N|S^N) = \prod_{n=1}^{N} p(y_n|S_n). \tag{2.43}$$

The state space model places no restrictions on the stochastic process $\{S_n\}$. We now define two classes of discrete state space channels with particular characteristics for $\{S_n\}$.

Finite-State Markov Channels

If the stochastic process $\{S_n\}$ is Markov with stationary transition probabilities, and its state space C is finite, then we call the state space model a Finite-State Markov channel. Let P be the matrix of transition probabilities for S, so

$$P_{km} = p(S_{n+1} = c_m|S_n = c_k), \tag{2.44}$$

and is independent of n. We also assume that the process $\{S_n\}$ is irreducible and aperiodic. Since the state space is finite, this implies that $\{S_n\}$ is also positive recurrent, ergodic, and

has a unique invariant distribution π_0. If the initial distribution of S_0 is π_0, then $\{S_n\}$ is also a stationary process. The Finite-State Markov Channel is illustrated in Figure 2.9. Since the state transitions are Markov, (2.40) becomes

$$p(S_{n+1}|S^n, x^n, y^n) = p(S_{n+1}|S_n). \tag{2.45}$$

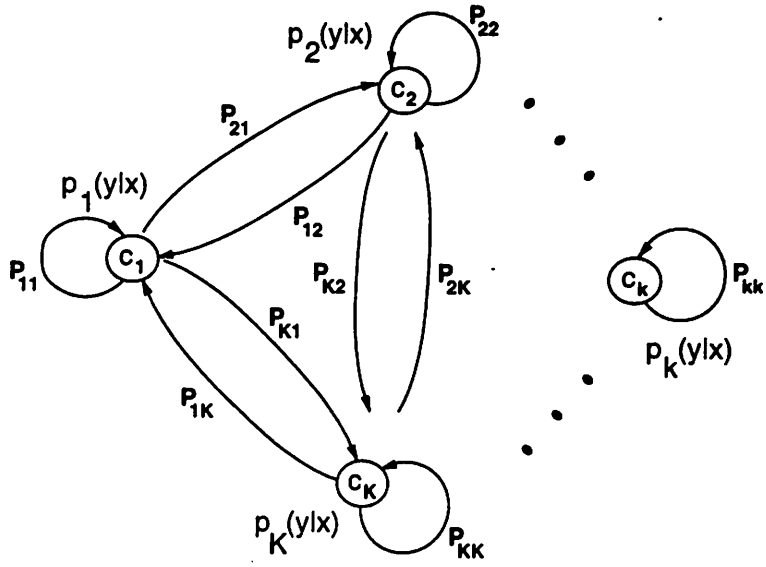

Figure 2.9: Finite-State Markov Channel.

Arbitrarily Varying Channels

The Arbitrarily Varying Channel is a state space channel where the stochastic structure $p(S_n|S^{n-1})$ of the S_ns is unknown for all n. It is also assumed that every state can reach every other state in one step. The channel output probability is thus given by

$$p(y^N|x^N, S^N) = \prod_{n=1}^{n} p(y_n|x_n, S_n), \tag{2.46}$$

where for each n, S_n is chosen at random from the set $\mathcal{C}$.

2.4.2 Continuous-Time Model

Variation of the continuous-time state space channel is governed by a continuous-time stochastic process $\{S_t, t > 0\}$ with state space $\mathcal{C}$. Each $c \in \mathcal{C}$ indexes a time-invariant

continuous-time channel, and S_t is called the channel state at time t. The channels in $\mathcal{C}$ need not be memoryless; however, even if they are memoryless, the time-varying channel still has memory due to the correlation of $\{S_t\}$. We now describe two examples of the continuous state space model which apply to wireless radio channels.

Narrowband Fading Channels

Narrowband fading can be modeled using the continuous-time state space model. Specifically, for an input $x(t)$, the channel output is given by

$$y(t) = S_t x(t) + n(t), \tag{2.47}$$

where $n(t)$ is an additive noise term which is independent of S_t. The channel gain at a particular time instant is determined by a stochastic process $\{S_t\}$ over the set of all positive real numbers, and the transition probabilities of S_t are determined by the autocorrelation of the fading statistics. This autocorrelation was given by (2.26) and (2.30) for Rayleigh and log-normal fading, respectively.

Impulse Response Channels

The continuous state space model also applies to channels with a time-varying impulse response. In this case, $c \in \mathcal{C}$ indexes a time-invariant impulse response $h_c(t)$ with additive noise. The channel response at time t to an impulse at time τ is given by

$$h(t,\tau) = \sum_{c\in\mathcal{C}} h_c(t-\tau)1[S_\tau = c]. \tag{2.48}$$

Thus, for an input $x(\tau)$, the channel output at time t is

$$y(t) = \int_{-\infty}^{t} x(\tau)h(t-\tau, S_\tau)d\tau + n(t), \tag{2.49}$$

where

$$h(t-\tau, S_\tau) \triangleq \sum_{c\in\mathcal{C}} h_c(t-\tau)1[S_\tau = c]. \tag{2.50}$$

The transition probabilities for S_t may be determined, for example, using a data-based model such as Turin's discrete-time wideband multipath channel model [6]. The model for the time-varying impulse response channel is illustrated in Figure 2.10.

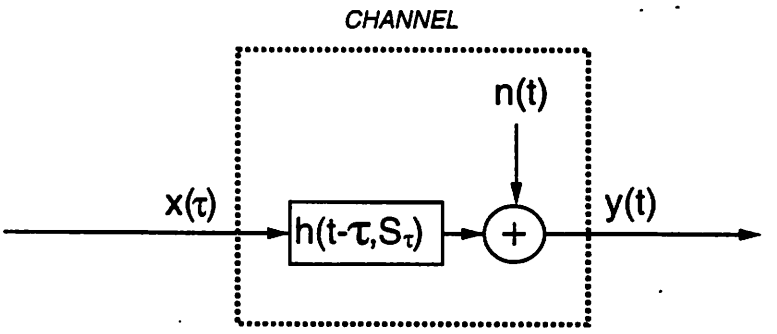

Figure 2.10: Time-Varying Impulse Response Channel.

2.5 Summary

We have outlined the main properties of several different types of time-varying channels. The first model, the additive noise channel, has been studied for quite some time; in the subsequence chapters we will include the effects of additive noise only as an additional impediment to other time-varying factors. The most significant impediments to reliable communication over radio channels are multipath, shadowing, and interference. Multipath has traditionally been characterized statistically with fast Rayleigh fading and slow log-normal shadowing. However, these statistical models break down in urban areas where the transmitter is placed below the building skyline. We therefore also described ray tracing techniques, which specifically calculate the attenuation and phase of each received signal path based on the geometrical configuration of the transmitter, receiver, and surrounding buildings.

Interference is introduced when different signals transmit within the same frequency band. In general, the spectrum is regulated to avoid this overlap. However, cellular systems deliberately introduce interference to reuse their available spectrum at spatially separated points, thereby accommodating more users. We described two types of urban cellular systems, macrocell systems and microcell systems, and showed that the propagation characteristics of both the data signals and the interference are different for each type.

We concluded the chapter with a more abstract model for channel variation: the state space channel. This model characterizes a channel that varies over a set of time-invariant channels, where the variation is governed by a stochastic process. The time-invariant channels, and the governing stochastic process, may be continuous or discrete. We

will see in the following chapters that for this channel model, knowledge of the stochastic process governing the channel variation can be used at the transmitter and receiver to increase the communication rate and reliability.

Chapter 3

Spectrally-Efficient Techniques for Time-Varying Feedback Channels

In this chapter we outline methods for communication over point-to-point time-varying channels, assuming that the channel can be estimated and this information fed back to the transmitter. Thus, the transmitter can adapt to the channel variation. We first derive the maximum spectral efficiency of these channels in terms of their Shannon capacity, and show that this maximum is achieved when three parameters are adapted to the channel variation: transmit power, data rate, and coding scheme. For time-varying impulse response channels, this optimal scheme can be interpreted as a "water-filling" in time and frequency. Variable rate, power, and coding is fairly complex to implement; we therefore compare the spectral efficiency of this optimal policy with that of the constant received power scheme currently proposed for fading cellular channels [38]. Our numerical results show that the constant power policy has a significantly lower spectral efficiency than the optimal policy. We also develop an adaptive trellis-coded modulation scheme for M-QAM, and calculate the spectral efficiency and maximum possible coding gain of this technique. We conclude the chapter with some discussion about the effects of estimation time and error.

3.1 Time-Varying Channel Capacity

The capacity of a time-invariant channel was defined by Shannon to be the mutual information between the channel input and output maximized over all possibly input

distributions [41]. The mutual information is defined as

$$I(X;Y) = E_{x,y} \log[p(x,y)/p(x)p(y)], \tag{3.1}$$

where $p(x,y)$ is the joint distribution of the channel input and output, and $p(x)$ and $p(y)$ denote the channel input and output distributions, respectively. Shannon also proved that, for any data rate below capacity, there exists a block code at that rate with an error probability that goes to zero with block length; however, the block code has no restriction on its code complexity or delay. In addition, no such coding scheme can achieve data rates above capacity with an arbitrarily small error probability.

It is somewhat surprising that the purely mathematical definition of channel capacity in terms of mutual information yields an upper bound on practical transmission rates for time-invariant channels. There is, however, no analogous mathematical definition of mutual information for time-varying channels, since the conditional input/output probabilities of the channel are time-dependent. Therefore, for time-varying channels we define the channel capacity to be the maximum achievable data rate with arbitrarily small probability of error without restriction on the code complexity or delay.

The motivation for determining this capacity in part is to see how close current modulation and coding techniques come to this maximum rate. To the author's knowledge, no coding schemes have been proposed specifically for time-varying channels with estimation and feedback, and the existence of a large gap between rates that are currently achievable and theoretically attainable might elicit more development. Moreover, the optimal code design might suggest effective practical techniques. Indeed, in §3.3 we use the capacity analysis to determine optimal power control, and in 3.5.2 we propose a coded-modulation technique based on the optimal code design which achieves rates approaching the capacity limit.

We now derive the capacity of the continuous-time state space channel described in §2.4.2, assuming that the channel variation S_t is known at time t by both the transmitter and receiver, and that there is an average power constraint on the input. We also assume that S_t is stationary and ergodic, and has a finite number of transitions in any finite time interval. The capacity analysis with these assumptions also applies to the discrete-time model of §2.4.1[1]. Finally, if the channels in $\mathcal{C}$ are not memoryless, we assume that if a state

[1]The discrete-time result without the power constraint is given by Theorem 4.6.1 of [39].

transition occurs at time t, then the input before t does not effect the channel output after time t. Equivalently, the channel for the duration of a particular state is a memoryless block. This assumption is always false if the channels in $\mathcal{C}$ are not memoryless. Let ρ_c denote the channel memory corresponding to state c. We can eliminate the effect of channel memory on subsequent channel states by using a guard band of duration ρ_c after a transition from state c, during which no data is transmitted. Of course, this guard band results in some capacity loss, since no data is transmitted during this period. We will assess the exact capacity loss of this guard band in §3.6.3. Based on these results, if ρ_c is small relative to the channel *latency* (the average amount of time between state transitions), then the capacity loss due to the guard band is small. In this sense, our memoryless block assumption is a reasonable approximation for slowly-varying channels.

With this memoryless block assumption, we can view the channel as a time division system with multiplexed input and demultiplexed output as in Figure 3.1. There are $M = |\mathcal{C}|$ pairs of input and output ports, one pair per channel state. When the time-varying channel is in state $c_i, i = 1, \ldots, M$, the ith pair of ports is connected through the time-invariant channel c_i. We may thus regard the single time-varying channel as M time-invariant channels in parallel, with the restriction that the ith channel can be operated only when the channel is in state c_i.

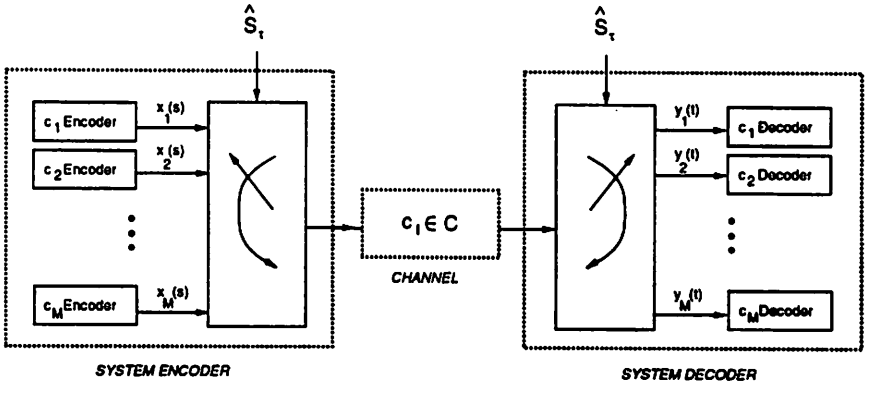

Figure 3.1: Time Diversity System.

A $(2^{RT}, \epsilon, T)$ *code* is a set $(x_1(t), \ldots, x_{2^{\lfloor RT \rfloor}}(t))$ of distinct input signals (codewords) over $[0, T]$, and a set $(\mathcal{Y}_1, \ldots, \mathcal{Y}_{2^{\lfloor RT \rfloor}})$ of disjoint sets in the output space such that $p(y(t) \notin \mathcal{Y}_j | x_j(t)) \leq \epsilon$. We will now define a set of codewords for the multiplexed channel, and in Theorem 3.1 we show that these codes achieve capacity. In Theorem 3.2 we show that no

channel input can do better than this set of channel codes.

Over a time interval $[0,T]$, let T_i denote the total time that $S_t = c_i$. Since S_t is stationary and ergodic,

$$\overline{\left[\frac{T_i}{T}\right]} = p[S_t = c_i] \triangleq \pi_i, \tag{3.2}$$

and

$$\lim_{T\to\infty} \frac{T_i}{T} = \pi_i. \tag{3.3}$$

Let $C_i(P_i)$ denote the Shannon capacity of the time-invariant channel $c_i \in \mathcal{C}$ with average power P_i. Then from [40], for any $R_i < C_i$ there exists a sequence of $(2^{R_i T}, \epsilon_T, T)$ codes that satisfies the power constraint with $\epsilon_T \to 0$ as $T \to \infty$. Fix T, and for each $c_i \in \mathcal{C}$, let $x_i^*(t)$ denote the code corresponding to the time interval $\overline{T_i}$. Over the interval $[0,T]$ the channel is in state c_i for duration T_i. We therefore require codewords corresponding to channel c_i of length T_i, not $\overline{T_i}$. Since $x_i^*(t)$ is defined on $[0,\overline{T_i}]$, we modify these codewords as follows: if $\overline{T_i} \le T_i$, let

$$x_i(t) \triangleq \begin{cases} x_i^*(t) & t \le \overline{T_i} \\ 0 & \overline{T_i} < t \le T_i \end{cases}, \tag{3.4}$$

and if $\overline{T_i} > T_i$, then $x_i(t) \triangleq x_i^*(t), 0 \le t \le T_i$.

Suppose now that during $[0,T]$, the channel is in state c_i for sub-intervals of duration T_{ij}, so $T_i = \sum_j T_{ij}$. The codeword $x_i(t), 0 \le t \le T_i$ can then be broken up into "fractional" codewords $x_{ij}(t)$ of duration T_{ij} corresponding to when the channel is in state c_i, as illustrated in Figure 3.2.

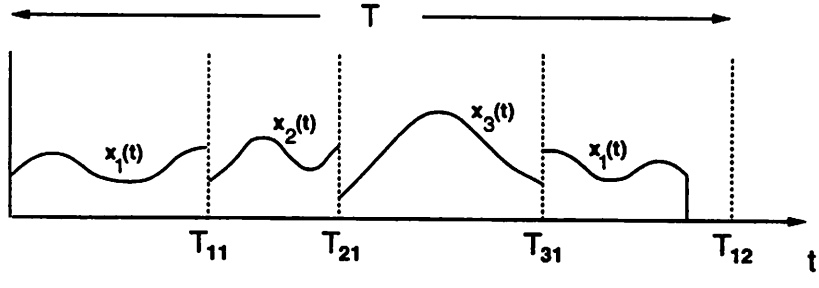

Figure 3.2: Fractional Codewords.

The fractional codewords for all M channels can be time-multiplexed to form a

single codeword $x(t)$ on $[0,T]$ with power

$$\frac{1}{T}\int_0^T |x(t)|^2 dt = \frac{1}{T}\sum_i \int_0^{T_i} |x_i(t)|^2 dt. \tag{3.5}$$

The received signal is demultiplexed and the received blocks for each of the i channels concatenated, which reduces the time-varying channel to M time-invariant channels of duration $T_i, i = 1,\ldots,M$.

The decoding delay of this multiplexed coding scheme will generally be much larger than in the time-invariant case, since to decode the signal corresponding to the ith channel, the decoder must wait for the *total* time that the channel spends in state i to equal the desired block length. The decoding delay corresponding to each of the M channels will also vary, since the dwell times[2] for each state will generally be different. Slowly-varying channels have long dwell times for each channel state; for these channels, the entire block code can generally be sent within one dwell time, so the decoding delay is the same as for the corresponding time-invariant channel. Although our capacity definition places no restriction on the decoding delay, these delays certainly impact practical code designs, especially for delay-constrained data.

Suppose codewords for the time-varying channel can have average signal power at most P. Let

$$\mathcal{P}^M \triangleq \{(P_1,\ldots,P_M) : P_i \geq 0, \sum \pi_i P_i \leq P\} \tag{3.6}$$

be the set of power allocation vectors over the M time-invariant channels. The capacity of a set of independent parallel channels with a mutual power constraint is the sum of the capacity of each channel maximized over the constraint. For our model, the capacity of each of the i channels must be weighted by T_i/T, which approaches π_i as $T \to \infty$. This motivates the following definition for the capacity of the time-varying channel.

$$C \triangleq \max_{P^M \in \mathcal{P}^M} \sum_i \pi_i C_i(P_i), \tag{3.7}$$

where $C_i(P_i)$ is the capacity of the time-invariant channel c_i with input power P_i. Theorem 3.1 shows that any rate $R < C$ can be achieved with arbitrarily small error probability.

Theorem 3.1 For any $R < C$ there exists a sequence of $(2^{RT}, \epsilon_T, T)$ codes with probability of error $\epsilon_T \to 0$ as $T \to \infty$.

[2]Average time before transitioning out of a state.

Proof We first show that we need only consider the case when M is finite. Indeed, since all the C_is have finite capacity and $\sum_i \pi_i = 1$, for all $\epsilon > 0$ there exists a finite $N_\epsilon \leq M$ such that

$$\sum_{i=N_\epsilon+1}^{\infty} \pi_i C_i(P) < \epsilon. \tag{3.8}$$

Let $\mathcal{P}^{N_\epsilon}$ denote the subset of $\mathcal{P}^M$ with $P_i = 0$ for all $i > N_\epsilon$, and define

$$C_{N_\epsilon} \triangleq \max_{P^M \in \mathcal{P}^{N_\epsilon}} \sum_i \pi_i C_i(P_i). \tag{3.9}$$

Combining (3.9) and (3.8), we see that $|C - C_{N_\epsilon}| \leq \epsilon$. Thus, we need only consider the case when M is finite.

Fix $P^M = (P_1, \ldots, P_M) \in \mathcal{P}^M$. Let the ith channel have impulse response $h_{c_i}(t)$. From the results for time-invariant channels [40, page 430], for $R_i < C_i(P_i)$ there exist $2^{\lfloor R_i T_i \rfloor}$ codewords of duration T_i and average power P_i which can be decoded with error probability $\epsilon_i \to 0$ as $T_i \to \infty$. The codewords for each of the M channels can be time-multiplexed in the manner described above to yield $\prod 2^{\lfloor R_i T_i \rfloor} = 2^{\sum \lfloor R_i T_i \rfloor}$ codewords of duration $\sum T_i$ and average power

$$\frac{1}{T} \sum_i \int_0^{T_i} |x_i(t)|^2 dt \leq \frac{1}{T} \sum_i \int_0^{T_i} |x_i^*(t)|^2 dt \to \sum \pi_i P_i \tag{3.10}$$

as $T \to \infty$, since by ergodicity $\frac{T_i}{T_i} \to 1$ and $\frac{T_i}{T} \to \pi_i$. So the new code satisfies the average power constraint in the limit as $T \to \infty$. The received signal is demultiplexed and concatenated for each channel as described above. By the memoryless block assumption, the concatenated output for the ith channel is the same as the response of the ith channel to the original codeword $x_i(t)$, and the decoding of each of the M channels is decoupled. The probability of error, ϵ_T, satisfies

$$\epsilon_T \leq \sum_{i=1}^{M} \epsilon_i \to 0 \quad \text{as} \quad T \to \infty, \tag{3.11}$$

since M is finite and $T \to \infty$ implies that $T_i \to \infty$ for all i. The rate of the new code is

$$R = \frac{1}{T} \sum R_i T_i \to \sum \pi_i R_i \quad \text{as} \quad T \to \infty, \tag{3.12}$$

so rates arbitrarily close to $\sum \pi_i C_i(P_i)$ are achievable. Since this is true for all $(P_1, \ldots, P_M) \in \mathcal{P}^M$, rates arbitrarily close to the capacity defined in (3.7) are achievable.

Theorem 3.2 Any sequence of $(2^{RT}, \epsilon_T, T)$ codes with $\epsilon_T \to 0$ must have $R \leq C$.

Proof Let W be uniformly distributed on $\{1, \ldots, \lfloor 2^{RT} \rfloor\}$. Consider a sequence of $(2^{RT}, \epsilon_T, T)$ codes $\{x_w(t), w = 1, \ldots, \lfloor 2^{RT} \rfloor\}$ with $\epsilon_T \to 0$ as $T \to \infty$ and average power P. Let $I(X;Y)$ denote the mutual information between the input and output of the time-varying channel on $[0,T]$, and $I(X_i;Y_i)$ denote the mutual information of the ith time-invariant channel on $[0,T_i]$. Define $P_i^T(w)$ to be the average power in code $x_w(t)$ which is transmitted while the channel is in state c_i:

$$P_i^T(w) \triangleq \frac{1}{T_i} \int_0^T |x_w(t)|^2 1[S_t = c_i] dt, \tag{3.13}$$

where $1[\cdot]$ is the indicator function and $w \in (1, \ldots, \lfloor 2^{RT} \rfloor)$. We then have

$$\begin{aligned}
RT &= H(W) \\
&= H(W|Y) + I(W;Y) \\
&\leq H(W|Y) + I(X;Y) \\
&\overset{a}{\leq} 1 + \epsilon_T RT + I(X;Y) \\
&\overset{b}{\leq} 1 + \epsilon_T RT + \sum_i I(X_i;Y_i) \\
&\overset{c}{=} 1 + \epsilon_T RT + \sum_i E_w I(X_i;Y_i|P_i^T(w)) \\
&\overset{d}{\leq} 1 + \epsilon_T RT + \sum_i E_w C_i(P_i^T(w)) T_i \\
&\overset{e}{\leq} 1 + \epsilon_T RT + \sum_i C_i(E_w[P_i^T(w)]) T_i,
\end{aligned} \tag{3.14}$$

where a follows from Fano's inequality, b follows from the memoryless property of the T_{ij} blocks, c and d follow from the definitions of mutual information and capacity, respectively, and e follows from Jensen's inequality.

Define $\overline{P}^T = (\overline{P}_1^T, \ldots, \overline{P}_M^T)$, where $\overline{P}_i^T = E_w[P_i^T(w)]$. By construction, $\overline{P}_i^T$ satisfies the average power constraint on $[0,T]$. Let $T_n \to \infty$ be a subsequence such that $\overline{P}^{T_n}$ converges:

$$\overline{P}^{T_n} = (\overline{P}_1^{T_n}, \ldots, \overline{P}_M^{T_n}) \to (P_1^\infty, \ldots, P_M^\infty). \tag{3.15}$$

Since $\overline{P}^{T_n}$ satisfies the average power constraint, it follows that

$$\lim_{n \to \infty} \sum_i \frac{T_i^n}{T_n} \overline{P}_i^{T_n} = \sum_i \pi_i P_i^\infty \leq P, \tag{3.16}$$

where the superscript n denotes the dependence of T_i on T_n. Dividing (3.14) by T, we get

$$R \leq \frac{1}{T} + \epsilon_T R + \sum_i \frac{T_i}{T} C_i(\overline{P}_i^T). \tag{3.17}$$

Taking the limit of the right-hand side of (3.17) along the subsequence T_n, we have

$$R \leq \lim_{n\to\infty} \sum_i \frac{T_i^n}{T_n} C_i(\overline{P}_i^{T_n}) = \sum_i \pi_i C_i(P_i^\infty) \leq C, \tag{3.18}$$

where the last inequality follows from (3.16) and the definition of C.

3.2 Water-Filling in Time and Frequency

In this section we use the capacity results of §3.1 to determine the optimal input spectrum for the time-varying impulse response channel of §2.4.2, where the channel variation is stationary and ergodic. The system model, shown in Figure 3.3, operates as follows. The receiver estimates the channel impulse response (perfectly in zero time) and feeds this information back to the transmitter. The transmitter then uses the multiplexing coding technique to adapt its output to the changing channel state.

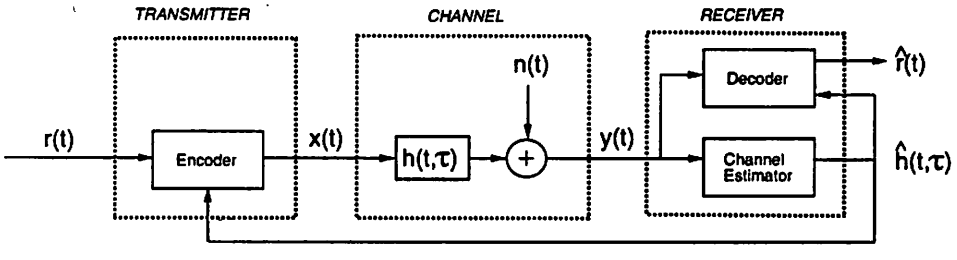

Figure 3.3: System Model for Impulse Response Channels.

An example of $h(t,\tau)$, given by (2.48), is plotted in Figure 3.4. The impulse response is constant for some random time period τ_1, at which point the channel state changes. The channel remains in the new state for the random time $\tau_2 - \tau_1$, then changes again, and so forth. The statistics of the transition times τ_i are determined by the transition probabilities of S_τ. Since we assume that channel has a finite number of transitions in a finite time interval, $\tau_i - \tau_{i-1}$ is strictly positive. In addition, within the dwell time $\tau_i - \tau_{i-1}$ the channel is assumed to be a memoryless block.

Taking the Fourier transform of $h(t,\tau)$ with respect to t yields

$$H(f,\tau) = \int_{-\infty}^{\infty} \left[\sum_{c\in C} h_c(t-\tau) 1[S_\tau = c]\right] e^{-j2\pi f t} dt = \sum_{c\in C} H_c(f) e^{-j2\pi f \tau} 1[S_\tau = c], \tag{3.19}$$

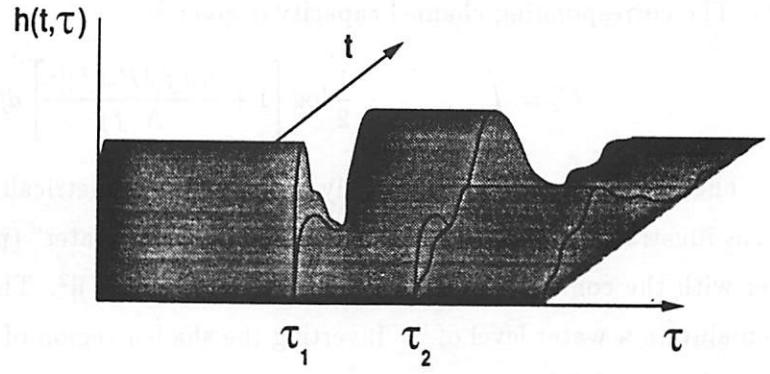

Figure 3.4: Time-Varying Impulse Response.

where $H_c(f)$ is the Fourier transform of $h_c(t)$. We plot $|H(f,\tau)|$ in Figure 3.5.

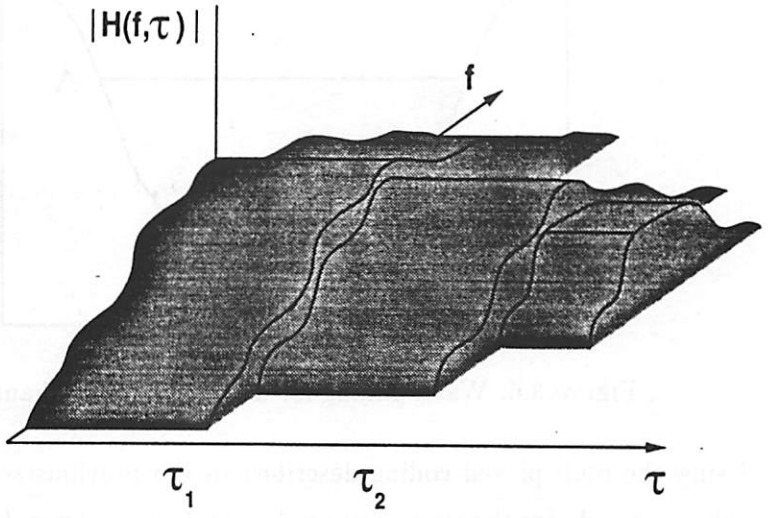

Figure 3.5: Fourier Transform of $h(t,\tau)$ Relative to t.

From Gallager [40], the capacity-achieving code of power P for a time-invariant additive Gaussian noise channel with impulse response $h_c(t)$ is a zero mean Gaussian stochastic process with spectrum

$$S_c(f) = \left[\Lambda - \frac{N_c(f)}{|H_c(f)|^2}\right]^+, \tag{3.20}$$

where $N_c(f)$ is the spectrum of the additive noise and Λ is chosen such that $S_c(f)$ does not exceed the power constraint. Since the noise $N_c(f)$ is usually introduced at the receiver, we will assume that its spectrum is the same for all c and denote this common spectrum

by $N(f)$. The corresponding channel capacity is given by

$$C_c = \int_{\frac{|H_c(f)|^2}{N(f)} \geq \frac{1}{\Lambda}} \frac{1}{2} \log \left[1 + \frac{S_c(f)|H_c(f)|^2}{N(f)} \right] df. \tag{3.21}$$

The spectrum $S_c(f)$ is generally interpreted geometrically using a *water-filling* analogy, as illustrated in Figure 3.6. A fixed amount P of "water" (power) is poured into a container with the container bottom defined by $N(f)/|H_c(f)|^2$. The water will distribute itself to maintain a water level of Λ. Inverting the shaded region of Figure 3.6 then yields the shape of the optimal input power spectrum.

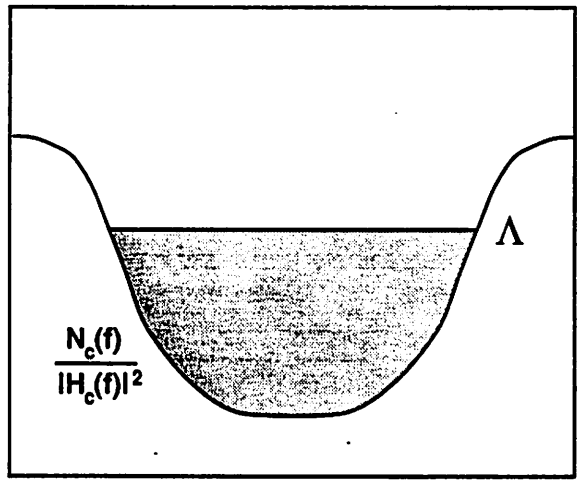

Figure 3.6: Water-Filling for Time-Invariant Channels.

Using the multiplexed coding described in the previous section, we see that the capacity-achieving code for the time-varying channel has spectrum $S_c(f)$ when the channel is in state c. Thus, the capacity-achieving code for the time-varying channel, $S(f,\tau)$, is the unique solution to the equation set

$$\begin{aligned} S(f,\tau) &= \left[\Lambda - \frac{N(f)}{|H(f,\tau)|^2} \right]^+, \\ P &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \int_{-\infty}^{\infty} S(f,\tau) df d\tau. \end{aligned} \tag{3.22}$$

The spectrum of the capacity-achieving code for the time-varying channel $h(t,\tau)$ can be interpreted geometrically as a water-filling in time and frequency. Specifically, the total input power P over all time and frequency is given by the shaded region under plane Λ in Figure 3.7. If we adjust the height of Λ such that the average power constraint is satisfied,

then an input power spectrum at time τ_0 and frequency f_0 of $\Lambda - S(f_0, \tau_0)$ achieves the time-varying channel capacity. This can be interpreted as water-filling in two dimensions, since $S(f, \tau)$ now defines the container bottom, and water is poured into the two dimensional container such that the time-average power equals P. Assuming that the region is connected, the water will distribute itself in such a way as to achieve capacity.

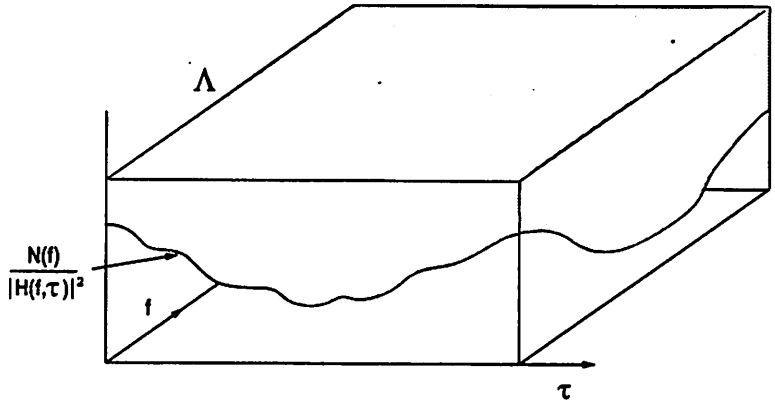

Figure 3.7: Water-Filling in Time and Frequency.

3.3 Power Control for Narrowband Fading Channels

When the transmitted signal is narrowband, multipath fading introduces a time-varying power gain $G(t)$, as described in §2.2.2. The system model for this case is shown in Figure 3.8, where we assume that $n(t)$ is AWGN. In this section we show that based on the results from §3.1, a policy which adapts the data rate, coding scheme, and power at the transmitter achieves the maximum zero-error spectral efficiency on a narrowband fading AWGN channel, where spectral efficiency is defined as the data rate per unit of bandwidth for a fixed error rate. We also compare the spectral efficiency of this optimal policy to that of two other policies which adapt only the transmit power. In order to compare these different power control schemes, we assume that the encoder output has unity average power, which is then multiplied by the power control value $P(t)$. The power control is subject to the average power constraint

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T P(t)dt \leq P. \tag{3.23}$$

We now determine the power control and coding policy which maximizes the spectral efficiency for this system.

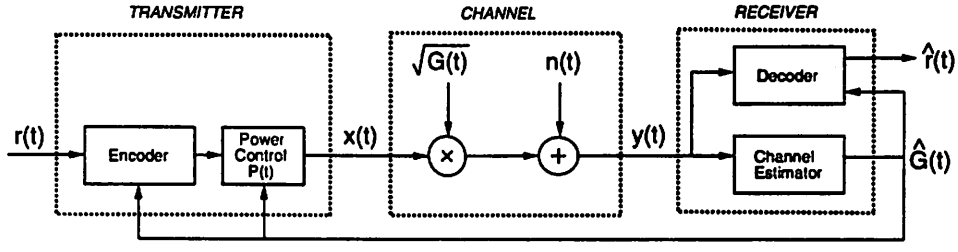

Figure 3.8: System Model for Narrowband Fading Channels.

3.3.1 Maximum Spectral Efficiency

The capacity of a time-invariant, bandlimited, AWGN channel with bandwidth B, gain G and power P in bits per second is [41, 42]

$$C = B\log_2\left[1 + \frac{PG}{N_oB}\right], \tag{3.24}$$

where $\gamma \triangleq PG/N_oB$ is the signal-to-noise ratio (SNR). If the channel gain G is time-varying, then the *instantaneous* SNR for constant transmit power P is $\gamma(t) \triangleq PG(t)/N_oB$. Since we assume perfect channel estimation in zero time, the transmitter knows $\gamma(t)$ at time t, and can adjust its power and code accordingly. Let $P(\gamma)$ denote the transmit power averaged over all times t such that $\gamma(t) = \gamma$, and let $\pi(\gamma)$ denote the distribution for γ, which is determined by the fading statistics (e.g. Rayleigh, log-normal, etc.).

Combining (3.7) and (3.24), we see that the maximum zero-error spectral efficiency for an AWGN channel with time-varying SNR $\gamma(t)$ and average power P is

$$\frac{C}{B} = \max_{P(\gamma)} \int \log_2\left[1 + \frac{P(\gamma)}{P}\gamma\right]\pi(\gamma)d\gamma, \tag{3.25}$$

where $P(\gamma)$ is subject to the power constraint

$$\int P(\gamma)\pi(\gamma)d\gamma = P. \tag{3.26}$$

Using Lagrange multipliers, it can be shown that the power control policy maximizing (3.25) is

$$\frac{P(\gamma)}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma} & \gamma \geq \gamma_0 \\ 0 & \gamma < \gamma_0 \end{cases} \tag{3.27}$$

for some "cutoff" value γ_0. If the received signal power is below this level, then no power is allocated to data transmission, so the *outage probability* for this policy is $p(\gamma < \gamma_0)$. This power control policy is depicted in Figure 3.9. Since γ is a function of $G(t)$, the maximizing power control policy is a "water-filling" formula in time that depends on the fading statistics only through the cutoff value γ_0.

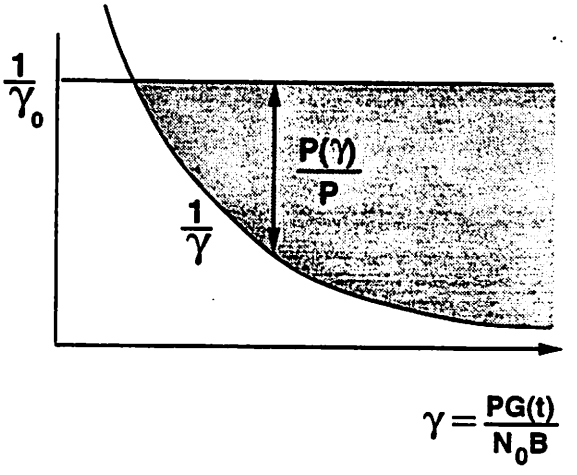

Figure 3.9: Optimal Power Control Policy.

Substituting (3.27) into (3.26), we can determine γ_0 by numerically solving

$$\int_{\gamma_0}^{\infty} \left(\frac{1}{\gamma_0} - \frac{1}{\gamma}\right) \pi(\gamma) d\gamma = 1. \tag{3.28}$$

Once γ_0 is known, we substitute (3.27) into (3.25) to get

$$\frac{C}{B} = \int_{\gamma_0}^{\infty} \log_2 \left(\frac{\gamma}{\gamma_0}\right) \pi(\gamma) d\gamma. \tag{3.29}$$

Equation (3.29) gives the maximum zero-error spectral efficiency of the narrowband fading channel, with no constraint on the delay or complexity of the coding strategy. Although the multiplexing code strategy of §3.2 suggests that the decoder delay is a random variable with distribution determined by the fading statistics, in §3.5 we develop a coding scheme with rates approaching the capacity limit where the decoder delay is fixed and independent of the fading correlation. We now consider constant power control policies, which avoid the use of variable-rate codes.

3.3.2 Constant Power Policies

In the previous section we derived a policy for maximizing spectral efficiency which adapts three parameters relative to the channel variations: the transmit power, data rate, and coding scheme. In this section we consider two policies which adapt only the transmit power to maintain a constant SNR at the receiver. Thus, the transmit power exactly compensates for the signal fading, as illustrated in Figure 3.10. Specifically, the constant power control policy is

$$P(\gamma)/P = P_R/\gamma, \tag{3.30}$$

where P_R equals the received signal-to-noise ratio. The channel then appears as a time-invariant AWGN channel with SNR = P_R. The constant P_R is determined by the transmit power constraint (3.26):

$$\int \frac{P_R}{\gamma}\pi(\gamma) = 1 \quad \Longrightarrow \quad P_R = \frac{1}{\mathrm{E}[1/\gamma]}. \tag{3.31}$$

The spectral efficiency with optimal coding for this policy (C_{cp}) is derived from the capacity of an AWGN channel with receive power P_R:

$$\frac{C_{cp}}{B} = \log_2[1 + P_R] = \log_2\left[1 + \frac{1}{\mathrm{E}[1/\gamma]}\right]. \tag{3.32}$$

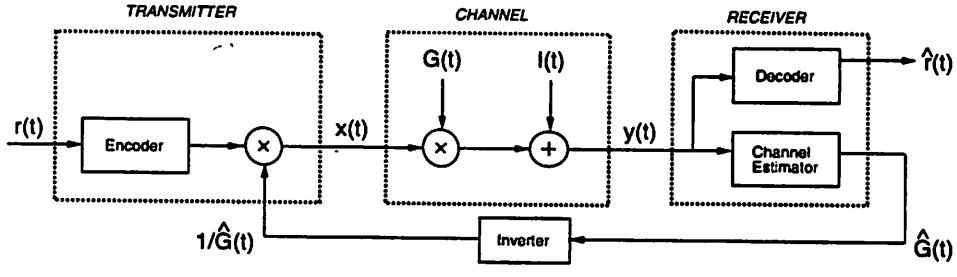

Figure 3.10: System Model for Constant Power Policy.

In severe fading conditions, the constant power policy of (3.30) allocates most of its power to compensate for deep fades. We therefore modify this policy to compensate for fading above a certain cutoff fade depth γ_0:

$$\frac{P(\gamma)}{P} = \begin{cases} \frac{P_R}{\gamma} & \gamma \geq \gamma_0 \\ 0 & \gamma < \gamma_0 \end{cases}. \tag{3.33}$$

Since the channel is only used when $\gamma \geq \gamma_0$, the power constraint (3.26) yields $P_R = 1/\mathrm{E}_{\gamma_0}[1/\gamma]$, where

$$\mathrm{E}_{\gamma_0}[1/\gamma] \triangleq \int_{\gamma_0}^{\infty} \frac{1}{\gamma}\pi(\gamma)d\gamma. \quad (3.34)$$

The spectral efficiency with this modified policy is then

$$\frac{C_{mcp}}{B} = \log_2\left[1 + \frac{1}{\mathrm{E}_{\gamma_0}[1/\gamma]}\right] p(\gamma \geq \gamma_0), \quad (3.35)$$

where $p(\gamma \geq \gamma_0) = \int_{\gamma_0}^{\infty} \pi(\gamma)d\gamma$. To get the maximum efficiency for the modified constant power policy, we must maximize (3.35) relative to γ_0. Alternatively, we can specify a particular outage probability p_{out}, and determine the cutoff γ_0 which satisfies $p(\gamma < \gamma_0) = p_{out}$.

3.3.3 Numerical Results

We now evaluate the spectral efficiency and outage probability of the power control policies in §3.3.1 – 3.3.2 for both log-normal and Rayleigh narrowband fading channels. Figure 3.11 shows the spectral efficiencies in log-normal fading with $\sigma = 8$dB for the optimal (3.29), constant power (3.30), and modified constant power (3.33) control policies, respectively. For the modified policy, we calculate the efficiency under two different criterion for the cutoff value γ_0: the value that maximizes the spectral efficiency for this policy, and the value that achieves the same outage probability as the optimal policy.

For Rayleigh fading, $\mathrm{E}[1/\gamma]$ is infinite. Thus, the spectral efficiency with the constant power policy is zero. Figure 3.12 shows the spectral efficiency of the other two policies in Rayleigh fading. There are two observations worth noting in these figures. First, the spectral efficiency of the modified constant power policy is close to the optimal policy's efficiency in both types of fading. Second, for log-normal fading at high SNRs, the constant power policies (3.30) and (3.33) perform almost the same.

The outage probability of the optimal and modified constant power policies is shown in Figures 3.13 for both log-normal and Rayleigh fading. The cutoff parameter for these calculations is the value which maximizes the spectral efficiency. In both types of fading, the outage probability with the optimal policy decreases exponentially. However, for the modified constant power policy, the outage probability becomes asymptotically constant for large SNRs. This behavior is explained by the cutoff values for each policy, which we plot in Figure 3.14 for both types of fading. The cutoff values for the optimal policy increase

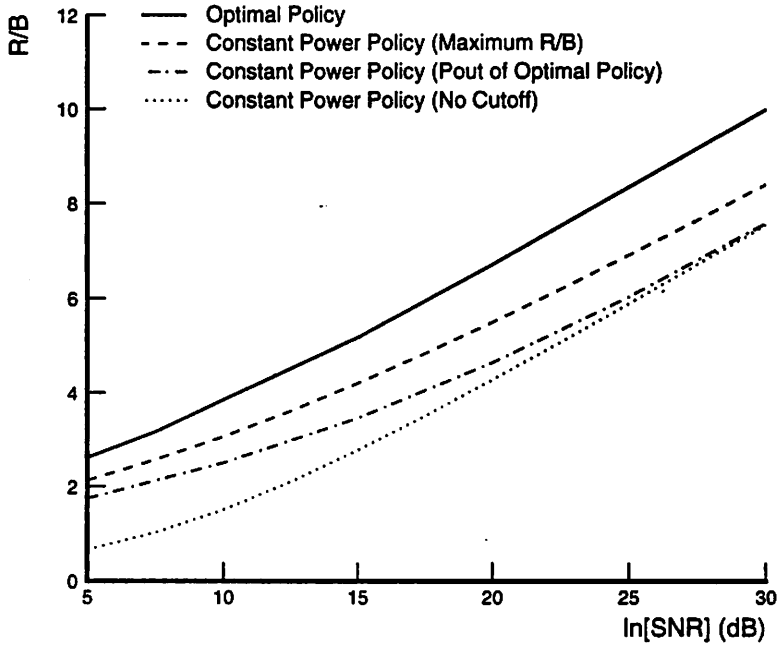

Figure 3.11: Spectral Efficiency in Log-Normal Fading.

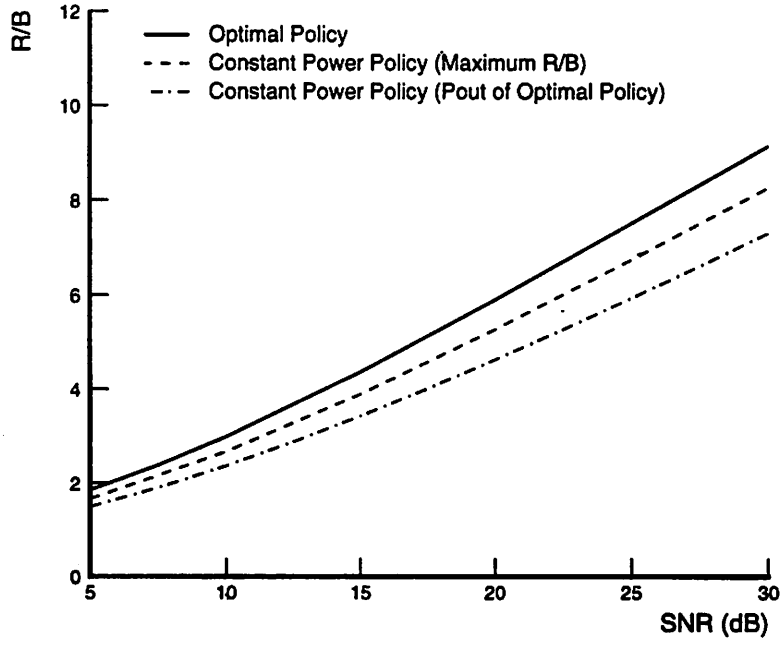

Figure 3.12: Spectral Efficiency in Rayleigh Fading.

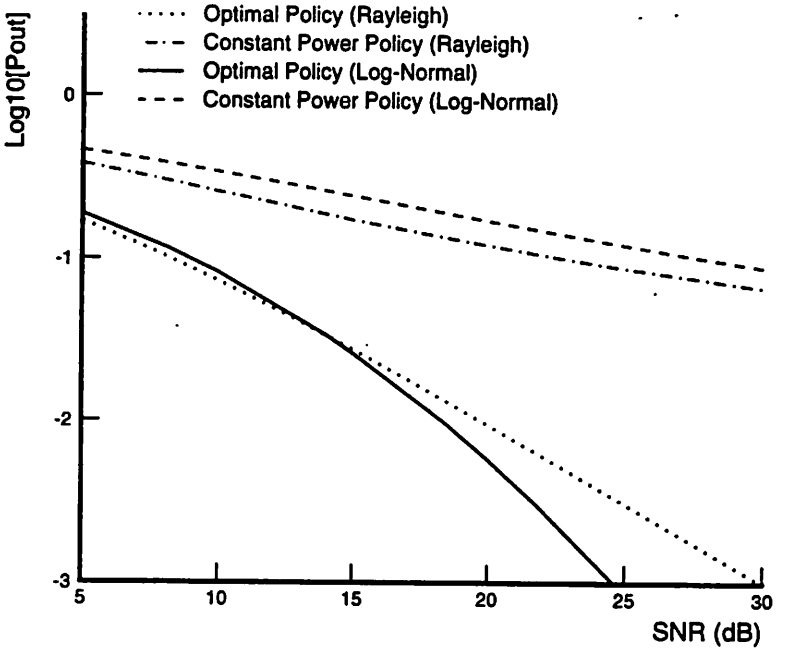

Figure 3.13: Outage Probability.

very slowly with SNR; thus, at high SNRs the probability of falling below the cutoff value is small. On the other hand, the cutoff values of the modified policy increase exponentially. Although the higher SNR increases the average value of γ, since the cutoff values are increasing proportionally, the outage probability remains approximately constant.

3.4 Uncoded Narrowband Modulation: Variable Rate M-QAM

The spectral efficiency calculated in 3.3.1 placed no constraints on the complexity or delay of the channel codes. We now consider spectral efficiency of uncoded M-QAM modulation with ideal Nyquist data pulses (sinc$[t/T]$). We will see that the spectral efficiency in this case depends on the allowable bit error rate (BER), and the policy to maximize efficiency adjusts the transmit power and the number of signal points in the M-QAM constellation. We also derive the maximum possible coding gain for M-QAM as a function of BER.

It has been shown [9, 43] that the BER for uncoded M-QAM can be approximated

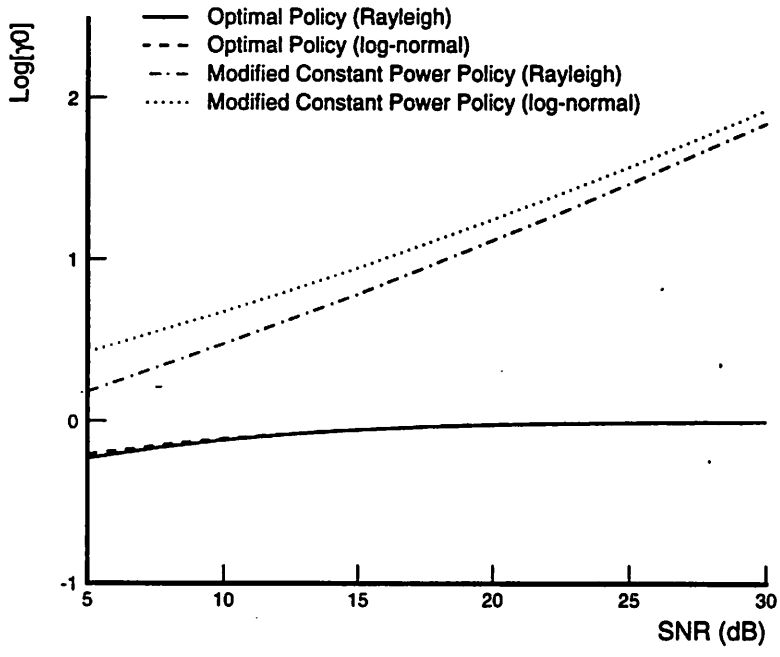

Figure 3.14: Cutoff Values.

by

$$\text{BER} \approx 2e^{-1.5(\overline{\gamma}/M-1)}, \tag{3.36}$$

where M is the number of M-QAM constellation points and $\overline{\gamma} = P/N_0B$ is the average received SNR. Let R denote the bit rate, and T denote the duration of each M-QAM symbol. The number of data bits per symbol is $\log_2 M$, and since the M-QAM pulses are Nyquist ($B = 1/T$), the spectral efficiency (R/B) is $\log_2 M/BT = \log_2 M$.

Let the fading parameter γ be as in §3.3.1, and let the power control policy adjust the transmit power to $P(\gamma)$ for fade level γ. For a fixed M, this increases the distance between points in the signal constellation, thereby reducing the BER. Specifically, the instantaneous BER is given by

$$\text{BER}(\gamma) \approx 2\exp\left[\frac{-1.5\gamma}{M-1}\frac{P(\gamma)}{P}\right]. \tag{3.37}$$

Suppose that in addition to adjusting the transmit power, we also adjust M to maintain a constant BER. Equivalently, we increase the size of the signal constellation while leaving the distance between points fixed. We can then rearrange (3.37) to get M in terms of the

BER, γ, P, and $P(\gamma)$:

$$M(\gamma) = 1 + \frac{1.5\gamma}{-\log(\text{BER}/2)}\frac{P(\gamma)}{P}. \tag{3.38}$$

The values of M may now be continuous. Constellations which transmit a non-integer number of bits per symbol are discussed in [44]. If we restrict the M-QAM to be a square constellation, then the following analysis, which assumes no restriction on M, yields optimistic results.

3.4.1 Maximum Spectral Efficiency

To maximize the spectral efficiency, we want to maximize

$$E[\log_2 M] = \int \log_2\left(1 + \frac{1.5\gamma}{-\log(\text{BER}/2)}\frac{P(\gamma)}{P}\right)\pi(\gamma)d\gamma, \tag{3.39}$$

subject to the power constraint

$$\int P(\gamma)\pi(\gamma)d\gamma = P. \tag{3.40}$$

The power control policy that maximizes (3.39) can be found using Lagrange multipliers, and is similar in form to (3.27):

$$\frac{P(\gamma)}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma \cdot K} & \gamma \geq \gamma_0/K \\ 0 & \gamma < \gamma_0/K \end{cases}, \tag{3.41}$$

where γ_0 is the cutoff fade depth, and

$$K \triangleq \frac{-1.5}{\log(\text{BER}/2)}. \tag{3.42}$$

Define $\gamma_K \triangleq \gamma_0/K$. Substituting (3.41) into (3.39), we get the spectral efficiency

$$\frac{R}{B} = \int_{\gamma_K}^{\infty} \log_2\left(\frac{\gamma}{\gamma_K}\right)\pi(\gamma)d\gamma. \tag{3.43}$$

By substituting (3.41) into (3.40), we can rewrite the power constraint in terms of γ_K:

$$\int_{\gamma_K}^{\infty}\left(\frac{1}{\gamma_K} - \frac{1}{\gamma}\right)\pi(\gamma)d\gamma = K. \tag{3.44}$$

The maximum spectral efficiency of uncoded M-QAM is the maximum of (3.43) subject to the constraint (3.44). This maximization problem is identical to the maximum efficiency with coding, defined by (3.29) and (3.28), with the transmit power constraint

reduced by K. Thus, there is a simple relationship between the spectral efficiencies of optimal coding and uncoded M-QAM modulation: uncoded M-QAM has an effective power reduction of K relative to optimal coding. Equivalently, K is the maximum possible coding gain for M-QAM. Equation (3.43) is the spectral efficiency with optimal power and rate adaptation; we next consider the constant power policy, which only adapts transmit power.

3.4.2 Constant Power Policies

If we again define the constant power policy as in (3.30), then the average power constraint requires that $P_R = 1/\mathrm{E}[1/\gamma]$. Substituting (3.30) into (3.39), we get the maximum spectral efficiency of the constant power policy:

$$\frac{R}{B} = \log_2\left(1 + \frac{-1.5}{\log(\mathrm{BER}/2)\mathrm{E}[1/\gamma]}\right). \tag{3.45}$$

Similarly, using the modified policy (3.33) and the corresponding power constraint $P_R = \mathrm{E}_{\gamma_0}[1/\gamma]$, the spectral efficiency (3.39) becomes

$$\frac{R}{B} = \log_2\left[1 + \frac{-1.5}{\log(\mathrm{BER}/2)\mathrm{E}_{\gamma_0}[1/\gamma]}\right] p(\gamma \geq \gamma_0). \tag{3.46}$$

Thus, the maximum spectral efficiency with the modified policy is (3.46), maximized relative to γ_0. As in the optimal coding case, we can also set γ_0 relative to a desired outage probability.

3.4.3 Numerical Results

We now evaluate the spectral efficiency and outage probability of these policies, and compare them with the coded cases in §3.3.3. Figure 3.15 shows the spectral efficiencies of the optimal (3.29), constant power (3.30), and modified constant power (3.33) control policies, respectively, for log-normal fading with σ = 8dB and a BER of 10^{-3}. For the modified policy, we determine γ_0 based on one of two criterion: maximizing the spectral efficiency, or matching the outage probability to that of the optimal policy. The efficiency of the constant power policy in Rayleigh fading is zero; in Figure 3.16 we plot the efficiency of the other two policies in Rayleigh fading. Figures 3.17 and 3.18 compare the efficiencies for the coded and uncoded cases. In these figures, the modified policy performance is derived for the γ_0 which maximizes spectral efficiency. Figures 3.19 and 3.20 compare the corresponding outage probabilities.

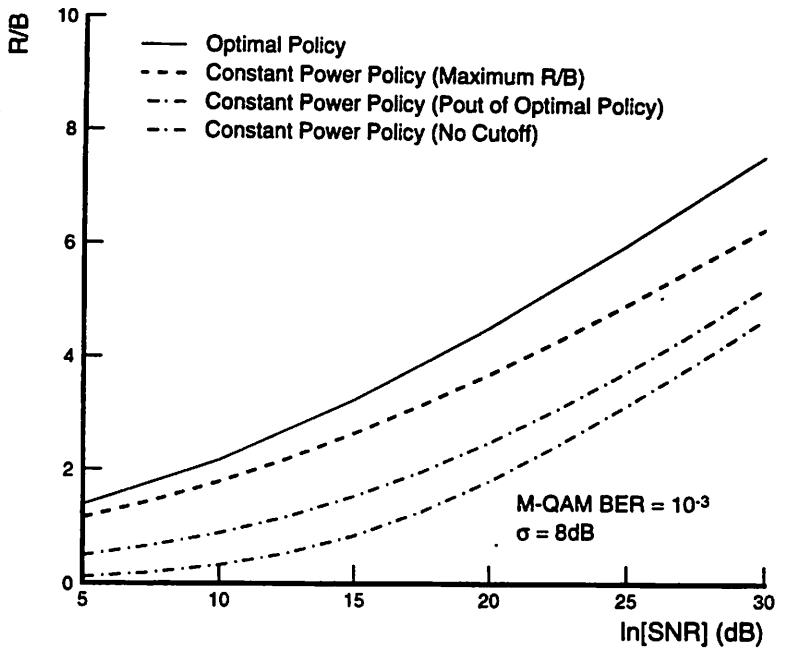

Figure 3.15: Efficiency of Uncoded M-QAM in Log-Normal Fading.

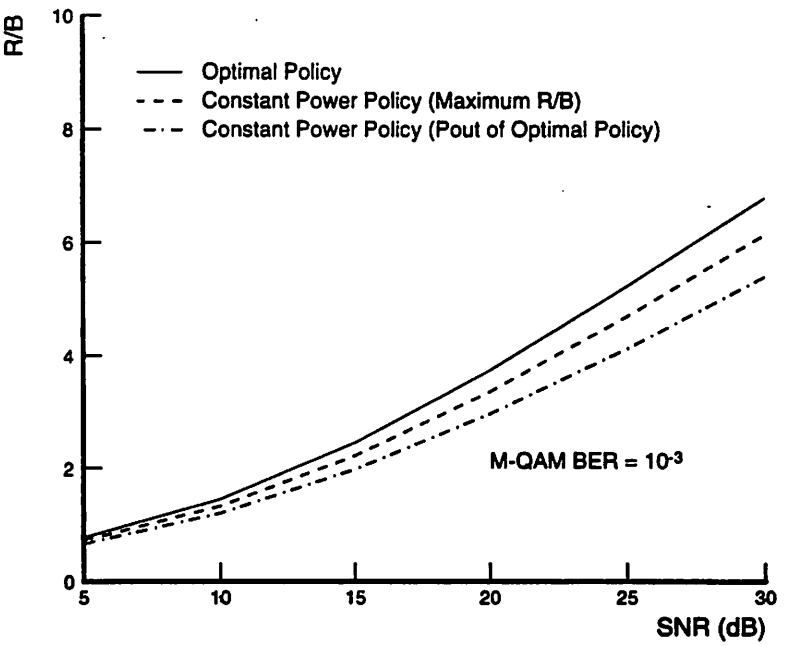

Figure 3.16: Efficiency of Uncoded M-QAM in Rayleigh Fading.

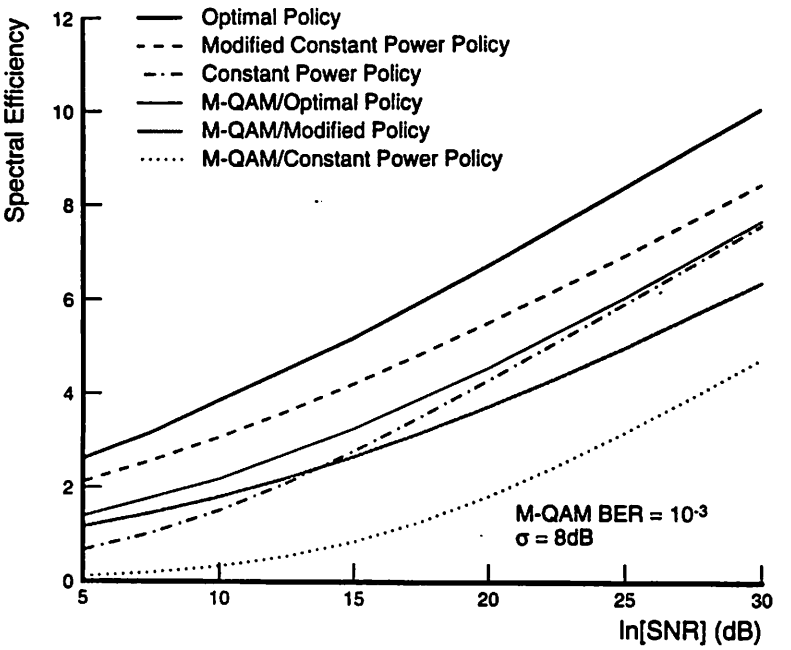

Figure 3.17: Coded and Uncoded Cases in Log-Normal Fading.

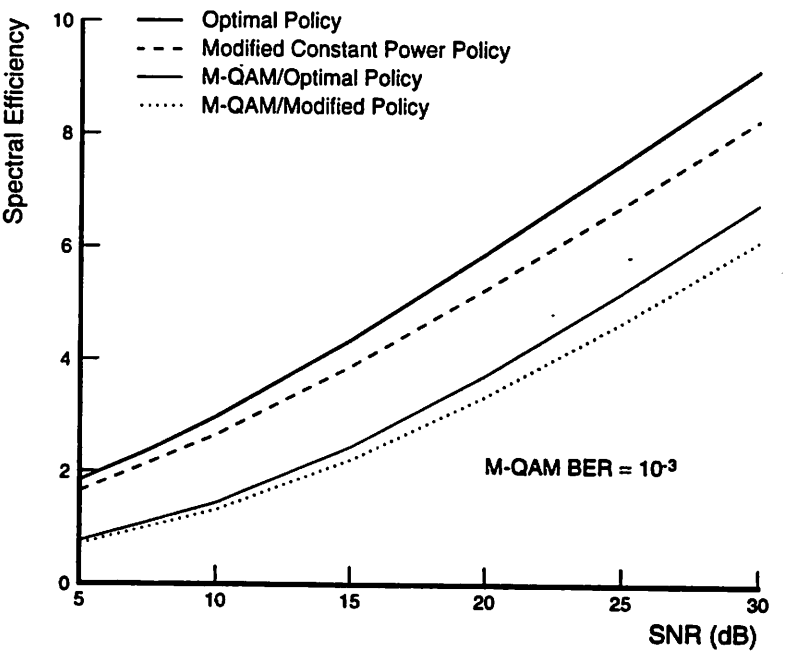

Figure 3.18: Coded and Uncoded Cases in Rayleigh Fading.

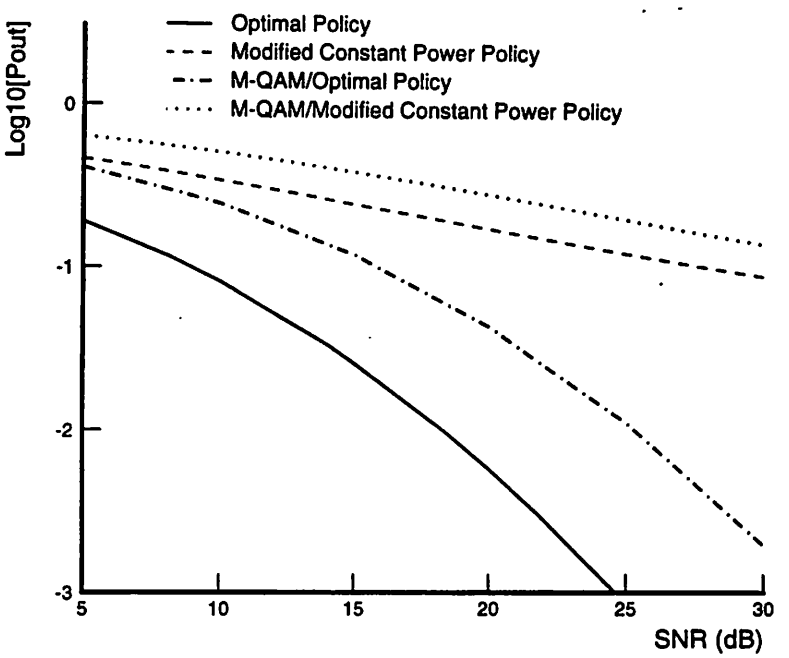

Figure 3.19: Outage Probabilities in Log-Normal Fading.

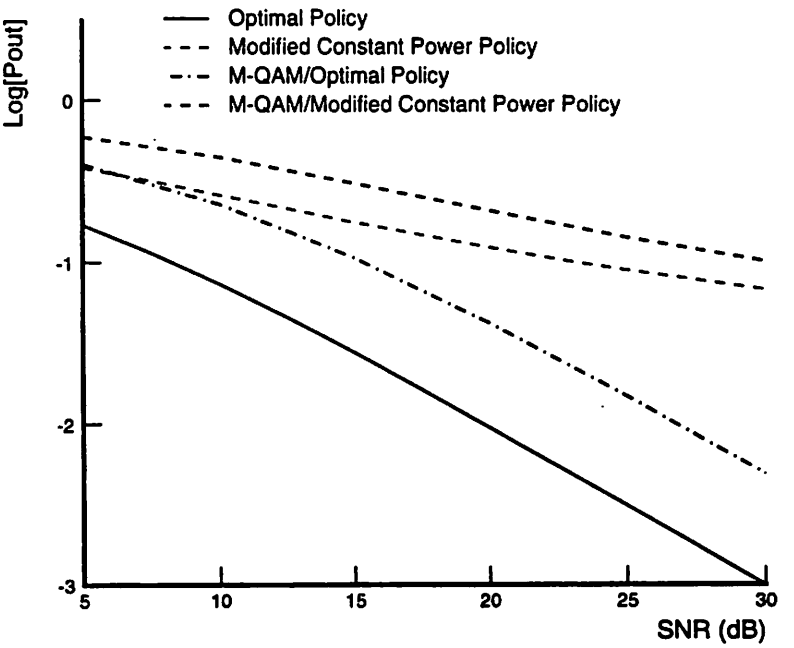

Figure 3.20: Outage Probabilities in Rayleigh Fading.

In Figure 3.21 we plot K, the maximum possible coding gain for uncoded M-QAM, as a function of BER. In the next section, we will discuss practical coding techniques to achieve some of this gain.

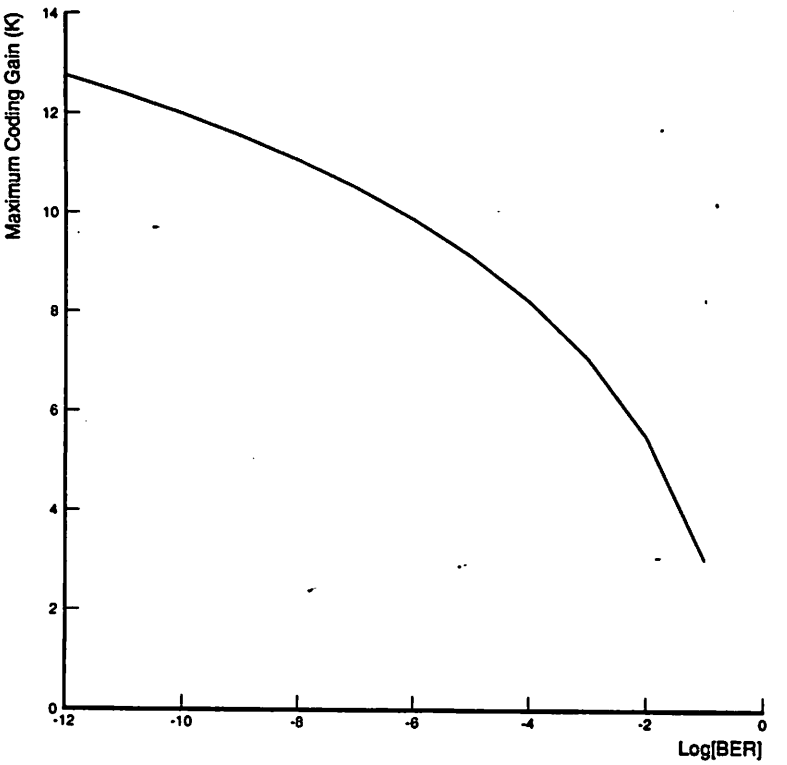

Figure 3.21: Maximum Coding Gain.

3.5 Coding

The coding strategy required to achieve the gain predicted in Figure 3.21 uses multiplexing of the capacity-achieving codes for the AWGN channel with bandwidth B and SNR γ. However, these capacity-achieving codes give little insight into practical code design [41]. We now discuss some of the recent advances in bandwidth-efficient coding for time-invariant channels. In particular, we review the basic ideas behind coded modulation, where the source and channel coding schemes are jointly optimized. We then propose an adaptive variable-rate coded-modulation technique for fading channels with estimation and transmitter feedback, and calculate the coding gain of this scheme relative to the uncoded

variable-rate M-QAM of the previous section.

3.5.1 Coded Modulation for Bandlimited AWGN Channels

Although Shannon proved the capacity theorem for AWGN channels in the late 1940s, it wasn't until recently that rates approaching the Shannon limit on bandlimited AWGN channels have been attained [44]. Shannon's theorem predicted the possibility of reducing both energy and bandwidth simultaneously through coding. However, traditional error-correction coding schemes, such as block and convolutional codes, reduce transmit power at the expense of increased bandwidth, since the added code bits increase the bit rate [45].

The spectrally-efficient coding breakthrough came when Ungerboeck [46] introduced a coded-modulation technique to jointly optimize both channel and source (modulation) coding. This joint optimization results in significant coding gains without bandwidth expansion. Ungerboeck's *trellis-coded* modulation, which uses multilevel/phase signal modulation and simple convolutional coding with mapping by set partitioning, has remained superior over more recent developments in coded modulation (coset and lattice codes), as well as more complex trellis codes [48]. We now outline the general principles of this coding technique. Comprehensive treatments of trellis, lattice, and coset codes can be found in [47, 44, 48], respectively.

The basic scheme for trellis and lattice coding, or more generally, any type of coset coding, is depicted in Figure 3.22. There are five elements required to generate the coded-modulation:

1. A conventional encoder E, block or convolutional, that operates on k uncoded data bits to produce $k + r$ coded bits.

2. A subset selector, which uses the coded bits to choose one of 2^{k+r} subsets from a partition of the N-dimensional signal constellation.

3. A point selector, which uses $n - k$ additional uncoded bits to choose one of the 2^{n-k} signal points in the selected subset.

4. A constellation map, which maps the selected point from N-dimensional space to a sequence of $N/2$ points in two-dimensional space.

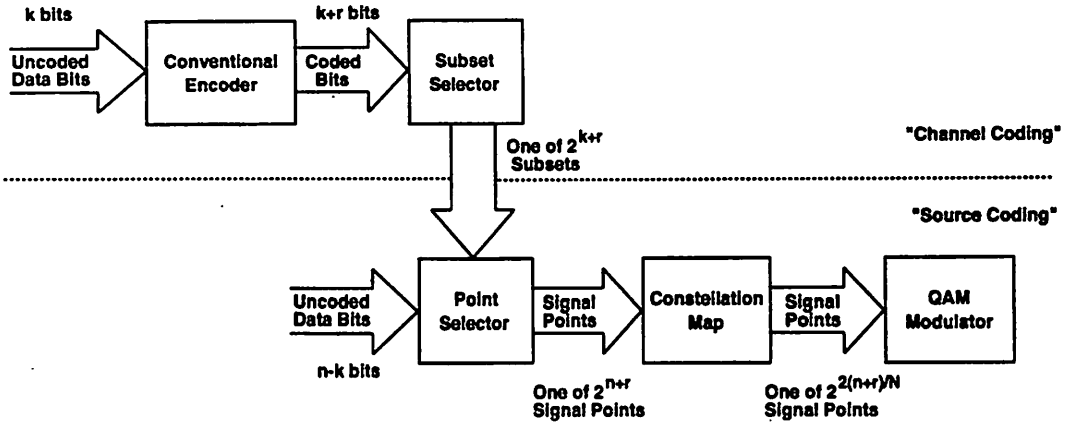

Figure 3.22: General Coding Scheme.

5. A QAM modulator.

The first two steps described above are referred to as *channel coding*, and the remaining steps are called *source coding* or modulation. The receiver essentially reverses the modulation and coding steps: after QAM demodulation and an inverse $2/N$ constellation mapping, decoding is done in essentially two stages: first, the points within each subset that are closest to the received signal point are determined; then, the maximum-likelihood subset sequence is calculated. When the encoder E is a convolutional encoder, this coded-modulation scheme is refered to as a trellis code; for E a block encoder, it is called a lattice (or block) code.

The steps described above essentially decouple the channel coding gain from the source (signal-shaping) gain. Specifically, the code distance properties, and thus the channel coding gain, are determined by the encoder (E) properties and the subset partitioning, which are essentially decoupled from the source coding. We will discuss the channel coding gain in more detail below. Optimal shaping of the signal constellation provides up to an additional 1.53 dB of shape gain (for asymptotically large N), independent of the channel coding scheme[3]. However, the performance improvement from shape gain is offset by the corresponding complexity of the constellation map, which grows exponentially with N. The size of the transmit constellation is determined by the average power constraint, and doesn't affect the source (or channel) coding gain.

The channel coding gain results from a selection of all possible sequences of signal

[3]A square constellation has 0dB of shape gain; a circular constellation, which is the geometrical figure with the least average energy for a given area, achieves the maximum shape gain for a given N [49].

points. If we consider a sequence of N input bits as a point in N-dimensional space (the *sequence space*), then this selection is used to guarantee some minimum distance d_{min} in the sequence space between possible input sequences. Errors generally occur when a sequence is mistaken for its closest neighbor, and in AWGN channels this error probability is a decreasing function of d_{min}^2. We can thus decrease the BER by increasing the separation between each point in the sequence space by a fixed amount ("stretching" the space). However, this will result in a proportional power increase, so no net coding gain is realized. The effective power gain of the channel code is, therefore, the minimum squared distance between allowable sequence points (the sequence points obtained through coding), multiplied by the density of the allowable sequence points. Specifically, if the minimum distance and density of points in the sequence space are denoted by d_0 and Δ_0, respectively, and if the minimum distance and density of points in the sequence space obtained through coding are denoted by d_{min} and Δ, respectively, then maximum-likelihood sequence detection yields a channel coding gain of

$$G_c = \left(\frac{d_{min}^2}{d_0^2}\right)\left(\frac{\Delta}{\Delta_0}\right). \tag{3.47}$$

The second bracketed term in this expression is also refered to as the *constellation expansion factor*, and equals 2^{-r} (per N dimensions) for a redundancy of r bits in the encoder E [48].

Some of the nominal coding gain in (3.47) is lost due to correct sequences having more than one nearest neighbor in the sequence space, which increases the possibility of incorrect sequence detection. This loss in coding gain is characterized by the *error coefficient*, which is tabulated for most common lattice and trellis codes in [48]. In general, the error coefficient is larger for lattice codes than for trellis codes with comparable values of G_c.

Channel coding is done using set partitioning of lattices. A *lattice* is a discrete set of vectors in real Euclidean N-space that forms a group under ordinary vector addition, so the sum or difference of any two vectors in the lattice is also in the lattice. A *sub-lattice* is a subset of a lattice that is itself a lattice. The sequence space for *uncoded* M-QAM modulation is just the N-cube[4], so the minimum distance between points is no different than in the two-dimensional case. By restricting input sequences to lie on a lattice in N-space that is denser than the N-cube, we can increase d_{min} while maintaining the same density (or equivalently, the same average power) in the transmit signal constellation; hence,

[4]The Cartesian product of two-dimensional rectangular lattices with points at odd integers.

there is no constellation expansion. The N-cube is a lattice, however for every $N > 1$ there are denser lattices in N-dimensional space. Finding the densest lattice in N dimensions is a well-known mathematical problem, and has been solved for all N for which the decoder complexity is manageable[5]. Once the densest lattice is known, we can form partioning subsets, or *cosets*, of the lattice through translation of any sublattice. The choice of the partitioning sublattice will determine the size of the partition, i.e. the number of subsets that the subset selector in Figure 3.22 has to choose from. Data bits are then conveyed in two ways: through the sequence of cosets from which constellation points are selected, and through the points selected within each coset. The density of the lattice determines the distance between points within a coset, while the distance between subset sequences is essentially determined by the binary code properties of the encoder E, and its redundancy r. If we let d_p denote the minimum distance between points within a coset, and d_s denote the minimum distance between the coset sequences, then the minimum distance code is $d_{min} = \min(d_p, d_s)$. The effective coding gain is given by

$$G_c = 2^{-2r/N} d_{min}^2, \tag{3.48}$$

where $2^{-2r/N}$ is the constellation expansion factor (in two dimensions) from the r extra bits introduced by the binary channel encoder.

Returning to Figure 3.22, suppose that we want to send $m = n + r$ bits per dimension, so an N sequence conveys mN bits. If we use the densest lattice in N space that lies within an N sphere, where the radius of the sphere is just large enough to enclose 2^{mN} points, then we achieve a total coding gain which combines the channel gain (resulting from the lattice density and the encoder properties) with the shape gain of the N sphere over the N rectangle. Clearly, the channel coding gain is decoupled from the shape gain. An increase in signal power would allow us to use a larger N sphere, and hence transmit more uncoded bits. We will use this idea in the next section to design a variable-rate coded-modulation technique for fading channels.

It is possible to generate maximum-density N-dimensional lattices for $N = 4$, 8, 16, and 24 using a simple partition of the two-dimensional rectangular lattice combined with either conventional block or convolutional coding. Details of this type of code construction, and the corresponding decoding algorithms, can be found in [44] for both lattice and trellis

[5]The complexity of the maximum-likelihood decoder implemented with the Viterbi algorithm is roughly proportional to N.

codes. For these constructions, an effective coding gain of approximately 1.5, 3.0, 4.5, and 6.0dB is obtained with lattice codes, for $N = 4$, 8, 16, and 24, respectively. Trellis codes exhibit higher coding gains with comparable complexity.

We conclude this section with an example of coded-modulation: the $N = 8$, 3dB gain lattice code proposed in [44]. First, the two-dimensional signal constellation is partitioned into four subsets as shown in Figure 3.23, where the subsets are represented by the points A_0, A_1, B_0, and B_1, respectively. From this subset partition, we form an 8-dimensional lattice by taking all sequences of four points in which all points are either A points or B points and moreover, within a four point sequence, the point subscripts satisfy the parity check $i_1 + i_2 + i_3 + i_4 = 0$ (so the sequence subscripts must be codewords in the (4,3) parity-check code, which has a minimum Hamming distance of two). Thus, three data bits and one parity check bit are used to determine the lattice subset. The minimum distance resulting from this subset partition is four times the minimum distance of the uncoded signal constellation, yielding a 6dB gain. However, the extra parity check bit expands the constellation by 3dB, so the net coding gain is $6 - 3 = 3$dB. The remaining data bits are used to choose a point within the selected subset, so for a data rate of m bits/symbol, the four lattice subsets must each have 2^{m-1} points[6].

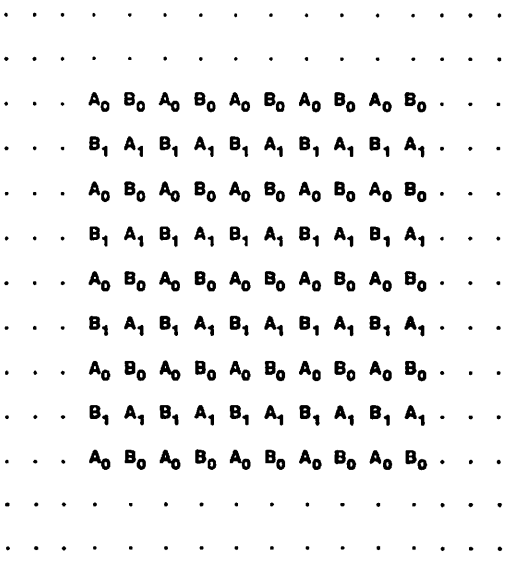

Figure 3.23: Subset Partition for an Eight-Dimensional Lattice.

[6]This yields $m - 1$ bits/symbol, with the additional bit/symbol conveyed by the channel code.

3.5.2 Variable-Rate Coded Modulation for Narrowband Fading Channels

We now propose a variable-rate coded-modulation technique which obtains some of the coding gain predicted by (3.42). We also calculate the spectral efficiency of this technique relative to the capacity limit and the uncoded case. The coded-modulation scheme is shown in Figure 3.24. The channel code design is the same as it would be for a time-invariant channel; thus, the lattice structure and conventional encoder are the same as those in Figure 3.22. From §3.5.1, the channel coding gain, G_c, is independent of the transmit signal constellation. We can therefore adjust the power and rate (number of levels or signal points) in the transmit constellation relative to the instantaneous SNR, as described in §3.4, without affecting the channel coding gain.

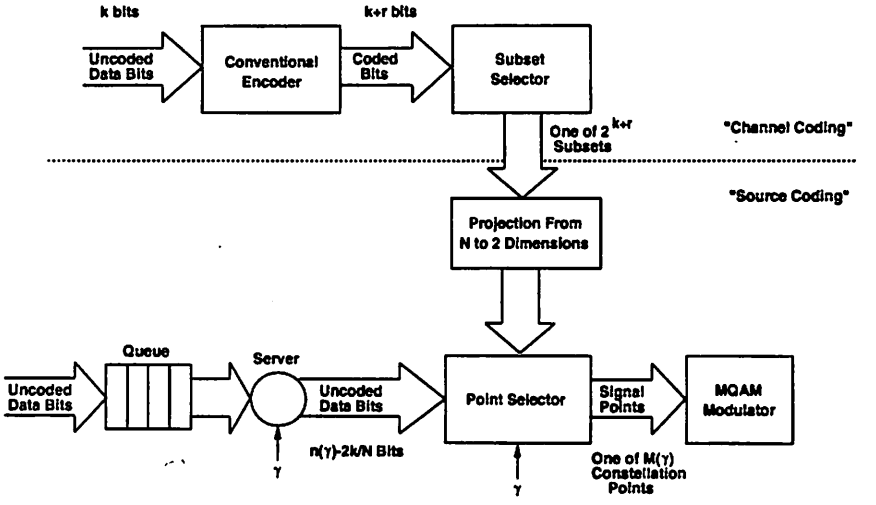

Figure 3.24: Variable-Rate Coded-Modulation Scheme.

The source coding (modulation) works as follows. The signal constellation is a square lattice with an adjustable number of constellation points M. Since we are using the N-cube for our signal constellation, the shape gain is zero. Therefore, we can move the constellation mapping before the point selection without changing the code performance, i.e., we project the chosen subset in N-dimensional space onto a sequence of $N/2$ subsets in two-dimensional space from which the M-QAM signal point is selected. The size of the M-QAM signal constellation is determined by the transmit power, which is adjusted relative to the instantaneous SNR and the desired BER, as in the uncoded case above.

Specifically, if the BER approximation (3.36) is adjusted for the coding gain, then

for a particular SNR=$\overline{\gamma}$,

$$\text{BER} \approx 2e^{-1.5(\overline{\gamma}G_c/M-1)}. \tag{3.49}$$

Therefore, the number of constellation points and the signal power can be adjusted relative to the instantaneous SNR to maintain a fixed BER:

$$M(\gamma) = 1 + \frac{1.5\gamma G_c}{-\log(\text{BER}/2)}\frac{P(\gamma)}{P}, \tag{3.50}$$

The number of uncoded bits required to select the coset point is $n(\gamma) - 2k/N = \log_2 M(\gamma) - 2(k+r)/N$. Since this value varies with time, these uncoded bits must be queued until needed, as shown in Figure 3.24.

The bit rate per transmission is $\log_2 M(\gamma)$, and the data rate is $\log_2 M(\gamma) - 2r/N$. Therefore, we maximize the data rate by maximizing $E[\log_2 M]$ relative to the power constraint (3.40). From this maximization, we obtain the optimal power control policy for this modulation scheme:

$$\frac{P(\gamma)}{P} = \begin{cases} \frac{1}{\gamma_0} - \frac{1}{\gamma \cdot K_c} & \gamma \geq \gamma_0/K_c \\ 0 & \gamma < \gamma_0/K_c \end{cases}, \tag{3.51}$$

where γ_0 is the cutoff fade depth, and $K_c = KG_c$. The optimal policy is the same "water-filling" as in the uncoded case, given by (3.44), with K replaced by K_c. Thus, the coded modulation increases the effective transmit power by G_c relative to the uncoded variable-rate M-QAM performance. The resulting spectral efficiency is

$$\frac{R}{B} = \int_{\gamma_{K_c}}^{\infty} \log_2\left(\frac{\gamma}{\gamma_{K_c}}\right)\pi(\gamma)d\gamma. \tag{3.52}$$

If the constellation expansion factor is not included in the coding gain G_c, then we must subtract $2r/N$ from (3.52) to get the data rate.

In Figure 3.25 we plot the spectral efficiency given by (3.52) in log-normal fading over a range of channel coding gains. Lattice codes which achieve these gains are described in [44]. For comparison we also plot the efficiency of uncoded modulation (3.43) and the capacity limit (3.29). From this figure, we see that a coding gain of 6dB comes reasonably close to the capacity limit, and the added complexity required to implement higher-gain channel codes is probably unwarranted for most applications. Figure 3.26 shows a similar comparison for Rayleigh fading, where 6dB of channel coding gain again yields close to optimal performance.

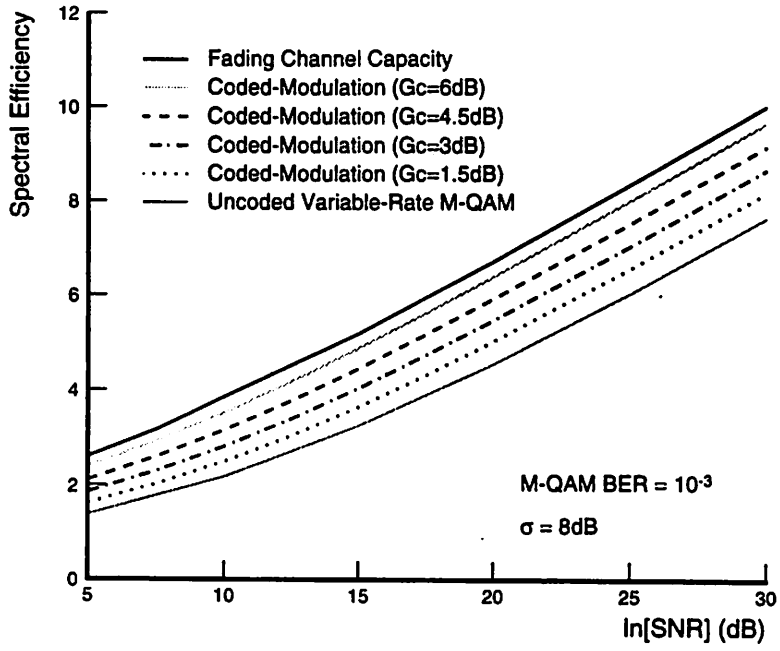

Figure 3.25: Efficiency in Log-Normal Fading with Variable-Rate Coding.

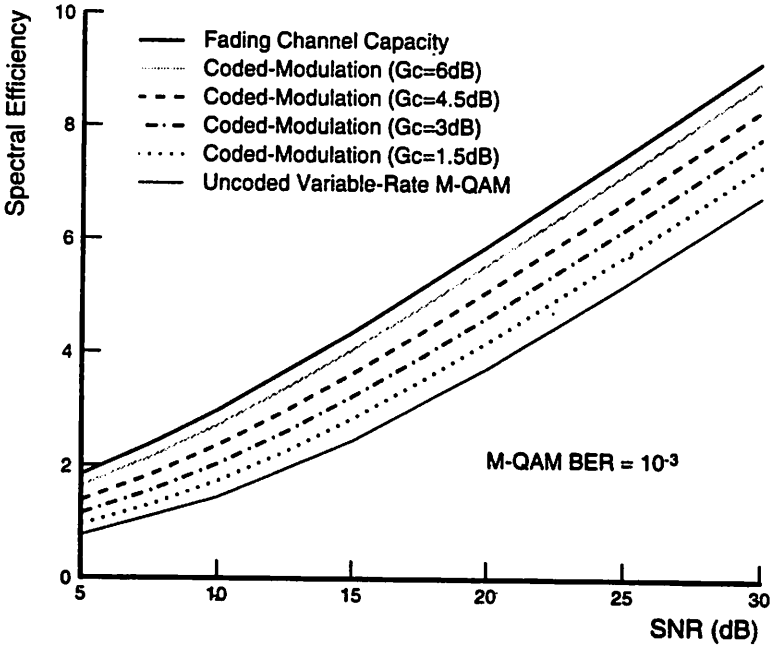

Figure 3.26: Efficiency in Rayleigh Fading with Variable-Rate Coding.

Most coded modulation techniques for fading channels do not assume channel state information at the receiver. Instead, they rely on built-in time diversity in the code to mitigate the effect of Rayleigh fading. Code designs of this type can be found in [50, 51]. Consider a built-in time diversity code of this type with coding gain G_c. The system BER with this code is determined by integrating (3.49) against the fading distribution of the SNR:

$$\text{BER} = \int 2e^{-1.5(\gamma G_c/M^2-1)}\pi(\gamma)d\gamma. \tag{3.53}$$

To calculate the spectral efficiency of built-in time diversity codes, we fix the BER and SNR, and determine the value of M which achieves this BER in (3.53). Figure 3.27 shows the resulting efficiency of time diversity codes with different coding gains, and compares their performance with that of the adaptive coded modulation. As expected, the adaptive technique is far superior. Thus, it appears that when channel state information is available at the transmitter, using this information for adaptive encoding yields a significant increase in system performance, as long as the additional complexity of adaptive constellation sizing is manageable.

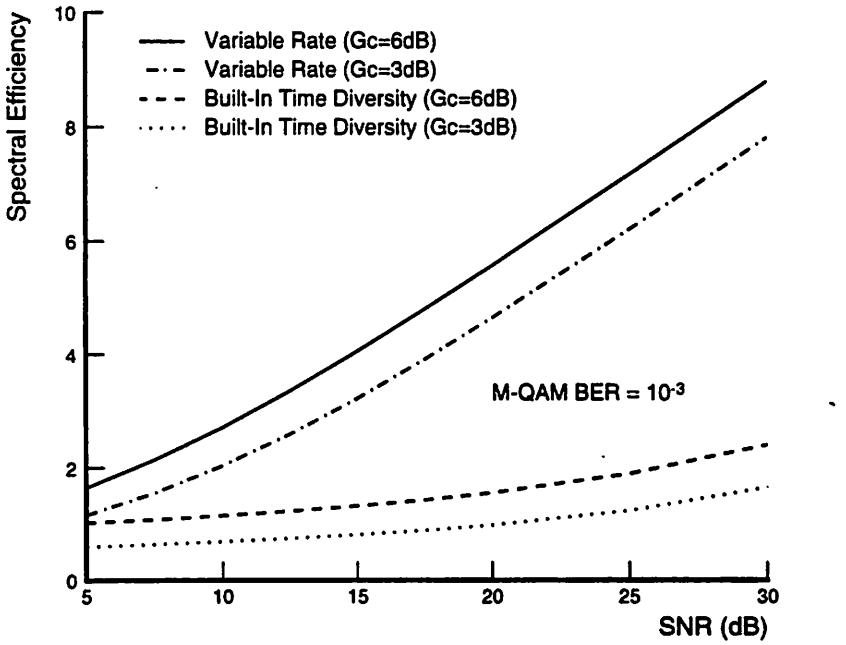

Figure 3.27: Variable-Rate and Time Diversity Codes in Rayleigh Fading.

3.6 Channel Estimation

We have assumed throughout this chapter that the channel variation is tracked perfectly at the receiver, and this information is sent to the transmitter via an error-free feedback path[7]. We now relax the assumption of perfect channel estimation, and study the impact of channel estimation errors. First we consider channel estimation in narrowband multipath channels with Rayleigh fading and log-normal shadowing. Since Rayleigh fading is usually too fast to measure accurately, we propose an estimation filter to minimize the dB error of the log-normal estimate while averaging out the Rayleigh fading. We analytically determine the statistics of the rms dB estimation error, and compute its value over a range of fading parameters. A more detailed study, which includes simulation of the estimation error, the effects of antenna diversity, and design of a fixed filter which is robust over a range of fading parameters can be found in [52].

Once the statistics of the estimation error is known, we can determine this error's effect on the power control, modulation, and coding techniques proposed in §§3.4 – 3.5. Specifically, we calculate the change in average power and data rate of our adaptive policies when the power estimate used for the adaptation is incorrect.

Although narrowband fading can be estimated concurrently with data detection, wideband channel variation is usually measured by sending a periodic training sequence known to both the transmitter and receiver [31]. Longer training sequences generally result in better channel estimates, but with a corresponding loss in data rate, since no data is transmitted during the training period. With respect to channel capacity, the periodic estimation is equivalent to turning the transmitter off periodically. This limits the input sequences that can be used for channel coding, thus reducing the channel capacity. We conclude this section by bounding this capacity loss for the periodically estimated, or *on-off* channel[8].

---

[7]We also assume that the feedback path has no delay. Delays in the feedback path will induce errors in the transmitter channel state information, and these errors will be proportional to the speed of the channel variation.

[8]The on-off channel can also be used to model a time division multiaccess system.

3.6.1 Optimal Filter for Power Estimation

The power control policies of §3.3 adapt the transmit power based on the instantaneous value of the channel fade level. We now consider estimation techniques for the received signal power. Assuming the statistical fading model of §2.2.2, the received power experiences two multiplicative forms of fading: rapid Rayleigh fading, and slower log-normal shadowing. In general the Rayleigh fluctuations are too quick to use for the power adaptation, so the goal is to track (and adapt to) shadow fading while averaging out the Rayleigh fading. Therefore, the low pass power measurement filter must be sufficiently narrowband to average out the Rayleigh fading, yet sufficiently wideband to track the shadow fading.

The received power $p(t)$ is given by the multiplicative form of (2.19), $p(t) = r(t)s(t)$, where r and s are, respectively, the Rayleigh and log-normal fade levels. We assume narrowband Rayleigh fading, so the multipath power has an exponential distribution. Since $\overline{p} = \overline{r}\,\overline{s}$, we can specify $\overline{r}$ to be one and use $\overline{s}$ to characterize the mean received power. The distribution of r is then

$$p(r) = e^{-r}; \quad r > 0. \tag{3.54}$$

Taking natural logs, we get that the distribution for $R \triangleq \log r$ is

$$p(R) = e^{R-e^R}. \tag{3.55}$$

We use natural logs throughout the analysis; to get results in dBs, we simply multiply the natural log results by $10/\log 10$.

From (2.26), the autocorrelation of r, with a normalized power of one, is given by

$$A_r(\tau) = J_0^2(2\pi v\tau/\lambda), \tag{3.56}$$

where v is the vehicle velocity and λ is the signal wavelength. If there are m independent samples of the process $r(t)$, as with m-branch space or time diversity receivers with uncorrelated branch signals, then the autocorrelation for the sample average of the m branches is $A_r(\tau)/m$.

For the shadow fading, we assume log-normal statistics, so $S = \log s$ has a Gaussian distribution. We denote the mean and standard deviation of S by μ and σ, respectively. From (2.30), the autocorrelation of S is given by

$$A_S(\tau) = \sigma^2 e^{-v\tau/X_c}.$$

Power Measurement Filter

The power measurement approach is shown in Figure 3.28. The received signal is passed through a square-law envelope detector and then amplified using a linear or log amplifier. We consider both types of amplifiers, since the statistics of the Rayleigh fading are simpler for a linear amplifier, while the statistics of the log-normal fading are simpler for a log amplifier. Moreover, the log amplifier reduces the estimation problem to classical parameter estimation in the presence of additive noise, while for the linear amplifier, the estimation is done for multiplicative noise. In both cases, the estimation filter $w(t)$ is designed to minimize the dB error of the shadow fading estimate, denoted by $\hat{s}$ and $\hat{S}$, for the linear and log amplifiers, respectively. When a log amplifier is used, the measurement method is refered to as the log-power method; when a linear amplifier is used, it is called the linear-power method.

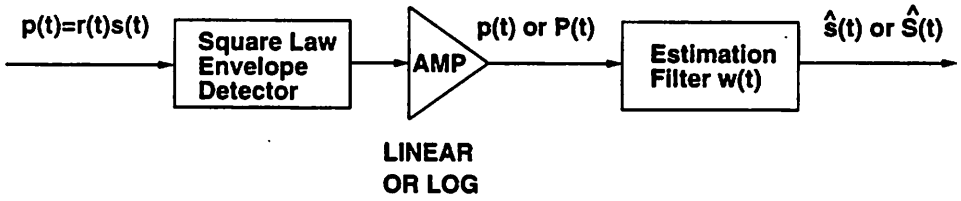

Figure 3.28: Power Measurement Technique.

Arbitrarily choosing the estimation time $t = 0$, we get the estimation values

$$\hat{s}(0) = \int_{-\infty}^{0} w(-t)r(t)s(t)dt; \quad \text{Linear-Power Method} \tag{3.57}$$

and

$$\hat{S}(0) = \int_{-\infty}^{0} w(-t)\left[R(t) + S(t)\right]dt; \quad \text{Log-Power Method.} \tag{3.58}$$

We consider two types of estimation filters: an integrate-and-dump (I&D) filter, and an RC filter. Thus, $w(t) = kg(t)$, where

$$g(t) = \begin{cases} \frac{1}{T_m}\text{Rect}\left[\frac{t}{T_m} - \frac{T_m}{2}\right]; & \text{I\&D filter} \\ \frac{1}{T_m}e^{-t/T_m}; & \text{RC filter} \end{cases} . \tag{3.59}$$

For both filter types, k represents the dc value of the filter's frequency response (since $g(t)$ has unit area), and T_m is the filter's effective averaging time. The filter design then reduces to optimizing k and T_m to minimize the dB estimation error, assuming that the fading parameters σ and X_c are known.

The estimation error is given by

$$\epsilon \triangleq \begin{cases} \log(\hat{s}(0)/s(0)); & \text{Linear-Power Method} \\ \hat{S}(0) - S(0); & \text{Log-Power Method} \end{cases}, \tag{3.60}$$

and its dB value is

$$\delta = (10/\log 10)\epsilon. \tag{3.61}$$

The estimation filter parameters should be set to minimize the value of δ. However, since both the Rayleigh and log-normal fading are stochastic processes, δ is a random variable. We will show that δ is approximately Gaussian distributed; therefore, to minimize δ, the optimal measurement filter should force the mean of δ to zero, and minimize its standard deviation.

Linear-Power Method

From (3.60), the value of ϵ for the linear-power method is

$$\epsilon \triangleq \frac{\hat{s}(0)}{s(0)} = \int_{-\infty}^{0} w(-t)r(t)\left[\frac{s(t)}{s(0)}\right] dt. \tag{3.62}$$

Generally speaking, the averaging time of the filter $w(t)$ should be large relative to the decorrelation time of $r(t)$, and small compared to that of $s(t)$. Therefore, the integral of (3.62) is approximately equal to a sum over several independent samples of the exponentially-distributed variable r, which yields an approximate Gamma distribution [31]. Since the Gamma distribution has the same general shape as the log-normal distribution, with appropriately chosen parameters we can approximate the distribution of $\hat{s}(0)/s(0)$ by a log-normal distribution.

Using this log-normal approximation, we get that $\epsilon = \log[\hat{s}(0)/s(0)]$ is Gaussian distributed. Let a and b denote, respectively, the mean and standard deviation of ϵ. Then

$$\overline{\hat{s}(0)/s(0)} = e^{a+.5b^2}, \tag{3.63}$$

and

$$\overline{(\hat{s}(0)/s(0))^2} = e^{2a+2b^2}. \tag{3.64}$$

Using (3.62), $w(t) = kg(t)$, and the independence of r and s yields

$$\overline{\hat{s}(0)/s(0)} = kW_1, \tag{3.65}$$

and

$$\overline{(\hat{s}(0)/s(0))^2} = k^2 W_2, \tag{3.66}$$

where

$$W_1 = \int_{-\infty}^{0} g(-t)\overline{s(t)/s(0)}dt \tag{3.67}$$

and

$$W_2 = \int_{-\infty}^{0}\int_{-\infty}^{0} g(-t)g(-t')A_r(t-t')\left[\frac{s(t)s(t')}{s^2(0)}\right]dtdt' + \int_{-\infty}^{0}\int_{-\infty}^{0} g(-t)g(-t')\left[\frac{s(t)s(t')}{s^2(0)}\right]dtdt'. \tag{3.68}$$

Since log-normality is preserved under multiplication and division, the variates $s(t)/s(0)$ and $s(t)s(t')/s^2(0)$ are log-normal, and their means and variances can be determined from $A_S(\tau)$ [9]. Given these means and variances, the values of W_1 and W_2 can be computed from (3.67) and (3.68), respectively.

Combining (3.63)-(3.66) yields

$$a = \log(kW_1^2), \tag{3.69}$$

and

$$b = \sqrt{\log(W_2/W_1^2)}. \tag{3.70}$$

From (3.69) and (3.70), setting the filter gain to

$$k = \sqrt{W_2}/W_1^2 \tag{3.71}$$

forces the mean of ϵ to zero without affecting the standard deviation. With this choice for k, the dB measurement error, δ, becomes unbiased and the rms dB error is simply the standard deviation of δ:

$$\Delta = (10/\log 10)\sqrt{\log(W_2/W_1^2)}. \tag{3.72}$$

It can be shown that the value of Δ depends only on three dimensionless parameters: the standard deviation σ of $\log s$, the ratio of shadow fading correlation distance to wavelength, X_c/λ, and the normalized measurement time vT_m/λ [52]. Since k is given by (3.71), minimization of the dB error reduces to minimizing Δ relative to T_m. A plot of this minimum Δ over a range of decorrelation distances and σ values is shown in Figure 3.29. The log-normal approximation for $\hat{s}(0)/s(0)$ (or equivalently, the Gaussian approximation for δ) and the rms dB errors of Figure 3.29 have all been verified by simulation in [52].

[9] Exact expressions for these terms can be found in [52].

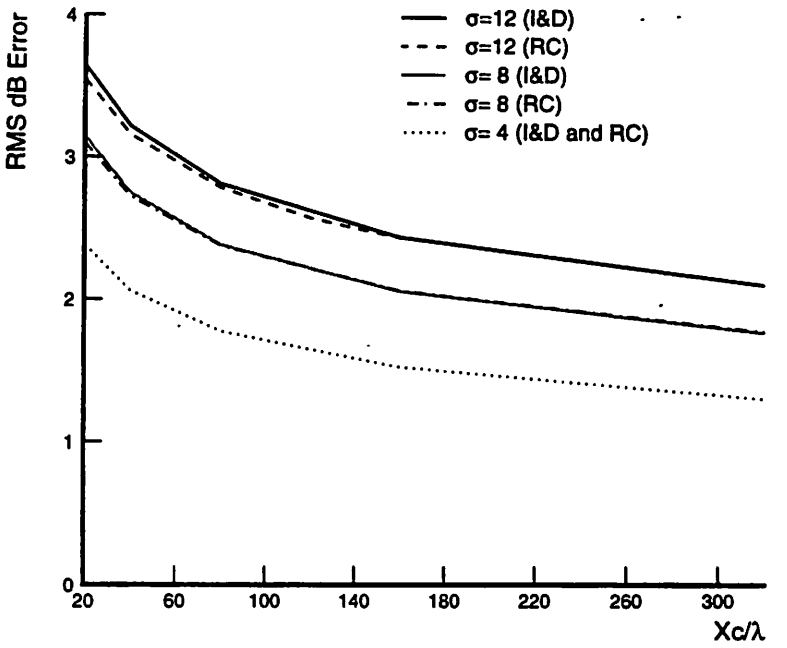

Figure 3.29: Rms dB Error for Linear-Power Method.

Log-Power Method

For the log-power method, $\epsilon \triangleq \hat{S}(0) - S(0)$ is given by

$$\begin{aligned} \hat{S}(0) - S(0) &= \int_{-\infty}^{0} w(-t)[R(t) + S(t)]dt - S(0) \\ &= \left[\int_{-\infty}^{0} w(-t)R(t)dt\right] + \left[\int_{-\infty}^{0} w(-t)S(t)dt - S(0)\right]. \end{aligned} \tag{3.73}$$

We now use (3.73) to approximate the distribution of ϵ. Since the Gaussian distribution is preserved under addition, the second bracketed term in (3.73) is Gauss-distributed. For the first bracketed term, the width of the estimation filter is large relative to the decorrelation time of $R(t)$. In addition, independent sums of random variables with distribution given by (3.55) converge rapidly to Gaussian. Therefore, the distribution of the first bracketed term in (3.73) is also approximately Gaussian, and hence so is ϵ.

Since S has mean μ and standard deviation σ, we can write $S(t)$ as the sum

$$S(t) = \mu + \sigma u(t), \tag{3.74}$$

where $u(t)$ is a zero-mean, unit-variance Gaussian process whose autocorrelation function

is $\epsilon^{-v|\tau|/X_c}$. Replacing $w(t)$ by $kg(t)$ in (3.73) and taking expectation yields

$$\overline{\epsilon} = k\overline{R} + (k-1)\mu. \tag{3.75}$$

Two steps are required to drive $\overline{\epsilon}$ to zero. First, since the mean of the shadow fading, μ, is not known, we set $k = 1$, which removes the second term in (3.75). Then, since we assume $\overline{r} = 1$, it can be shown that $\overline{R}$ equals Euler's Constant, and this quantity must be subtracted from the input to the low pass filter, which for $k = 1$ removes the first bias term in (3.75).

After these two steps, $\overline{\epsilon} = 0$, and the mean-square value of ϵ is

$$\overline{\epsilon^2} = \sigma^2[1 - 2W_1 + W_{2a}] + W_{2b}, \tag{3.76}$$

where, using (2.30) for $A_S(t)$, we get

$$W_1 = \int_{-\infty}^{0} g(-t)e^{-v|t|/X_c}dt, \tag{3.77}$$

$$W_{2a} = \int_{-\infty}^{0}\int_{-\infty}^{0} g(-t)g(-t')e^{-v|t-t'|/X_c}dtdt', \tag{3.78}$$

and

$$W_{2b} = \int_{-\infty}^{0}\int_{-\infty}^{0} g(-t)g(-t')A_R(t-t')dtdt'. \tag{3.79}$$

Computation of W_{2b} requires the autocorrelation function of $R(t)$, which is not available in closed form. However, it can be approximated with high accuracy by [52]

$$A_R(\tau) = \frac{\pi^2}{6}\left[.607J_0^2(2\pi v\tau/\lambda) + .393J_0^4(2\pi v\tau/\lambda)e^{-1.283v|\tau|}\right]. \tag{3.80}$$

Using this approximation, W_1, W_{2a}, and W_{2b} can be computed as functions of the fading and filter parameters. These values can be substituted into (3.76) to get $\overline{\epsilon^2}$, and the corresponding rms dB error of $\Delta = (10/\log 10)\sqrt{\overline{\epsilon^2}}$.

It can be shown that, as in the linear-power case, the value of Δ depends only on the parameters σ, X_c/λ, and vT_m/λ [52]. Since we require $k = 1$ for zero mean error, we can only adjust the value of T_m to minimize the value of Δ. A plot of this minimum Δ over a range of decorrelation distances and σ values is shown in Figure 3.30. This rms dB error was also confirmed by simulation in [52].

Comparing figures 3.29 and 3.30, we see that the rms dB errors for the log power method tend to be lower than those for the linear power method, but only by approximately

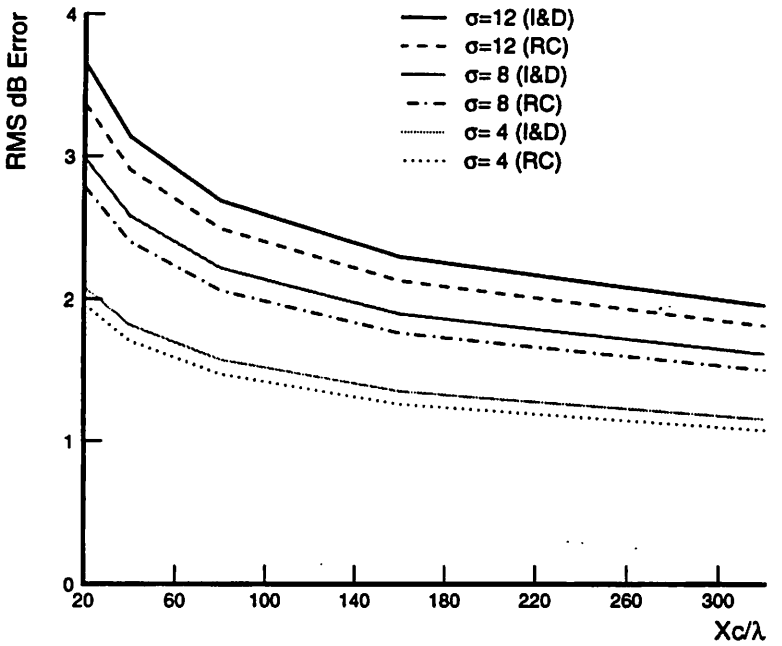

Figure 3.30: Rms dB Error for Log-Power Method.

.3dB or less. Therefore, the shape of the measurement filter has little effect on the rms dB error, as long as the filter parameters are optimized relative to the fading statistics. This analysis can be easily extended to include the effects of antenna diversity; in [52] it is shown that two antennas yield a reduction of at least 1dB in the rms dB error.

3.6.2 Estimation Error Effects

In §3.6.1 we analyzed power measurement filters to minimize the rms dB measurement error. We concluded that even when the fading parameters are known, the dB measurement error can be as high as 3dB. Moreover, since the fading parameters are not always known, and the estimation filters must work over a range of vehicle velocities and propagation environments, the rms dB error will generally exceed this nominal value. We now determine the effect that these estimation errors have on the variable rate M-QAM modulation and coding schemes of §§3.4 – 3.5.

We assume the multiplicative Rayleigh/log-normal fading model of the previous section, so the instantaneous power is given by $p(t) = r(t)s(t)$. Let γ denote the instan-

taneous value of the Rayleigh fading and $\overline{\gamma}$ denote its short term average (that is, the instantaneous shadow fading value). Since Rayleigh fading is relatively fast, power estimation techniques generally focus on determining the shadow fading value only (as in §3.6.1). We therefore assume that γ can be measured perfectly, but the short-term average $\overline{\gamma}$ cannot. We denote the estimate for $\overline{\gamma}$ by $\hat{\overline{\gamma}}$. Using the estimation error statistics derived in §3.6.1, we have

$$\hat{\overline{\gamma}} = 10^{\epsilon/10}\overline{\gamma}, \tag{3.81}$$

where ϵ is a zero-mean Gaussian variate with a standard deviation between one and four dB.

Since γ is known perfectly, we can still maintain a given BER by adjusting the power control policy $P(\gamma)$ and the number of constellation points M, as in (3.38):

$$M(\gamma) = 1 + \frac{1.5\gamma}{-\log(\mathrm{BER}/2)}\frac{P(\gamma)}{P}.$$

Recall from (3.41) that the power control policy maximizing the average rate, subject to the power constraint $\overline{(P(\gamma)/P)} \leq 1$, is

$$K\frac{P(\gamma)}{P} = \begin{cases} \frac{1}{\gamma_K} - \frac{1}{\gamma} & \gamma \geq \gamma_K \\ 0 & \gamma < \gamma_K \end{cases}, \tag{3.82}$$

where γ_K is the "cutoff" fade depth chosen to satisfy the power constraint, and $K = -1.5/\log(\mathrm{BER}/2)$ for uncoded modulation. With coding, $K = -1.5G_c/\log(\mathrm{BER}/2)$, where G_c is the coding gain.

As derived in (3.43), the maximum transmission rate equals

$$\frac{R}{B} = \int_{\gamma_K}^{\infty} \log_2\left(\frac{\gamma}{\gamma_K}\right)\pi(\gamma)d\gamma, \tag{3.83}$$

and from (3.44), the power constraint can be written as

$$\int_{\gamma_K}^{\infty} K\frac{P(\gamma)}{P} = \int_{\gamma_K}^{\infty}\left(\frac{1}{\gamma_K} - \frac{1}{\gamma}\right)\pi(\gamma)d\gamma = K. \tag{3.84}$$

The power control policy $P(\gamma)/P$ is optimal for *any* distribution on γ; however, the policy implicitly depends on $\pi(\gamma)$ through the cutoff value γ_K, which is determined by (3.84).

If the estimation error $\epsilon = 0$, then the distribution for γ is exponential with mean $\overline{\gamma}$. We will denote an exponential distribution for γ with mean μ by $\pi_e(\gamma|\mu)$. If $\epsilon \neq 0$, then the estimate $\hat{\overline{\gamma}}$ is known, but $\overline{\gamma}$ is not. The distribution of γ is then given by

$$\pi(\gamma) = \int_{\overline{\gamma}} \pi_e(\gamma|\overline{\gamma})p(\overline{\gamma}|\hat{\overline{\gamma}})d\overline{\gamma}, \tag{3.85}$$

where $p(\overline{\gamma}|\hat{\overline{\gamma}})$ is determined by inverting (3.81) to get $\overline{\gamma}$ as a function of $\hat{\overline{\gamma}}$ and ϵ, and then using the statistics of ϵ to get the distribution of $\overline{\gamma}$.

We can use one of two techniques to calculate the cutoff value when the estimate of $\overline{\gamma}$ is imperfect:

1. Calculate γ_K from (3.84) using $\pi(\gamma)$ given by (3.85).
2. Assume $\epsilon = 0$ and calculate γ_K as if $\hat{\overline{\gamma}} = \overline{\gamma}$.

From §§3.4–3.5, the first approach is optimal for maximizing the spectral efficiency under the given power constraint. However, this approach requires knowing the standard deviation of ϵ, which varies between one and four dB depending on the vehicle speed, estimation filter, and shadow fading statistics. Since these parameters are usually unknown, we will consider how the estimation errors impact the average transmit power and data rate under the second approach.

Using the second approach, the power control policy will use the cutoff value γ_K^* which satisfies

$$\int_{\gamma_K}^{\infty} \left(\frac{1}{\gamma_K} - \frac{1}{\gamma}\right) \pi_e(\gamma|\hat{\overline{\gamma}}) d\gamma = K. \tag{3.86}$$

Let γ_K denote the cutoff value which satisfies (3.86) when $\epsilon = 0$ (i.e. when $\overline{\gamma} = \hat{\overline{\gamma}}$). It is easily shown from (3.86) that if $\epsilon > 0$, then γ_K^* will be greater than γ_K. Using γ_K^* instead of γ_K in (3.84) and (3.83) with the true distribution of gamma ($\pi(\gamma) = \pi_e(\gamma|\overline{\gamma})$) yields the average transmit power and data rate under this policy. For $\gamma_K^* > \gamma_K$, both the average power and rate will be smaller than if γ_K had been used. Conversely, if $\epsilon < 0$, then γ_K^* will be less than γ_K, resulting in a larger average power and data rate. These effects are illustrated in Figures 3.31 and 3.32: Figure 3.31 shows the change in average transmit power as a function of the estimation error ϵ, and Figure 3.32 shows the corresponding average data rate.

We can also consider the same estimation error effects on the modified constant power policy of §3.3.2, which compensates for fading above a certain cutoff fade depth γ_0:

$$\frac{P(\gamma)}{P} = \begin{cases} \frac{P_R}{\gamma} & \gamma \geq \gamma_0 \\ 0 & \gamma < \gamma_0 \end{cases}, \tag{3.87}$$

where

$$P_R = \left[\int_{\gamma_0}^{\infty} \frac{1}{\gamma} \pi(\gamma) d\gamma\right]^{-1}. \tag{3.88}$$

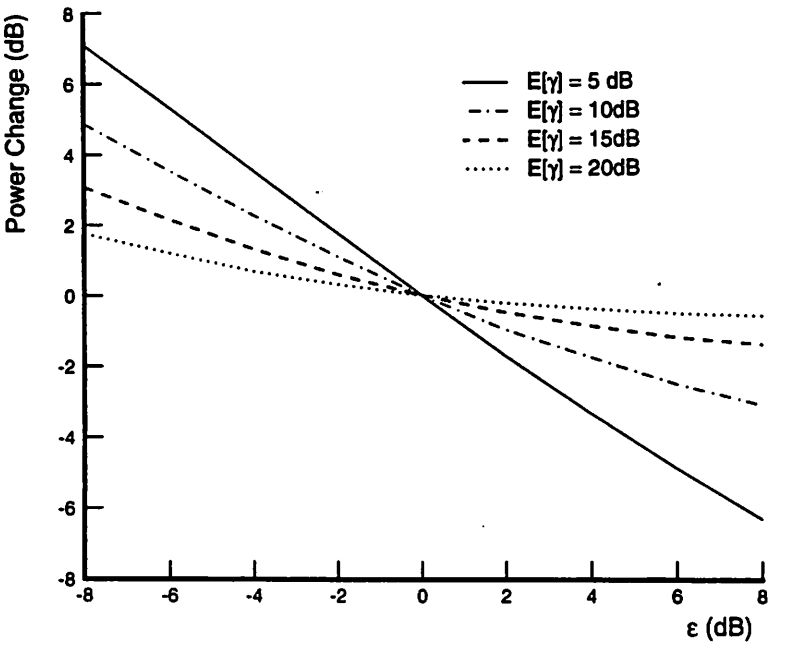

Figure 3.31: Average Transmit Power versus ϵ.

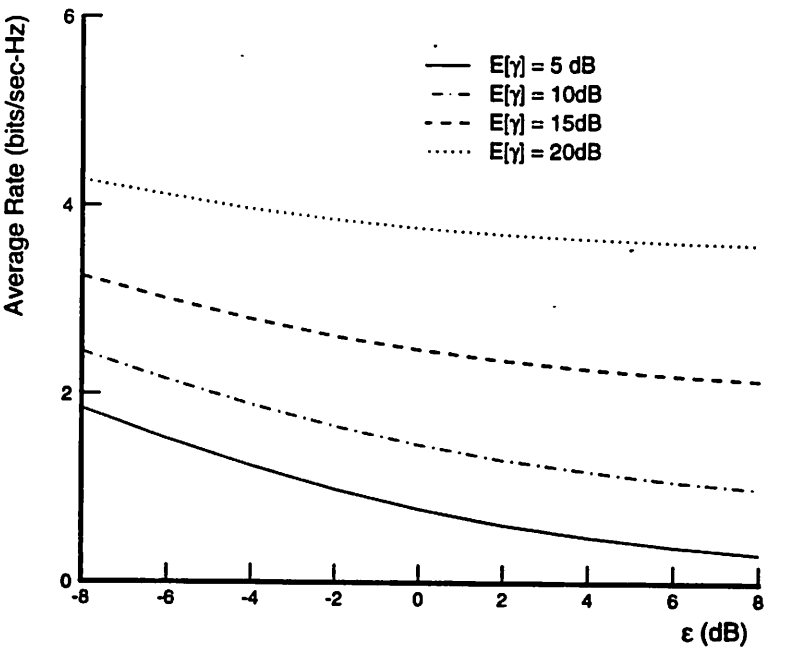

Figure 3.32: Average Data Rate versus ϵ.

The average data rate with this policy, assuming a coding gain of G_c, is

$$\frac{R}{B} = \log_2[1 + G_c P_R] \int_{\gamma_0}^{\infty} \pi_e(\gamma|\overline{\gamma}) d\gamma, \tag{3.89}$$

and this expression is maximized relative to γ_0 to get the maximum data rate[10]. If the estimate of $\overline{\gamma}$ is in error, then the modified policy will determine γ_0 by maximizing

$$\log_2[1 + G_c P_R^*] \int_{\gamma_0}^{\infty} \pi_e(\gamma|\hat{\overline{\gamma}}) d\gamma \tag{3.90}$$

relative to γ_0, where

$$P_R^* = \left[\int_{\gamma_0}^{\infty} \frac{1}{\gamma} \pi_e(\gamma|\hat{\overline{\gamma}}) d\gamma\right]^{-1}. \tag{3.91}$$

Let γ_0^* denote this maximizing cutoff value. The average transmit power using γ_0^*, P_R^*, and the true distribution of γ is

$$\overline{\left(\frac{P(\gamma)}{P}\right)} = \int_{\gamma_0^*}^{\infty} \frac{P_R^*}{\gamma} \pi_e(\gamma|\overline{\gamma}) d\gamma, \tag{3.92}$$

and the corresponding data rate is

$$\frac{R}{B} = \log_2[1 + G_c P_R^*] \int_{\gamma_0^*}^{\infty} \pi_e(\gamma|\overline{\gamma}) d\gamma. \tag{3.93}$$

We plot the average transmit power and data rate of the modified constant power policy with estimation errors, given by (3.92) and (3.93), respectively, in Figures 3.33 and 3.34. These curves show global maximum values for both data rate and power, regardless of how large or small the estimation error may be. This behavior in the average transmit power is good from an interference and power conservation perspective, since regardless of the estimation error, the transmit power will not deviate above this global maximum.

3.6.3 Periodic Estimation: The On/Off Channel

The estimation techniques outlined above are for instantaneous power estimation of narrowband fading channels. These techniques do not apply to wideband channels with nonzero delay spread, since the channel impulse response cannot be estimated instantaneously. For wideband channels, a sequence of bits known to both the transmitter and receiver can be used to learn the channel [31]. This bit sequence is generally refered to

[10]This rate is based on a particular log-normal shadowing value $\overline{\gamma}$. Since $\overline{\gamma}$ will vary slowly over time, the data rate (3.89) will change with this slow variation.

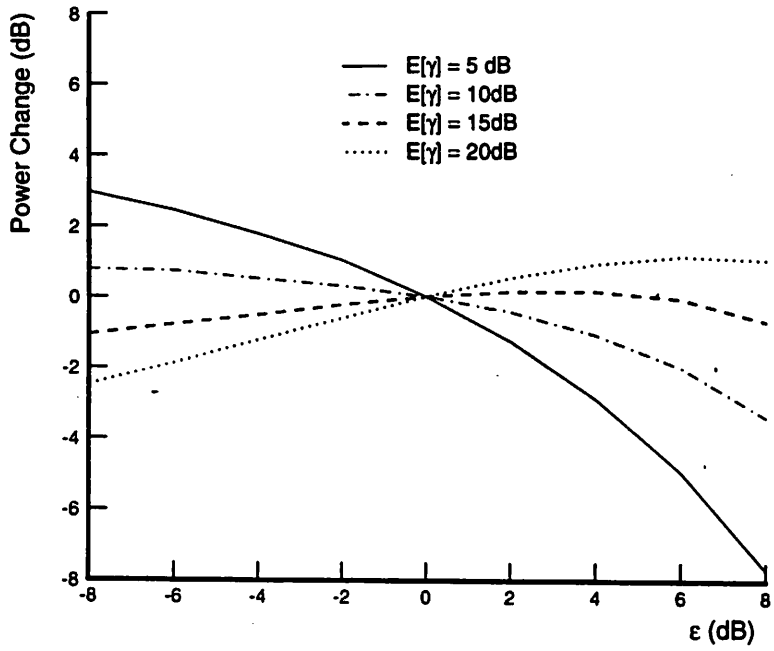

Figure 3.33: Average Transmit Power for Modified Constant Power Policy.

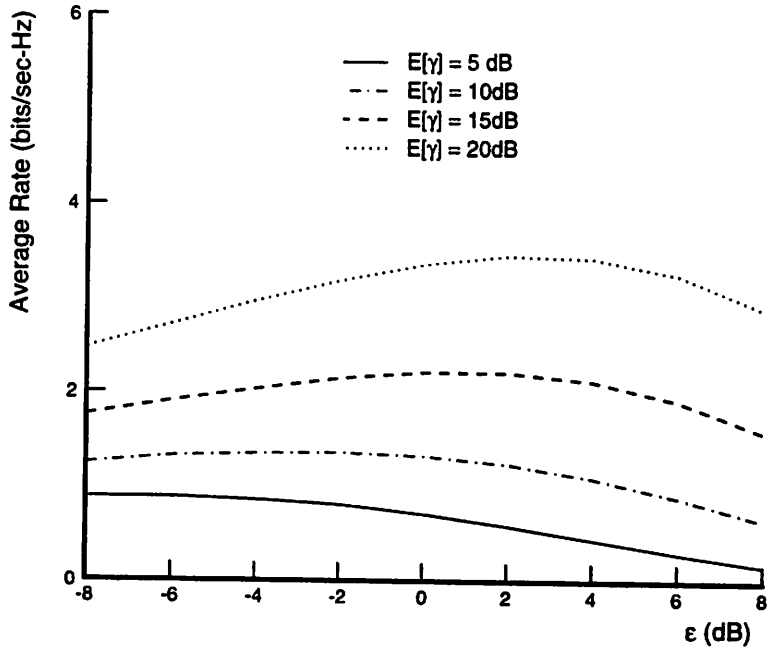

Figure 3.34: Average Data Rate for Modified Constant Power Policy.

as a training sequence. After initial training, the channel estimate can be updated by either using data decisions to modify the channel estimate (decision-feedback equalization), or sending the training sequence periodically to update the channel estimate. The former approach has the advantage that after the initial training sequence, data transmission need not be interrupted to update the channel estimate. However, this approach gives rise to decision- feedback errors, which may cause the channel estimate to diverge. Moreover, for time division systems, where users share the same frequency band using periodic time slots, the channel corresponding to each user slot is different, and therefore the channel must be re-estimated during each time slot. Periodic channel estimation introduces some capacity loss, since data transmission is turned off at periodic intervals. We will now precisely bound this capacity loss, relative to the capacity derived in § 3.1, for the continuous-time state space channel of § 2.4.2.

In our model, a data sequence is transmitted over the channel for time T and then a known training sequence is transmitted over the channel for time T_e. This is equivalent, relative to the data rate, to turning the transmitter off for T_e seconds after every T seconds of data transmission. We therefore refer to it as the *on/off* channel. This model also applies to a time division system, where the off time for a given user, user A, equals the time slots occupied by all the other users.

We assume that the channel state is constant while the data is being transmitted, only changing during the estimation period. The effect of continuously changing channel parameters, and the resulting channel estimation errors, are quantified in [53]; we will not address this issue. We also assume that the channel memory is less than the estimation time T_e. The second assumption implies that data transmitted before the training sequence does not affect data received after the training sequence. Combining this with the first assumption, we can model the periodically-estimated time-varying channel using the time diversity model of Figure 3.1, with the channel input multiplied by a periodic rectangular wave r(t), as in Figure 3.35. For this channel model, the estimate $\hat{S}$ is derived during the training period T_e, and is assumed to equal the true channel state. We will let C^o denote the capacity of the time-varying channel depicted in Figure 3.35. We also assume a transmit power constraint of P.

Consider now a time-invariant channel c_i with impulse response $h_{c_i}(t)$, average power P_i, and periodic off time T_e over T seconds of data transmission, as shown in Figure 3.36. Let $C_i^o(P_i)$ denote the Shannon capacity of this channel. Then by the same

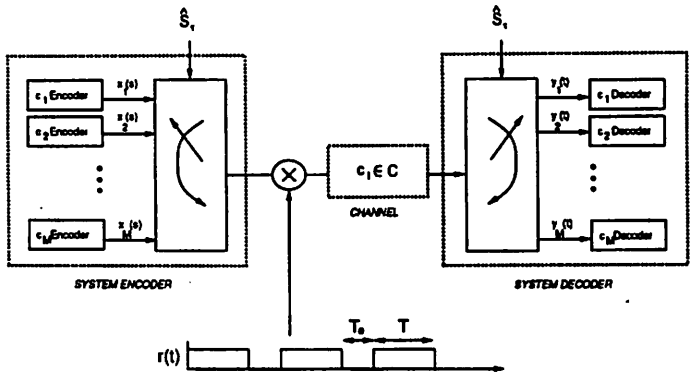

Figure 3.35: Time Diversity System with Periodic Estimation.

capacity argument in §3.1, the capacity of the time diversity system in Figure 3.35 equals the weighted sum of the periodically estimated channels C_i^o:

$$C^o = \max_{P^M \in \mathcal{P}^M} \sum_i \pi_i C_i^o(P_i), \tag{3.94}$$

where $\mathcal{P}^M$ and π are defined by (3.6) and (3.2), respectively. We now derive upper and lower bounds for C_i^o; upper and lower bounds for C^o are then easily derived by using either all upper or all lower bounds for C_i^o in (3.94).

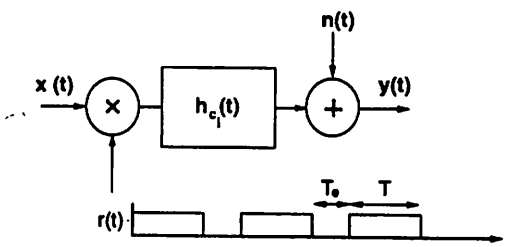

Figure 3.36: On/Off Time-Invariant Channel.

To calculate an upper bound for C_i^o, we introduce a random delay after the multiplier of Figure 3.36. This is illustrated in Figure 3.37, where δ is the random delay uniformly distributed on $[0, T+T_e]$. If $x(t)$ is wide-sense stationary (WSS), the signal after the random delay, $v(t)$, is also WSS with spectrum

$$V(f) = X(f) * \hat{S}(f), \tag{3.95}$$

where

$$X(f) = \mathcal{F}[Ex(t)x(t-\tau)], \tag{3.96}$$

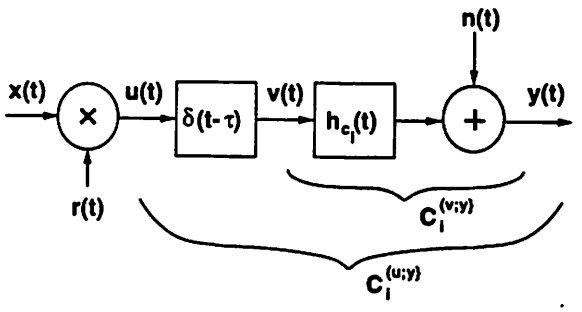

Figure 3.37: On/Off Channel with Random Delay.

and

$$\hat{S}(f) = \mathcal{F}[Es(t-\xi)s(t-\tau-\xi)]. \tag{3.97}$$

The bracketed term in (3.97) is periodic in τ, as shown in Figure 3.38.

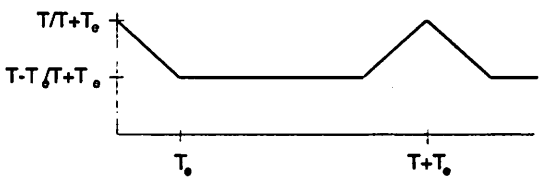

Figure 3.38: $E_\xi[s(t-\xi)s(t-\tau-\xi)]$.

The capacity of the time-invariant on/off channel is the maximum mutual information between $u(t)$ and $y(t)$ in Figure 3.37, which we denote by $C_i^{u;y}$. By the data processing inequality [42], this is less than $C_i^{v;y}$, the maximum mutual information between $v(t)$ and $y(t)$. Since $v(t)$ is WSS, $C_i^{v;y}$ is given by Gallager's result for time-invariant channels [40, page 424]. Thus, for average power constraint P_i and spectral noise density $N(f)$, the capacity $C_i^o = C_i^{u;y}$ is bounded above by

$$C_i^o \le C_i^{v;y} = \max_{X(f):\int X(f)\le P_i} \int \frac{1}{2}\log\left[1+\frac{|H(f)|^2[X(f)*\hat{S}(f)]}{N(f)}\right]df. \tag{3.98}$$

We now derive a lower bound, using a fixed set of codewords and a specific encoding and decoding scheme. Let $h_{c_i}(t)$ be the estimated channel impulse response and $x(t)$ be the corresponding codeword that achieves capacity for the time-invariant channel h_{c_i} without estimation. Assume $T > T_e$. We define a new codeword $\hat{x}(t)$ as follows.

$$\hat{x}(t) = \begin{cases} x(t) & t \le T \\ 0 & T < t \le T + T_e \\ x(t - 2T_e) & T + T_e < t \le 2T + T_e \\ 0 & 2T + T_e < t < 2(T + T_e) \\ \vdots & \vdots \\ x(t - 2nT_e) & n(T + T_e) < t \le (n+1)T + nT_e \\ 0 & (n+1)T + nT_e < t < (n+1)(T + T_e) \\ \vdots & \vdots \end{cases} \tag{3.99}$$

Basically, the new codeword repeats the T_e seconds of the original codeword immediately proceeding the off time just after the off time, and doesn't transmit anything during the off time. This repeating allows the receiver to concatenate the received signal such that it is equivalent to the original codeword transmitted on the time-invariant channel h_{c_i}.

If $\hat{y}(t)$ is the response of the on/off channel to $\hat{x}(t)$, then we process $\hat{y}(t)$ to get $y(t)$ as follows.

$$\begin{aligned} y(t) \;=\; & \hat{y}(t)1[t \le T] + \hat{y}(t - 2T_e)1[T + 2T_e < t \le 2T + T_e] + \ldots + \\ & \hat{y}(t - 2nT_e)1[nT + (n+1)T_e) < t \le (n+1)T + nT_e] + \ldots \end{aligned} \tag{3.100}$$

The concatenated output $y(t)$ is equivalent to the output of a channel with impulse response $h_{c_i}(t)$ to the input signal $x(t)$. Denote the capacity of the unestimated time-invariant channel corresponding to $h_{c_i}(t)$ with average power P_i as $C_i(P_i)$. Due to the concatenation, the on/off channel requires $T + T_e$ seconds to receive $(T - T_e)C_i$ bits, so the attainable rate for the on/off channel must be weighted accordingly. On the other hand, if we assume that the repeated portions of the codeword have the same average power as the entire codeword, then if P_i is the average power of the codeword x, the corresponding codeword $\hat{x}(t)$ has average power $P_iT/T + T_e$. We can thus increase the power of the original codeword x to $P_i(T + T_e)/T$ without violating the average power constraint for $\hat{x}$. Combining these results we get the following lower bound for C_i^o:

$$C_i^o \ge \frac{T - T_e}{T + T_e} C_i \left(\frac{P_i(T + T_e)}{T} \right). \tag{3.101}$$

In the limit as $T_e/T \to 0$, the upper and lower bounds in (3.98) and (3.101) both approach the capacity of the time-invariant channel C_{h_i}. Thus, the effect of the off time on

channel capacity becomes small when the off time is negligible relative to the on time, as would be the case for estimation of a slowly-varying channel.

3.7 Summary

We have presented several techniques for increasing spectral efficiency on time-varying channels, where the channel can be estimated and this estimate fed back to the transmitter. We first calculated the capacity of a general time-varying channel assuming perfect channel information at the transmitter; this capacity specifies the maximum spectral efficiency of a channel for an arbitrarily small error probability, with no restriction on the complexity or delay of the encoder or decoder. We then applied this result to channels with a time-varying impulse response, and showed that the optimal input power spectrum for this channel is derived from a water-filling in time and frequency of the channel impulse response.

Next, we applied the capacity results to narrowband fading channels. We found in this case that spectral efficiency is maximized when transmit power, data rate, and coding are all adapted relative to the channel fading. Moreover, the optimal scheme is intuitive in the sense that it increases power and data rate to take advantage of favorable channels. We compared this optimal scheme with a common power control policy which inverts the channel fading; numerical results show that our optimal policy is significantly better in terms of both spectral efficiency and outage probability. The capacity analysis also suggested a variable-rate M-QAM modulation technique for fading channels. We determined the spectral efficiency of this technique, and found a closed-form expression for its maximum coding gain relative to Shannon capacity. Finally, we proposed a variable-rate coded modulation scheme for M-QAM constellations, and proposed specific coding strùctures to achieve near-capacity rates with moderate coding complexity.

All of these techniques assumed perfect channel estimation in zero time. We then analyzed the impact of channel estimation on capacity. We first proposed a power measurement filter for narrowband fading channels, and evaluated its rms dB error. We then determined the effect of this estimation error on our optimal power control and adaptive coded modulation scheme. Finally, we bounded the capacity loss from periodic channel estimation, where no data is transmitted during the estimation sequence. There are several obvious extensions to the estimation analysis. Power control algorithms which use the

statistics of the estimation error should be considered. In addition, the spectral efficiency of joint channel estimation and data transmission should be compared with that of periodic estimation. Finally, the impact on spectral efficiency of wideband channel estimation errors should be evaluated. There is an obvious tradeoff between the amount of time spent estimating the channel and the corresponding estimation error. We have bounded the capacity loss resulting from the estimation time; if we also determine the impact on capacity of estimation error, then combining these two results would yield the optimal estimation time, relative to channel capacity, of a time-varying channel.

Chapter 4

Spectrally-Efficient Techniques for Time-Varying Nonfeedback Channels

The adaptive techniques proposed in the previous chapter assume that the channel is estimated at the receiver, and this estimate fed back to the transmitter. However, a reliable feedback path is not always available. Moreover, the feedback path will generally exhibit a nonzero delay; thus, the channel estimate may be outdated by the time it reaches the transmitter, especially for rapidly-varying channels. For these reasons, nonfeedback approaches must also be considered. Therefore, we now explore signal processing and coding techniques to increase spectral efficiency on time-varying channels without feedback.

We use the discrete-time finite-state Markov channel (FSMC) model of §2.4.1. First we calculate the capacity of this channel. We then propose a low-complexity decision-feedback decoder, which uses the Markov transition probabilities for maximum-likelihood sequence detection. We also calculate the decoder performance for a two-state variable noise channel.

If the correlation properties of the channel variation are not known, then the channel memory can be dispersed through interleaving to remove burst errors, and memoryless channel encoding can be used. These techniques are discussed in §4.3. An alternate approach uses unequal error protection codes, described in §4.5. This type of coding prioritizes the transmitted bit stream, and allows some loss of low-priority data when the channel is

bad, thus providing robust communication of some data even under adverse channel conditions. We will examine both the theoretical performance limits of such coding techniques, as well as some practical code designs.

4.1 Performance Limits for Finite-State Markov Channels

Our capacity results are an extension of the analysis by Mushkin and Bar-David [54] for the Gilbert-Elliot channel to the more general Finite-State Markov channel (FSMC) of §2.4.1.1. The Gilbert-Elliot channel is a stationary two-state Markov chain, where each state is a binary symmetric channel (BSC), as in Figure 4.1. The transition probabilities between states are g and b respectively, and the crossover probabilities for the "good" and "bad" BSCs are p_G and p_B respectively, where $p_G < p_B$.

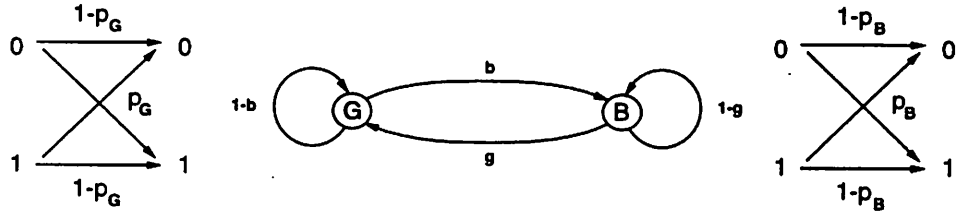

Figure 4.1: Gilbert-Elliot Channel

Let $x_n \in \{0,1\}$, $y_n \in \{0,1\}$, and $z_n = x_n \oplus y_n$ denote respectively the channel input, channel output, and channel error on the nth transmission. In [54], the capacity of the Gilbert-Elliot channel is derived as

$$C = \lim_{n\to\infty} 1 - \mathrm{E}[h(q_n)] = \lim_{n\to\infty} 1 - \mathrm{E}[h(q_n^*)], \tag{4.1}$$

where $h(p) = p\log p + (1-p)\log(1-p)$, $q_n = p(z_n = 1|z^{n-1})$, $q_n^* = p(z_n = 1|z^{n-1}, S_0)$, and S_0 is the initial channel state.

The FSMC is a more general model, since the channels are not necessarily BSCs, and the input/output alphabets are only required to be finite. If the transmitter and receiver have perfect state information, then from §3.1, the capacity of the FSMC is just the statistical average over all states of the corresponding channel capacity. On the other hand, with no information about the channel state or its transition structure, capacity is reduced to that of the Arbitrarily Varying Channel [55]. We consider the intermediate case, where the channel transition probabilities of the FSMC are known.

4.1.1 Conditional State Distribution

The conditional channel state distribution is the key to determining the capacity of the FSMC. It is also a sufficient statistic for the input given all past inputs and outputs, thus allowing for the reduced complexity of the maximum-likelihood decoder we propose in §4.3. We now show that the state distribution conditioned on past input/output pairs can be calculated using a recursive formula. Thus, it is a Markov chain. A similar formula is derived for the state distribution conditioned on past outputs alone, under the assumption of independent channel inputs.

Let K be the size of the channel state space. We denote the conditional channel state distributions by the K dimensional random vectors $\pi_n = (\pi_n(1), \ldots, \pi_n(K))$ and $\rho_n = (\rho_n(1), \ldots, \rho_n(K))$, where

$$\rho_n(k) = p(S_n = c_k | y^{n-1}), \tag{4.2}$$

and

$$\pi_n(k) = p(S_n = c_k | x^{n-1}, y^{n-1}). \tag{4.3}$$

The following recursive formula for π_n is derived in Appendix 4.A.1:

$$p(S_{n+1} = c_l | x^n, y^n) = \frac{\sum_{j \in K} p(y_n | S_n = c_j, x_n) p(S_n = c_j | x^{n-1}, y^{n-1}) P_{jl}}{\sum_{k \in K} p(y_n | S_n = c_k, x_n) p(S_n = c_k | x^{n-1}, y^{n-1})}. \tag{4.4}$$

This expression can be written in the following vector form,

$$\pi_{n+1} = \frac{\pi_n D(x_n, y_n) P}{\pi_n D(x_n, y_n) \underline{1}} = f(x_n, y_n, \pi_n), \tag{4.5}$$

where $D(x_n, y_n)$ is a diagonal $K \times K$ matrix with kth diagonal term $p_k(y_n | x_n)$, and $\underline{1} = (1, \ldots, 1)^T$ is a K dimensional vector. Equation (4.5) defines a recursive relation for π_n. Thus, π_n is a Markov chain with state space $\Delta = \{\alpha \in R^K | \alpha_i \geq 0, \sum \alpha_i = 1\}$. The chain has initial value $\pi_0 = (p(S_0 = c_1), \ldots, p(S_0 = c_K))$, and transition probabilities

$$p(\pi_{n+1} = \alpha | \pi_n = \beta) = \sum_{\substack{x_n \in \mathcal{X} \\ y_n \in \mathcal{Y}}} 1[(x_n, y_n) : f(x_n, y_n, \beta) = \alpha] p(y_n | \pi_n = \beta, x_n) p(x_n). \tag{4.6}$$

When the initial state is known ($S_0 = c_i$ for some i), the distribution π_0 is a "delta" function,

$$\pi_0(j) = \begin{cases} 0 & j \neq i \\ 1 & j = i \end{cases}. \tag{4.7}$$

In this case we denote the state distribution π_n by π_n^i, so $\pi_n^i \triangleq p(S_n|x^{n-1}, y^{n-1}, S_0 = c_i)$.

For independent inputs, we can obtain a similar recursive formula for ρ_n:

$$\rho_{n+1} = p(S_{n+1} = c_l|y^n) = \frac{\sum_{j\in K} p(y_n|S_n = c_j)p(S_n = c_j|y^{n-1})P_{jl}}{\sum_{k\in K} p(y_n|S_n = c_k)p(S_n = c_k|y^{n-1})}. \tag{4.8}$$

The derivation is similar to that of Appendix 4.A.1, using (2.43) instead of (2.41) and removing all x terms. We can also write (4.8) in vector form:

$$\rho_{n+1} = \frac{\rho_n B(y_n)P}{\rho_n B(y_n)\underline{1}} = \hat{f}(y_n, \rho_n), \tag{4.9}$$

where $B(y_n)$ is a diagonal $K \times K$ matrix with kth diagonal term $p(y_n|S_n = c_k)$[1]. Thus, ρ_n is a Markov chain with initial value $\rho_0 = \pi_0$ and transition probabilities

$$p(\rho_{n+1} = \alpha|\rho_n = \beta) = \sum_{y_n\in\mathcal{Y}} 1[y_n : \hat{f}(y_n, \beta) = \alpha]p(y_n|\rho_n = \beta). \tag{4.10}$$

When the initial state is known, we denote ρ_n by ρ_n^i, where $\rho_n^i \triangleq p(S_n|y^{n-1}, S_0 = c_i)$.

We next show that under some mild constraints on C, the Markov chains π_n and ρ_n converge in distribution when the inputs are i.i.d., and the resulting limit distributions are independent of the initial states. Moreover, these limit distributions are continuous functions of the input distribution $p(x)$.

4.1.2 Convergence of the State Distribution

To obtain the weak convergence of π_n and ρ_n, we assume that the channel inputs are i.i.d., then apply convergence results for partially observed Markov chains [56]. Consider the new stochastic process $U_n \triangleq (S_n, y_n, x_n)$ defined on the state space $\mathcal{U} = \mathcal{C} \times \mathcal{Y} \times \mathcal{X}$. Since S_n is stationary and ergodic, and x_n is i.i.d., U_n is stationary and ergodic. It is easily checked that U_n is Markov.

Let $(S, y, x)_j$ denote the jth element of $\mathcal{U}$, and $J \triangleq |\mathcal{U}|$. To specify the individual components of the vector $\mathcal{U}$, we use the notation

$$(S_{(j)}, y_{(j)}, x_{(j)}) \triangleq (S, y, x)_j.$$

The $J \times J$ probability transition matrix for U, P^U, is

$$P_{kj}^U = p[(S_{n+1}, y_{n+1}, x_{n+1}) = (S, y, x)_j|(S_n, y_n, x_n) = (S, y, x)_k)], \tag{4.11}$$

[1]Note that $B(y_n)$ has an implicit dependence on the distribution of x_n.

independent of n. The initial distribution of U, π_0^U, is given by

$$\pi_0^U \triangleq p(S_0 = c_k, y_0 = y, x_0 = x) = \pi_0(k)p_k(y_0|x_0)p(x_0). \tag{4.12}$$

Let $g_{y,x} : \mathcal{U} \to \mathcal{Y} \times \mathcal{X}$ and $g_y : \mathcal{U} \to \mathcal{Y}$ be the projections $g_{y,x}(S_n, y_n, x_n) = (y_n, x_n)$ and $g_y(S_n, y_n, x_n) = (y_n)$, respectively. These projections form the new processes $W_n = g_{y,x}(U_n)$ and $V_n = g_y(U_n)$. We regard W_n and V_n as partial observations of the Markov chain U_n; the pairs (U_n, W_n) and (U_n, V_n) are referred to as partially observed Markov chains. We denote the distribution of U_n conditioned on W^n by $\pi_n^U = (\pi_n^U(1), \ldots, \pi_n^U(J))$, where

$$\pi_n^U(j) = p(U_n = (S, y, x)_j | W^n). \tag{4.13}$$

Similarly, $\rho_n^U = (\rho_n^U(1), \ldots, \rho_n^U(J))$ denotes the distribution of U_n conditioned on V^n, where

$$\rho_n^U(j) = p(U_n = (S, y, x)_j | V^n). \tag{4.14}$$

Note that

$$\begin{aligned} \pi_n^U(j) &= p(U_n = (S, y, x)_j | W^n) \\ &= p(S_n = S_{(j)} | x^n, y^n) 1[x_n = x_{(j)}, y_n = y_{(j)}] \\ &= \pi_n(k) 1[x_n = x_{(j)}, y_n = y_{(j)}], \end{aligned} \tag{4.15}$$

where $S_{(j)} = c_k$. Thus if π_n^U converges in distribution, π_n must also converge in distribution. Similarly, ρ_n converges in distribution if ρ_n^U does.

We will use the following definition for subrectangular matrices in the subsequent theorem.

Definition Let $D = (D_{ij})$ denote a square matrix. If $D_{i_1,j_1} \neq 0$ and $D_{i_2,j_2} \neq 0$ implies that also $D_{i_1,j_2} \neq 0$ and $D_{i_2,j_1} \neq 0$, then D is called a subrectangular matrix.

We can now state the convergence theorem, due to Kaijser [56], for the distribution of a Markov chain conditioned on partial observations.

Theorem 4.1.2.1 Let U_n be a stationary and ergodic Markov chain with transition matrix P^U and state space $\mathcal{U}$. Let g be a function with domain $\mathcal{U}$ and range $\mathcal{Z}$. Define a

new process $Z_n = g(U_n)$. For $z \in \mathcal{Z}$ and $U_{(j)}$ the jth element of $\mathcal{U}$, define matrix $M(z)$ by

$$M_{i,j}(z) = \begin{cases} P_{ij}^U & \text{if } g(U_{(j)}) = z \\ 0 & \text{otherwise} \end{cases} . \tag{4.16}$$

Suppose that P^U and g are such that there exists a finite sequence $z_1, \ldots, z_m$ of elements in $\mathcal{Z}$ that yield a nonzero subrectangular matrix for the matrix product $M(z_1) \ldots M(z_m)$. Then $p(U_n|Z^n)$ converges in distribution and moreover, the limit distribution is independent of the initial distribution of U.

We first apply this theorem to π_n^U.

Assumption 1 Assume that there exists a finite sequence $\{(y_n, x_n)\}_{n=1}^m$, such that the matrix product $M(y_1, x_1) \ldots M(y_m, x_m)$ is nonzero and subrectangular, where

$$M_{i,j}(y,x) = \begin{cases} P_{ij}^U & \text{if } g_{y,x}[(S,y,x)_j] = (y,x) \\ 0 & \text{otherwise} \end{cases} . \tag{4.17}$$

With this assumption we can apply Theorem 4.1.2.1 to π_n^U; thus, π_n^U converges in distribution to a limit which is independent of its initial distribution. By (4.15), this implies that π_n also converges in distribution, and its limit distribution is independent of π_0. We thus get the following lemma.

Lemma 4.1.2.1 For any bounded continuous function f, the following limits exist and are equal for all i:

$$\lim_{n\to\infty} E[f(\pi_n)] = \lim_{n\to\infty} E[f(\pi_n^i)]. \tag{4.18}$$

The subrectangularity condition on M is satisfied if for some input $x \in \mathcal{X}$ there exists a $y \in \mathcal{Y}$ such that $p_k(y|x) > 0$ for all k. It is also satisfied if all the elements of the matrix P are nonzero.

From (4.5) and (4.6), the limit distribution of π_n is a function of the i.i.d. input distribution $p(x)$. Let $\mathcal{P}(\mathcal{X})$ denote the set of all possible distributions on $\mathcal{X}$. The following lemma, proved in Appendix 4.A.2, shows that the limit distribution of π_n is continuous on $\mathcal{P}(\mathcal{X})$.

Lemma 4.1.2.2 Let μ^θ denote the limit distribution of π_n as a function of the i.i.d. distribution $\theta \in P(\mathcal{X})$. Then μ^θ is a continuous function of θ, so $\theta_m \to \theta$ implies that $\mu^{\theta_m} \to \mu^\theta$.

We now consider the convergence and continuity of ρ_n's distribution. Define the matrix N by

$$N_{i,j}(y) = \begin{cases} P_{ij}^U & \text{if } g_y[(S,y,x)_j] = y \\ 0 & \text{otherwise} \end{cases}, \tag{4.19}$$

and note that for any $y \in \mathcal{Y}$ and $x \in \mathcal{X}$,

$$M_{i,j}(y,x) = N_{i,j}(y) I(x_{(j)} = x). \tag{4.20}$$

To apply Theorem 4.1.2.1 to ρ_n^U, we must find a sequence $y_1, \ldots, y_l$ which yields a nonzero and subrectangular matrix for the product $N(y_1) \ldots N(y_l)$. Consider the projection onto $\mathcal{Y}$ of the sequence $\{(y_n, x_n)\}_{n=1}^m$ from Assumption 1. Let $\{y_n\}_{n=1}^m$ denote this projection, and define the matrices $M \triangleq M(y_1,x_1) \ldots M(y_m,x_m)$ and $N \triangleq N(y_1) \ldots N(y_m)$. Combining (4.20) and the fact that all the elements of M are nonnegative, it is easily shown that if $M_{i,j}$ is nonnegative for a particular i and j, then $N_{i,j}$ is nonnegative also. From this we deduce that if M is nonzero and subrectangular, then N must also be nonzero and subrectangular.

We can now apply Theorem 4.1.2.1 to ρ_n^U, which yields the convergence in distribution of ρ_n^U and thus ρ_n. Moreover, the limit distributions of these random vectors are independent of their initial states. Thus, we get the following result.

Lemma 4.1.2.3 For any bounded continuous function f, the following limits exist and are equal for all i:

$$\lim_{n\to\infty} E[f(\rho_n)] = \lim_{n\to\infty} E[f(\rho_n^i)]. \tag{4.21}$$

From (4.9) and (4.10), the limit distribution of ρ_n is also a function of $p(x)$. The following lemma, also proved in Appendix 4.A.2, shows that this limit distribution is also continuous on $\mathcal{P}(\mathcal{X})$.

Lemma 4.1.2.4 Let ν^θ denote the limit distribution of ρ_n as a function of the i.i.d. distribution $\theta \in \mathcal{P}(\mathcal{X})$. Then ν^θ is a continuous function of θ, so $\theta_m \to \theta$ implies that $\nu^{\theta_m} \to \nu^\theta$.

4.1.3 Entropy, Mutual Information, and Capacity

We now derive the capacity of the FSMC based on the distributions of π_n and ρ_n. We also obtain some additional properties of the entropy and mutual information when the channel inputs are i.i.d.

By definition, the Markov chain S_n is aperiodic and irreducible over the finite state space, so the effect of its initial state dies away exponentially with time [57]. Thus, the FSMC is an indecomposable channel [40, page 105]. From [40], the capacity of an indecomposable channel is independent of the initial state, and is given by

$$C = \lim_{n\to\infty} \max_{\mathcal{P}(\mathcal{X}^n)} \frac{1}{n} I(X^n; Y^n), \tag{4.22}$$

where $I(\cdot;\cdot)$ denotes mutual information and $\mathcal{P}(\mathcal{X}^n)$ denotes the set of all input distributions on $\mathcal{X}^n$. The mutual information can be written as

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n), \tag{4.23}$$

where $H(Y) = \mathbf{E}[-\log p(y)]$, and $H(Y|X) = \mathbf{E}[-\log p(y|x)]$. It is easily shown [42] that

$$H(Y^n) = \sum_{m=1}^{n} H(Y_m|Y^{m-1}) \tag{4.24}$$

and

$$H(Y^n|X^n) = \sum_{m=1}^{n} H(Y_m|X_m, Y^{m-1}, X^{m-1}). \tag{4.25}$$

Lemma 4.1.3.1

$$H(Y_n|X_n, X^{n-1}, Y^{n-1}) = \mathbf{E}\left[-\log \sum_{k=1}^{K} p(y_n|x_n, S_n = c_k)\pi_n(k)\right] = H(Y_n|X_n, \pi_n), \tag{4.26}$$

and

$$H(Y_n|Y^{n-1})) = \mathbf{E}\left[-\log \sum_{k=1}^{K} p(y_n|S_n = c_k)\rho_n(k)\right] = H(Y_n|\rho_n). \tag{4.27}$$

Proof We have

$$\begin{aligned} H(Y_n|X_n, X^{n-1}, Y^{n-1}) &= \mathbf{E}[-\log p(y_n|x_n, x^{n-1}, y^{n-1})] \\ &= \mathbf{E}\left[-\log \sum_{k=1}^{K} p(y_n|x_n, S_n = c_k)p(S_n = c_k|x^{n-1}, y^{n-1})\right] \\ &= \mathbf{E}\left[-\log \sum_{k=1}^{K} p(y_n|x_n, S_n = c_k)\pi_n(k)\right] \\ &= \mathbf{E}[-\log p(y_n|x_n, \pi_n)] \\ &= H(Y_n|X_n, \pi_n). \end{aligned} \tag{4.28}$$

The argument for (4.27) is the same, with all the x terms removed and π_n replaced by ρ_n. □

Using this Lemma in (4.24) and (4.25), and substituting into (4.23) yields the capacity in terms of the distributions of π_n and ρ_n; we summarize this in the following theorem.

Theorem 4.1.3.1 The capacity of the FSMC is given by

$$C = \lim_{n\to\infty} \max_{\mathcal{P}(\mathcal{X}^n)} \frac{1}{n} \sum_{i=1}^{n} \left[\mathbf{E}\left[-\log \sum_{k=1}^{K} p(y|S=c_k)\rho_i(k) \right] - \mathbf{E}\left[-\log \sum_{k=1}^{K} p(y|x,S=c_k)\pi_i(k) \right] \right], \tag{4.29}$$

where the dependence on $\theta \in \mathcal{P}(\mathcal{X}^n)$ of the distributions for π_i, ρ_i, and y is implicit.

Using Lemma 4.1.3.1, we can also express the capacity as

$$C = \lim_{n\to\infty} \max_{\mathcal{P}(\mathcal{X}^n)} \frac{1}{n} \sum_{i=1}^{n} [H(Y_i|\rho_i) - H(Y_i|X_i,\pi_i)]. \tag{4.30}$$

Although Gallager's theorem [40, page 109] guarantees the convergence of (4.29), the random vectors π_n and ρ_n do not necessarily converge in distribution for general input distributions. We proved this convergence in §4.1.2 for i.i.d. inputs. We now derive some additional properties of the entropy and mutual information under this input restriction.

Lemma 4.1.3.2 When the channel inputs are stationary,

$$\begin{aligned} H(Y_n|X_n,X^{n-1},Y^{n-1}) &\geq H(Y_{n+1}|X_{n+1},X^n,Y^n) \\ &\geq H(Y_{n+1}|X_{n+1},X^n,Y^n,S_0) \\ &\geq H(Y_n|X_n,X^{n-1},Y^{n-1},S_0). \end{aligned} \tag{4.31}$$

Proof We first note that the conditional entropy $H(Y|X)$ is a concave function of $p(y|x)$ for $p(x)$ fixed [42]. To show the first inequality, let f denote any concave function. Then

$$\begin{aligned} f(p[y_n|x_n,x^{n-1},y^{n-1}]) &\stackrel{a}{=} f(p[y_{n+1}|x_{n+1},x_2^n,y_2^n]) \\ &\stackrel{b}{=} f(\mathbf{E}(p[y_{n+1}|x_{n+1},x^n,y^n]|x_{n+1},x_2^n,y_2^n)) \\ &\stackrel{c}{\geq} \mathbf{E}(f(p[y_{n+1}|x_{n+1},x^n,y^n])|x_{n+1},x_2^n,y_2^n) \\ &\stackrel{d}{=} f(p[y_{n+1}|x_{n+1},x^n,y^n]), \end{aligned} \tag{4.32}$$

where a follows from the stationarity of the channel and the inputs, b and d follow from properties of conditional expectation [57], and c is a consequence of Jensen's inequality.

The second inequality results from the fact that conditioning on an additional random variable, in this case the initial state S_0, always reduces the entropy [42]. The proof of the third inequality is similar to that of the first:

$$\begin{aligned} f(p[y_{n+1}|x_{n+1},x^n,y^n,S_0]) &\stackrel{a}{=} f(\mathbf{E}(p[y_{n+1}|x_{n+1},x^n,y^n,S_1]|x_{n+1},x^n,y^n,S_0)) \\ &\stackrel{b}{=} f(\mathbf{E}(p[y_{n+1}|x_{n+1},x_2^n,y^{i_2},S_1]|x_{n+1},x^n,y^n,S_0)) \\ &\stackrel{c}{\geq} \mathbf{E}(f(p[y_{n+1}|x_{n+1},x_2^n,y_2^n,S_1])|x_{n+1},x^n,y^n,S_0) \\ &\stackrel{d}{=} f(p[y_{n+1}|x_{n+1},x_2^n,y_2^n,S_1]) \\ &\stackrel{e}{=} f(p[y_n|x_n,x^{n-1},y^{n-1},S_0]), \end{aligned} \tag{4.33}$$

where a and d follow from properties of conditional expectation, b follows from (2.42), c follows from Jensen's inequality, and e follows from the channel and input stationarity. □

Lemma 4.1.3.3 For i.i.d. input distributions, the following limits exist and are equal:

$$\lim_{n\to\infty} H(Y_n|X_n,X^{n-1},Y^{n-1}) = \lim_{n\to\infty} H(Y_n|X_n,X^{n-1},Y^{n-1},S_0). \tag{4.34}$$

Proof From Lemma 5.1,

$$\lim_{n\to\infty} H(Y_n|X_n,X^{n-1},Y^{n-1}) = \lim_{n\to\infty} \mathbf{E}\left[-\log\sum_{k=1}^{K} p(y|x,S=c_k)\pi_n(k)\right]. \tag{4.35}$$

Similarly,

$$\lim_{n\to\infty} H(Y_n|X_n,X^{n-1},Y^{n-1},S_0) = \lim_{n\to\infty} \mathbf{E}\left[-\log\sum_{k=1}^{K} p(y|x,S=c_k)\pi_n^*(k)\right], \tag{4.36}$$

where $\pi_n^* \stackrel{\Delta}{=} \pi_n^i$ for some i. Applying Lemma 4.1 to (4.35) and (4.36) completes the proof. □

We now consider the entropy in the output alone. The following lemma is proved using essentially the same argument as in Lemma 4.1.3.2 with all the x terms removed from (4.32) and (4.33); the details can be found in Appendix 4.A.3.

Lemma 4.1.3.4 For stationary inputs,

$$H(Y_n|Y^{n-1}) \geq H(Y_{n+1}|Y^n) \geq H(Y_{n+1}|Y^n,S_0) \geq H(Y_n|Y^{n-1},S_0). \tag{4.37}$$

Finally, we prove the analog of Lemma 4.1.3.3 for $H(Y_n|Y^{n-1})$.

Lemma 4.1.3.5 For i.i.d. input distributions, the following limits exist and are equal:

$$\lim_{n\to\infty} H(Y_n|Y^{n-1}) = \lim_{n\to\infty} H(Y_n|Y^{n-1}, S_0). \tag{4.38}$$

Proof Following a similar argument as in Lemma 4.1.3.3, we have that

$$\lim_{n\to\infty} H(Y_n|Y^{n-1}) = \lim_{n\to\infty} \mathbf{E}\left[-\log \sum_{k=1}^{K} p(y|S=c_k)\rho_n(k)\right], \tag{4.39}$$

and

$$\lim_{n\to\infty} H(Y_n|Y^{n-1}, S_0) = \lim_{n\to\infty} \mathbf{E}\left[-\log \sum_{k=1}^{K} p(y|S=c_k)\rho_n^*(k)\right], \tag{4.40}$$

where $\rho_n^* \stackrel{\Delta}{=} \rho_n^i$ for some i. Applying Lemma 4.1.2.3 to (4.39) and (4.40) completes the proof. □

Having established the basic properties of the entropies with i.i.d. inputs, we now evaluate I_{iid}.

Lemma 4.1.3.6 The mutual information maximized over all i.i.d. input distributions $P(\mathcal{X})$ is

$$\begin{aligned} I_{iid} &\stackrel{\Delta}{=} \lim_{n\to\infty} \max_{P(\mathcal{X})} \frac{1}{n} I(X^n;Y^n) \\ &= \lim_{n\to\infty} \max_{P(\mathcal{X})} [H(Y_n|\rho_n) - H(Y_n|X_n,\pi_n)] \\ &= \lim_{n\to\infty} \max_{P(\mathcal{X})} \left[\mathbf{E}\left[-\log \sum_{k=1}^{K} p(y|S=c_k)\rho_n(k)\right] - \mathbf{E}\left[-\log \sum_{k=1}^{K} p(y|x,S=c_k)\pi_n(k)\right]\right]. \end{aligned} \tag{4.41}$$

Proof For $\theta \in P(\mathcal{X})$ fixed,

$$H(Y^n|X^n) = \sum_{m=1}^{n} H(Y_m|X_m, Y^{m-1}, X^{m-1}) \tag{4.42}$$

by (4.25), and the terms of the summation are nonnegative and monotonically decreasing in m by Lemma 4.1.3.2. Thus

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} H(Y_m|X_m, Y^{m-1}, X^{m-1}) = \lim_{n\to\infty} H(Y_n|X_n, X^{n-1}, Y^{n-1}). \tag{4.43}$$

Similary, from (4.24),

$$H(Y^n) = \sum_{m=1}^{n} H(Y_m|Y^{m-1}), \tag{4.44}$$

and by Lemma 4.1.3.4, the terms of this summation are nonnegative and monotonically decreasing in m. Hence

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} H(Y_m|Y^{m-1}) = \lim_{n\to\infty} H(Y_n|Y^{n-1}). \tag{4.45}$$

The limits of (4.43) and (4.45) exist by Lemmas 4.1.3.3 and 4.1.3.5. Moreover, since $I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n)$, we can combine (4.42)-(4.45) to get that for any $\epsilon > 0$, there exists an N such that for all $n > N$,

$$\left| \max_{\mathcal{P}(\mathcal{X})} \left[\frac{1}{n} I(X^n;Y^n) \right] - \max_{\mathcal{P}(\mathcal{X})} \left[H(Y_n|Y^{n-1}) - H(Y_n|X_n, X^{n-1}, Y^{n-1}) \right] \right| < \epsilon. \tag{4.46}$$

The lemma follows by taking the limit of (4.46) as $n \to \infty$, and applying Lemma 4.1.3.1. □

Finally, the following theorem uses Lemmas 4.1.2.2 and 4.1.2.4 to interchange the limit and maximization in (4.41). Thus, we get I_{iid} in terms of the limit distributions on π and ρ.

Theorem 4.1.3.2

$$I_{iid} = \max_{\theta\in\mathcal{P}(\mathcal{X})} \left[\int_{\rho\in\Delta} \sum_{y\in\mathcal{Y}} \left(-\log p^\theta(y|\rho) \right) p^\theta(y|\rho) \nu^\theta(d\rho) - \right.$$

$$\left. \int_{\pi\in\Delta} \sum_{\substack{y\in\mathcal{Y} \\ x\in\mathcal{X}}} (-\log p(y|x,\pi)) p(y|x,\pi) \theta(x) \mu^\theta(d\pi) \right], \tag{4.47}$$

where ν^θ and μ^θ are the limit distributions of ρ and π, respectively, for input distribution θ, $p^\theta(y|\rho) = \sum_{k=1}^{K} \sum_{x\in\mathcal{X}} p(y|x, S = c_k)\theta(x)\rho(k)$, and $p(y|x,\pi) = \sum_{k=1}^{K} p(y|x, S = c_k)\pi(k)$.

Proof For an i.i.d. input distribution of θ, let $\nu_n^\theta = p(\rho_n)$ and $\mu_n^\theta = p(\pi_n)$. Using the notation of (4.47), we can then rewrite (4.41) as

$$I_{iid} = \lim_{n\to\infty} \max_{\theta\in\mathcal{P}(\mathcal{X})} \left[\int_{\rho\in\Delta} \sum_{y\in\mathcal{Y}} \left(-\log p^\theta(y|\rho) \right) p^\theta(y|\rho) \nu_n^\theta(d\rho) - \right.$$

$$\left. \int_{\pi\in\Delta} \sum_{\substack{y\in\mathcal{Y} \\ x\in\mathcal{X}}} (-\log p(y|x,\pi)) p(y|x,\pi) \theta(x) \mu_n^\theta(d\pi) \right]. \tag{4.48}$$

For any integer m, let $\mathcal{P}_m(\mathcal{X}) \subset P(\mathcal{X})$ be a finite subset defined as follows: For any distribution $\psi \in \mathcal{P}(\mathcal{X})$, $\psi \in \mathcal{P}_m(\mathcal{X})$ iff for all $x \in \mathcal{X}$ there exists a set of integers $k_x \leq m$ such that $\psi(x) = k_x/m$. Clearly, $\lim_{m\to\infty} \mathcal{P}_m(\mathcal{X}) = \mathcal{P}(\mathcal{X})$. Substituting this into (4.48), we get

$$I_{iid} = \lim_{n\to\infty} \lim_{m\to\infty} \max_{\theta_m \in \mathcal{P}_m(\mathcal{X})} \left[\int_{\rho\in\Delta} \sum_{y\in\mathcal{Y}} \left(-\log p^{\theta_m}(y|\rho)\right) p^{\theta_m}(y|\rho) \nu_n^{\theta_m}(d\rho) - \right.$$

$$\left. \int_{\pi\in\Delta} \sum_{\substack{y\in\mathcal{Y} \\ x\in\mathcal{X}}} (-\log p(y|x,\pi)) p(y|x,\pi) \theta_m(x) \mu_n^{\theta_m}(d\pi) \right]. \quad (4.49)$$

Let $\theta^* \in P(\mathcal{X})$ be the distribution that achieves the maximum in (4.47), and $\theta_m^* \in \mathcal{P}_m(\mathcal{X})$ be the distribution that achieves the maximum in (4.49). Then $\lim_{m\to\infty} \theta_m^* = \theta^*$. With this notation, we can rewrite (4.49) as follows.

$$I_{iid} = \lim_{n\to\infty} \lim_{m\to\infty} \left[\int_{\rho\in\Delta} \sum_{y\in\mathcal{Y}} \left(-\log p^{\theta_m^*}(y|\rho)\right) p^{\theta_m^*}(y|\rho) \nu_n^{\theta_m^*}(d\rho) - \right.$$

$$\left. \int_{\pi\in\Delta} \sum_{\substack{y\in\mathcal{Y} \\ x\in\mathcal{X}}} (-\log p(y|x,\pi)) p(y|x,\pi) \theta_m^*(x) \mu_n^{\theta_m^*}(d\pi) \right]. \quad (4.50)$$

But $\lim_{n\to\infty} \mu_n^{\theta_m^*} = \mu^{\theta_m^*}$ by definition of μ_n and μ, and $\lim_{m\to\infty} \nu^{\theta_m^*} = \nu^{\theta^*}$ by Lemma 4.1.2.4. Moreover, by the triangle inequality,

$$|\mu_n^{\theta_m^*} - \mu^{\theta^*}| \leq |\mu_n^{\theta_m^*} - \mu^{\theta_m^*}| + |\mu^{\theta_m^*} - \mu^{\theta^*}|, \quad (4.51)$$

so

$$\lim_{n,m\to\infty} \mu_n^{\theta_m^*} = \mu^{\theta^*}. \quad (4.52)$$

Similarly, using Lemma 4.1.2.4 we get that

$$\lim_{n,m\to\infty} \nu_n^{\theta_m^*} = \nu^{\theta^*}. \quad (4.53)$$

Finally, $p(y)$ is linear in $p(x)$, so

$$\lim_{m\to\infty} p^{\theta_m^*}(y|\rho) = p^{\theta^*}(y|\rho). \quad (4.54)$$

Since both bracketed terms in (4.50) are bounded by $\log|\mathcal{Y}|$, we can bring the limits (4.52), (4.53). and (4.54) inside the integral and summation, yielding

$$I_{iid} = \left[\int_{\rho \in \Delta} \sum_{y \in \mathcal{Y}} \left(-\log p^{\theta^*}(y|\rho) \right) p^{\theta^*}(y|\rho) \nu^{\theta^*}(d\rho) - \right.$$

$$\left. \int_{\pi \in \Delta} \sum_{\substack{y \in \mathcal{Y} \\ x \in \mathcal{X}}} (-\log p(y|x,\pi)) p(y|x,\pi) \theta^*(x) \mu^{\theta^*}(d\pi) \right]. \qquad (4.55)$$

The theorem then follows from the definition of θ^*. □

Note that by definition, I_{iid} is a lower bound for the capacity C given by (4.29).

4.2 Uniformly Symmetric Variable Noise Channels

The Gilbert-Elliot channel has two features which facilitate a closed-form solution to its capacity: its conditional entropy $H(Y^n|X^n)$ is independent of the input distribution, and it is a symmetric channel, so a uniform input distribution induces a uniform output distribution. We now define two classes of FSMCs, uniformly symmetric channels, and variable noise channels, which each have one of these features. The mutual information and capacity of these channel classes have additional properties which we outline in the lemmas below. We also show that for the class of FSMCs with both of these features, called uniformly symmetric variable noise channels, I_{iid} equals the channel capacity. Moreover, we will see in the next section that the decision-feedback decoder achieves capacity for uniformly symmetric variable noise FSMCs. FSMCs with symmetric PSK inputs and variation due to amplitude fading or quantized variable-power additive white noise are contained in this channel class.

Definition: For a discrete memoryless channel, let M denote the matrix of input/output probabilities, where $M_{ij} \triangleq p(y = j|x = i)$, $j \in \mathcal{Y}$, $i \in \mathcal{X}$. A discrete memoryless channel is *output symmetric* if the rows of M are permutations of each other, and the columns of M are permutations of each other[2].

Definition: A FSMC is *uniformly symmetric* if every channel $c_k \in \mathcal{C}$ is output symmetric.

---

[2]Symmetric channels, defined in [40, p. 94], are a more general class of memoryless channels: an output symmetric channel is a symmetric channel with a single output partition.

Lemma 4.2.1 For uniformly symmetric FSMCs, $H(Y_n|\rho_n)$, $H(Y_n|\rho_n^*)$, $H(Y_n|\pi_n)$, and $H(Y_n|\pi_n^*)$ are all maximized for $p(x^n)$ uniform and i.i.d., and these maximum values equal $\log|\mathcal{Y}|$.

Proof From [42], $H(Y_n|\rho_n) \leq H(Y_n) \leq \log|\mathcal{Y}|$ and similarly $H(Y_n|\rho_n^*) \leq H(Y_n) \leq \log|\mathcal{Y}|$. But since each $c_k \in \mathcal{C}$ is output symmetric, for each k the columns of M^k are permutations of each other. Thus, if the marginal $p(x_n)$ is uniform, then $p(y_n|S_n = c_k)$ is also uniform, so $p(y_n|S_n = c_k) = 1/|\mathcal{Y}|$. We therefore have that for any $\rho_n \in \Delta$,

$$p(y_n|\rho_n) = \sum_{k=1}^{K} p(y_n|S_n = c_k)\rho_n(k) = \frac{1}{|\mathcal{Y}|}\sum_{k=1}^{K}\rho_n(k) = \frac{1}{|\mathcal{Y}|}, \tag{4.56}$$

and similarly, $p(y_n|\rho_n^*) = 1/|\mathcal{Y}|$. Thus,

$$\begin{aligned} H(Y_n|\rho_n) &= \sum_{\rho_n\in\Delta}\sum_{y_n\in\mathcal{Y}} p(\rho_n)p(y_n|\rho_n)\left[-\log p(y_n|\rho_n)\right] \\ &= \sum_{\rho_n\in\Delta} p(\rho_n)\sum_{y_n\in\mathcal{Y}} p(y_n|\rho_n)\left[-\log p(y_n|\rho_n)\right] \\ &= \sum_{\rho_n\in\Delta} p(\rho_n)\sum_{y_n\in\mathcal{Y}}\frac{1}{|\mathcal{Y}|}\log|\mathcal{Y}| \\ &= \log|\mathcal{Y}|, \end{aligned} \tag{4.57}$$

and similarly. $H(Y_n|\rho_n^*) = \log|\mathcal{Y}|$. Since (4.57) only requires that $p(x_n)$ is uniform for each n, an i.i.d. uniform input distribution achieves this maximum. Substituting π for ρ in the above argument yields the result for $H(Y_n|\pi_n)$ and $H(Y_n|\pi_n^*)$. □

Definition: Let X_n and Y_n denote the input and output, respectively, of a FSMC. We say that a FSMC is a *variable noise channel* if there exists a function f such that for $Z_n = f(X_n, Y_n)$, Z^n is a sufficient statistic[3] for S^n, and $p(Z^n|X^n) = p(Z^n)$.

If Z^n is a sufficient statistic for S^n, then

$$\pi_n = p(S_n|X^{n-1}, Y^{n-1}, Z^{n-1}) = p(S_n|Z^{n-1}). \tag{4.58}$$

Using (4.58), and replacing the pairs (X_n, Y_n) with Z_n in the derivation of Appendix 4.A.1, we get the recursive relation

$$\pi_{n+1} = \frac{\pi_n D(z_n)P}{\pi_n D(z_n)\underline{1}} \triangleq f(z_n, \pi_n), \tag{4.59}$$

[3] $Z = f(X, Y)$ is a sufficient statistic for S if S is independent of X and Y given Z [42].

where $D(z_n)$ is a diagonal $K \times K$ matrix with kth diagonal term $p(z_n|S_n = c_k)$. From (4.59), π_n is a Markov chain with state space Δ and transition probabilities

$$p(\pi_{n+1} = \alpha|\pi_n = \beta) = \sum_{z_n \in \mathcal{Z}} 1[(z_n) : f(z_n, \beta) = \alpha] p(z_n|\pi_n = \beta). \tag{4.60}$$

Lemma 4.2.2 For uniformly symmetric variable noise FSMCs, $H(Y_n|X_n, \pi_n)$ and $H(Y_n|X_n, \pi_n^*)$ don't depend on $p(x^n)$.

Proof We consider only $H(Y_n|X_n, \pi_n)$, since the same argument applies to $H(Y_n|X_n, \pi_n^*)$. Since each $c_k \in \mathcal{C}$ is output symmetric, the sets $\{\{p_k(y|x) : y \in \mathcal{Y}\}; x \in \mathcal{X}\}$ are permutations of each other. Thus,

$$\begin{aligned} H(Y_n|X_n, \pi_n) &= \sum_{\pi_n} \sum_{x_n} \sum_{y_n} \left(-\log \sum_k p_k(y_n|x_n)\pi_n(k) \right) \sum_k p_k(y_n|x_n)\pi_n(k)p(x_n)p(\pi_n) \\ &= \sum_{\pi_n} \sum_{y_n} \left(-\log \sum_k p_k(y_n|x_n)\pi_n(k) \right) p_k(y_n|x_n)p(\pi_n). \end{aligned} \tag{4.61}$$

So $H(Y_n|X_n, \pi_n)$ depends only on the distribution of π_n. But by (4.60), this distribution depends only on the distribution of Z^{n-1}. The proof then follows from the fact that $p(Z^n|X^n) = p(Z^n)$. □

Consider a FSMC where each $c_k \in \mathcal{C}$ is an additive white noise channel with noise n_k. If we let $Z = Y - X$, then it is easily checked that this is a variable noise channel. For such channels, however, the output alphabet $\mathcal{Y}$ is infinite. In general, the output of an additive white noise channel is quantized to the nearest symbol in a finite output alphabet; we call this the quantized additive white noise (Q-AWN) channel.

If the Q-AWN channel has a symmetric multiphase input alphabet and output phase quantization [58, page 80], then it is easily checked that $p_k(y|x)$ depends only on $p_k(|y - x|)$, which in turn depends only on the noise n_k; thus, it is a variable noise channel[4]. We show in Appendix 4.A.4 that variable noise Q-AWN channels with the same input and output alphabets are also uniformly symmetric. Uniformly symmetric variable noise channels have the property that I_{iid} equals the channel capacity, as we show in the following

[4]If the input alphabet of a Q-AWN channel is not symmetric or the input symbols have different amplitudes, then the distribution of $Z = |Y - X|$ will depend on the input. To see this, consider a Q-AWN channel with a 16-QAM input/output alphabet (so the output is quantized to the nearest input symbol). There are four different sets of $Z = |Y - X|$ values, depending on the amplitude of the input symbol. Thus, the distribution of Z over all its possible values (the union of all four sets) will change, depending on the amplitude of the input symbol.

lemma.

Lemma 4.2.3 Capacity of uniformly symmetric variable noise channels is achieved with uniform i.i.d. inputs, so $C = I_{iid}$. Moreover, $C = \lim_{n\to\infty} C_n = \lim_{n\to\infty} C_n^*$, where

$$C_n \triangleq \max_{P(\mathcal{X}^n)} H(Y_n|\rho_n) - H(Y_n|X_n, \pi_n) \tag{4.62}$$

increases with n, and

$$C_n^* \triangleq \max_{P(\mathcal{X}^n)} H(Y_n|\rho_n^*) - H(Y_n|X_n, \pi_n^*) \tag{4.63}$$

decreases with n.

Proof From Lemmas 4.2.1 and 4.2.2, C_n, C_n^*, and C are all maximized with uniform i.i.d. inputs. With this input distribution, $C_n = \log|\mathcal{Y}| - H(Y_n|X_n, \pi_n)$ and $C_n^* = \log|\mathcal{Y}| - H(Y_n|X_n, \pi_n^*)$. Applying Lemmas 4.1.3.2 and 4.1.3.3, we get that $H(Y_n|X_n, \pi_n)$ decreases with n, $H(Y_n|X_n, \pi_n^*)$ increases with n, and both converge to the same limit, which completes the proof. □

The BSC is equivalent to a binary input Q-AWN channel with binary quantization [58]. Thus, a FSMC where c_k indexes a set of BSCs with different crossover probabilities is a uniformly symmetric variable noise channel. Therefore, Proposition 4 of [54] is a corollary of Lemma 4.2.3. Moreover, Lemma 4.2.3 holds for FSMCs where $\mathcal{C}$ consists of any finite number of BSCs.

4.3 Decision-Feedback Decoding

In principle, communication over a finite-state channel is possible at any rate below the channel capacity. However, good maximum-likelihood coding strategies for channels with memory are difficult to determine, and the decoder complexity grows exponentially with memory length. Thus, a common strategy for channels with memory is to disperse the memory using an interleaver; if the span of the interleaver is long, then the cascade of the interleaver, channel, and deinterleaver can be considered memoryless, and coding techniques for memoryless channels may be used [58, page 115]. However, this cascaded channel has a lower inherent Shannon capacity than the original channel, since one is restricted to memoryless channel codes.

The complexity of maximum-likelihood decoding can be reduced significantly without this capacity degradation by implementing a *decision-feedback decoder*. Figure 4.2 shows a block diagram for a system with decision-feedback decoding. The system is composed of a conventional encoder for memoryless channels, block interleaver, FSMC, deinterleaver, and a decision-feedback decoder. Figure 4.3 outlines the decision-feedback decoder design, which consists of a channel state estimator followed by a maximum-likelihood decoder. We now show that, if we ignore error propagation and decoding delay, a system employing this decision-feedback decoding scheme on uniformly symmetric variable noise channels is *information lossless*: it has the same capacity as the original FSMC, given by (4.41). Moreover, we will see that the output of the state estimator is a sufficient statistic for the deinterleaver output, given all past inputs and outputs. Therefore, the maximum-likelihood decoder input (y_n, π_n), conditioned on x_n, is independent of x^{n-1}. We can thus determine the maximum-likelihood input sequence on a symbol-by-symbol basis, eliminating the complexity and delay of sequence decoders. We will also calculate the capacity penalty of the decision-feedback decoder for general FSMCs (ignoring error propagation), and the system cutoff rate.

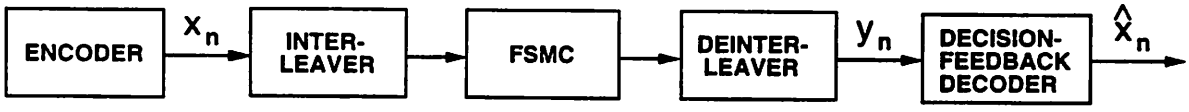

Figure 4.2: System Model

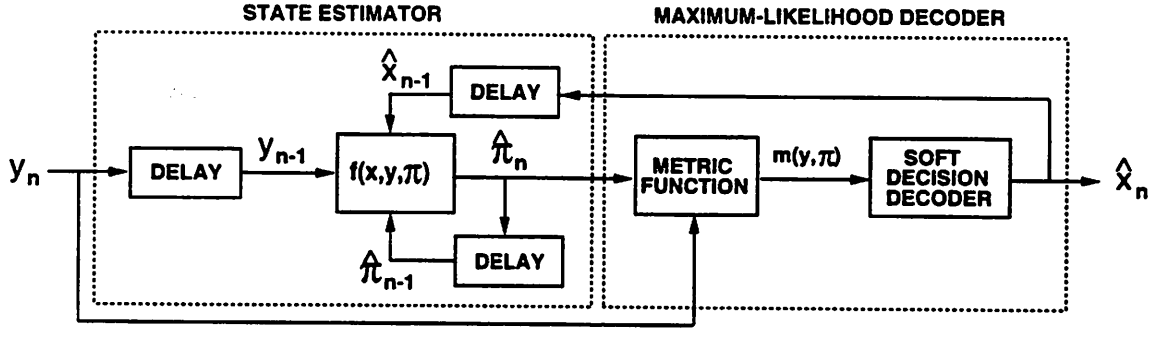

Figure 4.3: Decision-Feedback Decoder

The interleaver works as follows. The output of the encoder is stored row by row in a $J \times L$ interleaver, and transmitted over the channel column by column. The

deinterleaver performs the reverse operation. Because the effect of the initial channel state dies away, the received symbols within any row of the deinterleaver become independent as J becomes infinite. However, the symbols within any column of the interleaver are received from consecutive channel uses, and are thus dependent. This dependence is called the *latent* channel memory, and the state estimator enables the maximum-likelihood decoder to make use of this memory.

Specifically, the state estimator uses the recursive relationship of (4.5) to estimate π_n. It will be shown below that the maximum-likelihood decoder operates on a memoryless system, and can therefore determine the maximum-likelihood input sequence on a per symbol basis. The input to the maximum-likelihood decoder is the channel output y_n and the state estimate $\hat{\pi}_n$, and its output is the x_n which maximizes $\log p(y_n, \hat{\pi}_n | x_n)$, assuming equally likely input symbols[5]. The soft decision decoder uses conventional Viterbi decoding techniques with branch metrics

$$m(y, \pi) \triangleq \log p(y, \pi | x). \tag{4.64}$$

If the input sequence is coded, then there will be some delay in the soft-decision decoder's calculation of $\hat{x}_n$, so the decision will not be immediately available to feed back to the state estimator. The identical problem affects decision-feedback equalizers (DFEs) [59]. Recent DFE designs which alleviate this problem include parallel DFEs, which keep track of all possible symbol decisions [60], and interleaver/deinterleaver pairs, which rearrange the order of received symbols prior to decoding such that delayed reliable decisions can be used for feedback [61]. We assume that the same techniques can be applied to our decision-feedback decoder, thus we ignore decoding delay in this analysis. Error propagation analysis of DFEs may also help to determine the effect of wrong decisions in our decoder performance [62], which we ignore in this study.

We now evaluate the information, capacity, and cutoff rates of a system using the decision-feedback decoder, assuming $\hat{\pi}_n = \pi_n$ (i.e., ignoring error propagation). We will use the notation $y_{jl} \triangleq y_n$ to explicitly denote that y_n is in the jth row and lth column of the deinterleaver. Similarly $\pi_{jl} \triangleq \pi_n$ and $x_{jl} \triangleq x_n$ denote, respectively, the state estimate and interleaver input corresponding to y_{jl}. Assume now that the state estimator is reset every J iterations, so for each l, the state estimator goes through j recursions of (4.5) to

[5] If the x_n are not equally likely, then $\log p(x_n)$ must be added to the decoder metric.

calculate π_{jl}. By (4.6), this recursion induces a distribution $p(\pi_{jl})$ on π_{jl} that depends only on $p(\mathcal{X}^j)$. Thus, the system up to the output of the state estimator is equivalent to a set of parallel *π-output* channels, where the π-output channel is defined, for a given j, by the input x_{jl}, the output pair (y_{jl}, π_{jl}), and the input/output probability

$$p(y_{jl}, \pi_{jl}|x_{jl}) = \sum_k p_k(y_{jl}|x_{jl})\pi_{jl}(k)p(\pi_{jl}). \tag{4.65}$$

For each j, the π-output channel is the same for $l = 1, 2, \ldots, L$, and therefore there are J different π-output channels, each used L times. The first π-output channel ($j = 1$) is equivalent to the FSMC with interleaving and memoryless channel encoding, since the estimator is reset and therefore $\pi_{1l} = \pi_0, \quad 1 \leq l \leq L$.

The jth π-output channel is discrete, since x_{jl} and y_{jl} are taken from finite alphabets, and since π_{jl} can have at most $|\mathcal{X}|^j|\mathcal{Y}|^j$ different values. It is also asymptotically memoryless with deep interleaving (large J), which we prove in Appendix 4.A.5. Finally, we show in Appendix 4.A.6 that for $p(\mathcal{X}^J)$ fixed, the J π-output channels are independent, and the average mutual information of the parallel channels is

$$I_J = \frac{1}{J} I(Y^J, \pi^J; X^J) = \frac{1}{J} \sum_{j=1}^{J} H(Y_j|\pi_j) - H(Y_j|X_j, \pi_j). \tag{4.66}$$

Let

$$C_J \triangleq \max_{P(\mathcal{X}^J)} \frac{1}{J} \sum_{j=1}^{J} H(Y_j|\pi_j) - H(Y_j|X_j, \pi_j) = \max_{P(\mathcal{X}^J)} \frac{1}{J} \sum_{j=1}^{J} C_j, \tag{4.67}$$

where

$$C_j \triangleq H(Y_j|\pi_j) - H(Y_j|X_j, \pi_j), \tag{4.68}$$

for the maximizing distribution $p(\mathcal{X}^J)$. The capacity of the decision-feedback decoding system is then

$$C_{df} \triangleq \lim_{J \to \infty} C_J \tag{4.69}$$

Comparing (4.69) to (4.30), we see that the $H(Y|X, \pi)$ terms are common to C and C_{df}. Therefore, an upper bound for the capacity penalty of the decision-feedback decoder is

$$C - C_{df} \leq \lim_{J \to \infty} \max_{P(\mathcal{X}^J)} \frac{1}{J} \sum_{j=1}^{J} [H(Y_j|\rho_j) - H(Y_j|\pi_j)]. \tag{4.70}$$

Let $I_{iid(df)}$ denote the mutual information I_J of the decision-feedback decoder for i.i.d. inputs. Then

$$I_{iid} - I_{iid(df)} \leq \lim_{J\to\infty} \max_{P(\mathcal{X})} \frac{1}{J} \sum_{j=1}^{J} [H(Y_j|\rho_j) - H(Y_j|\pi_j)]. \tag{4.71}$$

By Lemmas 4.1.3.3 and 4.1.3.5, $H(Y_j|\pi_j)$ and $H(Y_j|\rho_j)$ converge for i.i.d. inputs. Moreover, $H(Y_j|\rho_j)$ is monotonically decreasing in j by Lemma 4.1.3.4, and an argument similar to that of Lemma 4.1.3.2 shows that $H(Y_j|\pi_j)$ is also. Thus, the limit in (4.71) can be moved inside the summation, yielding the following upper bound for the rate penalty of a decision-feedback decoder with i.i.d. inputs:

$$I_{iid} - I_{iid(df)} \leq \max_{\theta\in P(\mathcal{X})} \int_{\rho\in\Delta} \left[\sum_{y\in\mathcal{Y}} \left(-\log p^{\theta}(y|\rho)\right) p^{\theta}(y|\rho) \right] \left(\nu^{\theta}(d\rho) - \mu^{\theta}(d\rho)\right), \tag{4.72}$$

where ν^{θ} and μ^{θ} are as defined in Theorem 4.1.3. Finally, by Lemma 4.2.1, I_{iid} and $I_{iid(df)}$ are maximized with uniform input distributions. Thus, the bracketed summation in (4.71) equals $\log|\mathcal{Y}|$ for any ρ. Therefore, the right side of (4.72) vanishes for uniformly symmetric channels. Moreover, since uniformly symmetric variable noise channels have $C = I_{iid}$, the decision-feedback decoder preserves the inherent capacity of such channels.

Although capacity gives the maximum data rate for any maximum-likelihood encoding scheme, established coding techniques generally operate at or below the channel cutoff rate [58]. Since the π-output channels are independent for fixed $p(\mathcal{X}^J)$, the random coding exponent for the parallel set is

$$\mathbf{E}_o(1, p(\mathcal{X}^J)) = \sum_{j=1}^{J} R_j, \tag{4.73}$$

where

$$R_j = -\log \sum_{y,\pi} \left[\sum_{x} p(x_j) \sqrt{\sum_{k=1}^{K} p(y|x, S = c_k)\pi_j(k)p(\pi_j)} \right]^2. \tag{4.74}$$

The cutoff rate of the decision-feedback decoding system is

$$R_{df} \triangleq \lim_{J\to\infty} \max_{P(\mathcal{X}^J)} \frac{1}{J} \sum_{j=1}^{J} R_j. \tag{4.75}$$

We show in Appendix 4.A.7 that for uniformly symmetric variable noise channels, the maximizing input distribution in (4.75) is uniform and i.i.d., the resulting value of R_j is increasing in j, and the cutoff rate R_{df} becomes

$$R_{df} = \lim_{j\to\infty} R_j = -\log \sum_{y,\pi\in\Delta} \left[\sum_{x} \frac{1}{|\mathcal{X}|} \sqrt{\sum_{k=1}^{K} p(y|x, S = c_k)\pi(k)\mu(\pi)} \right]^2, \tag{4.76}$$

where μ is the invariant distribution for π under i.i.d. uniform inputs.

4.4 Capacity and Cutoff Rates for a Two-State Variable Noise Channel

We now compute the capacity and cutoff rate of a two-state Q-AWN channel with variable SNR, Gaussian noise, and 4-PSK modulation. The variable SNR can represent different fading levels in a multipath channel, or different noise and/or interference levels. The model is shown in Figure 4.4. The input to the channel is a 4-PSK symbol, to which noise of variance n_G or n_B is added, depending on whether the channel is in state G (good) or B (bad). We assume that the SNR is 10dB for channel G, and -5dB for channel B. The channel output is quantized to the nearest input symbol, so from sections 4.2 and 4.3, the capacity and cutoff rates are achieved with uniform i.i.d. inputs. The state transition probabilities are depicted in Figure 4.4. We assume a stationary initial distribution of the state process, so $p(S_0 = G) = g/(g+b)$ and $p(S_0 = B) = b/(g+b)$.

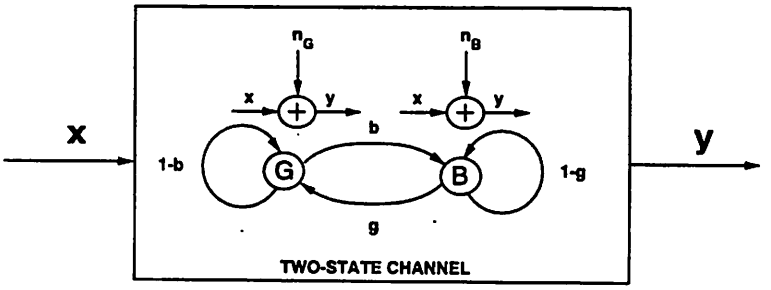

Figure 4.4: Two-State Fading Channel

Figure 4.5 shows the iterative calculation of (4.6) for $p(\pi_n(G) = \alpha)$, where $\pi_n(G) = p(S_n = G|x^{n-1}, y^{n-1})$. In this example, the difference of subsequent distributions after 15 recursions is below the quantization level ($d\alpha = .01$) of the graph. Figure 4.6 shows the capacity (C_j) and cutoff rate (R_j) of the jth π-output channel, given by (4.68) and (4.74) respectively. Note that $C_{j=1}$ and $R_{j=1}$ in this figure are the capacity and cutoff rate of the FSMC with interleaving and memoryless channel encoding; thus, the difference between the initial and final values of C_j and R_j indicate the performance improvement of the decision-feedback decoder over conventional techniques.

For this two-state model, the channel memory can be quantified by the parameter $\mu \triangleq 1 - g - b$, since for $\sigma \in \{G, B\}$ [54],

$$p(S_n = \sigma|S_0 = \sigma) - p(S_n = \sigma|S_0 \neq \sigma) = \mu^n. \tag{4.77}$$

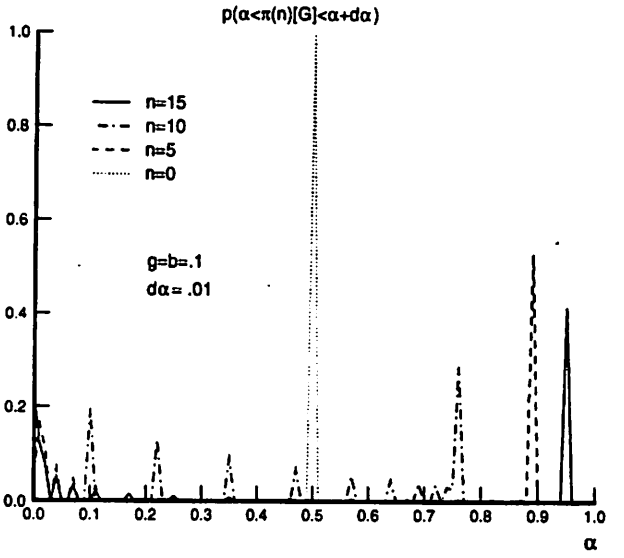

Figure 4.5: Recursive Distribution of π_n

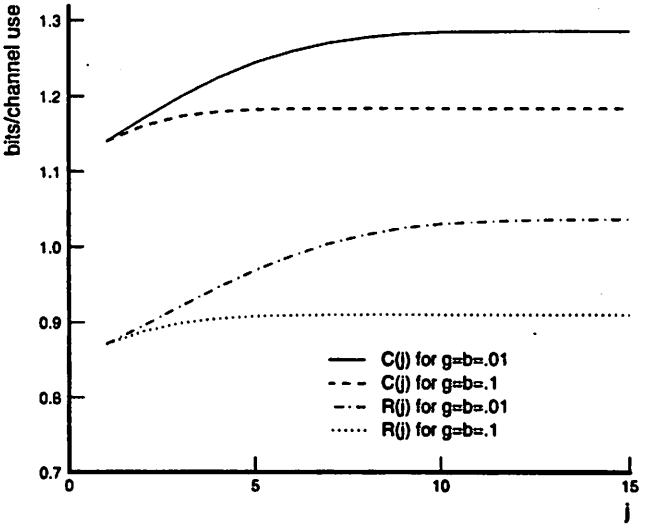

Figure 4.6: Capacity and Cutoff Rate for jth π-Output Channel

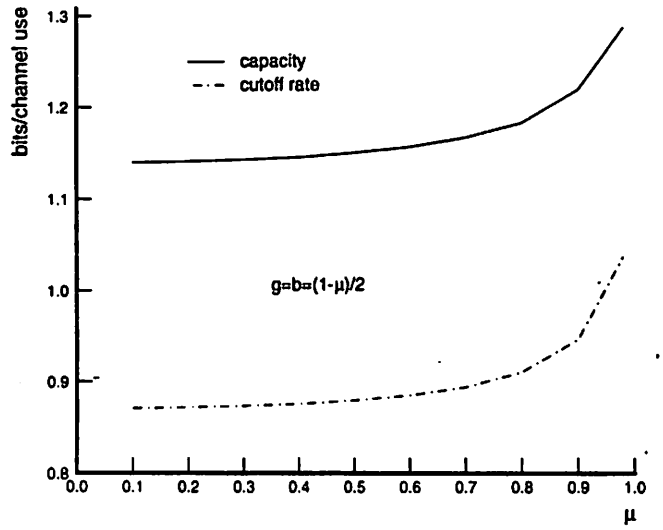

Figure 4.7: Decoder Performance versus Channel Memory

In Figure 4.7, we show the decision-feedback decoder's capacity and cutoff rate (C_{df} and R_{df} respectively) as a function of μ. We expect these performance measures to increase as μ increases, since more latency in the channel should improve the accuracy of the state estimator; Figure 4.7 confirms this hypothesis. Finally, in Figure 4.8 we show the decision-feedback decoder's capacity and cutoff rates as a function of g. The parameter g is inversely proportional to the average number of consecutive bad channel states (which correspond to the 15dB fading channel); thus, Figure 4.8 can be interpreted as the relationship between the maximum transmission rate and the average fade duration.

4.5 Unequal Error Protection Codes for Fading Channels

We now consider the case where the correlation structure of the channel variation is unknown. In this case, the channel state varies *arbitrarily* over its state space. The capacity of such Arbitrarily Varying Channels (AVCs) was first studied by Blackwell, Breiman and Thomasian [63]; more recent treatments can be found in [55, 64], and the references therein. In general, the capacity-achieving code of an AVC assumes the worst-case channel state for probability of error calculation. Similarly, practical code designs for time-varying channels typically specify a (maximum or average) error probability for *all* received data

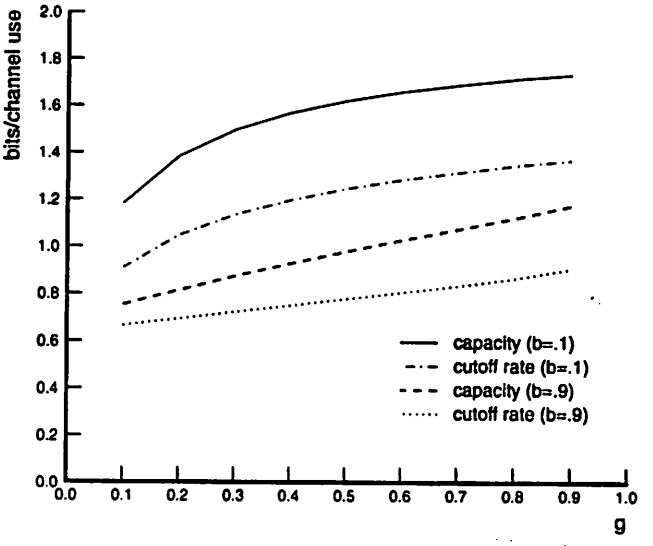

Figure 4.8: Decoder Performance versus g

bits. A higher data rate can be achieved if the input sequence is prioritized into high- and low-priority bit streams, where the error probability of the high-priority stream is lower than that of the low-priority stream. Then, even under worst-case channel conditions, the high-priority bits will get through. This type of channel coding requires bit prioritization by the source encoder, which is inherent to some voice and video compression schemes, such as sub-band coding [65]. It can also be applied to heterogeneous traffic streams with different BER criterion, like voice and data. We now explore some of these Unequal Error Protection (UEP) techniques for fading channels. We first derive the maximum average data rate of a narrowband fading channel with optimal UEP coding. We then describe two practical implementations of UEP coding: time-multiplexing of coded bit streams, and coded modulation with multiresolution codes.

4.5.1 Performance Limits

We assume that the fading channel under consideration is constant for the duration of a symbol transmission, and that the range of the received SNR can be partitioned into a finite number of intervals[6]. Thus, the channel can be modeled as a discrete-time state space

[6]This model was used to characterize Rayleigh fading in [67].

channel, where each state is an AWGN channel with a different SNR, as in the two-state variable noise channel of §4.4.

Let K denote the number of channel states, and n_i denote the noise power associated with state c_i, where the n's are increasing ($n_i < n_j$ for $i < j$). Define the *incremental* noise power by

$$\nu_i \triangleq \begin{cases} n_i & i = 1 \\ n_i - n_{i-1} & i > 1 \end{cases}, \tag{4.78}$$

so $n_i = \sum_{j=1}^{i} \nu_i$. Since the AWGN channel c_i has the noise power $\sum_{j=1}^{i} \nu_j$, the set of channels $c_1, \ldots, c_K$ with common input can be represented by the incremental noise channel shown in Figure 4.9.

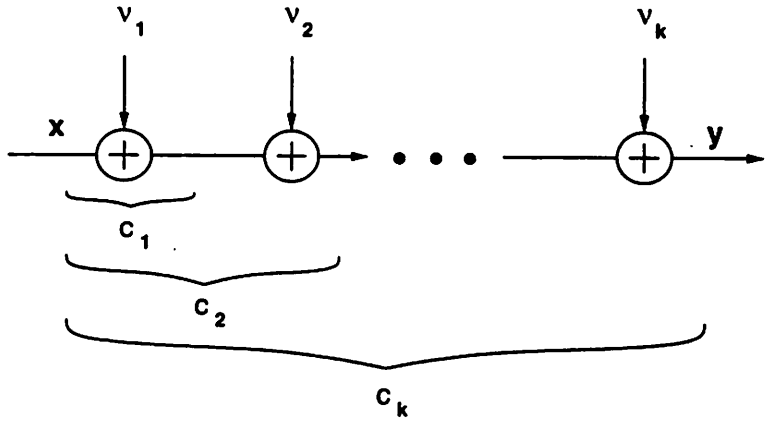

Figure 4.9: Incremental Noise Channel

The fading channel can be considered as an incremental noise channel with only one of the c_i channels active on each transmission (the c_i corresponding to the current fade level). We will use this fact below to determine the average rate of the fading channel from the rate region of the incremental noise channel, which we now obtain. The maximum rate of the incremental noise channel can be considered in the more general framework of degraded broadcast channels [68]. The degraded broadcast channel models a system with one transmitter and many receivers sharing a particular frequency band, where the channel quality between the transmitter and each receiver is different, as would generally be the case when the receivers are in different locations. The goal of the transmitter is to send as much information as possible to each of the receivers. Thus, the incremental noise channel models a system of one transmitter and multiple receivers, where receiver i obtains

the transmitted signal plus AWGN of power n_i. Assume for now that the transmitter sends independent information to each receiver. The rate vector $(R_1, \ldots, R_K)$ that can be achieved simultaneously on the set of incremental noise channels $c_1, \ldots, c_K$ is refered to as its *rate region*, and the maximum rate region is called the *capacity region* [42]. The capacity region of the channel of Figure 4.9 was determined by Bergmans [69] to be the convex hull of all rate vectors $(R_1, \ldots, R_K)$, where R_i is given by

$$R_i = \log\left(1 + \frac{\alpha_i P}{n_i B + \sum_{j<i} \alpha_j P}\right) \tag{4.79}$$

for any set of α_is that satisfy $\sum_{i=1}^{K} \alpha_i = 1$, where B and P are, respectively, the total channel bandwidth and power allocated to the channel set, and α_i is the fraction of power allocated to channel c_i.

The intuitive explanation for (4.79) is the following [69]. Since $n_i < n_j$ for $i < j$, user i correctly receives all the data transmitted to user j. Therefore, user i can correctly decode and then subtract out user j's message, then decode its own message. However, user j cannot decode the message intended for user i, since it has a less-favorable channel; thus, user i's message, with power $\alpha_i P$, contributes an additional noise term to user j's received message. This explains the additional noise terms in the denominator of (4.79).

This capacity region is achieved by *superposition* codes, which form the theoretical basis for the multiresolution coded modulation described in the next section. Superposition codes are constructed by using multiple codebooks to generate the coded data [42]. There is a codebook associated with each of the channels in the channel set $c_1, \ldots, c_K$. The general idea behind superposition codes is to have a refinement in the code structure, so that the receiver associated with each channel can determine the coarse code structure (the *code clouds*), while more favorable channels can determine some of fine code structure (the codewords within the clouds). The best channel can distinguish all of the coarse and fine code structure.

Suppose now that we remove the assumption of independent information for each receiver. Since user i automatically receives the information sent to all receivers with $j > i$, if we assume that this information is also desired by user i, then we can include this as an additional component to user i's information rate. The capacity region with this common information is then

$$(R_1', \ldots, R_K') \triangleq \left(\sum_{i=1}^{K} R_i, \sum_{i=2}^{K} R_i, \ldots, R_K\right). \tag{4.80}$$

This common information could be the high-priority bits of a source encoding scheme. Thus, all channels $(c_1, \ldots, c_K)$ would receiver the high-priority bits; the best channel c_1 would receive all high and low-priority bits, and the number of lost low-priority bits would increase with i.

We now return to the fading channel model. The capacity region in (4.80) gives the simultaneous rates achievable for all channels $(c_1, \ldots, c_K)$. For the fading channel, however, only one channel c_i, with rate R_i', is realized on each symbol transmission. Thus, the data rate R on each symbol transmission is a random variable with distribution

$$p(R = r) = \begin{cases} p(c = c_i) & r = R_i' \\ 0 & \text{else} \end{cases}, \tag{4.81}$$

where c denotes the channel state for that transmission. The average transmission rate of the incremental noise channel is thus

$$\overline{R} = \sum_{i=1}^{K} p(c = c_i) R_i'. \tag{4.82}$$

With this transmission scheme, if the fade level realized on a particular transmission corresponds to channel c_i, then $\sum_{j=i+1}^{K} R_j'$ bits of information will be lost on this transmission.

With channel estimation and transmitter feedback, however, the expected transmission rate $\overline{R}$ equals the actual transmission rate. This is because the transmitter knows how much information was lost on each transmission, and can retransmit this data on a subsequent transmission. We illustrate this process for a $K = 2$ incremental noise channel. Let $\mathcal{R} = (R_1 + R_2, R_2)$ be the capacity region for the channel. Using the terminology of Bergmans [68], on each transmission we will have 2^{R_2} cloud centers corresponding to the information transmitted to the second channel and 2^{R_1} satellite codewords appended to each cloud center corresponding to the additional information transmitted to the first channel. If the channel realization for this transmission is the first channel, then the receiver will successfully decode the $2^{R_1+R_2}$ codewords. The feedback mechanism informs the transmitter that the first channel was realized, hence the transmitter knows that all the transmitted information was successfully decoded. If the second channel is realized, then only the 2^{R_2} cloud centers can be successfully decoded. The feedback mechanism informs the transmitter that the satellite codewords were lost, and these satellite codewords are then appended to the next set of cloud centers to be transmitted. Thus, no information is lost, and rate R_2 is achieved when the second channel is realized, rate $R_1 + R_2$ when the

first channel is realized. The fraction of time that the ith channel is realized is $p(c = c_i)$, so the actual rate asymptotically approaches the average rate. As expected, this average rate (4.82) equals the capacity of a time-varying feedback channel (3.7) that was derived in the previous chapter.

4.5.2 Multilevel Coding Techniques

Practical implementation of a multilevel code was first studied by Imai and Hirakawa [70]. Binary UEP codes were later considered both for combined speech and channel coding [65], and combined image and channel coding [71]. These implementations use traditional (block or convolutional) error-correction codes, so coding gain is directly proportional to bandwidth expansion. More recently, two bandwidth-efficient implementations for UEP have been proposed: time-multiplexing of bandwidth-efficient coded modulation [72], and the coded-modulation techniques of §3.5.1 applied to both uniform and nonuniform signal constellations [66, 73, 74]. All of these multilevel codes can be designed for either AWGN or fading channels, depending on the distance criterion of the code, which will be discussed in more detail below. We now briefly summarize these UEP techniques; specifically, we describe the principles behind multilevel coding and multistate decoding, and the more complex bandwidth-efficient implementations.

A block diagram of a general multilevel encoder is shown in Figure 4.10. The source encoder first divides the information sequence into M parallel bit streams of decreasing priority. The channel encoder consists of M different binary error-correcting codes $C_1, \ldots, C_M$ with decreasing codeword distances. For AWGN channels, the binary encoder should maximize the Euclidean distance between codewords; for fading channels, the Hamming distance should be maximized [75]. The ith priority bit stream enters the ith encoder, which generates the coded bits s_i. If the 2^M points in the signal constellation are numbered from 0 to $2^M - 1$, then the point selector chooses the constellation point s corresponding to

$$s = \sum_{i=1}^{M} s_i \times 2^{i-1}. \tag{4.83}$$

For example, if $M = 3$ and the signal constellation is 8PSK, then the chosen signal point will have phase $2\pi s/8$.

Optimal decoding of the multilevel code uses a maximum-likelihood decoder, which determines the input sequence that maximizes the received sequence probability. The

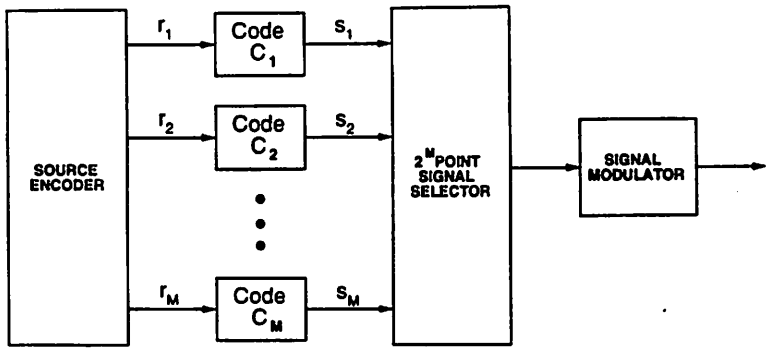

Figure 4.10: Multilevel Encoder

maximum-likelihood decoder must therefore jointly decode the code sequences $s_1, \ldots, s_m$. Hence, if the encoder memories are of length $\mu_1, \ldots, \mu_M$, the number of states in the optimal decoder is $2^{\mu_1+\cdots+\mu_M}$. This leads to very high complexity in the optimal decoder, even if the memories of the individual encoders $C_1, \ldots, C_M$ are small. Due to this complexity, the suboptimal technique of multistage decoding, introduced in [70], is used for most implementations. Multistage decoding is accomplished by decoding the component codes sequentially. First, the most robust code, C_1, is decoded, then C_2, and so forth. Once the code sequence corresponding to encoder C_i is estimated, it is assumed correct for code decisions on the less robust code sequences.

The binary encoders of this multilevel code require extra code bits to achieve their coding gain, thus they are not bandwidth-efficient. An alternative approach recently proposed in [73] uses time-multiplexing of the bandwidth-efficient coset codes described in §3.5.1. In this approach, different conventional coded modulation schemes, such as lattice or trellis codes, with different coding gains are used for each priority class of input data. The transmit signal constellations corresponding to each encoder may differ in size (number of signal points), but the average power of each constellation is the same. The signal points output by each of the individual encoders are then time-multiplexed together for transmission over the channel, as shown in Figure 4.11 for two different priority bit streams. Let R_i denote the bit rate of encoder C_i in this figure, for $i = 1, 2$. If T_1 equals the fraction of time that the high-priority C_1 code is transmitted, and T_2 equals the fraction of time that the C_2 code is transmitted, then the total bit rate is $(R_1T_1 + R_2T_2)/(T_1 + T_2)$, with the high-priority bits comprising $R_1T_1/(R_1T_1 + R_2T_2)$ percent of this total.

The optimal coding results for the degraded broadcast channel suggest that the

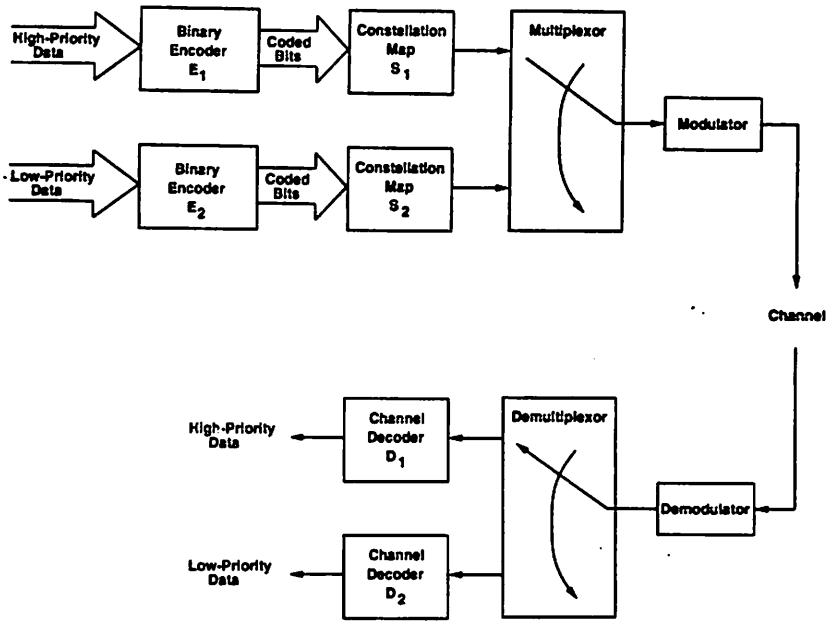

Figure 4.11: Transceiver for Time-Multiplexed Coded Modulation

time-multiplexed coding method yields a higher gain if the constellation maps S_1 and S_2 of Figure 4.11 are designed jointly. This revised scheme is shown in Figure 4.12 for 2 encoders, where the extension to M encoders is straightforward. In fact, the coded modulation of Figure 3.22 in Chapter 3 can be considered as a two-level code of this type. Recall that in this scheme, bits are encoded to select the lattice subset, and uncoded bits choose the constellation point within the subset. The binary encoder properties reduce the BER for the encoded bits only; the BER for the uncoded bits is determined by the separation of the constellation signal points. We can easily modify this scheme to yield two levels of coding gain, where the high-priority bits are encoded as in Figure 3.22 to choose the lattice subset, and the low-priority bits are encoded using a binary encoder, whose output selects the constellation signal point.

More complex multilevel code designs use non-uniform signal constellations. For example, in [73], the nonuniform 32-QAM signal constellation of Figure 4.13 is considered. In this scheme, the high-priority bits are encoded with an eight state trellis encoder, yielding two coded bits per transmission, and the low-priority bits are encoded using an eight state trellis encoder and two uncoded bits, resulting in three coded bits per transmission. The two high-priority coded bits are used to determine the quadrant of the transmitted signal

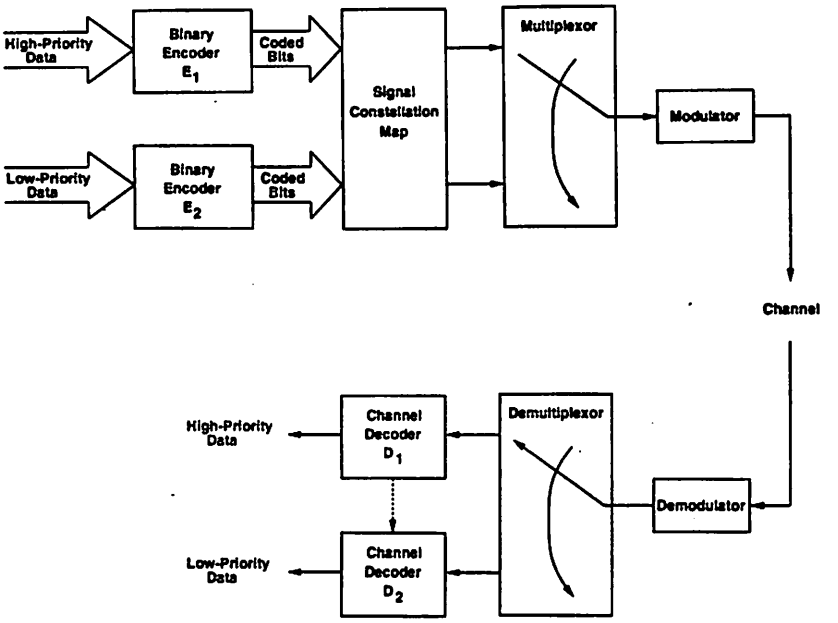

Figure 4.12: Joint Optimization of Signal Constellation

point, or equivalently, one of the four constellation *superpoints* shown in Figure 4.13. The three low-priority coded bits are then used to select one of the 8-PSK points centered around the superpoint. Coding gains for this scheme, for different percentages of high-priority bits and different spacings between the superpoints and between the 8-PSK points, are calculated in [73] and compared with those of the time-multiplexing technique depicted in Figure 4.11. This comparison shows that the time-multiplexing scheme performs better when the percentage of high-priority bits is small; otherwise, coded-modulation with nonuniform signal modulation is better. This result is somewhat surprising, since the capacity analysis for degraded broadcast channels predicts that coded modulation with nonuniformly spaced codewords should always outperform time-multiplexing [42]. This discrepancy between theory and practice may result from the fact that the theoretical results do not consider code complexity, or that they rely on random coding schemes, rather than specific code designs.

4.6 Summary

We have examined techniques for spectrally-efficient communication on time-varying channels without feedback. We first considered finite-state Markov channels, where the channel variation is governed by a Markov process with statistics known to both trans-

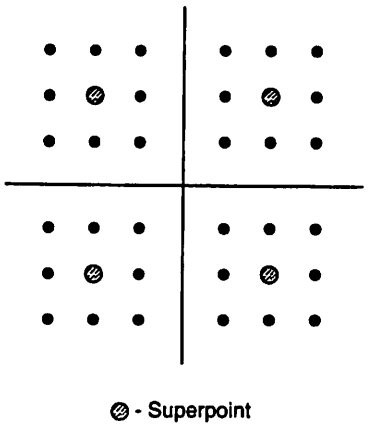

Figure 4.13: Nonuniform 32-QAM with embedded 4-PSK

mitter and receiver. After deriving the Shannon capacity of this channel, we proposed a decision-feedback maximum-likelihood decoder, which uses the Markov transition probabilities to estimate the channel state distribution. This estimate allows the decoder to make maximum-likelihood decisions on a symbol-by-symbol basis, even though the channel memory is infinite. We defined a class of channels for which the decision-feedback decoder achieves channel capacity, and bounded the capacity loss of our scheme for general channels. The capacity and cutoff rate of our decoding scheme was then compared to those of conventional memoryless encoding methods for variable noise channels. We found that the decision-feedback decoding method, for a small increase in complexity, yields a significant capacity increase which is most pronounced on slowly-varying channels.

When the channel varies arbitrarily, multilevel codes can be be used to maintain high-priority data transfer even under worst-case channel conditions. This type of coding prioritizes the transmitted bit stream into data classes; this data prioritization is already inherent to many speech and video source coding techniques. We first determined the average data rate possible with optimal multilevel coding for variable noise channels. We then discussed some practical implementations of this type of coding. Surprisingly, when high-priority data comprises a large percentage of the transmitted bit stream, a simple multiplexing scheme, which is theoretically inferior to multilevel codes with constellation optimization, in practice performs better.

Appendix 4.A.1

In this section, we derive the recursive formula (4.4) for π_n. First, we have

$$\begin{aligned} p(S_n|x^n,y^n) &\stackrel{a}{=} \frac{p(x_n,y_n|S_n,x^{n-1},y^{n-1})p(S_n,x^{n-1},y^{n-1})}{p(x^n,y^n)} \\ &\stackrel{b}{=} \frac{p(y_n|S_n,x_n,x^{n-1},y^{n-1})p(x_n|S_n,x^{n-1},y^{n-1})p(S_n,x^{n-1},y^{n-1})}{p(x^n,y^n)} \\ &\stackrel{c}{=} \frac{p(y_n|S_n,x_n)p(x_n|x^{n-1})p(S_n,x^{n-1},y^{n-1})}{p(x^n,y^n)} \\ &\stackrel{d}{=} \frac{p(y_n|S_n,x_n)p(x_n|x^{n-1})p(S_n|x^{n-1},y^{n-1})p(x^{n-1},y^{n-1})}{p(x^n,y^n)}, \end{aligned} \tag{4.84}$$

where a, b, and d follow from Bayes rule and c follows from (2.42). Moreover,

$$\begin{aligned} p(x^n,y^n) &= \sum_{k\in K} p(x^n,y^n,S_n=c_k) \\ &= \sum_{k\in K} p(x_n,y_n|S_n=c_k,x^{n-1},y^{n-1})p(S_n=c_k,x^{n-1},y^{n-1}) \\ &= \sum_{k\in K} p(y_n|S_n=c_k,x_n,x^{n-1},y^{n-1})p(x_n|S_n,x^{n-1},y^{n-1})p(S_n=c_k,x^{n-1},y^{n-1}) \\ &= \sum_{k\in K} p(y_n|S_n=c_k,x_n)p(x_n|x^{n-1})p(S_n=c_k|x^{n-1},y^{n-1})p(x^{n-1},y^{n-1}), \end{aligned} \tag{4.85}$$

where we again use Bayes rule and the last equality follows from (2.41). Substituting (4.85) in the denominator of (4.84), and canceling the common terms $p(x_n|x^{n-1})$ and $p(x^{n-1},y^{n-1})$ yields

$$p(S_n|x^n,y^n) = \frac{p(y_n|S_n,x_n)p(S_n|x^{n-1},y^{n-1})}{\sum_{k\in K} p(y_n|S_n=c_k,x_n)p(S_n=c_k|x^{n-1},y^{n-1})}, \tag{4.86}$$

which, for a particular value of S_n, becomes

$$p(S_n=c_l|x^n,y^n) = \frac{p(y_n|S_n=c_l,x_n)p(S_n=c_l|x^{n-1},y^{n-1})}{\sum_{k\in K} p(y_n|S_n=c_k,x_n)p(S_n=c_k|x^{n-1},y^{n-1})}. \tag{4.87}$$

Finally, from (2.40),

$$p(S_{n+1}=c_l|x^n,y^n) = \sum_{j\in K} p(S_n=c_j|x^n,y^n)P_{jl}. \tag{4.88}$$

Substituting this into (4.87) yields the desired result.

Appendix 4.A.2

We must show that for all $\theta_m,\theta \in P(\mathcal{X})$, if $\theta_m \to \theta$, then $\mu^{\theta_m} \to \mu^{\theta}$, and $\nu^{\theta_m} \to \nu^{\theta}$. We first show the convergence of ν^{θ_m}. From [57, page 346], in order to show that $\nu^{\theta_m} \to \nu^{\theta}$,

it suffices to show that $\{\nu^{\theta_m}\}$ is a tight sequence of probability measures[7] , and that any subsequence of ν^{θ_m} which converges weakly converges to ν^{θ}.

Tightness of the sequence $\{\nu^{\theta_m}\}$ follows from the fact that Δ is a compact set. Now suppose there is a subsequence $\nu^{\theta_{m_k}} \triangleq \nu^{\theta_k}$ which converges weakly to ψ. We must show that $\psi = \nu^{\theta}$, where ν^{θ} is the unique invariant distribution for ρ under the transformation (4.10) with input distribution $p(x) = \theta$. Thus, it suffices to show that for every bounded, continuous, real-valued function ϕ on Δ,

$$\int_{\Delta} \phi(\gamma)\psi(d\gamma) = \int_{\Delta}\int_{\Delta} \phi(\alpha)\psi(d\beta)p^{\theta}(d\alpha|\beta), \tag{4.89}$$

where $p^{\theta}(\alpha|\beta) \triangleq p(\rho_{n+1} = \alpha|\rho_n = \beta)$ is given by (4.10) under the input distribution θ. Applying the triangle inequality, we get that for any k,

$$\left|\int_{\Delta} \phi(\gamma)\psi(d\gamma) - \int_{\Delta}\int_{\Delta} \phi(\alpha)\psi(d\beta)p^{\theta}(d\alpha|\beta)\right|$$

$$\leq \left|\int_{\Delta} \phi(\gamma)\psi(d\gamma) - \int_{\Delta} \phi(\gamma)\nu^{\theta_k}(d\gamma)\right| \tag{4.90}$$

$$+ \left|\int_{\Delta} \phi(\gamma)\nu^{\theta_k}(d\gamma) - \int_{\Delta}\int_{\Delta} \phi(\alpha)\nu^{\theta_k}(d\beta)p^{\theta}(d\alpha|\beta)\right| \tag{4.91}$$

$$+ \left|\int_{\Delta}\int_{\Delta} \phi(\alpha)\nu^{\theta_k}(d\beta)p^{\theta}(d\alpha|\beta) - \int_{\Delta}\int_{\Delta} \phi(\alpha)\psi(d\beta)p(d\alpha|\beta)\right|. \tag{4.92}$$

Since this inequality holds for all k, in order to show (4.89), we need only show that the three terms (4.90), (4.91), and (4.92) all converge to zero as $k \to \infty$. But (4.90) converges to zero since ν^{θ_k} converges weakly to ψ. Moreover, (4.91) equals zero for all k, since ν^{θ_k} is the invariant ρ distribution under the transformation (4.10) with input distribution θ_k. Substituting (4.10) for $p^{\theta}(\alpha|\beta)$ in (4.92) yields

$$\left|\int_{\Delta}\int_{\Delta} \phi(\alpha)\nu^{\theta_k}(d\beta)p^{\theta_k}(d\alpha|\beta) - \int_{\Delta}\int_{\Delta} \phi(\alpha)\psi(d\beta)p^{\theta}(d\alpha|\beta)\right| =$$

$$\left|\sum_{y\in\mathcal{Y}}\int_{\Delta} \phi(f^{\theta_k}(y,\beta))p^{\theta_k}(y|\beta)\nu^{\theta_k}(d\beta) - \sum_{y\in\mathcal{Y}}\int_{\Delta} \phi(f^{\theta}(y,\beta))p^{\theta}(y|\beta)\psi(\beta)d\beta\right|, \tag{4.93}$$

where f^{θ} is given by (4.9) with $p(x) = \theta$, and

$$p^{\theta}(y|\beta) = \sum_{x\in\mathcal{X}}\sum_{k=1}^{K} p(y|x, S = c_k)\beta(k)\theta(x). \tag{4.94}$$

---

[7] A sequence of probability measures $\{\nu_m; m \geq 1\}$ is tight if for all $\epsilon > 0$ there exists a compact set K such that $\nu(K) > 1 - \epsilon$ for all ν_m.

Since $\mathcal{Y}$ is a finite set, (4.93) converges to zero if for every $y \in \mathcal{Y}$,

$$\left| \int_\Delta \phi(f^{\theta_k}(y,\beta)) p^{\theta_k}(y|\beta) \nu^{\theta_k}(d\beta) - \int_\Delta \phi(f^{\theta}(y,\beta)) p^{\theta}(y|\beta) \psi(d\beta) \right| \to 0. \tag{4.95}$$

Fix an arbitrary $y \in \mathcal{Y}$. Applying the triangle inequality to (4.95) yields

$$\left| \int_\Delta \phi(f^{\theta_k}(y,\beta)) p^{\theta_k}(y|\beta) \nu^{\theta_k}(d\beta) - \int_\Delta \phi(f^{\theta}(y,\beta)) p^{\theta}(y|\beta) \psi(d\beta) \right|$$

$$\leq \left| \int_\Delta \phi(f^{\theta_k}(y,\beta)) p^{\theta_k}(y|\beta) \nu^{\theta_k}(d\beta) - \int_\Delta \phi(f^{\theta}(y,\beta)) p^{\theta}(y|\beta) \nu^{\theta_k}(d\beta) \right| \tag{4.96}$$

$$+ \left| \int_\Delta \phi(f^{\theta}(y,\beta)) p^{\theta}(y|\beta) \nu^{\theta_k}(d\beta) - \int_\Delta \phi(f^{\theta}(y,\beta)) p^{\theta}(y|\beta) \psi(d\beta) \right|. \tag{4.97}$$

But for any fixed y and β, $\theta_k \to \theta$ implies that $f^{\theta_k}(y,\beta) \to f^{\theta}(y,\beta)$, since from (4.9), the numerator and denominator of f are linear functions of θ, and the denominator is nonzero. Similarly, $\theta_k \to \theta$ implies that for fixed y and β, $p^{\theta_k}(y|\beta) \to p^{\theta}(y|\beta)$, since $p^{\theta}(y|\beta)$ is linear in θ. Since ϕ is continuous, this implies that for fixed y and β, $\phi(f^{\theta_k}(y,\beta)) p^{\theta_k}(y|\beta) \to \phi(f^{\theta}(y,\beta)) p^{\theta}(y|\beta)$. Since ϕ is also bounded on Δ, (4.96) converges to zero by the dominated convergence theorem [57]. Moreover, for fixed y and θ, $f^{\theta}(y,\beta)$ and $p^{\theta}(y|\beta)$ are linear in β, so $\phi(f^{\theta}(y,\beta)) p^{\theta}(y|\beta)$ is a bounded continuous functions of β. Thus, (4.97) converges to zero by the weak convergence of ν^{θ_k} to ψ.

Since the $\{\mu^{\theta_m}\}$ sequence is also tight, the proof that $\mu^{\theta_m} \to \mu^{\theta}$ follows if the limit of any convergent subsequence of $\{\mu^{\theta_m}\}$ is the invariant distribution for π under (4.6). This is shown with essentially the same argument as above for $\nu^{\theta_k} \to \nu^{\theta}$, using (4.6) instead of (4.10) for $p(\alpha|\beta)$, $p^{\theta}(y|x,\beta)$ instead of $p^{\theta}(y|\beta)$, and summations over $\mathcal{X} \times \mathcal{Y}$ instead of $\mathcal{Y}$. The details are omitted.

Appendix 4.A.3

To prove Lemma 5.4, we must show that

$$H(Y_n|Y^{n-1}) \geq H(Y_{n+1}|Y^n) \geq H(Y_{n+1}|Y^n, S_0) \geq H(Y_n|Y^{n-1}, S_0). \tag{4.98}$$

For the first inequality, let f denote any concave function. Then

$$\begin{aligned} f(p[y_n|y^{n-1}]) &\stackrel{a}{=} f(p[y_{n+1}|y_2^n]) \\ &\stackrel{b}{=} f(\mathbf{E}(p[y_{n+1}|y^n]|y_2^n)) \\ &\stackrel{c}{\geq} \mathbf{E}(f(p[y_{n+1}|y^n])|y_2^n) \\ &\stackrel{d}{=} f(p[y_{n+1}|y^n]), \end{aligned} \tag{4.99}$$

where a follows from the stationarity of the inputs and channel, b and d follow from properties of conditional expectation [57], and c is a consequence of Jensen's inequality.

The second inequality results from the fact that conditioning on an additional random variable reduces entropy. Finally, for the third inequality, we have

$$\begin{aligned} f(p[y_{n+1}|y^n, S_0]) &\stackrel{a}{=} f(\mathbf{E}(p[y_{n+1}|y^n, S_1]|y^n, S_0)) \\ &\stackrel{b}{=} f(\mathbf{E}(p[y_{n+1}|y^{i_2}, S_1]|y^n, S_0)) \\ &\stackrel{c}{\geq} \mathbf{E}(f(p[y_{n+1}|y_2^n, S_1])|y^n, S_0) \\ &\stackrel{d}{=} f(p[y_{n+1}|y_2^n, S_1]), \\ &\stackrel{e}{=} f(p[y_n|y^{n-1}, S_0]), \end{aligned} \tag{4.100}$$

where a and d follow from properties of conditional expectation, b follows from (2.42), c follows from Jensen's inequality, and e follows from the channel and input stationarity.

Appendix 4.A.4

We consider a Q-AWN channel where the output is quantized to the nearest input symbol and the input alphabet consists of symmetric PSK symbols. We want to show that for any k, $P_{ij}^k \stackrel{\Delta}{=} p_k(y = j|x = i)$ has rows which are permutations of each other and columns which are permutations of each other. The input/output symbols are given by

$$y_m = x_m = A \exp^{j2\pi m/M}, \quad m = 1, \ldots, M. \tag{4.101}$$

Define the $M \times M$ matrix Z by $Z_{ij} = |y_i - x_j|$ and let $q_k(Z_{ij})$ denote the distribution of the quantized noise, which is determined by n_k, A, M. By symmetry of the input/output symbols and the noise, the rows of Z are permutations of each other, and the columns are also permutations of each other.

If M is odd, then

$$p_k(y|x) = \begin{cases} q_k(|y-x|) & |y-x| = 0 \\ q_k(|y-x|)/2 & \text{else} \end{cases}, \tag{4.102}$$

and if M is even,

$$p_k(y|x) = \begin{cases} q_k(|y-x|) & |y-x| = 0 \text{ or } |y-x| = 2A \\ q_k(|y-x|)/2 & \text{else} \end{cases}. \tag{4.103}$$

Thus, P_{ij}^k depends only on the value of Z_{ij}; the rows of P_{ij}^k are therefore permutations of each other, and so are the columns.

Appendix 4.A.5

We will show that the π-output channel is asymptotically memoryless as $J \to \infty$. Indeed, since the FSMC is indecomposable and stationary, for asymptotically large J,

$$p(S_{n+J}, S_n) = p(S_{n+J})p(S_n) \tag{4.104}$$

for any n, and thus also

$$p(\pi_{n+J}, \pi_n) = p(\pi_{n+J})p(\pi_n). \tag{4.105}$$

Therefore, since π_{jl} and $\pi_{j(l-1)}$ are J iterations apart, π_{jl} and $\pi_{j(l-1)}$ are asymptotically independent for large J.

In order to show that the π-output channel is memoryless, we must show that for any j and L,

$$p(y^{jL}, \pi^{jL}|x^{jL}) = \prod_{l=1}^{L} p(y_{jl}, \pi_{jl}|x_{jl}). \tag{4.106}$$

We can decompose $p(y^{jL}, \pi^{jL}|x^{jL})$ as follows:

$$p(y^{jL}, \pi^{jL}|x^{jL}) = \prod_{l=1}^{L} p(y_{jl}, \pi_{jl}|x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}x^{j(l-1)}). \tag{4.107}$$

Thus we need only show that the lth factor in the right hand side of (4.107) equals $p(y_{jl}, \pi_{jl}|x_{jl})$ in the limit as $J \to \infty$. This result is proved in the following lemma.

Lemma 4.A.5.1 For asymptotically large J,

$$p(y_{jl}, \pi_{jl}|x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}x^{j(l-1)}) = p(y_{jl}, \pi_{jl}|x_{jl}). \tag{4.108}$$

Proof

$$\begin{aligned}
&p(y_{jl}, \pi_{jl}|x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}x^{j(l-1)}) \\
&\quad = p(y_{jl}|\pi_{jl}, x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}, x^{j(l-1)})p(\pi_{jl}|x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}, x^{j(l-1)}) \\
&\quad = p(y_{jl}|\pi_{jl}, x_{jl})p(\pi_{jl}|y^{j(l-1)}, \pi^{j(l-1)}, x^{j(l-1)}) \\
&\quad = p(y_{jl}|\pi_{jl}, x_{jl})p(\pi_{jl}|\pi_{(j+1)(l-1)}) \\
&\quad = p(y_{jl}|\pi_{jl}, x_{jl})p(\pi_{jl}) \\
&\quad = p(y_{jl}, \pi_{jl}|x_{jl}),
\end{aligned} \tag{4.109}$$

where the second equality follows from (2.40) and (2.41), the third equality follows from (4.5), and the fourth equality follows from (4.105) in the asymptotic limit of deep interleaving.

Appendix 4.A.6

The π-output channels are independent if

$$p(y^J, \pi^J | x^J) = \prod_{j=1}^{J} p(y_j, \pi_j | x_j). \tag{4.110}$$

This is shown in the following string of equalities.

$$\begin{aligned} p(y^J, \pi^J | x^J) &= \prod_{j=1}^{J} p(y_j, \pi_j | x_j, y^{j-1}, \pi^{j-1}, x^{j-1}) \\ &= \prod_{j=1}^{J} p(y_j | \pi_j, x_j, y^{j-1}, \pi^{j-1}, x^{j-1}) p(\pi_j | x_j, y^{j-1}, \pi^{j-1}, x^{j-1}) \\ &= \prod_{j=1}^{J} p(y_j | \pi_j, x_j) p(\pi_j | x_j, y^{j-1}, \pi^{j-1}, x^{j-1}) \\ &= \prod_{j=1}^{J} p(y_j | \pi_j, x_j) p(\pi_j), \end{aligned} \tag{4.111}$$

where the third equality follows from (2.41) and the last equality follows from the fact that we ignore error propagation, so x^{j-1}, y^{j-1}, and π^{j-1} are all known at time j.

We now determine the average mutual information of the parallel π-output channels for fixed $p(\mathcal{X}^J)$. The average mutual information of the parallel set is

$$I_J = \frac{1}{J} I(Y^J, \pi^J; X^J). \tag{4.112}$$

From above, the parallel channels are independent, and each channel is memoryless with asymptotically deep interleaving. Thus, we obtain (4.66) as follows:

$$\begin{aligned} \frac{1}{J} I(Y^J, \pi^J; X^J) &= H(Y^J, \pi^J) - H(Y^J, \pi^J | X^J) \\ &= H(Y^J | \pi^J) + H(\pi^J) - (H(Y^J | \pi^J, X^J) + H(\pi^J | X^J)) \\ &= H(Y^J | \pi^J) - (H(Y^J | \pi^J, X^J) \\ &= \sum_{j=1}^{J} H(Y_j | Y^{j-1}, \pi^J) - H(Y_j | Y^{j-1}, \pi^J, X^J) \end{aligned}$$

$$= \sum_{j=1}^{J} H(Y_j|\pi_j) - H(Y_j|\pi_j, X_j), \tag{4.113}$$

where the third equality follows from the fact that $p(\pi^J|x^J) = p(\pi^J)$, and the last inequality follows from the fact that

$$H(Y_j|Y^{j-1}, \pi^J) = H(Y_j|\rho_j, \pi^J) = H((Y_j|\pi_J), \tag{4.114}$$

since $\rho_j = \mathbf{E}_{x_j}\pi_J$.

Appendix 4.A.7

In this section, we examine the cutoff rate for uniformly symmetric variable noise channels. The first four lemmas show that for these channels, the maximizing distribution of (4.75) is uniform and i.i.d. We then determine that R_j, as given by (4.74), is monotonically increasing in j, and use this to get a simplified formula for R_{df} in terms of the limiting value of R_j.

Lemma 4.A.7.1 For all j, R_j depends only on $p(\mathcal{X}_j)$.

Proof From the proof of Lemma 6.2, π_j is a function of Z^{n-1}, and is independent of X^{n-1}. So $p(\pi_j)$ doesn't depend on the input distribution. The result then follows from the definition of R_j. □

Lemma 4.A.7.2 An independent input distribution achieves the maximum of R_{df}.

Proof Let p^* denote the maximizing distribution of R_{df}, and assume that under p^*, the inputs are not independent. Define the independent input distribution $\hat{p}$ by $\hat{p}(x_j) = p^*(x_j)$. Since by the previous lemma, $R_j, j = 1, 2, \ldots$ is the same for inputs governed by p^* or $\hat{p}$, the distribution $\hat{p}$ must also achieve R_{df}. □

Lemma 4.A.7.3 For a fixed input distribution $p(\mathcal{X}^J)$, the J corresponding π-output channels are all symmetric [40, page 94].

Proof We must show that for any $j < J$, the set of outputs for the jth π-output channel can be partitioned into subsets such that the corresponding submatrices of transition probabilities has rows which are permutations of each other and columns which are permutations of each other. We will call such a matrix *row/column permutable.*

Let $n_j \leq |\mathcal{X}|^j|\mathcal{Y}|^j$ be the number of points $\delta \in \Delta$ with $p(\pi_j = \delta) > 0$, and let $\{\delta_i\}_{i=1}^{n_j}$ explicitly denote this set. Then we can partition the output into n_j sets, where the ith set consists of the pairs $\{(y, \delta_i) : y \in \mathcal{Y}\}$. We want to show that the transition probability matrix associated with each of these output partitions is row/column permutable, i.e. that for all i, $1 \leq i \leq n_j$, the $|\mathcal{X}| \times |\mathcal{Y}|$ matrix

$$P^i \triangleq p(y_j = y, \pi_j = \delta_i | x_j = x), \quad x \in \mathcal{X}, y \in \mathcal{Y} \tag{4.115}$$

has rows which are permutations of each other, and columns which are permutations of each other.

Since the FSMC is a variable noise channel, $p_k(y|x)$ depends only on $z = f(x, y)$ for all k, $1 \leq k \leq K$. Therefore, if for some k', $p_{k'}(y|x) = p_{k'}(y'|x')$, then $f(x, y) = f(x', y')$. But since $z = f(x, y)$ is the same for all k, this implies that

$$p_k(y|x) = p_k(y'|x') \ \ \forall k, 1 \leq k \leq K. \tag{4.116}$$

Fix k'; then by definition of uniform symmetry, $p_{k'}(y|x)$ is row/column permutable. Using (4.116), we get that the $|\mathcal{X}| \times |\mathcal{Y}|$ matrix

$$P_\Sigma = \sum_{k=1}^{K} p_k(y|x), \quad x \in \mathcal{X}, y \in \mathcal{Y} \tag{4.117}$$

is also row/column permutable. Moreover, multiplying a matrix by any constant will not change the permutability of its rows and columns, hence the matrix

$$P_\Sigma^i = \left[\sum_{k=1}^{K} p_k(y|x)\right] \delta_i p(\pi_j = \delta_i), x \in \mathcal{X}, y \in \mathcal{Y} \tag{4.118}$$

is also row/column permutable. But this completes the proof, since

$$p(y_j = y, \pi_j = \delta_i | x_j = x) = \sum_{k=1}^{K} p_k(y_j = y | x_j = x) \delta_i p(\pi_j = \delta_i). \tag{4.119}$$

□

Lemma 4.A.7.4 For i.i.d. uniform inputs, R_j is monotonically increasing in j.

Proof For i.i.d. inputs,

$$R_j = -\log \frac{1}{|\mathcal{X}|^2} \sum_{\pi_j \in \Delta} p(\pi_j) \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y|x, S = c_k)\pi_j(k)} \right]^2 . \tag{4.120}$$

Let

$$f(\pi_j) \triangleq \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y|x, S = c_k)\pi_j(k)} \right]^2 , \tag{4.121}$$

so $R_j = -\log \frac{1}{|\mathcal{X}|^2} \mathbf{E}[f(\pi_j)]$. We must show $-\log \frac{1}{|\mathcal{X}|^2} \mathbf{E}[f(\pi_j)] \leq -\log \frac{1}{|\mathcal{X}|^2} \mathbf{E}[f(\pi_{j+1})]$, or equivalently, $\mathbf{E}[f(\pi_j)] \geq \mathbf{E}[f(\pi_{j+1})]$. Following an argument similar to that of Lemma 5.2, we have

$$\begin{aligned}
f(\pi_j) &= \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y|x, S = c_k)\pi_j(k)} \right]^2 \\
&= \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y|x, S = c_k) p(S_j = c_k | x^{n-1}, y^{n-1})} \right]^2 \\
&\stackrel{a}{=} \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y|x, S = c_k) p(S_{j+1} = c_k | x_2^n, y_2^n)} \right]^2 \\
&= \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y|x, S = c_k) \mathbf{E}[p(S_{j+1} = c_k | x^n, y^n) | x_2^n, y_2^n)]} \right]^2 \\
&\stackrel{b}{\geq} \sum_{y \in \mathcal{Y}} \left[\mathbf{E} \sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y|x, S = c_k) p(S_{j+1} = c_k | x^n, y^n)} \right]^2 \\
&= \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y|x, S = c_k) p(S_{j+1} = c_k | x^n, y^n)} \right]^2 \\
&= f(\pi_{j+1}),
\end{aligned} \tag{4.122}$$

where a follows from stationarity and b follows from Jensen's inequality. Taking expectation of both sides in (4.122) yields the desired result. □

Lemma 4.A.7.5 For uniformly symmetric variable noise channels, a uniform i.i.d. input distribution maximizes R_{df}. Moreover,

$$R_{df} = \lim_{j \to \infty} R_j. \tag{4.123}$$

Proof From Lemma 4.A.7.2, the maximizing distribution for R_{df} is independent. Moreover, from Lemma 4.A.7.3, each of the π-output channels are symmetric, therefore from [40, page 144], a uniform distribution for $p(\mathcal{X}_j)$ maximizes R_j for all j, and therefore it maximizes R_{df}. Moreover, by Lemmas 4.A.7.4 and 4.1, for i.i.d. uniform inputs, R_j is monotonically increasing in j and converges to a limit independent of the initial channel state. Therefore,

$$R_{df} = \lim_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} R_j = \lim_{j \to \infty} R_j. \tag{4.124}$$

Chapter 5

Multiuser Systems

The previous chapters dealt with increasing spectral efficiency on single-user channels; we now consider the case where multiple users share the same channel. There are several methods for dividing the channel frequency spectrum among many users: the most common are time division (TDMA), frequency division (FDMA), code division (CDMA), and hybrid combinations of these methods. Currently, there are four standards with different spectrum-sharing techniques for digital cellular phone systems alone: one for Europe, one for Japan, and two for North America. The debate among cellular and personal communication standards committees and equipment providers over which approach to use has led to countless analytical studies claiming superiority of one technique over the other. In many cases the a priori assumptions used in these analyses bias the results in favor of one technique over the other alternatives; usually the technique that is of some economic interest to the authors of the study. In this chapter we provide an unbiased evaluation of the different spectrum-sharing techniques for both time-invariant and time-varying broadcast and multiple access channels.

We begin with a summary of the capacity and achievable rate regions for time-invariant AWGN channels. We consider only broadcast and multiple access channels, which model two-way transmission in systems where many users are communicating with a single transceiver, as in cellular, satellite, TV broadcast, and packet radio systems. We will see that CDMA (with interference cancellation and no power control) and FDMA techniques both achieve the maximum total rate for multiaccess channels, and CDMA achieves the maximum rate region for broadcast and multiaccess channels. In addition, if power control is used to equalize received power in a broadcast system, then CDMA, FDMA, and TDMA

all have the same rate regions. Finally, without interference cancellation CDMA is generally inferior to both FDMA and TDMA. Although CDMA with interference cancellation is always at least as good as the other techniques, it also requires more complexity in both the transmitter and receiver, which may preclude its use in low-power mobile receivers [76]. We will also summarize the derivation by Cheng and Verdú of the capacity region for wideband time-invariant ISI channels [77]. For these channels, FDMA and CDMA with interference cancellation have the same rate regions for equal user priorities, and CDMA is superior when the user priorities are not equal.

We then extend the time-invariant analyses to time-varying memoryless channels. For FDMA and TDMA spectrum sharing, the users are orthogonal, and the time-varying capacity results of Chapter 3 can be applied. We also show that the time-varying capacity region of CDMA with interference cancellation dominates both time and frequency division techniques, and we discuss the capacity region of CDMA without interference cancellation.

The capacity of cellular systems cannot be evaluated using the methods outlined above, since spatial reuse is not incorporated into the multiuser channel model. Reusing frequencies at spatially-separated cells allows more efficient use of the frequency spectrum, however it also introduces intercell interference, which reduces the capacity of all users. The tradeoff between increased spectrum efficiency and decreased user capacity is quantified by the *area spectral efficiency*, defined as the data rate/Hz/unit area of all users in the system. We calculate this efficiency as a function of reuse distance for FDMA with a very simple signal and interference model. Optimization of power control and reuse distance to maximize this efficiency for more complicated models is also discussed. We conclude the chapter with some interference mitigation techniques.

5.1 Rate Regions for Memoryless AWGN Channels

When several users share the same channel, the channel capacity can no longer be characterized by a single number. At the extreme, if all but one user occupies the channel, then the single-user capacity results of Chapter 3 apply. However, since there is an infinite number of ways to "divide" the channel between many users, the multiuser channel capacity is characterized by a *rate region*, where each point in the region is a vector of achievable rates that can be maintained by all the users simultaneously. The set of all achievable rates is called the *capacity region* of the multiuser system. In this section we analyze two time-

invariant memoryless AWGN channels: the broadcast channel and the multiaccess channel. We examine rate regions for these channels using CDMA with and without interference cancellation, TDMA, and FDMA spectrum-sharing techniques. The maximum rate region, achieved using CDMA with interference cancellation, relies on the concept of superposition codes and successive decoding, as described in §4.5. Specifically, the user with the highest priority decodes all lower priority messages and subtracts them before decoding his message; the lower priority users treat higher priority messages as noise. We will elaborate on this technique for the two channel models under consideration in the following sections. We will also show that the rate region of CDMA without interference cancellation is inferior to all the other spectrum-sharing techniques. As in the single-user case, the capacity region gives the maximum set of rates without constraint on the complexity and delay of the coding, decoding, and spectrum-sharing method.

5.1.1 Broadcast Channels

The broadcast channel consists of one transmitter sending information to many receivers over a common channel, as shown in Figure 5.1. The transmitter must encode information meant for the different receivers into a common signal. The capacity region of the broadcast channel characterizes how much information can be conveyed to the different receivers simultaneously.

We consider rate regions for a two-user discrete AWGN broadcast channel only; the extension to multiple users is straightforward [78]. Thus, there is one sender of power P, and two distant receivers, each with AWGN of power n_i, $i = 1, 2$. We also assume that the data pulses are Nyquist, so the signal bandwidth $B = 1/T$, where T denotes the length of each data pulse. We can order the channels relative to the noise powers without loss of generality, so that $n_1 \leq n_2$, i.e., receiver 1's channel is less noisy than receiver 2's. If we denote the transmitted signal by X, then user 1 receives the signal $Y_1 = X + N_1$, and user 2 receives the signal $Y_2 = X + N_2$, where N_i denotes the noise sample of the ith receiver. The transmitter wishes to send independent messages to receivers 1 and 2 at rates R_1 and R_2, respectively.

We encountered this broadcast channel model in §4.5 when we analyzed unequal error protection codes for fading channels. We now consider the capacity region of this

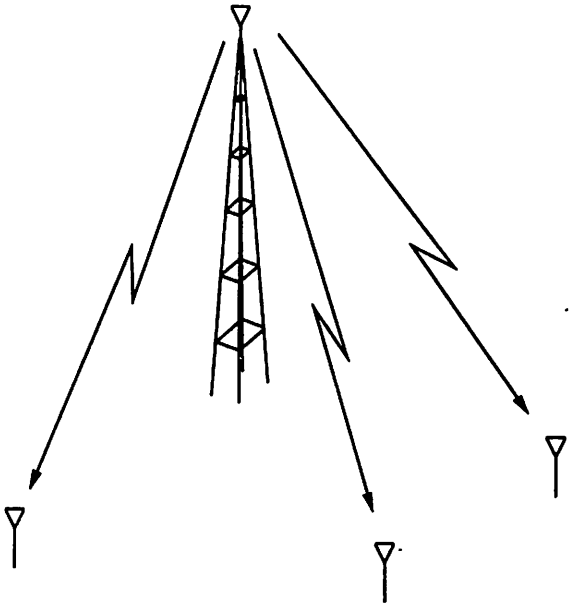

Figure 5.1: Broadcast Channel.

channel model in the multiuser context, as analyzed by Bergmans [69][1]. If we denote the total power and bandwidth allocated to both users by P and B, respectively, then the single-user capacity C_i of receiver i's channel is given by:

$$C_i = \frac{B}{2} \log \left[1 + \frac{P}{n_i B} \right]. \tag{5.1}$$

If the transmitter allocates all the power and bandwidth to one of the users, then the other user receives no data; therefore, the set of simultaneously achievable rates (R_1, R_2) includes the pairs $(C_1, 0)$ and $(0, C_2)$. These two rate pairs bound the multiuser capacity region. We now consider rate pairs in the interior of the region, which are achieved using more equitable methods of dividing the channel resources.

One scheme for dividing the bandwidth and power between the two users is time division, where the full power and bandwidth is allocated to user 1 for a fraction τ of the total transmission time, and then to user 2 for the remainder of the transmission. This time division scheme achieves any rate pair $(\tau C_1, (1-\tau) C_2)$, so a straight line connecting the points $(C_1, 0)$ and $(0, C_2)$, as shown in Figure 5.2, bounds the rate region achievable through time division. A problem with this scheme is delay: since the transmissions to each

[1] Bergmans' results were for continuous-time channels, however it can be shown that the same formulas hold for discrete-time channels with Nyquist data pulses [42].

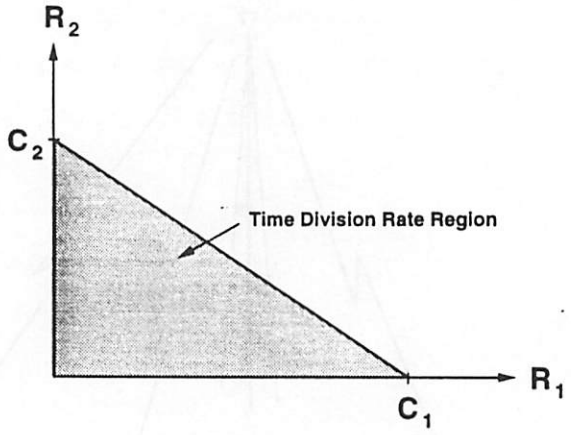

Figure 5.2: Rate Region with Time Division.

user take place sequentially, the second user must wait to receive data until after the first user finishes with the channel. The more common method of time multiplexing alleviates this delay problem, since the channel is periodically alternated between the users. However, multiplexing introduces some capacity loss due to code restriction; this loss was bounded in §3.6.3.

An alternative approach for spectrum-sharing is frequency division. In this method the ith user is allocated power P_i and bandwidth B_i of the total, so $P_1 + P_2 = P$ and $B_1 + B_2 = B$. The set of achievable rates, for fixed P_i and B_i, is then given by

$$\begin{aligned} R_1 &= \frac{B_1}{2} \log \left[1 + \frac{P_1}{n_1 B_1} \right], \\ R_2 &= \frac{B_2}{2} \log \left[1 + \frac{P_2}{n_2 B_2} \right]. \end{aligned} \tag{5.2}$$

It was shown by Bergmans [69] that, for n_1 strictly less than n_2 and any fixed frequency division (B_1, B_2), there exists a range of power allocations (P_1, P_2) whose corresponding rate pairs dominate a segment of the time division rate region, as illustrated by the shaded region in Figure 5.3. Moreover, the rate regions achievable through time division can always be exceeded by optimizing both the frequency and power division in (5.2). Finally, the frequency division rate region boundary intersects the time division line at the point where the power allocation P_i is proportional to the bandwidth B_i. This intersection point Π has a negative derivative with respect to α_1, and so there must be another intersection point for a smaller value of α_1, as shown in Figure 5.3. However, these rate region properties are valid only when $n_1 \neq n_2$. When the noise powers are equal there is no performance

difference among these spectrum-sharing techniques: time division and frequency division have the same rate regions [68].

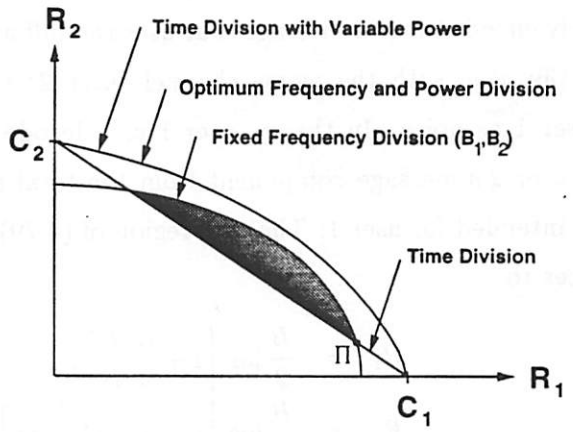

Figure 5.3: Rate Region with Frequency Division.

We can also view the broadcast channel with time division as a time-varying channel with two channel states, where each state is an AWGN channel of power n_i. If we allow one user to use more average power than the other, then we can achieve the same capacity with time division as with frequency division. To see this, let P_1 and P_2 denote the power allocated to users 1 and 2 respectively, where the channel is occupied by user 1 for a fraction τ_1 of the total transmission time, and by user 2 for the remaining time fraction $\tau_2 = 1 - \tau_1$. To satisfy the total power constraint we must have $\tau_1 P_1 + \tau_2 P_2 = P$. The set of variable-power time division rates is then given by

$$\begin{aligned} R_1 &= \tau_1 \frac{B}{2} \log\left[1 + \frac{P_1}{n_1 B}\right], \\ R_2 &= \tau_2 \frac{B}{2} \log\left[1 + \frac{P_2}{n_2 B}\right], \end{aligned} \tag{5.3}$$

where B denotes the total channel bandwidth. Define $B_i \triangleq \tau_i B$, and $\pi_i \triangleq \tau_i P_i$, so the power constraint becomes $\pi_1 + \pi_2 = P$. Making these substitutions in (5.3) yields

$$\begin{aligned} R_1 &= \frac{B_1}{2} \log\left[1 + \frac{\pi_1}{n_1 B_1}\right], \\ R_2 &= \frac{B_2}{2} \log\left[1 + \frac{\pi_2}{n_2 B_2}\right]. \end{aligned} \tag{5.4}$$

Comparing this with (5.2), we see that with appropriate choice of P_i and τ_i, any point in the frequency division rate region can also be achieved through time division with variable power.

The degraded broadcast channel rate region, given by (4.79), was shown in [79] to strictly dominate the regions achievable through either time or frequency division, when $n_1 < n_2$. This region is achieved through superposition codes, where the message to each receiver is jointly encoded into a message that uses the full available bandwidth and power. For decoding, the user with the worse channel (user 2) treats the message component intended for user 1 as noise. In theory, user 1 can decode user 2's message perfectly; it then subtracts user 2's message component from the total received message, leaving only the component intended for user 1. The rate region of (4.79) for superposition coding with two users reduces to

$$\begin{aligned} R_1 &= \frac{B}{2}\log\left[1+\frac{\alpha_1 P}{n_1 B}\right], \\ R_2 &= \frac{B}{2}\log\left[1+\frac{\alpha_2 P}{n_2 B+\alpha_1 P}\right], \end{aligned} \tag{5.5}$$

where α_i denotes the fraction of total power allocated to user i, so $\alpha_1 = 1 - \alpha_2$. The rate regions for all the spectrum-sharing methods, and the superiority of (5.5), is shown in Figure 5.4. Moreover, Bergmans shows in [79] that (5.5) defines the capacity region, i.e., the maximum achievable set of rate pairs. However, superposition coding is superior only when $n_1 \neq n_2$; otherwise, all the spectrum-sharing methods we have described have the same rate region [68]. Therefore, if the constant power policy of §3.3.2 is used to equalize the received SNR of all the users, then each of the spectrum-sharing techniques yields the same performance.

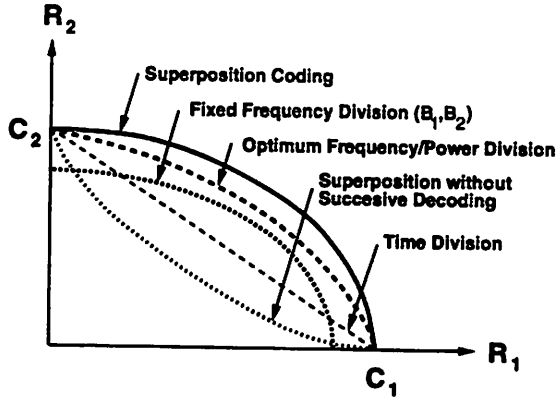

Figure 5.4: Superposition Rate Region.

In practice, successive decoding of superposition codes adds complexity and delay

in the decoding process, as well as the potential for feedback errors when user 2's message is not decoded properly. Superposition coding is mostly done using spread spectrum techniques [38], and successive decoding for this implementation is generally too complex to build into a low-power portable device [76]. We will discuss practical implementations for successive decoding in more detail in §5.4.2. Most commercial spread spectrum receivers don't use successive decoding; they treat all messages intended for other users as noise, resulting in the two-user rate region

$$\begin{aligned} R_1 &= \frac{B}{2}\log\left[1+\frac{\alpha_1 P}{n_1 B+\alpha_2 P}\right], \\ R_2 &= \frac{B}{2}\log\left[1+\frac{\alpha_2 P}{n_2 B+\alpha_1 P}\right], \end{aligned} \tag{5.6}$$

where $\alpha_1+\alpha_2=1$. By taking second derivatives of R_1 and R_2 with respect to α_1, we see that (R_1,R_2) as a function of α_1 is convex, with end points C_1 and C_2, as shown in Figure 5.4. Therefore, *both time division and frequency division always dominate superposition coding without successive decoding.* The fixed frequency division scheme also dominates this suboptimal technique over some range of rate regions, in particular the shaded region shown in Figure 5.3.

5.1.2 Multiaccess Channels

The multiaccess channel consists of K transmitters sending information to one receiver over a common channel of bandwidth B, as shown in Figure 5.5. The transmitters must encode their individual signals such that they can be determined from the received signal, which consists of the sum of signals from each transmitter. The rate region of the multiaccess channel characterizes how much information can be received simultaneously from all the transmitters.

The multiaccess model consists of several transmitters, each with power P_i, sending to a receiver which is corrupted by AWGN of power n. If we denote the ith transmitted signal by X_i, then the received signal is given by

$$Y=\sum_{i=1}^{K} X_i + N, \tag{5.7}$$

where N is an AWGN sample of power n. The two-user capacity region of this channel was determined by Cover to be the closed convex hull of all vectors (R_1,R_2) satisfying [42]

$$R_i \le \frac{B}{2}\log\left[1+\frac{P_i}{nB}\right],$$

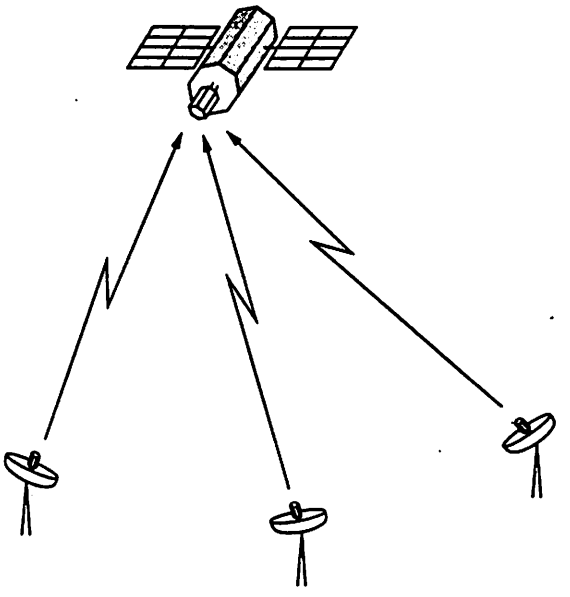

Figure 5.5: Multiaccess Channel.

$$R_1 + R_2 \leq \frac{B}{2}\log\left[1 + \frac{P_1 + P_2}{nB}\right]. \tag{5.8}$$

This region is shown in Figure 5.6, where C_i and C_i^* are given by

$$C_i = \frac{B}{2}\log\left[1 + \frac{P_i}{nB}\right], \; i = 1, 2, \tag{5.9}$$

$$C_1^* = \frac{B}{2}\log\left[1 + \frac{P_1}{nB + P_2}\right], \tag{5.10}$$

and

$$C_2^* = \frac{B}{2}\log\left[1 + \frac{P_2}{nB + P_1}\right]. \tag{5.11}$$

The point $(C_1, 0)$ is the achievable rate vector when transmitter 1 is sending at its maximum rate and transmitter 2 is silent, and the opposite scenario achieves the rate vector $(0, C_2)$. The corner points (C_1, C_2^*) and (C_1^*, C_2) are achieved using the successive decoding technique described above for superposition codes. Specifically, let the first user operate at the maximum data rate C_1. Then its signal will appear as noise to user 2; thus, user 2 can send data at rate C_2^* which can be decoded at the receiver with arbitrarily small error probability. If the receiver then subtracts out user 2's message from its received signal, the remaining message component is just users 1's message corrupted by noise, so rate C_1 can be achieved with arbitrarily small error probability. Hence, (C_1, C_2^*) is an achievable rate vector. A similar argument with the user roles reversed yields the rate point (C_1^*, C_2).

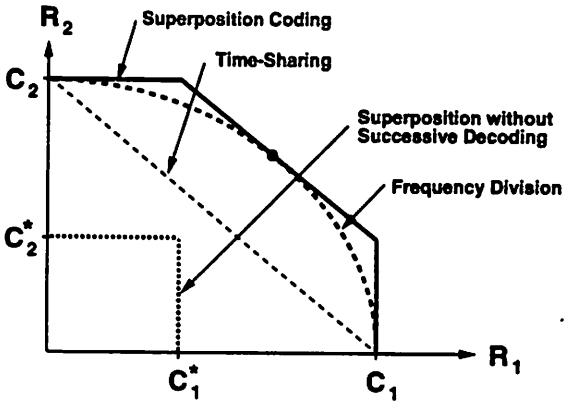

Figure 5.6: Multiaccess Channel Capacity Region.

Time division between the two transmitters operating at their maximum rates, given by (5.9), yields any rate vector on the straight line connecting C_1 and C_2. With frequency division, the rates depend on the fraction of the total bandwidth that is allocated to each transmitter. Letting B_1 and B_2 denote the bandwidth allocated to each of the two users, we get the following rate region:

$$\begin{aligned} R_1 &\leq \frac{B_1}{2}\log\left[1+\frac{P_1}{nB_1}\right], \\ R_2 &\leq \frac{B_2}{2}\log\left[1+\frac{P_2}{nB_2}\right]. \end{aligned} \tag{5.12}$$

Clearly this region dominates time division, since setting $B_1 = \tau B$ and $B_2 = (1-\tau)B$ in (5.12) yields a higher rate region than $(\tau C_1, (1-\tau)C_2)$. Varying the values of B_1 and B_2 subject to the constraint $B_1 + B_2 = B$ yields the frequency division curve shown in Figure 5.6. It can be shown [42] that this curve touches the rate region boundary at one point, and this point corresponds to the rate vector which maximizes the sum $R_1 + R_2$. To achieve this point, the bandwidths B_1 and B_2 must be proportional to their corresponding powers P_1 and P_2.

As with the broadcast multiuser channel, we can achieve the same rate region with time division as with frequency division by efficient use of the transmit power. If we take the constraints P_1 and P_2 to be average power constraints, then since user i only uses the channel τ_i percent of the time, its average power over that time fraction can be increased to P_i/τ_i. The rate region achievable through time division is then given by

$$C_i = \tau_i \frac{B}{2}\log\left[1+\frac{P_i}{n\tau_i B}\right], \; i = 1,2, \tag{5.13}$$

and substituting $B_i \triangleq \tau_i B$ in (5.13) yields the same rate region as in (5.12).

Superposition codes without successive decoding can also be used. With this approach, each transmitter's message acts as noise to the others. Thus, the maximum achievable rate in this case cannot exceed (C_1^*, C_2^*), which is clearly dominated by frequency division for some bandwidth allocations, in particular the allocation that intersects the rate region boundary. More work is needed to determine when, if ever, this suboptimal technique achieves better rates than time or frequency division.

5.2 Rate Regions for Wideband Multiaccess Channels

We now describe the capacity rate region of the Gaussian wideband multiaccess channel. This section is mainly a summary of a paper by Cheng and Verdú [77]. Recall from §3.2 that the capacity of a single-user time-invariant additive Gaussian noise channel was achieved with spectrum $S_c(f)$ given parametrically by (3.20). We first consider the two-user multiaccess channel where both channels have the same frequency response $H_1(f) = H_2(f) = H(f)$ and power constraints P_1 and P_2, respectively. In this case, the Karhunen-Lòeve expansion can be used to decompose the channel into a set of independent parallel AWGN channels with different noise levels, as in the proof of the capacity theorem for single-user wideband channels [40]. The capacity region of the wideband channel is then given by the sum of capacity regions corresponding to the individual memoryless channels. The memoryless multiaccess channel capacity region was given by (5.8); the two-user capacity region for the wideband channel, which is a sum of these memoryless regions, is [77]:

$$R_i \leq \frac{1}{2\pi} \int_0^\pi \log \left[1 + \frac{S_i(f)|H(f)|^2}{N(f)} \right] df, \tag{5.14}$$

$$R_1 + R_2 \leq \frac{1}{2\pi} \int_0^\pi \log \left[1 + \frac{S_{12}(f)|H(f)|^2}{N(f)} \right], \tag{5.15}$$

where $S_i(f), i = 1,2$ is the transmit power spectrum of user i's transmission with total power less than or equal to P_i, and $S_{12}(f) = S_1(f) + S_2(f)$ is the joint spectrum of the two users. Note that the spectrum $S_{12}(f)$ maximizing the rate sum $R_1 + R_2$ is determined by water-filling as if there was a single user on the channel with power $P_1 + P_2$:

$$S_{12}(f) = [\Lambda_{12} - N(f)/|H(f)|^2]^+, \tag{5.16}$$

where Λ_{12} is chosen such that the total power in $S_{12}(f)$ equals $P_1 + P_2$.

The capacity region of (5.15) forms a pentagon, as shown in Figure 5.7. The point $(C_1, 0)$ is achieved when only user 1 occupies the channel, which reduces it to a single-user channel; therefore, C_1 equals the capacity of the single-user channel $H(f)$ with power P_1. The transmit spectrum for user 1 which achieves this capacity is $S_1(f) = [\Lambda_1 - N(f)/|H(f)|^2]^+$, where Λ_1 is chosen such that the total power in $S_1(f)$ equals P_1. The same argument with the user roles reversed achieves the rate point $(0, C_2)$.

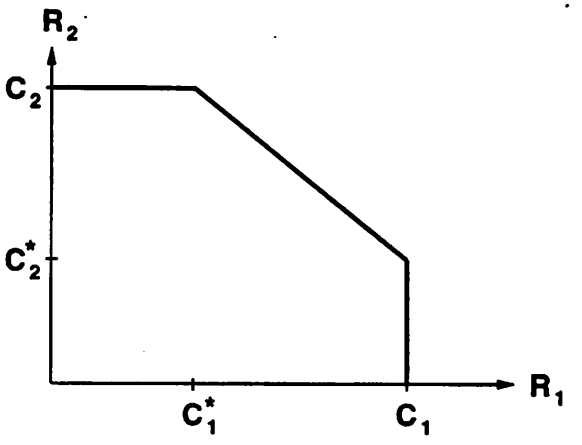

Figure 5.7: Capacity Region for $H_1(f) = H_2(f)$.

The value of C_1^* in Figure 5.7 is given by

$$C_1^* = \frac{1}{2\pi} \int_0^{\pi} \log \left[1 + \frac{S_1^*(f)|H(f)|^2}{N(f)} \right] df, \tag{5.17}$$

where $S_1^*(f) = S_{12}(f) - S_2(f)$ for $S_{12}(f)$ given by (5.16) and $S_2(f) = [\Lambda_2 - N(f)/|H(f)|^2]^+$ has total power P_2. The geometric interpretation for $S_1^*(f)$ is shown in Figure 5.8. Intuitively, the point (C_1^*, C_2) is achieved when the sum of the spectra for users 1 and 2 equals $S_{12}(f)$, and user 2's spectrum is optimal for the single-user channel $H(f)$ with power P_2. The value of C_2^* is obtained in a similar manner by reversing the roles of the two users. The line connecting points (C_1^*, C_2) and (C_1, C_2^*) is achieved through time division.

In general, it is unlikely that different users will have the same channel impulse response. When $H_1 \neq H_2$, there is no common Karhunen-Lòeve kernel that can decompose both H_1 and H_2 into sets of independent channels. However, using circular channel methods of [80, 81], an orthogonal decomposition of the channel can be found that is independent of the channel impulse response [77]. Using this decomposition and the capacity region formula derived in [82], Cheng and Verdú obtained the following expression for the capacity

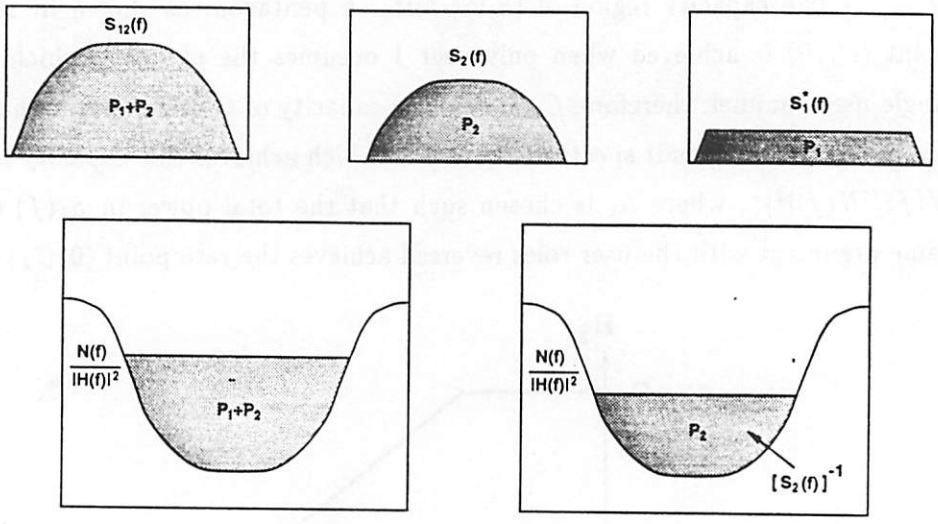

Figure 5.8: Transmit Spectra for Achieving the Rate Point (C_1^*, C_2).

region of a two-user Gaussian multiaccess channel with $H_1 \neq H_2$:

$$R_i \leq \frac{1}{2\pi} \int_0^{\pi} \log \left[1 + \frac{S_i(f)|H_i(f)|^2}{N(f)} \right] df; \quad i = 1, 2, \tag{5.18}$$

$$R_1 + R_2 \leq \frac{1}{2\pi} \int_0^{\pi} \log \left[1 + \frac{S_1(f)|H_1(f)|^2}{N(f)} + \frac{S_2(f)|H_2(f)|^2}{N(f)} \right], \tag{5.19}$$

where $S_i(f)$, the input spectrum of the ith user, is any nonnegative real-valued function with total power less than or equal to P_i.

For each i, let C_i denote the single-user capacity for channel H_i with power P_i, so C_i equals the right side of (5.18) with $S_i(f)$ obtained from the water-filling equation (3.20). The rate point $(C_1, 0)$ is then achieved when only user 1 occupies the total channel bandwidth with a transmit spectrum of power P_1 that is optimized to the single-user channel H_1. A similar argument for user 2 achieves the rate point $(0, C_2)$, and time division yields any point on the straight line connecting $(C_1, 0)$ and $(0, C_2)$. Moreover, if the channels H_1 and H_2 do not overlap in bandwidth, then the rate region (C_1, C_2) can be achieved since the users are orthogonal, and can therefore optimize their transmit spectra independently. If the channels H_1 and H_2 do overlap, then the overlapping portion of their spectra can be divided between the two users using frequency division; the optimal transmit spectra for the orthogonal frequency bands is then obtained independently via water-filling, as shown in Figure 5.9. Alternatively, the overlapping portion of the channels H_1 and H_2 can be shared

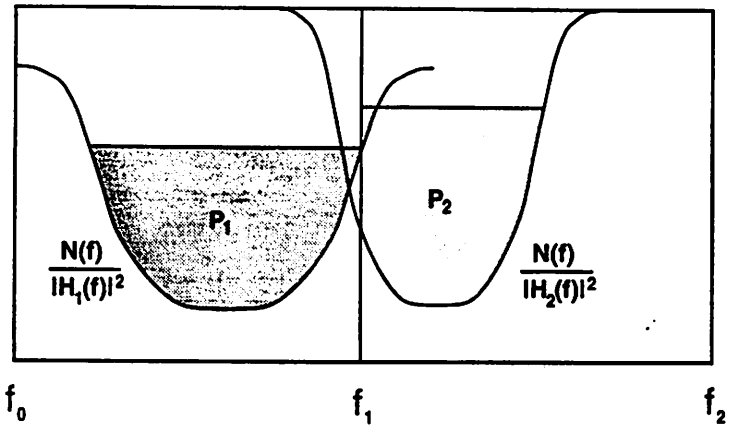

Figure 5.9: Frequency Division for $H_1 \neq H_2$.

between the two users with a superposition coding scheme. It turns out that frequency division maximizes the rate sum $R_1 + R_2$[2]. This fact was derived in [77] using equivalent channel models, where the equivalent channel is a scaled version of $H_i, i = 1, 2$, as shown in Figure 5.10. If the input power is multiplied by the scale factor k_i, then the capacity of the equivalent channel is the same as the original channel capacity, and is achieved with the input spectrum of the original channel multiplied by k_i.

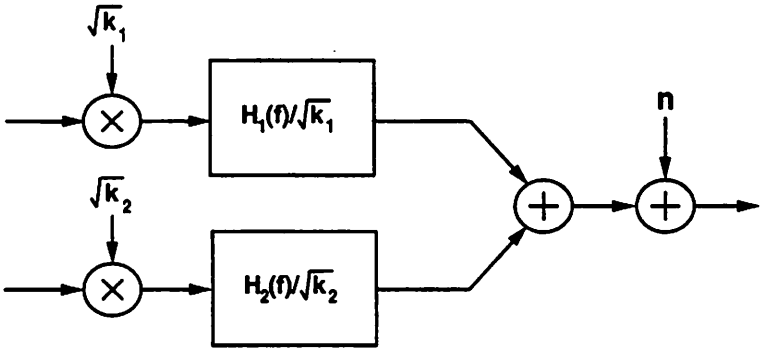

Figure 5.10: Equivalent Channel Model.

Appropriate choice of k_1 and k_2 allows the input spectral densities that maximize the rate sum to be derived via water-filling. The scaling is required since in general, the water-filling on each individual channel results in a different water level, therefore the optimal division of the channel bandwidth and power cannot be determined from a single

[2]Recall that this was also the case for multiaccess channels without ISI (§5.1.2).

diagram. Suppose, however, that we fix the water level to be 1, and plot the two curves $k_1 N(f)/|H_1(f)|^2$ and $k_2 N(f)/|H_2(f)|^2$ on the same diagram. The optimal spectrum for each user is then determined by adjusting the parameters k_1 and k_2 such that the total amount of water in the joint water-filling diagram, $k_1 P_1 + k_2 P_2$, equals one, and the amount of water in the region where $k_1/|H_1(f)|^2 \leq k_2/|H_2(f)|^2$ equals $k_1 P_1$, as in Figure 5.11.

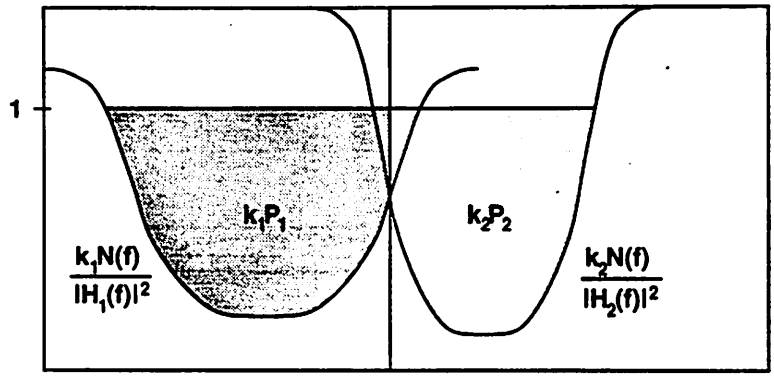

Figure 5.11: Spectral Densities for Equivalent Channel Model.

This combined water-filling maximizes the rate sum for the following reason. We want to maximize the combined rate of the two users. By scaling the two channels, we effectively reduce the equivalent two-user channel to a single-user channel with spectrum $H_{eq}(f) = \max[H_1(f)/k_1, H_2(f)/k_2]$. The spectral density which maximizes the rate on this single-user channel, S_{eq}, is determined by water-filling. But this optimal spectrum must equal the sum of the two users' equivalent channel spectra: $S_{eq}(f) = k_1 S_1(f) + k_2 S_2(f)$. Since for a particular frequency f_0, one of the two equivalent channels has a less noisy impulse response, all of the power in $S_{eq}(f_0)$ is assigned to that more favorable channel. Specifically, the optimizing spectrum for each user is given by

$$k_i S_i(f) = S_{eq}(f) 1\,[\max[H_1(f)/k_1, H_2(f)/k_2] = H_i/k_i]\,. \tag{5.20}$$

Suppose we are interested in the wideband capacity region at points other than the maximum rate sum. Then starting from the maximum rate sum point (R_1^*, R_2^*), user i can increase its rate above R_i^*, and then user $j \neq i$ must lower its rate below R_j^*. We say that the user which increases its rate has *user priority*. To achieve points on the rate region other than the maximum rate sum, the superposition coding and subsequent decoding methods of the previous section are used. The message of the low-priority user is decoded perfectly and subtracted out of the total message, so that the spectrum of the low-priority user does

not affect the high-priority message. However, the spectrum of the high-priority message is treated as noise to the low-priority user; therefore, the spectrum of both users must be designed jointly. The joint design of the spectra uses a similar technique as the equivalent channel scaling, with an offset in the water-filling formulas that is proportional to the user priority. Details of this technique can be found in [77].

5.3 Time-Varying Rate Regions

In the previous two sections, we analyzed the rate regions for multiuser time-invariant channels. We now consider the maximum achievable rates for time-varying multiuser channels with channel estimation and transmitter feedback. The rate regions for such channels combine the superposition coding ideas of the previous two sections with the single-user power control techniques outlined in §§3.1 – 3.3.

5.3.1 Narrowband Broadcast AWGN Channels

The two-user time-varying narrowband broadcast channel has one sender of average power P and bandwidth B, and two receivers with AWGN of time-varying power $n_i(t), i = 1, 2$. We assume that the set of values over which $n_i(t)$ varies is finite, and that the noise variation follows the discrete-time model of §2.4.1. The receivers have perfect channel estimation and error-free delayless transmitter feedback, so at time t the transmitter has perfect estimates of $n_1(t)$ and $n_2(t)$. The transmitter can vary its instantaneous power $P(t)$ relative to the noise samples, subject only to the average power constraint $\overline{P(t)} = P$.

We first consider the time division method of sharing the common channel bandwidth. In this case, we allocate average transmit power P and bandwidth B to the first user over the time interval $[0, \tau T]$ and to the second user over the time interval $[\tau T, T]$. This method reduces the two-user channel to a single-user channel corresponding to user 1's channel over the first time interval, and user 2's channel over the second interval. Therefore, we can maximize each user's rate over their respective time intervals independently, and the total rate region is just the sum of these maximum rates weighted by the fraction of time each user occupies the channel. Since the channel variation is stationary, the maximum rate of the second user can be calculated for transmission over the time interval $[0.(1-\tau)T]$. In the limit as $T \to \infty$, the maximum rate of each user becomes the single-user

time-varying capacity derived in §3.1. Thus, we can achieve any rate point

$$(R_1, R_2) = (\tau C_1(P), (1-\tau)C_2(P)), \quad 0 \le \tau \le 1, \tag{5.21}$$

where $C_i(P), i = 1,2$, is the capacity of the time-varying channel with average power P, bandwidth B, and time-varying noise $n_i(t)$. Applying the time-varying capacity formula (3.7), we get

$$C_i(P) = \max_{\{\Phi_{j_i}\}} \sum_{n_{j_i}: p(n_i(t)=n_{j_i}>0)} \pi_{j_i} C_{j_i}(\Phi_{j_i}), \tag{5.22}$$

where $\pi_{j_i} = p(n_i(t) = n_{j_i})$, $C_{j_i}(\Phi_{j_i})$ equals the capacity of a time-invariant AWGN channel with noise power n_{j_i}, bandwidth B, and average signal power Φ_{j_i}, and the Φ_{j_i}s are subject to the single-user power constraint

$$\sum \pi_{j_i} \Phi_{j_i} \le P. \tag{5.23}$$

Combining (5.21) with (5.22) yields the following expression for the time division rate region:

$$\begin{aligned} R_1 &\le \max_{\Phi_{j_1}} \sum_{n_{j_1}} \pi_{j_1} \tau C_{j_1}(\Phi_{j_1}), \\ R_2 &\le \max_{\Phi_{j_2}} \sum_{n_{j_2}} \pi_{j_2} (1-\tau) C_{j_2}(\Phi_{j_2}), \end{aligned} \tag{5.24}$$

where the maximum is subject to the power constraint (5.23). The time-varying power of each user can be optimized independently, since time division renders the two users orthogonal.

We can also consider (5.24) as a weighted sum of time-invariant capacity rate regions by letting the noise variances n_{j_1} and n_{j_2} represent a set of channel states. Since there is only a finite number of values for n_{j_1} and n_{j_2}, there is also a finite number of values for the variance pairs (n_{j_1}, n_{j_2}), and these variance pairs characterize the two-user channel at any point in time. Let K denote the number of distinct variance pairs (n_{j_1}, n_{j_2}), $N_k \triangleq (n_{k_1}, n_{k_2})$ denote the kth of these distinct pairs, and π_k denote the probability of the pair N_k. Suppose we allocate power Φ_{k_i} to user i when the channel is in state N_k. Since the average power of each user equals P, the Φ_{k_i}s are subject to the power constraint

$$\sum_{k=1}^{K} \pi_k \Phi_{k_i} \le P, \quad i = 1,2. \tag{5.25}$$

Let $C_{k_i}(\Phi_{k_i})$ denote the capacity of a time-invariant AWGN channel with noise power n_{k_i}, bandwidth B, and signal power Φ_{k_i}:

$$C_{k_i}(\Phi_{k_i}) = \frac{B}{2}\log\left[1+\frac{\Phi_{k_i}}{n_{k_i}B}\right]. \tag{5.26}$$

With this notation, we can express (5.24) as a weighted sum of these time-invariant rate regions with spectrum-sharing through time division. The set of achievable rates is thus

$$\begin{aligned}(R_1,R_2) &= \sum_{k=1}^{K}\pi_k(\tau C_{k_1}(\Phi_{k_1}),(1-\tau)C_{k_2}(\Phi_{k_2})) \\ &= \left(\tau\sum_{k=1}^{K}\pi_k C_{k_1}(\Phi_{k_1}),(1-\tau)\sum_{k=1}^{K}\pi_k C_{k_2}(\Phi_{k_2})\right),\end{aligned} \tag{5.27}$$

where the Φ_{k_i}s are subject to the constraint (5.25). Optimizing (5.27) subject to (5.25) defines a straight line connecting the points $C_1(P)$ and $C_2(P)$, where

$$C_i(P) = \max_{\{\Phi_{k_i}:\sum_k \pi_k\Phi_{k_i}=P\}}\sum_k \pi_k C_{k_i}(\Phi_{k_i}). \tag{5.28}$$

Fixed frequency division, which divides the total channel bandwidth B into nonoverlapping segments B_1 and B_2, also reduces the two-user channel to independent single-user channels. The total average power P can be divided between the two users in any way such that their power sum $P_1 + P_2 = P$. For P_1 and P_2 fixed, the rate region is given by

$$\begin{aligned}R_1 &\le \max_{\Psi_{k_1}}\sum_k \pi_k C_{k_1}(\Psi_{k_1},B_1), \\ R_2 &\le \max_{\Psi_{k_2}}\sum_k \pi_k C_{k_2}(\Psi_{k_2},B_2),\end{aligned} \tag{5.29}$$

where $C_{k_i}(\Psi_{k_i},B_i)$ denotes the capacity of a time-invariant AWGN channel with noise power n_{k_i}, average signal power Ψ_{k_i}, and bandwidth B_i, and the maximization is subject to the power constraint

$$\sum_k \pi_k\Psi_{k_i} = P_i. \tag{5.30}$$

As in the time-invariant case, the time division rate region will dominate the fixed frequency division rate region over some range of power allocations P_1 and P_2, in particular when all of the power is allocated to one of the frequency bands (e.g. $P_1 = P, P_2 = 0$). We now show that the fixed frequency division rate region intersects the time division line at the point where the power allocation between the two channels is proportional to the bandwidth. This was also true in the time-invariant case.

Let $(\Phi_{k_1}, \Phi_{k_2})_{k=1}^{K}$ denote the maximizing power set in (5.25) for time division, and let (B_1, B_2) denote a bandwidth partition for frequency division. The distance from a given frequency division rate point (R_1, R_2) to the time division line $(\tau C_1, (1-\tau)C_2)$ is given by

$$d = \frac{R_1}{C_1} + \frac{R_2}{C_2} - 1. \tag{5.31}$$

This distance is positive for points above the time division line, and negative for points below the time division line, as illustrated in Figure 5.12. Substituting (5.29) and (5.24) into (5.31) yields

$$d = \frac{\sum \pi_k C_{k_1}(\Psi_{k_1}, B_1)}{\sum \pi_k C_{k_1}(\Phi_{k_1})} + \frac{\sum \pi_k C_{k_2}(\Psi_{k_2}, B_2)}{\sum \pi_k C_{k_2}(\Phi_{k_2})} - 1, \tag{5.32}$$

where Ψ_{k_i} is the power allocated to frequency band B_i when the channel is in state N_k.

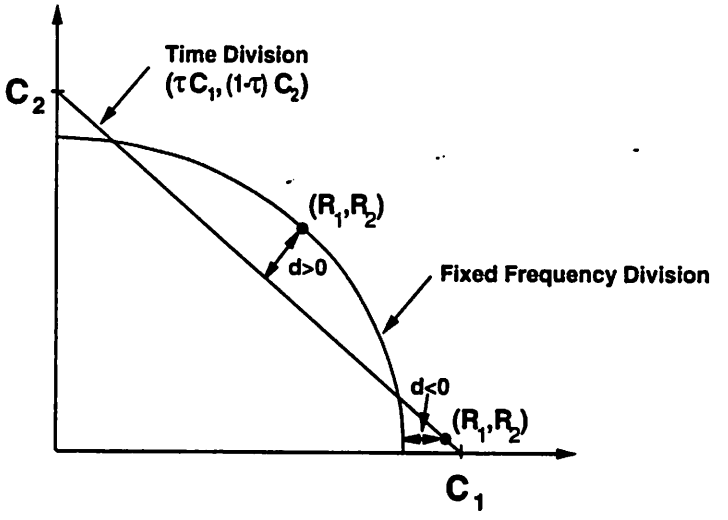

Figure 5.12: Distance Between (R_1, R_2) and the Time Division Line.

Let $\alpha_i \triangleq B_i/B, i = 1,2$ define the fraction of bandwidth allocated to user i. Suppose also that when the channel is in state N_k we let $\Psi_{k_i} = \alpha_i \Phi_{k_i}$. This power allocation satisfies the average power constraint, since

$$P_{total} = \sum_k \pi_k (\Psi_{k_1} + \Psi_{k_2}) = \sum_k \pi_k (\alpha_1 \Phi_{k_1} + \alpha_2 \Phi_{k_2}) = P. \tag{5.33}$$

The frequency division and time division rates for the ith user are given by

$$C_{k_i}(\Psi_{k_i}, B_i) = \alpha_i B \log\left[1 + \frac{\alpha_i \Phi_{k_i}}{n_{k_i} B_i}\right] = \alpha_i B \log\left[1 + \frac{\Phi_{k_i}}{n_{k_i} B}\right], \tag{5.34}$$

and

$$C_{k_i}(\Phi_{k_i}) = B \log\left[1 + \frac{\Phi_{k_i}}{n_{k_i} B}\right], \tag{5.35}$$

respectively. Substituting (5.34) and (5.35) into (5.32) yields

$$d = \frac{\alpha_1 \sum \pi_k B \log\left[1 + \frac{\Phi_{k_1}}{n_{k_1} B}\right]}{\sum \pi_k B \log\left[1 + \frac{\Phi_{k_1}}{n_{k_1} B}\right]} + \frac{\alpha_2 \sum \pi_k B \log\left[1 + \frac{\Phi_{k_2}}{n_{k_2} B}\right]}{\sum \pi_k B \log\left[1 + \frac{\Phi_{k_2}}{n_{k_2} B}\right]} - 1 = \alpha_1 + \alpha_2 - 1 = 0. \tag{5.36}$$

Thus, the frequency division and time division regions intersect at this point. We've seen in the time-invariant case that frequency division dominates a portion of the time division line over some range of power allocations. Since the values of τ for which frequency division dominates will be different for the different N_k channels, it is not obvious that this will be true in the time-varying case, although we conjecture that it is.

However, if we allow both the power and the bandwidth partition to vary for each channel N_k, then the resulting rate region dominates both fixed frequency division and time division. The achievable rates in this case are given by

$$(R_1, R_2) = \max_{(\Psi_{k_1}, \Psi_{k_2}, B_{k_1}, B_{k_2})} \sum_k \pi_k (C_{k_1}(\Psi_{k_1}, B_{k_1}), C_{k_2}(\Psi_{k_2}, B_{k_2})), \tag{5.37}$$

where the Ψ_{k_i}s satisfy the power constraint

$$\sum \pi_k (\Psi_{k_1} + \Psi_{k_2}) = P, \tag{5.38}$$

and $B_{k_1} + B_{k_2} = B$ for all k. Both the power and bandwidth allocations are optimized jointly to achieve the maximum in (5.37), so the two users are no longer independent. Clearly, any rate point achievable with fixed frequency division can also be achieved with this scheme. To show that (5.37) also dominates time division, let $(\Phi_{k_1}, \Phi_{k_2})_{k=1}^K$ denote an arbitrary set of power allocations for the time division rate region. Choose an arbitrary time division parameter τ. From §5.1.1, for a given channel N_k we can find a bandwidth partition $\{(B_{k_1}, B_{k_2}); B_{k_1} + B_{k_2} = B\}$ and power partition $\{(\Psi_{k_1}, \Psi_{k_2}); \Psi_{k_1} + \Psi_{k_2} = \Phi_{k_1} B_{k_1}/B + \Phi_{k_2} B_{k_2}/B\}$ such that the frequency division rates achieved with these parameters dominate the time division rates. Therefore, the weighted average of the frequency division rates will dominate the weighted average of the time division rates.

The idea of reallocating bandwidth as the channel varies is closely related to dynamic channel allocation, where each user measures the noise (and interference) in a particular frequency band, and only occupies the frequency band if the noise is below some

threshold [83]. Suppose two users want to access the same frequency band, and the noise level is below threshold for both, but lower for one of the users. The frequency allocation of (5.37) suggests that instead of using a threshold level to determine which user should occupy the channel (which for this example would not differentiate between the two users), the channel is allocated to the user which gets the most capacity from it.

We now consider superposition coding, where both users occupy the full channel bandwidth over all time. The achievable rates are again weighted averages of the achievable rates on each channel N_k:

$$(R_1, R_2) = \sum \frac{\pi_k B}{2} \left(\log \left[1 + \frac{\Gamma_{k_1}}{n_1 B + \Gamma_{k_2} 1[n_{k_1} \geq n_{k_2}]} \right], \log \left[1 + \frac{\Gamma_{k_2}}{n_2 B + \Gamma_{k_1} 1[n_{k_2} > n_{k_1}]} \right] \right), \tag{5.39}$$

where $1[\cdot]$ is the indicator function and the Γ_{k_i}s must satisfy the power constraint

$$\sum \pi_k (\Gamma_{k_1} + \Gamma_{k_2}) = P. \tag{5.40}$$

Since superposition codes dominate time and frequency division in the time-invariant case, we expect this to be true for time-varying channels as well. Indeed, consider any achievable rate point in the frequency division rate region (5.37). Associated with that point will be a set of frequency divisions $(B_{k_1}, B_{k_2})_{k=1}^K$ and a set of transmit power values $(\Psi_{k_1}, \Psi_{k_2})_{k=1}^K$ for each of the K channel states. Let $\Psi_k = \Psi_{k_1} + \Psi_{k_2}$. From §5.3.1 there exists a superposition code with total power Ψ_k that dominates the frequency division code on channel N_k. Since we can find such a dominating code for all k, the weighted sum of the superposition rates dominates the frequency division sum.

5.3.2 Narrowband Multiaccess AWGN Channels

The two-user time-varying narrowband multiaccess channel has two transmitters with average power P_1 and P_2, respectively, and one receiver with bandwidth B and AWGN of time-varying power $n(t)$. We assume that $n(t)$ varies over a finite set of values $n_1, \ldots, n_K$, so n_k characterizes the channel state with probability $\pi_k \triangleq p(n(t) = n_k)$. We also assume that at time t both transmitters have perfect estimates of $n(t)$. The transmitters may vary their instantaneous transmit power $P_i(t)$ relative to $n(t)$, subject only to the average power constraint $\overline{P_i(t)} = P_i$ for $i = 1, 2$.

We first consider spectrum sharing through time division. With this technique we can achieve any point

$$(R_1, R_2) = \sum_{k=1}^{K} \pi_k \left(\tau C_k(\Phi_{k_1}), (1-\tau) C_k(\Phi_{k_2})\right), \tag{5.41}$$

where

$$C_k(\Phi_{k_i}) \triangleq \frac{B}{2} \log\left[1 + \frac{\Phi_{k_i}}{n_k B}\right], \tag{5.42}$$

and Φ_{k_i}, the power allocated to the ith user for channel state n_k, is subject to the average power constraint

$$\sum_k \pi_k \Phi_{k_i} = P_i. \tag{5.43}$$

The Φ_{k_i}s in (5.41) can be optimized independent of each other, since under time division the two users are orthogonal. Optimizing (5.41) subject to (5.43) therefore defines a straight line connecting the points $C_1(P_1)$ and $C_2(P_2)$, where

$$C_i(P_i) = \max_{\{\Phi_{k_i}: \sum_k \pi_k \Phi_{k_i} = P_i\}} \sum_k \pi_k C_k(\Phi_{k_i}). \tag{5.44}$$

Fixed frequency division partitions the total bandwidth B into nonoverlapping segments B_1 and B_2, which are then allocated to the respective transmitters. Since the bandwidths are separate, the users are independent, and they can allocate their time-varying power independently, subject only to the total power constraint P_i. The fixed frequency division rate region is thus given by

$$\begin{aligned} R_1 &\leq \max_{\Psi_{k_1}} \sum_k \pi_k C_k(\Psi_{k_1}, B_1), \\ R_2 &\leq \max_{\Psi_{k_2}} \sum_k \pi_k C_k(\Psi_{k_2}, B_2), \end{aligned} \tag{5.45}$$

where

$$C_k(\Psi_{k_i}, B_i) = \frac{B_i}{2} \log\left[1 + \frac{\Psi_{k_i}}{n_k B_i}\right], \tag{5.46}$$

and the Ψ_{k_i}s satisfy the power constraint

$$\sum \pi_k \Psi_{k_i} = P_i. \tag{5.47}$$

We now show that fixed frequency division dominates time division. Suppose we use the power allocations Φ_{k_i} which achieve the maximum time division rate in (5.44) for a fixed frequency division scheme. This power allocation is included in the set over which

(5.44) is maximized, so we need only show that frequency division dominates time division in this case. Using these Φ_{k_i}s, we achieve the frequency division rates

$$(R_1, R_2) = \sum_k \pi_k (C_k(\Phi_{k_1}, B_1), C_k(\Phi_{k_2}, B_2)). \tag{5.48}$$

The rate vector (5.48) is a linear combination of fixed frequency division rate vectors for the n_k channels. From §5.1.2, varying the bandwidth partition (B_1, B_2) yields a convex capacity region for each channel n_k. Therefore, varying B_1 and B_2 in (5.48) over a range of values for which $B_1 + B_2 = B$ yields a linear combination of convex regions, so the resulting capacity region is convex. Moreover, since the power allocations are the same, the endpoints of the regions defined by (5.48) and (5.44) are also the same (i.e., allocating all the transmission time to user i with time division is the same as allocating all the bandwidth to user i with frequency division). Since the time division boundary is linear, the frequency division boundary is concave, and the two boundaries have the same endpoints, fixed frequency division strictly dominates time division with this power allocation.

We conclude by showing the dominance of superposition codes over frequency division. Consider any point on the boundary of the frequency division rate region, as given by (5.48). Corresponding to that point will be a bandwidth partition (B_1, B_2) and a set of transmit power values $(\Phi_{k_1}, \Phi_{k_2})_{k=1}^K$ for each of the K channel states. Then from §5.1.2, a superposition code can achieve any rate point

$$\begin{aligned} R_i &= \frac{B}{2} \log \left[1 + \frac{\Phi_{k_i}}{nB} \right], \\ R_1 + R_2 &= \frac{B}{2} \log \left[1 + \frac{\Phi_{k_1} + \Phi_{k_2}}{n_k B} \right], \end{aligned} \tag{5.49}$$

and within this region there is a rate point which dominates frequency division on channel n_k. Thus, a linear combination of (5.49) dominates (5.48).

5.4 Interference in Cellular Systems

The capacity results above assume multiple users sharing the same frequency band through either an orthogonal (FDMA/TDMA) or semi-orthogonal (CDMA) partition of the spectrum. As was discussed in §2.3, the spectral efficiency over a large geographical area for any of these partition techniques can generally be increased by reusing the same frequency, time slot, or code at spatially separated cells, where the power falloff with distance reduces

the effect of the intercell interference. The magnitude of the intercell interference depends on both the distance between the interfering transmitters and the intended receiver as well as the propagation laws governing the interferers' transmissions.

The interference distribution is generally assumed to be Gaussian. This is a reasonable assumption for CDMA systems, where there are many intracell and intercell interferers, so the Gaussian distribution follows from the law of large numbers. With FDMA or TDMA, however, there is usually only a few dominant interferers[3], so the white noise assumption is generally not valid. For capacity calculations, Gaussian interference is a worst-case noise assumption [84], and under this assumption the capacity-achieving transmit spectrum for all users (i.e. signal and interference) is Gaussian. Most cellular systems are *interference limited*, meaning that the receiver noise power is generally much less than the interference power, and can hence be neglected.

In the following sections, we first define the *area spectral efficiency*, which quantifies the effect of in-cell and out-of-cell interference on cellular system capacity. We then outline some methods of interference mitigation including antenna sectorization, voice activity monitoring, and interference cancellation. We also discuss the effects of power control on intracell and intercell interference, and conclude with a proposal for a hybrid power control algorithm which adapts to the system traffic load, channel characteristics, and performance specifications of each user.

5.4.1 Reuse Distance and Area Efficiency

Let the radius of a cell be normalized to one, and define the *reuse distance* R_D to be the minimum distance between any two base stations that use the same code, frequency, or time slot. The reuse distance is illustrated in Figure 5.13 for frequency division. Let the area spectral efficiency of a cell be defined as the total bit rate/Hz/unit area that is supported by a cell's base station. Since a code, time slot, or frequency slot is reused at a distance R_D, the area covered by one of these partitions is roughly $\pi(.5R_D)^2$. The area spectral efficiency is therefore approximated by

[3]The interference comes from the closest ring of cells (Figure 2.7). On the forward link, one or two mobiles which are close to the cell boundaries will generally dominate the interference. On the reverse link, there are at most six interfering base stations for hexagonal cells.

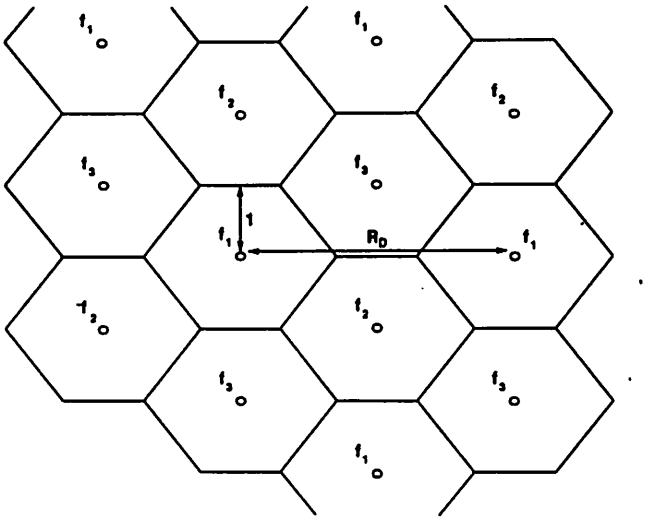

Figure 5.13: Reuse Distance.

$$A_e = \frac{\sum_{j=1}^{N} R_j/B}{\pi(.5R_D)^2}, \tag{5.50}$$

where N is the total number of users per cell, R_j is the data rate of the jth user, and B is the bandwidth occupied by each user.

If we can shrink the reuse distance without changing the R_js, then the area efficiency can be increased. However, decreasing the reuse distance increases intercell interference (since the interference travels a shorter distance), thereby reducing the S/I of each user. Since R_j is an increasing function of S/I, the numerator and denominator of (5.50) are both increasing functions of R_D. Therefore, in order to maximize the area efficiency relative to R_D, we must first determine R_j for all j as a function of R_D, then maximize (5.50) relative to R_D.

As a simple example, consider an FDMA multiple access channel where the signal-to-interference power of each user is exponentially distributed (as in Rayleigh fading). With FDMA there is only one user per cell, so $N = 1$ in (5.50), and we only need to find the rate R of this one user as a function of R_D. Assume that we have hexagonal cells of diameter one. If the signal is transmitted from a midpoint between the base station and the cell boundary, then the signal travels a distance of .25. If we assume a power falloff with

distance of d^{-2}, the average received signal power is $\overline{S} = P_t(.25)^{-2}$, where P_t is the average transmit power for both the signal and the interferers. We assume a pessimistic interference model, with six interferers at the boundaries of the closest adjacent cells using the same frequency band. The distance that each interferer travels is therefore $R_D - .5$, the frequency reuse distance minus the cell radius. Since intercell interference generally travels a much farther distance than the signal, we assume an interference power falloff with distance of d^{-4}, so the average interference power is $\overline{I} = 6P_t(R_D - .5)^{-4}$. We use this model to obtain the average signal-to-interference $\overline{\gamma} = \overline{S}/\overline{I}$ as a function of reuse distance, then calculate R from the single-user time-varying capacity formula (3.25). This calculation yields the following table of efficiency values as a function of R_D.

R_D	$\overline{\gamma}(dB)$	R	A_e
1	-7.78	.34	.43
2	11.30	3.34	1.06
3	20.18	5.95	.84
4	26.02	7.80	.62

The table suggests that for this simple model, $R_D = 2$ maximizes area efficiency.

The calculation of R_j from (3.25) will depend on the distribution of the jth user's signal-to-interference ratio γ_j. In the previous example we assumed a Rayleigh distribution, independent of the power control policy; in general this distribution will depend on the power control policy of both the signal and the interferers. Therefore, the power control policy that maximizes a single user's data rate may not maximize the area efficiency, since increasing the signal power of one user increases that user's interference to everyone else. Determining the power control policy that maximizes area efficiency is a complex optimization problem which will depend on the spectrum partitioning technique, propagation characteristics, system layout, and the number of users. This optimization may be too complex for analysis, and therefore suboptimal techniques must be considered. We propose such a scheme, which combines some of the benefits of both the constant power and the water-filling power control policies, in §5.4.4.

If we fix the power control policy, and assume a particular set of system parameters, then the distribution of γ_j can be determined either analytically or via simulation. The distribution of γ_j for CDMA systems (i.e., with both intracell and intercell interference), assuming Gaussian interference and the constant power control policy, has been determined analytically in [38, 85, 86], and via simulation in [87, 88]. The distribution of γ_j for CDMA

under other power control policies, and for FDMA and TDMA under any form of power control, has not yet been determined. With these distributions, a comprehensive comparison of area efficiency under different power control policies and spectrum partitioning methods could be done using the methods described above. However, even if such a study finds that water-filling achieves the highest area efficiency, it still might not be the best power control policy to use, since the user data rates change with the channel variation, and therefore cannot be guaranteed over any given time interval. In §5.4.4 we will discuss some of the performance tradeoffs other than area efficiency which must be considered in the design of power control policies.

5.4.2 Interference Mitigation

The area efficiency for any of the three spectrum-sharing techniques will be increased if interference can be reduced while maintaining the same number of users per cell and the same reuse distance. Several techniques have been proposed to accomplish this, including speech gating, sectorization of the base station antennas, and interference cancellation. We now describe each of these techniques in somewhat more detail.

Speech gating takes advantage of the fact that in duplex voice transmission, each speaker is only active approximately 40% of the time [89]. If voice activity is monitored, and transmission suppressed when no voice is present, then overall interference caused by the voice transmission is reduced. If we denote the average percentage of time that voice is active by ρ, then through speech gating the average power of both intracell and intercell interference is reduced by ρ.

Antenna sectorization refers to the use of directional transmit and receive antennas at the base station. For example, if the 360^o omni base station antenna is divided into three sectors to be covered by three directional antennas of 120^o beamwidths, then the interferers seen by each directional antenna is one third the number that would be seen by the omni. If N_S denotes the number of directional antennas used to cover the 360^o beamwidth then, on average, antenna sectorization reduces the total interference power by a factor of N_S.

Another method of mitigating interference in CDMA systems is multiuser detection. The received CDMA signal is a superposition of each user's signal, where user i modulates its data sequence with a unique spreading code. The multiuser detector for such a received signal jointly detects the data sequences of all users: if the data sequences of

the interference is known, then it can be subtracted out from the desired signal, as in the superposition coding techniques described above. The optimal receiver for CDMA joint detection was derived by Verdú in [90]; it uses a bank of matched filters and the Viterbi algorithm to determine either the maximum-likelihood set of received signal sequences or the set of signal sequences with minimum error probability. However, the complexity of this optimal receiver structure is exponential in the number of interfering users, making the receiver impractical for systems with many interferers. The detection algorithm also requires knowledge of the signal energies, which is not always available.

Several suboptimal multidetection schemes which are more practical to implement have also been developed. A multiuser decorrelator for joint detection which does not require knowledge of the user energies and with complexity that is only linear in the number of users was proposed in [91] and [92] for synchronous and asynchronous users, respectively. Multistage detectors [93, 94] decode the users' signals sequentially in decreasing order of their received power. Specifically, the highest-power signal is detected using a conventional CDMA receiver (i.e., all interference signals are treated as noise). This signal is then subtracted from the total received signal, and then the highest-power remaining signal is detected. This successive interference cancellation is done until all signals have been estimated. The decision-feedback detector, proposed in [95], uses both forward and feedback filters to remove multiuser interference. As with decision-feedback equalization, this approach suffers from error propagation. The multistage detectors generally yield better performance than the decorrelator and decision-feedback detectors at a cost of increased complexity (although still linear in the number of users). These detectors were designed for AWGN channels, while more recent studies have looked at multiuser detection in fading channels [96, 97].

A common interference problem for CDMA systems with conventional detectors is the "near-far" effect on the forward link, which refers to a signal having a poor transmission path to the base station and the interferers having strong paths. In this case, the interference power is still quite large even after despreading. Power control can reduce the near-far effect, as we will discuss in the following section. However, multiuser detection schemes inherently compensate for the near-far effect, since they are designed to detect all signals jointly. Since strong interferers are easily detected and subtracted out, the multiuser detection schemes generally work best when the received signals are at different power levels. Thus, the water-filling power control policy might be well-suited for a CDMA system with multiuser

detection, since the data rates of some users can be increased without causing degradation to the weaker users.

5.4.3 Power Control Impact on Interference

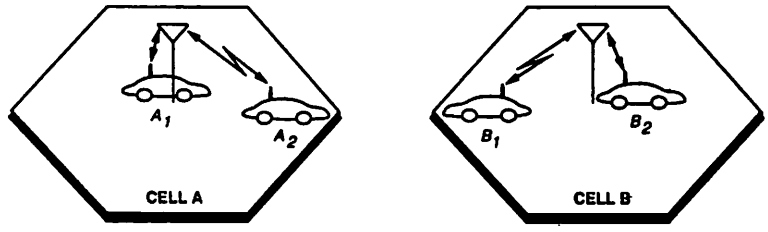

Figure 5.14: Interference Effects.

In this section we describe the impact of power control on intracell and intercell interference, including the near-far effect discussed above. Consider first the case of intracell interference on the forward link (mobile to base station), where two users A_1 and A_2 are transmitting to the same base station, as shown in Figure 5.14. Recall that intracell interference only occurs in CDMA systems, since with FDMA or TDMA only one user is assigned to each frequency or time slot in the cell. If both A_1 and A_2 transmit at the same power level, then the signal received by the base station from A_1 will generally be stronger than the signal received by A_2. Therefore, the interference caused by A_1 to A_2 will be strong even after despreading. This difference in received signal strength is called the near-far effect. To compensate for this effect, the constant power policy of §3.3.2 is used to equalize the receive power of all users within a cell. With this policy, the received power of users A_1 and A_2 at the base station is the same, regardless of their individual path losses, so the signal-to-interference power after receiver processing equals the spreading gain. The water-filling policy of §3.3.1 has the opposite effect: since A_1 has a good signal path it will increase its transmit power, while A_2 has a bad signal path, so it will decrease its signal power. Moreover, this policy has a recursive effect: A_1 increasing its power causes A_2 to have an even worse channel, so A_2 will lower its power. This decreases the interference to A_1, so A_1 increases its power further, and so on. Roughly speaking, the constant power policy equalizes the performance of all users in the cell, while water-filling tends to *remove* all users from the cell except the one with the most favorable channel. Therefore, if we consider only intracell interference effects, the water-filling policy is unacceptable when all

the users within a cell require a guaranteed rate at all times. However, it may have a higher throughput in a system where the users within a cell can tolerate long periods of no transmission with an occasional burst of very high-rate data, as in packet radio systems. This assumes that all the users within a cell will eventually have the best signal path to the base station.

The effect of these two power control policies on intercell interference is quite different. Again refering to Figure 5.14, suppose we have intercell interferers B_1 and B_2 from cell B coupling into cell A. Without power control, the interference power from B_1 will be strong, since it is close to the boundary of cell A, while the interference from B_2 has much farther to travel to the base station of cell A, and will therefore be weaker. With the constant power policy, B_1 will transmit at a high power since it is far from its base station, and this will cause a higher level of interference in cell A. Since B_2 reduces power with this policy, and it is far from cell A's base station, the constant power policy has the effect of magnifying the power of interferers near cell B's boundary while reducing the power of interferers close to cell B's base station. Conversely, the water-filling power control will cause B_1 to lower its power and B_2 to increase its power, so that the intercell interferers in cell B have approximately the same amount of power coupling into cell A's base station, regardless of their location in cell B. Since the dominant intercell interferers are generally near the cell boundaries, water-filling will significantly reduce intercell interference on the forward link.

For the reverse link, the intracell interference and signal are both transmitted from the base station, so their path loss at any point within cell A is the same. Therefore, no power control is required to equalize the received signal strength of the signal and interference (equivalently, the constant power policy for the reverse link is achieved with no power control). Water-filling power control has the same recursive effect as in the forward link: since A_1 has a good path, the base station transmits to A_1 at a high power, which will cause interference to A_2, so transmit power to A_2 is reduced, and so on. Hence, the effect of these two power controls policies on intracell interference is roughly the same for both the forward link and the reverse link.

For intercell interferers, if the base station is sending to B_1 and B_2 at the same power level, then the location of B_1 and B_2 will not affect the amount of power coupling in to cell A. With water-filling, the base station will send at a higher power to B_2 and a lower power to B_1, but these interference signals have the same path loss to the mobiles in

cell A. Therefore, it is difficult to say which power control policy will cause worse intercell interference on the reverse link.

5.4.4 Hybrid Power Control

We now propose a hybrid power control scheme that incorporates the benefits of both constant power control (guaranteed performance for all users) and water-filling (increased efficiency when the channel is favorable). This scheme has the capability to accommodate different performance specifications for each user in the system, and to determine when additional users can access the system.

The "best" power control policy, as part of the overall system design, will depend on the performance criterion of each user, as well as on the overall system requirements. These criteria may include average efficiency, minimum guaranteed data rate, outage probability, delay constraint, total system throughput, maximum/minimum number of users, and the overall system complexity. Many of these criteria require tradeoffs; for example, we've seen in the single user case that the constant power control policy is fair to all users in the system but has a lower average efficiency than water-filling.

It is also desirable that the system accommodate heterogeneous traffic with different performance criteria, for example a high-rate user with delay-tolerant data and a user with low rate delay-constrained voice traffic. Moreover, if the overall traffic on the system is low, then additional users should be able to access the system. We now propose a power control and adaptive data rate scheme which combines both water-filling and constant power control to achieve these two goals. The basic idea is to provide a higher level of performance to users with favorable channels while maintaining a minimum performance threshold for all users.

The algorithm requires global system knowledge in the base stations. Specifically, we assume that each base station knows the transmit and receive power level of mobiles within its cell, and in the adjacent interfering cells. Equivalently, the base stations know the transmit power level and path loss of all intracell and intercell mobiles. Typically, a base station only knows the transmit power level of mobiles within its own cell, and the total interference level. The additional information we require can be obtained by an information exchange between the base stations of their mobiles' transmit power levels, and through transmission of base station pilot tones to determine the path loss values for

intercell mobiles [98].

There are basically three steps to the algorithm, which we summarize below and then describe in more detail.

1. Determine the power control vector for all users to maintain the threshold received SNR required for minimum performance. This defines the feasibility region.

2. Increase the power of the users with the best channels by some increment. Confirm that the new vector is feasible. Continue the process until a user drops below its minimum required SNR.

3. As conditions change and/or new users request access to the system, the power control vector is updated.

We initially assume that all users are transmitting at a power level such that their received S/I is sufficient to maintain the minimum performance specification (data rate and BER) specified by each user. The S/I of user i in cell j is given by

$$S_{ij} = \frac{G_{i_j} P_{i_j}}{\sum_{m=1}^{M} \sum_{k=1, k \neq i}^{K_m} G_{k_m}^{j} P_{k_m} B}, \tag{5.51}$$

where G_{i_j} is the path loss from the ith user in the jth cell to its base station, P_{i_j} is the transmit power of the ith user in the jth cell, M is the total number of cells, K_m is the number of users in the kth cell, $G_{k_m}^{j}$ is the path loss of the kth user in the mth cell to the base station in the jth cell, and B is the spreading gain if the system is CDMA (otherwise $B = 1$). Since all these parameters are known, it is easy to see if the minimum S/I specifications for each user are met for a particular set of transmit power levels, or power vector. A power vector which satisfies this specification lies in the *feasibility region* of all possible power vectors. We assume that new users are only allowed on to the system if a set of transmit powers can be found which lies in this feasibility region. Necessary and sufficient conditions for the existence of feasible power vectors for a TDMA system have been derived in [98].

Suppose we have a vector in the feasibility region, so that all users are operating at their minimum S/I specification. If there is excess capacity available in the system, then how should it be divided among all the users? One method would be to increment the power of all users equally by an amount such that the power vector still lies in the feasibility region.

Such an increase would allow all the users to either increase their data rates or reduce their BERs. However, some of the users might not be able to increase their data rate or gain much from a decreased BER; for example, users with only voice traffic. Moreover, a slight increase in power of one mobile might cause a large increase in interference to other mobiles, so that this power increase is not being used efficiently. Based on the water-filling policy, we therefore propose that the excess capacity be allocated to the users with the best channels (high S/I), since they can gain the most capacity increase by a small power increase. This is the philosophy behind the second step in the power control algorithm: the users with the best channels are allowed to increase their power by some increment until the S/I of all users lies on the boundary of the feasibility region. This also allows a single user to have heterogeneous traffic, so that its minimum S/I can be specified for constant rate (voice and data) traffic, and when a good channel is available, packetized high-rate data can be sent. The intercell interference caused by this power increase is small, since generally the mobiles with good channels will be close to their base stations (and therefore far away from other base stations). However, intracell interference would be increased.

Since the system is dynamic, the path loss factors will be constantly changing, which will require adjustments to the transmit power values such that they remain in the feasibility region. Suppose now that we have completed the second step of the algorithm, and the transmit power vector is on the feasibility region boundary. If a new user requests access to the system at a particular minimum S/I, then clearly some of the power levels must be lowered to accommodate it. Since some of the users are operating above their minimum S/I ratios, these users can return to these minimum levels, and then the feasibility of the system with the new user can be confirmed. If there is no such feasible power vector, then the new user is either denied access, or granted access at a lower S/I which can achieve a feasible vector.

The advantages of this algorithm are its adaptability to changing conditions, its efficient allocation of excess system capacity, its ability to accommodate heterogeneous users with different data requirements, and its built-in capability to process new access requests. Obviously much work remains to specify the exact details of the algorithm and determine its performance.

5.5 Summary

In this chapter we first reviewed the achievable rate regions for multiuser time-invariant AWGN channels under TDMA, FDMA, and CDMA spectrum-sharing techniques. We consider both broadcast and multiaccess channels, which model the reverse and forward links, respectively, of a cellular system. We show that for both these channels, TDMA and FDMA are equivalent if the TDMA power can be varied, and both these techniques are inferior to CDMA with interference cancellation. We also show that without interference cancellation, CDMA is inferior to the other techniques. We then combine the time-varying single-user analysis of §3.3 with these rate region results to obtain the multiuser time-varying rate region for narrowband broadcast and multiaccess channels. In general, the relative performance of TDMA, FDMA, and CDMA is the same in this case as in the time-invariant case.

These rate region results cannot be applied directly to cellular systems, since frequency reuse is not taken into account. We therefore define the area efficiency as the data rate/Hz/unit area, with interference effects included in the data rate calculation. We compute the area efficiency for a simple interference model, and show area efficiency can be used to determine the optimal frequency reuse distance. We also discuss some methods of interference mitigation such as antenna sectorization, voice gating, and multiuser detection.

Power control is commonly used in CDMA systems to equalize interference within a cell, however this aggravates the intercell interference. We analyze the impact of the water-filling and constant power control policies on both intracell and intercell interference. We then use the general conclusions of this analysis to propose a hybrid power control policy which exploits the advantages of both policies. This hybrid scheme is also adaptive to changes in channel conditions, user requirements, and overall system loading.

Chapter 6

Wireless Networks

The wireless networking vision for the next decade, as shown in Figure 6.1, is to provide high-speed, high-quality mobile voice and data communication anywhere and any time. The previous chapters addressed techniques for the single-user and multiuser wireless communication links that support these applications. We now consider *internetworking* of various wireless subnetworks. The network infrastructure must be able to support different applications with very different data types, coverage requirements, and system specifications. In addition, this infrastructure must provide seamless communication between the different wireless applications, as well as interconnection to the backbone wireline network supporting the Public Switched Telecommunications Network (PSTN) and Integrated Services Data Network (ISDN), as well as the Internet.

We begin this chapter by outlining the various wireless applications currently in demand. We then examine the infrastructure necessary to interconnect these different applications. In Figure 6.1 internetworking is accomplished via a wireless gateway which connects the wireless subnets to each other and to a high-speed fiber backbone. However, this infrastructure creates an enormous bottleneck at the gateway. A hierarchical infrastructure alleviates this problem, and also provides more flexibility to accommodate the different requirements of the various subnetworks. We propose such an infrastructure in §6.2.

Another major design element in the network is mobility management and routing. Specifically, the network must be able to locate and route data between hundreds of millions of mobiles located over a very large geographical area in an efficient manner. Existing techniques for location of mobile units are paging and registration. The paging technique is very wasteful of bandwidth if the mobiles are located over a large geographical area, so it

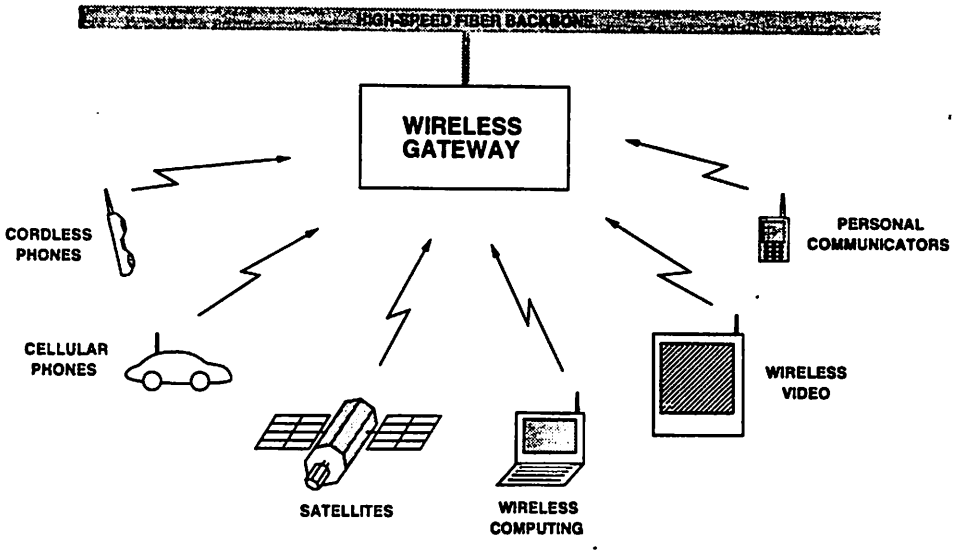

Figure 6.1: Wireless Vision.

is certainly impractical for routing between the subnetworks of Figure 6.1. The registration technique currently requires a mobile to register its location at a central database whenever it is outside its home location. The central database will quickly fill up for a large number of users, suggesting that distributed databases for mobile location must be used. In §6.3 we propose mobility management techniques similar to those used for call-forwarding in the wireline Intelligent Network [99] and roaming in cellular systems. Routing strategies through both the wireless and wireline infrastructure are also discussed in this section.

We conclude with a brief discussion about some other issues in the wireless network design, including network security, pricing, and control.

6.1 Wireless Applications

In this section we describe the various wireless applications that are currently in demand. We will also reference existing or proposed systems that meet these demands. More details on these systems can be found in [100, 101] and the references therein.

1. Voice communication in or near the home. This demand is partially met by existing cordless phone technology. Coverage of these systems is currently limited to within close range of the wireless base. Second-generation cordless phones (DECT, CT-2) aim to increase coverage by allowing the wireless headset to access many base stations

within a given area, where the base stations are coordinated by a central switch. The headsets are assumed to be stationary or slowly-moving, so there is no handover between base stations. These base stations will also be set up in public areas like airports and shopping centers.

2. Voice communication in offices. The second-generation cordless phones will also act as wireless PBXs to provide voice communication throughout office buildings. Some handover in these systems may be required as people move between floors or down long corridors.

3. Voice communication in vehicles. This demand is currently being met in the U.S. with the analog cellular system AMPS. Second-generation systems are digital, providing greater capacity and voice quality. Several different standards have been proposed for these systems, including the GSM standard for Europe, the JDC standard for Japan, and the IS-54 and IS-95 standards for the United States. None of these standards are compatible. Further increases in capacity will be achieved by shrinking the cell size from its current one to five mile radius (macrocell) to a one thousand foot radius (microcell).

4. Ubiquitous low-speed data and voice devices - the personal communicator. These devices are targeted for use in homes and offices, as well as outdoors in residential and urban areas. The devices must be "pocket-sized", hence low-power. These communicators are similar in concept to pagers, except that they allow two-way communication and real-time voice. No products have yet been developed for this application.

5. High-speed data in buildings. These systems are oriented towards replacing the Ethernet with more easily configurable wireless network. Existing products include Motorola's Altair and NCR's WaveLAN, both operating around 5 Mbps. Higher-rate systems are still in the research stage. Many of the proposed high-rate systems are asymmetric, with a very high-speed (10-100Mbps) broadcast channel transmitting to low-power portable devices that return data at much slower rate. Due to the high-speed requirements of the base station, and the power restriction in the portable devices, the coverage area of these systems is small, on the order of several meters.

6. Global low-speed packet data. This need is partially met through current satellite paging systems, which are almost exclusively one-way. Satellite systems providing

global full-duplex store-and-forward packet data services include Qualcomm's Omni-TRACKS system at 5-15 Kbps and Geostar at 1.2 Kbps.

7. Global and regional voice and data. Geostationary and low-earth orbit satellite systems which provide voice and data (5-20 Kbps) transmission with global or regional coverage include Inmarsat, MOBILESAT, and MSAT.

The ultimate goal of the wireless revolution is ubiquitous high-rate data and voice communication through a single low-power portable device. However, due to the different requirements and coverage areas involved in such a system, it is unlikely that, given foreseeable technology, this goal will be met any time soon. Therefore, we will concentrate in this chapter on the wireless subsystems designed to meet the needs enumerated above.

6.2 Network Architectures

Existing network architectures can be divided into basically two categories: circuit-switched networks designed for voice and packet-switched networks designed for data. In this section we will first review several existing architectural paradigms for these two types of networks. We then use these examples as a baseline to sketch a wireless network infrastructure.

6.2.1 Circuit-Switched Network Architecture

In circuit-switched networks, two users that wish to communicate must first establish a dedicated transmission path between them which is held throughout the duration of their transmission. The dedicated line insures sequential arrival of the data, and the data delay consists of the time necessary to establish the connection. These features make circuit-switching the preferred technology for voice telephony. In this section we will discuss the architecture for the PSTN, and its extension to cellular and cordless phone architectures.

The circuit-switched architecture for the PSTN is shown in Figure 6.2. The traffic is generated from either telephones or data sets (such as modems), which are connected with dedicated telephone lines to a local exchange office. If the call destination is not directly connected to the local exchange, then the exchange determines the next local office on the route to the final destination, and requests a connection to this office on the trunk (set of multiple lines) connecting them. The trunk transmission may be via copper wire, fiber, or

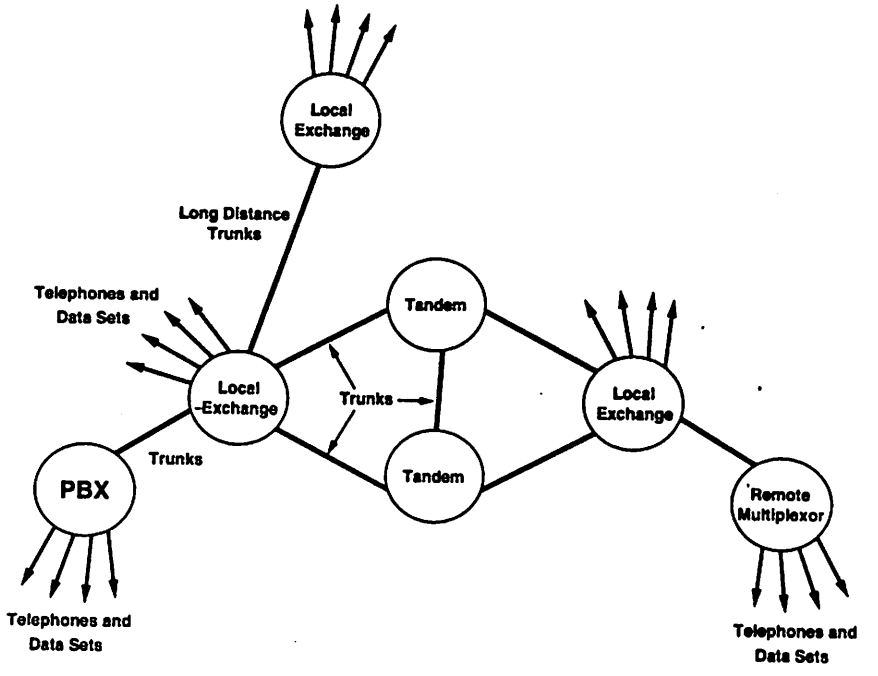

Figure 6.2: PSTN Architecture.

satellite radio. All trunks are full-duplex connections, so communication can take place in either direction. If a trunk is available, it is reserved for this call. Each local exchange office on the route to the final destination determines the next office along the route, and establishes a trunk connection with this office. When the connection to the local exchange office at the final destination is made, a dedicated line between the call initiator and final destination is established through these reserved trunks, and data transfer can commence. The route can either go through several other local exchange offices, or along a long-distance trunk line. Local exchange offices which do not generate local traffic but merely serve as a connection between other offices are called tandem offices. A private branch exchange, or PBX, is a privately owned switch connected to the public network. The PBX is similar to the local exchange office, except that it is privately owned. A remote multiplexor is used to multiplex remote users via one transmission facility to a local office. It performs the same function as a PBX, but is part of the public network rather than being privately owned.

The cellular and cordless phone architectures use the PSTN as a backbone infrastructure, as shown in Figure 6.3. The cordless phone local base station communicates with a wireless headset via a duplex radio connection. More sophisticated headsets have several channels available for the radio connection, with channel selection based on the amount

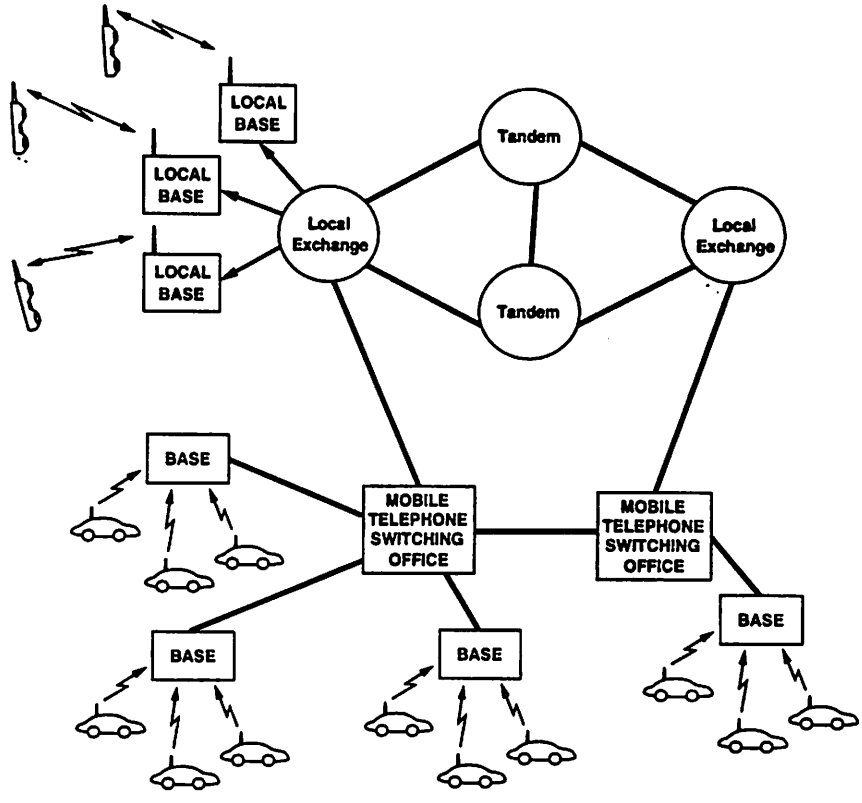

Figure 6.3: Cordless and Cellular Extension to the PSTN.

of interference and noise measured on the available channels. The local base emulates the standard telephone connection to the PSTN.

The main restriction of this architecture is the small coverage area of the local base station and the susceptibility of the wireless headset-base connection to interference and interception. The first of these restrictions is being corrected in second-generation designs by changing the cordless phones from stand-alone consumer items to elements of a geographically dispersed network. The wireless terminals will be able to access base stations at thousands of public locations which connect directly to the PSTN [102]. In business environments, cordless phones will have access to several base stations that hand off the user between them as it moves from one location to another.

Cellular systems have a similar structure with an intermediate mobile telephone switching office (MTSO) to control the base stations. Each base station services a subset of the geographical area covered by the MTSO. The base stations are essentially dumb terminals which transfer the wireless data from the mobiles via a (wireline) trunk to the

MTSO. The MTSO monitors signal strength of each mobile on all the base stations, determines the base station which should service calls to a particular mobile, and controls the handoff and channel allocation between base stations and mobiles. Calls between mobiles within the MTSO's service area are directly routed between the appropriate base stations by the MTSO. Calls intended for PSTN destinations are routed through the local exchange connected to the MTSO. To service subscribers in locations remote from their home service areas, proprietary communication links are established between MTSOs to exchange mobile location information and transfer calls.

There are several problems with this architecture, including limited capacity, centralized control, and poor mobility tracking and intersystem handoff. The capacity is limited by the number of subscribers that can be serviced with each base station; therefore, the total system capacity can be increased by shrinking the size of the cells, as was discussed in §2.3. However, this increases the processing burden on the MTSO in two ways: it must monitor more mobiles within a given geographical area, and it must hand off mobiles between base stations more often due to the smaller cell size. The trend for future cellular architectures is to distribute these control and monitoring functions among the base station, mobile, and MTSO. Moreover, roaming and intersystem handoff will be managed with standardized signaling systems linking MTSOs and databases. We will discuss this further in §6.2.3.

6.2.2 Packet-Switched Network Architecture

In packet-switched networks, the data stream is first decomposed into packets (smaller data strings whose length varies according to the network), and each packet is labeled with the address of its destination and a sequence number. Packet switches use the destination address to determine the next packet switch to which it should send the packet. There may be several valid routes to the final destination. Packets share the link and switch facilities with other packets routed through the network, so switches must generally queue packets until they can be forwarded, as shown in Figure 6.4. There are essentially two types of packet-switching: datagram packet-switching and virtual-circuit switching. With datagram packet-switching, packets from a given source are routed independent of each other, and since some routes may take longer than others, the packet sequence numbers must be used to order the packets sequentially at their final destination. With virtual-circuit routing, the packets follow the same route through the network, as in the case of

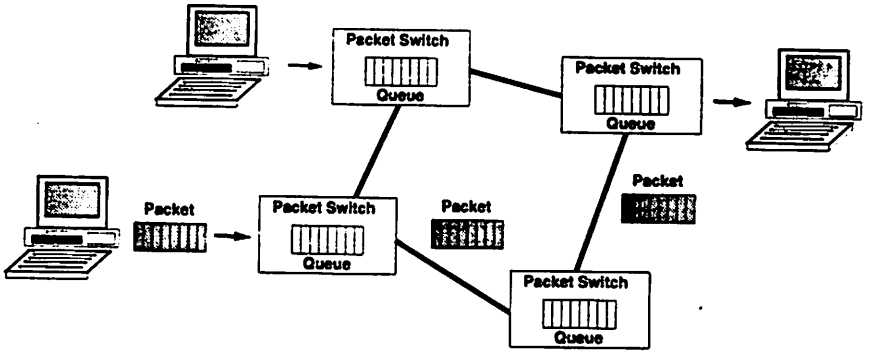

Figure 6.4: Packet-Switched Network.

circuit-switching, except that they don't have a dedicated path: some segment of the route may be shared with packets traveling a different path..

Packet-switched data networks tend to fall into three categories based on the geographical distance that they span: Local Area Networks (LANs), Metropolitan Area Networks (MANs), and Wide Area Networks (WANs). The wireline LANs typically have a diameter of a few kilometers, a total data rate of at least several Mbps, and are generally owned by a single organization. The most common LANs are ALOHA packet radio, Ethernet, token bus, token ring, FDDI, and DQDB. Details of these network designs can be found in [103]. By contrast, WANs typically span entire countries, have much lower data rates, and are generally owned by several organizations, including the ubiquitous PSTN, since most WANS use leased lines of the PSTN for their backbone communications infrastructure. WAN examples include IBM's SNA, DECnet, and Siemen's TRANSDATA, among others. A MAN is a network which generally spans an entire region, like a city or university campus, but uses essentially LAN technology or interconnected LANs. Details on these network architectures and protocols can be found in [103, 104]

Many computers today are linked to one of the networks described above. Therefore, computers connected to the same network can exchange information between them. In order to build a global communications network connecting all computers, it is necessary to *internetwork* the LANs, MANs, and WANs described above. ATM is emerging as the standard protocol for information transfer between heterogeneous data networks [103]. The most prevalent architecture for this network interconnection, which is likely to provide the backbone infrastructure for the information superhighway of both wireline and wireless networks, is the Internet. A segment of the basic Internet architecture is shown in Figure 6.5.

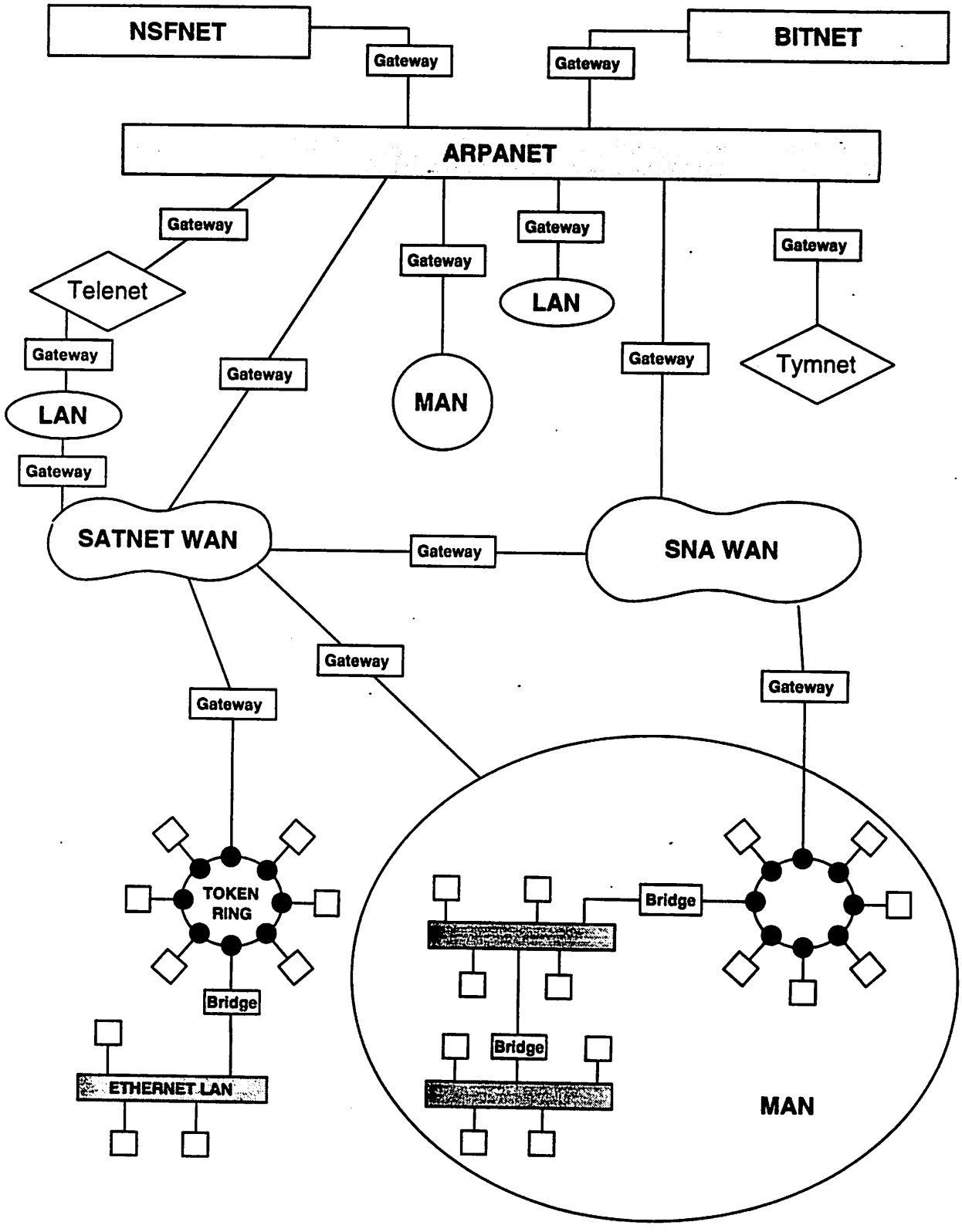

Figure 6.5: Internet Architecture.

Descriptions of the specific networks shown in this figure can be found in [103].

The Internet is primarily a hierarchical network centered around the backbone ARPANET. At the lowest level of the hierarchy are LANs, which are generally private networks to connect computers in one office building or university department. These LANs can be connected by bridges, which basically store and forward frames between different LANs [104]. A linked set of LANs may span a group of offices or a university campus. In order for a LAN or MAN to communicate with a host some distance away, it needs to transfer its data across a WAN. A gateway, which connects dissimilar networks, is used to connect LANs and WANs. The gateway is similar in function to a bridge, however it must generally make more changes in the data structures to make the networks that it transfers between compatible (for example, a gateway can convert between different addressing formats). Finally, the WANs connect via a gateway to the ARPANET, which has worldwide network nodes. Other global networks have recently been incorporated into the Internet to expand access, including BITNET and NSFNET, among others. The Internet therefore connects millions of globally dispersed computers through these sequences of network connections. However, the Internet is designed for transfer of data, and does not guarantee any minimum delay or maximum rate for data transfer. Therefore, it cannot accommodate a large number of users with delay-constrained data, like voice or video.

6.2.3 A Proposed Architecture for Hybrid Wireless Networks

In order to support voice, video, and data traffic, the wireless network infrastructure will require a combination of circuit and packet-switched architectures. It will also likely interconnect with the evolving PSTN/ISDN network and the Internet. The wireless subsystems described in §6.1 have analogies with wireline networks relative to their coverage areas: cordless phone systems and high-speed indoor data systems cover roughly the same area as wireline LANs, current cellular systems and second-generation cellular and cordless technologies will span distances of wireline MANs, and satellite systems cover the large geographical regions of WANs, in fact many of the WANs on the Internet already use satellite communication links. These analogies suggest that a hierarchical structure similar to that of the Internet will provide the most flexibility in the network architecture, as well as backward compatibility with existing wireless subsystems. A proposed hierarchical structure for the evolving network is shown in Figure 6.6.

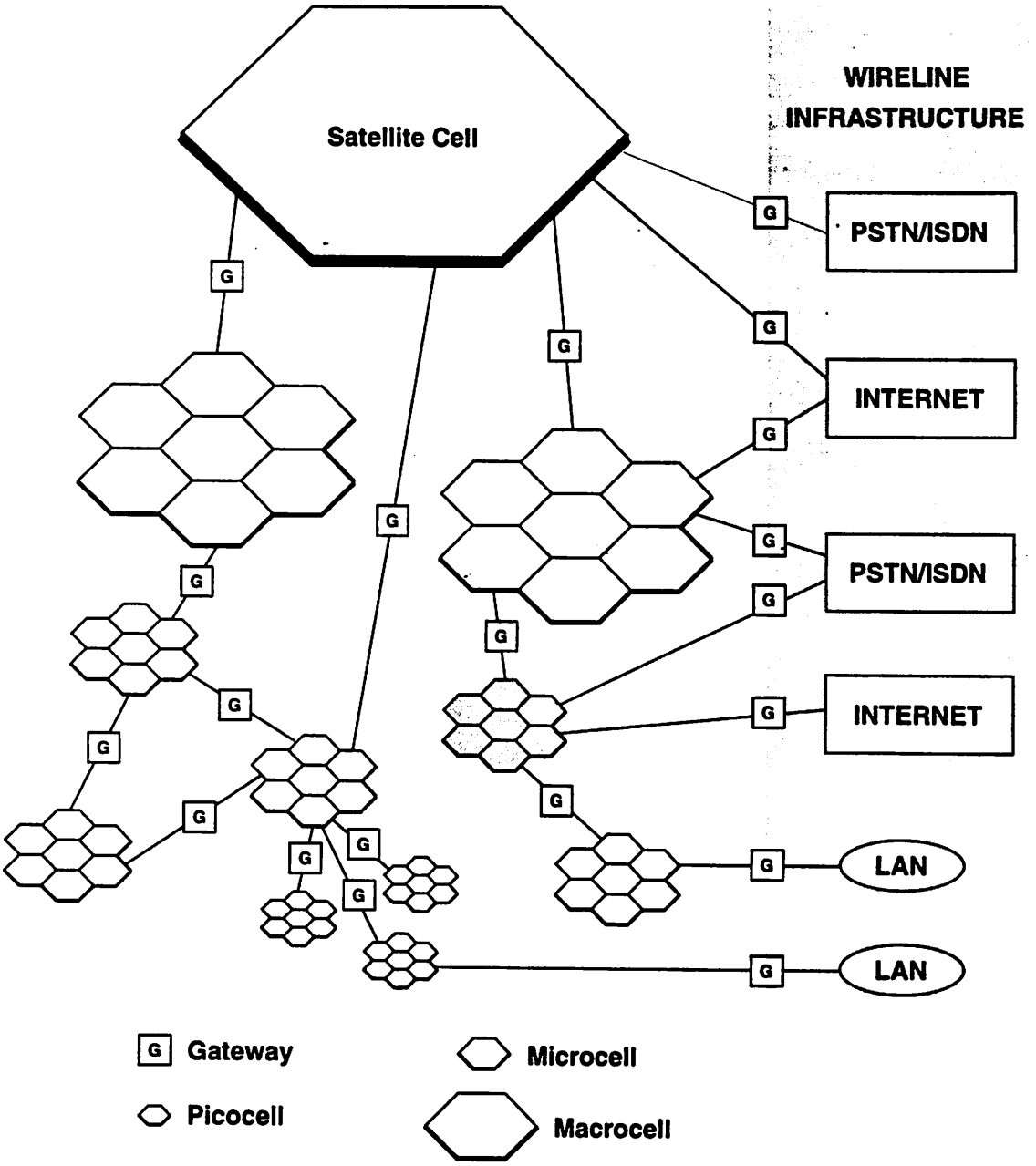

Figure 6.6: Wireless Network Architecture.

The cellular clusters in this figure represent self-contained wireless networks, for example an AMPS cellular system, or a second-generation cordless phone system. We assume that routing, handover, and control functions within these cellular clusters are self-contained, although some of these functions could be better optimized by relying on higher levels within the hierarchy, particularly for handover [105]. The picocells of this figure represent applications spanning approximately five to ten meters in diameter, and therefore provide coverage within a small office, cubicle, or classroom. Microcells cover applications over kilometer distances, and provide coverage over a city block or a floor within a large office. Macrocells span roughly ten kilometers, and cover large areas within cities or suburban areas. Finally, the satellite cell generally spans very large distances, although directional spot beams can be used to provide coverage within the smaller cell sizes of macrocells and microcells. Thus, satellite beams can be used to relieve congestion within cellular systems that are below it in the network hierarchy.

The gateways of Figure 6.6 will perform the same function as gateways on the Internet: converting protocols between networks to make them compatible. The wireless-to-wireless gateways require multimode transceiver hardware, for example satellite and cellular phone capabilities. To reduce transceiver size, it is desirable to use many of the same hardware components for the different transceiver functions. However, this design goal has proved difficult to accomplish in current dual-mode analog AMPS and digital IS-95 cellular transceivers, whose modes of operations have much more in common than the modes of a satellite/cellular phone transceiver. In any event, building light-weight handheld multimode transceivers will become easier through technological advances in component size and power reduction.

Communication between cellular systems within the hierarchy can be strictly through the wireless infrastructure, strictly through the backbone wireline network, or through some combination of the two. It is likely that the reliance on the wireline infrastructure will persist for quite some time, for several reasons. First of all, current cellular and cordless phone systems use the PSTN for routing, and wireless LANs use the Internet. New generations of these systems will require backward-compatibility with their older predecessors. Moreover, the cost of building a completely wireless infrastructure may not be justified by current or future demand. Therefore, the existing wireline infrastructure allows for the introduction of new wireless services at an increment cost. Its also not clear that a wireless infrastructure could ever fully support the demand for wireless services without

a wireline backbone, given the capacity limitations of the wireless channel. Finally, the exchange of backbone control and routing information between base stations within and between cellular subsystems may be more effective using wireline network technology.

In the wireless communication services described in 6.1, the coverage area and data rates for each system are inversely proportional. This is due mainly to the fact that high data rates require high power, and power dissipates with distance. Moreover, the cost of implementing a system with a large coverage area (e.g., launching a satellite or laying fiber throughout a city), is generally much greater than the costs associated with a small coverage system. Therefore, the bandwidth of the large coverage systems must be divided among many users to recoup this cost. This inverse data rate-coverage area relationship implies that, moving up the hierarchy of Figure 6.6, systems provide greater mobility but less bandwidth. Hence, within the global communication infrastructure, there will be low-mobility high-bandwidth communication islands connected by high-mobility low-bandwidth bridges. In this context, the terms "high" and "low" are relative to the level within the hierarchy of Figure 6.6.

The frequency spectrum available for wireless services is scarce, and thus many of the systems within the wireless network hierarchy will operate in the same frequency band. Spectrum-sharing techniques were discussed in the previous chapter in the context of a single system. However, it's not clear if these multiuser spectrum-sharing results apply to users with different coverage areas, power levels, and propagation characteristics.

6.3 Mobility Management and Routing

In the PSTN and the Internet, terminals (or ports) are assigned identification numbers associated with their physical location. However, within the wireless infrastructure of Figure 6.6, the network must be able to locate and transmit data to an end user based only on a personal identification number (PIN) which is independent of the user's location or communication device. The PIN is required to deliver true mobility to the user, since it separates the user's logical address from the physical address of the port used to access the network. It also allows different applications to send to data a user's address, rather than a device address, eliminating the need for one user to have a separate address for each wireless device (e.g., computer, cellular phone, fax machine). The network must manage the association between the user's PIN and current physical address in order to route traffic,

regardless of the user's location, type of wireless device, or layer of the network hierarchy through which it is connected. Of course, the type of data that can be transferred is ultimately restricted by the wireless device and network capabilities.

The process of locating a user and routing a call are somewhat separate, since once a user's location is known, there will generally be many possible ways of routing calls to it. There are several methods for locating mobile users. One technique, currently used in cellular systems, is for PINs to be assigned to a unique home location gateway. This is the gateway through which the user generally connects to the wireless or wireline network. If a user is not connected through this home gateway, the address of the gateway through which the user is connected, (its *visiting gateway*), is sent to a home location database (HLD) within the home gateway's cellular cluster. For a call initiator to determine the location of a particular user, it need only query the home gateway. These roaming mobiles that are away from their home gateways will also be registered in the visitor location database (VLD) of their visiting gateways. A particular cellular system may have multiple home gateways, since it may connect to a higher level in the wireless network, the PSTN/ISDN network, and the Internet, as shown in Figure 6.6 for the shaded cellular cluster. All three of these home gateways can access the HLD and forward information about a user's location. The main disadvantage of this technique is the amount of control traffic necessary to keep updating the location databases for highly mobile users. Moreover, if the network latency is high relative to user mobility, the location information may be outdated by the time it is received by the query intiator.

The location information may also be stored at databases located higher up in the wireless network hierarchy than the home location gateway. For example, the location database for users within a picocell could be stored in a microcell or macrocell cluster above the picocell in the network hierarchy. This reduces the amount of traffic associated with location queries and updates, since these messages would not have to traverse as many levels in the network hierarchy to reach the location databases. One disadvantage of this method is the increased size and complexity of the location databases. It would also be more difficult for wireline networks which connect directly to a user's home gateway to get that user's location, since the information is stored higher up in the wireless hierarchy. A hybrid solution would be to have location databases distributed at several levels throughout the hierarchy. The most efficient means of distributing location information in these databases would depend on the particulars of the wireless and wireline network interconnections.

Routing through the network will be of two types depending on the data constraints: circuit-oriented routing which guarantees sequential data arrival within a delay constraint for real-time data, and packet-oriented routing which has no constraint on the time-arrival of the data packets. For circuit-switching, the network would need to locate the mobile destination, and dedicate wireless or wireline links through the hierarchy between the sender and destination. Packet-oriented routing could be done using techniques based on the Internet routing protocol with some slight modifications for packet-forwarding as mobiles relocate [106].

Since the wireline network already has established routing procedures, and will eventually convert to fiber which has a much higher bandwidth and reliability than wireless technology, it would appear that the most efficient routing schemes would connect into the wireline network as soon as possible. However, it may be prudent to route circuit-oriented data through one level higher up in the wireless network hierarchy than necessary. With this technique, when a mobile passes between different systems at a particular hierarchy level, call handover between these two systems can be managed by the higher-level network, reducing the chance of call interruption [105]. Moreover, it's not clear that going through the wireline network is the most efficient routing scheme between two wireless systems, especially if the systems have dual mode gateways which would allow them to talk to each other directly. Developing and analyzing routing protocols for the emerging wireless network infrastructure is an important area of research that has received little attention to date, despite the fact that it will ultimately determine the performance of ubiquitous communication between mobiles.

6.4 Other Issues

Network Security - Wireless data transmission raises questions about network security and privacy, since anyone with a monopole antenna and simple radio can intercept conversations, or attempt to access the network. This has been a major problem for analog cellular systems. Conversion to digital technology on second-generation cordless and cellular systems will allow encryption and authentication more readily than on their analog predecessors. However, it's not clear whether these techniques will be applicable to a general wireless infrastructure of nonhomogeneous networks, with network control distributed throughout the system.

Network Pricing - Pricing for data transfer across a range of nonhomogeneous networks will affect both user demand, and possibly routing strategies. The cost of routing data will depend on the type of service guarantees required, which may determine whether a call is routed through the wireless or wireline infrastructure. The cost of current cellular technology has not dropped as was once anticipated, and if the price of services within the wireless network are not competitive with wireline services then demand may not be sufficient to support an interconnected wireless infrastructure. Pricing for different services traversing wireline nonhomogeneous networks is an area of current research, and it would seem that the addition of wireless services will only make the problem more difficult.

Local and Global Control - Network control functions include fault detection and correction, performance monitoring, network topology monitoring, traffic monitoring and billing, and security. These functions are generally assumed at both the local and global levels within the network. Many of the control requirements for the wireless network infrastructure are similar to those in wireline networks, and current proposals for wireline network management will be applicable to wireless networks also. However, the changing topology of wireless networks will require many of these control functions to be performed more frequently, and may change the level within the hierarchy where certain functions are best performed.

6.5 Summary

After outlining some of the wireless applications currently in demand, we discuss the implementation of a wireless network supporting these applications. We first propose an architecture to interconnect various wireless and wireline subnetworks with different coverage areas and requirements. We then discuss a few schemes for locating mobile units, regardless of their physical location in the network. The routing of different types of data is also discussed. We conclude by addressing network security, pricing, and control issues. The topics in this chapter are still very much in the research stage, and the discussion throughout is not meant to provide definitive proposals for the design of the global wireless network, but rather to outline the various design issues that must be addressed to ultimately connect all the wireless subsystems currently under development.

Chapter 7

Conclusions and Future Work

The wireless communication vision of high-speed high-quality information exchange between portable devices located anywhere in the world faces many technical hurdles. In this thesis we mainly focussed on techniques to improve the quality and achievable data rates of single and multiple users over time-varying communication links. In particular, we derived the capacity of a single-user time-varying channel with channel state information available at the transmitter, and proposed a variable-rate coded modulation scheme that achieved data rates approaching this capacity limit. We also developed a reduced-complexity maximum-likelihood sequence detector for the case when only the distribution of the channel variation is known. Multiuser rate regions for narrowband time-varying channels under different spectrum-sharing methods were also evaluated. Finally, an infrastructure to support nonhomogeneous wireless applications was proposed.

7.1 Conclusions

Two important conclusions can be drawn from Chapter 2: the wireless communications link has many impairments, and these impairments vary greatly depending on the characteristics and topology of the region over which the signal propagates. In particular, changing the distance the signal propagates or the height of the transmitting or receiving antennas fundamentally changes the *model* for signal propagation. Therefore, analysis of a system designed for a large coverage area will generally not apply to the system's performance over a small coverage area.

From Chapter 3 we conclude that optimizing the transmit signal spectrum to the

channel variation maximizes the achievable data rate on time-varying channels. In particular, Shannon's theorem of maximizing efficiency on fixed wideband channels through a "water-filling" in frequency of the transmit power spectrum extends to a two-dimensional water-filling in both frequency and time. Applying these results to narrowband channels, we conclude that the policy which maximizes the average data rate transmits more power and data when the channel is good and less power and data when the channel is bad. This may seem intuitive, but it is the exact opposite of power control policies being implemented in current cellular systems. However, this optimal policy does not take into account interference to other users or guaranteed data rate and delay requirements. Since the data rate fluctuates with the channel variation, this may not be acceptable for applications with delay-constrained data, like voice or video. The capacity analysis also leads to the design of a variable-rate coded modulation scheme which achieves near-capacity rates.

Time-varying Markov channels, where the channel variation is not known but its statistics are, were considered in Chapter 4. We found in this chapter that, even though these channels have infinite memory, the channel variation statistics can be used for maximum-likelihood sequence estimation without a significant increase in complexity or delay over the conventional method of interleaving and memoryless channel encoding. Moreover, this maximum-likelihood detection scheme achieves channel capacity for a particular channel class. Finally, this scheme shows a significant capacity increase over the conventional technique, and the increase is most pronounced on slowly-varying channels.

In Chapter 5 we looked at spectrum-sharing techniques for multiuser systems. We found that TDMA and FDMA are equivalent if the transmit power can be varied, and we also found that CDMA with interference cancellation is superior to FDMA/TDMA, and inferior without the cancellation. However, these conclusions apply to spectrum-sharing within a single cell of a cellular system, and don't take into account intercell interference. Including intercell interference in the achievable data rate calculation requires a new definition - the area efficiency. We define this quantity and use it to obtain the optimal reuse distance for a simple interference model under an FDMA spectrum-sharing scheme. Determining the spectrum-sharing technique which maximizes area efficiency requires more analysis and/or simulation to obtain the distribution of signal power under various power control policies. We conclude this chapter by proposing a hybrid power control policy with the benefits of both the constant power policy (guaranteed data rates) and the water-pouring policy (increased data rates under good propagation conditions). This scheme also accom-

modates different user specifications and channel access requests. The main conclusion to draw from this chapter is that there is probably no "best" method of spectrum-sharing and power control. Therefore, nontraditional and hybrid methods should be considered along with the more traditional approaches.

The main conclusion to draw from Chapter 6 is that internetworking the heterogeneous wireless subnetworks will be quite challenging, and the protocols and network infrastructure for this internetworking should be addressed at a global level in the near future. Second- and third-generation cellular and cordless phone systems are already being built to adhere to a particular networking structure and set of protocols, which are based on emerging PSTN/ISDN technology. Emerging wireless computing devices will likely adhere to the Internet or ATM standards. Therefore, although the divide between communications and computers will continue to blur as their respective devices become multimedia wireless terminals, the networking protocols for these devices are likely to differ significantly, given the disparate networking philosophy between communication and computer engineers today. For this reason, global standards for interconnection of all these devices should precede their development in order to make them compatible. Issues of routing, mobility management, network security, pricing, and control of the wireless network may borrow from the standards of wireline networks, but must also take into account the unique character of terminal mobility.

7.2 Future Work

Much work remains to be done in the design and analysis of high-speed wireless communication networks. Extensions to this thesis fall into four main categories: communication link techniques, channel estimation and feedback, power control and spectrum sharing, and wireless networks.

7.2.1 Communication Link Techniques

The variable-rate modulation and coded-modulation techniques of §§3.4 – 3.5 should be verified via simulations to determine their feasibility. The effects of channel estimation error and delay should also be quantified. Different trellis and lattice structures for the variable-rate coded-modulation technique should be considered, and their relative performance determined both analytically and via simulation.

Simulation of the decision-feedback decoder proposed in §4.3 would verify its performance under different coding schemes. We must also examine the techniques outlined in §4.3 to eliminate the effects of decoding delay in the decoder design. Analytic results or bounds on the effect of error propagation would also be useful.

We found that when channel state information is available at the transmitter then variable-rate coding achieves good performance, and when this information is not available, then unequal error protection codes are effective. Perhaps when the channel state is known with some uncertainty, some combination of these techniques could be used, resulting in variable-rate codes with unequal error protection. This type of coding merits further investigation.

7.2.2 Channel Estimation and Feedback

There is a dichotomy in communication over time-varying channels relative to how much time should be spent estimating the channel, and how much time should be spent transmitting data. Channel estimates can be used at both the receiver and the transmitter (if there is feedback) to increase data rates or decrease BER. Intuitively, the better the channel estimate, the more it can improve performance. However, a good channel estimate requires a long estimation sequence, which reduces the data rate. Therefore, there should be some *optimal* estimation time which maximizes data rate for a given BER and set of channel parameters. In §3.6.3 we determined the reduction in channel capacity as a function of estimation time. If we could determine the combined effects of estimation error and estimation time on channel capacity, then we could obtain the optimal estimation time relative to channel capacity. A related topic is when to use feedback in time-varying communication links. If the channel is changing very rapidly, then by the time the channel is estimated and fed back to the transmitter, the estimate may no longer be valid. In addition, the feedback communication link is neither error-free nor delayless, as we assumed in our analysis. Therefore, a valuable topic for further investigation is to determine, under more realistic system constraints, when full or partial transmitter feedback improves system performance.

7.2.3 Power Control and Spectrum Sharing

Determining the "best"method of spectrum sharing and power control will depend on the performance criteria of the individual users and the system. As described in §5.4.1, the area efficiency under different spectrum-sharing techniques and power control policies quantifies the most efficient technique relative to system throughput, and calculating this quantity under different power control and spectrum-sharing policies would be a valuable addition to the debate on CDMA/TDMA/FDMA spectrum sharing. The hybrid power control policy proposed in §5.4.4 should be evaluated both analytically for simple cases and via simulation. Other hybrid power control and spectrum-sharing schemes may prove superior to anything proposed thus far. Therefore, it is important to move beyond the CDMA/FDMA/TDMA debate and look at other solutions relative to the specific wireless application.

7.2.4 Wireless Networks

Design and analysis of a wireless infrastructure to support existing and pending wireless subnetworks and connect them to the wireline infrastructure is critical for ultimately achieving the wireless communications vision. Once this infrastructure has been defined, research on routing, mobility management, security, control, and service pricing will be needed. The protocols for these functions will borrow heavily from those of existing wireline technology. However, terminal mobility introduces the need for adaptability far greater than in fixed wireline structures. Thus, it is not clear if modification of existing protocols will suffice, or a completely new outlook is necessary. In addition, the wireless radio link introduces increased flexibility in interconnection, since all wireless networks can communicate directly with each other if they are within transmission range and have the appropriate hardware. This flexibility should be incorporated into the protocol suite developed for the wireless infrastructure.

Bibliography

[1] W.C. Jakes, Jr., *Microwave Mobile Communications.* New York: Wiley, 1974.

[2] J.W. McKown and R.L. Hamilton, Jr., "Ray tracing as a design tool for radio networks," *IEEE Network* , Vol. 5, No. 6, pp. 27–30, Nov. 1991.

[3] N. Amitay, "Modeling and computer simulation of wave propagation in lineal line-of-sight microcells," *IEEE Trans. Vehic. Technol.*, Vol VT-41, No. 4, pp. 337–342, Nov. 1992.

[4] K. Schaubach, N.J. Davis IV, and T.S. Rappaport, "A ray tracing method for predicting path loss and delay spread in microcellular environments," *Vehic. Technol. Conf. Rec.*, pp. 932–935, May 1992.

[5] D.C. Cox. "910 MHz urban mobile radio propagation: Multipath characteristics in New York City," *IEEE Trans. Commun.*, Vol. COM-21, No. 11, pp. 1188–1194, Nov. 1973.

[6] G.L. Turin. "Introduction to spread spectrum antimultipath techniques and their application to urban digital radio," *IEEE Proceedings*, Vol. 68, No. 3, pp. 328–353, March 1980.

[7] R.S. Kennedy. *Fading Dispersive Communication Channels.* New York: Wiley, 1969.

[8] W.D. Rummler, "More on the multipath fading channel model," *IEEE Trans. Commun.*, Vol. COM-29, No. 3, pp. 346–352, March 1981.

[9] W.C. Wong and L.J. Greenstein, "Multipath fading models and adaptive equalizers in microwave digital radio," *IEEE Trans. Commun.*, Vol. COM-32, No. 8, pp. 928–934, Aug. 1984.

[10] A.J. Rustako, Jr., N. Amitay, G.J. Owens, and R.S. Roman, "Radio propagation at microwave frequencies for line-of-sight microcellular mobile and personal communications," *IEEE Trans. Vehic. Technol.*, Vol VT-40, No. 1, pp. 203–210, Feb. 1991.

[11] W.C.Y. Lee, *Mobile Communications Engineering.* New York: McGraw-Hill, 1982.

[12] J.-E. Berg, R. Bownds, and F. Lotse, "Path loss and fading models for microcells at 900 MHz," *Vehic. Technol. Conf. Rec.*, pp. 666–671, May 1992.

[13] P. Harley, "Short distance attenuation measurements at 900 MHz and 1.8 GHz using low antenna heights for microcells," *IEEE J. Selected Areas Commun.*, Vol. SAC-7, No. 1, pp. 5–11, Jan. 1989.

[14] J.-F. Wagen, "Signal strength measurements at 881 MHz for urban microcells in downtown Tampa," *Globecom Conf. Rec.*, pp. 1313–1317, Dec. 1991.

[15] R.J.C. Bultitude and G.K. Bedal, "Propagation characteristics on microcellular urban mobile radio channels at 910 MHz," *IEEE J. Selected Areas Commun.*, Vol. SAC-7, No. 1, pp. 31–39, Jan. 1989.

[16] J.H. Whitteker, "Measurements of path loss at 910 MHz for proposed microcell urban mobile systems," *IEEE Trans. Vehic. Technol.*, Vol VT-37, No. 6, pp. 125–129, Aug. 1988.

[17] H. Börjeson, C. Bergljung, and L.G. Olsson, "Outdoor microcell measurements at 1700 MHz.," *Vehic. Technol. Conf. Rec.*, pp. 927–931, May 1992.

[18] A.J. Goldsmith and L.J. Greenstein, "A measurement-based model for predicting coverage areas of urban microcells," *IEEE J. Selected Areas Commun.*, Vol. SAC-11, No. 7, pp. 1013–1023, Sept. 1993.

[19] F. Ikegami, S. Takeuchi, and S. Yoshida, "Theoretical prediction of mean field strength for urban mobile radio," *IEEE Trans. Antennas Propagat.*, Vol. AP-39, No. 3, pp. 299–302, March 1991.

[20] M.C. Lawton and J.P. McGeehan, "The application of GTD and ray launching techniques to channel modeling for cordless radio systems," *Vehic. Technol. Conf. Rec.*, pp. 125–130, May 1992.

[21] R.J. Luebbers, "Finite conductivity uniform GTD versus knife edge diffraction in prediction of propagation path loss," *IEEE Trans. Antennas Propagat.*, Vol. AP-32, No. 1, pp. 70–76, Jan. 1984.

[22] C. Bergljung and L.G. Olsson, "Rigorous diffraction theory applied to street microcell propagation," *Globecom Conf. Rec.*, pp. 1292–1296, Dec. 1991.

[23] G.K. Chan, "Propagation and coverage prediction for cellular radio systems," *IEEE Trans. Vehic. Technol.*, Vol VT-40, No. 4, pp. 665–670, Nov. 1991.

[24] K.C. Chamberlin and R.J. Luebbers, "An evaluation of Longley-Rice and GTD propagation models," *IEEE Trans. Antennas Propagat.*, vol AP-30, No. 11, pp. 1093–1098, Nov. 1982.

[25] M.I. Skolnik, *Introduction to Radar Systems.* 2nd Ed. New York: McGraw-Hill, 1980.

[26] S.Y. Seidel, T.S. Rappaport, S. Jain, M.L. Lord, and R. Singh, "Path loss, scattering, and multipath delay statistics in four European cities for digital cellular and microcellular radiotelephone," *IEEE Trans. Vehic. Technol.*, Vol VT-40, No. 4, pp. 721–730, Nov. 1991.

[27] S.T.S. Chia, "1700 MHz urban microcells and their coverage into buildings," *IEE Antennas Propagat. Conf. Rec.*, pp. 504–511, York, U.K., April 1991.

[28] M. Schwartz, W. Bennett, and S. Stein, *Communication Systems and Techniques.* New York: McGraw-Hill, 1966.

[29] S. Stein, "Fading channel issues in system engineering," *IEEE J. Selected Areas Commun.*, Vol. SAC-5, No. 2, pp. 68–89, Feb. 1987.

[30] S.O. Rice, "Mathematical analysis of random noise," *Bell System Tech. J.*, Vol. 23, No. 7, pp. 282–333, July 1944, and Vol. 24, No. 1, pp. 46–156, Jan. 1945.

[31] J.G. Proakis, *Digital Communications.* New York: McGraw-Hill, 1983.

[32] R.C. Dixon, *Spread Spectrum Systems.* 2nd Ed. New York: Wiley, 1984.

[33] H. Hashemi, "Simulation of the urban radio propagation channel," *IEEE Trans. Vehic. Technol.*, Vol VT-28, No. 3, pp. 213–225, Aug. 1979.

[34] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. Vehic. Technol.*, Vol VT-29, No. 3, pp. 317–325, Aug. 1980.

[35] J.M. Kahn, *Private communication.*

[36] M. Gundmundson, "Correlation model for shadow fading in mobile radio systems," *Electr. Lttrs.*, Vol. 27, pp. 2145–2146, Nov. 7, 1979.

[37] W.C.Y. Lee, "Overview of cellular CDMA," *IEEE Trans. Vehic. Technol.*, Vol. VT-40, No. 2, pp. 291–302, May 1991.

[38] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, Jr., and C. E. Wheatley III, "On the capacity of a cellular CDMA system," *IEEE Trans. Vehic. Technol.*, Vol. VT-40, No. 2, pp. 303–312, May 1991.

[39] J. Wolfowitz, *Coding Theorems of Information Theory.* 2nd Ed. New York: Springer-Verlag, 1964.

[40] R.G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[41] C. E. Shannon and W. Weaver, *A Mathematical Theory of Communication.* Urbana, IL: Univ. Illinois Press, 1949.

[42] T. Cover and J. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[43] G. J. Foschini and J. Salz, "Digital communications over fading radio channels," *Bell System Tech. J.*, Vol. 62, No. 2, pp. 429–456, Feb. 1983.

[44] G.D. Forney, Jr., R.G. Gallager, G.R. Lang, F.M. Longstaff, and S.U. Quereshi, "Efficient modulation for band-limited channels," *IEEE J. Selected Areas Commun.*, Vol. SAC-2, No. 5, pp. 632–647, Sept. 1984.

[45] S. Lin and D.J. Costello, Jr., *Error Control Coding: Fundamentals and Applications.* Englewood Cliffs, NJ: Prentice Hall, 1983.

[46] G. Ungerboeck. "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, Vol. IT-28, No. 1, pp. 55–67, Jan. 1982.

[47] G. Ungerboeck. Trellis-coded modulation with redundant signal sets, Part I: Introduction and Part II: State of the art. *IEEE Commun. Mag.*, Vol. 25, No. 2, pp. 5–21, Feb. 1987.

[48] G.D. Forney, Jr., "Coset codes - Part I: Introduction and geometrical classification," *IEEE Trans. Inform. Theory*, Vol. IT-34, No. 5, pp. 1123–1151, Sept. 1988.

[49] G.D. Forney, Jr., and L.-F. Wei, "Multidimensional constellations - Part I: Introduction, figures of merit, and generalized cross constellations," *IEEE J. Selected Areas Commun.*, Vol. SAC-7, No. 6, pp. 877–892, Aug. 1989.

[50] L.-F. Wei, "Coded M-DPSK with built-in time diversity for fading channels," *IEEE Trans. Inform. Theory*, Vol. IT-39, No. 6, pp. 1820–1839, Nov. 1993.

[51] D. Divsalar and M.K. Simon, "The design of trellis coded MPSK for fading channels: Set partitioning for optimum code design," *IEEE Trans. Commun.*, Vol. COM-36, No. 9, pp. 1013–1021, Sept. 1988.

[52] A.J. Goldsmith, L.J. Greenstein, and G.J. Foschini, "Error statistics of real-time power measurements in cellular channels with multipath and shadowing," *IEEE Vehic. Technol. Conf. Rec.*, pp. 108–111, May 1993. Also to appear in the *IEEE Trans. Vehic. Technol.*.

[53] R.A. Ziegler and J.M. Cioffi, "Estimation of time-varying digital radio channels," *IEEE Trans. Vehic. Technol.*, Vol. VT-41, No. 2, pp. 134–151, May 1992.

[54] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliot channel," *IEEE Trans. Inform. Theory*, Vol. IT-35, No. 6, pp. 1277–1290, Nov. 1989.

[55] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels.* New York: Academic Press, 1981.

[56] T.Kaijser, "A limit theorem for partially observed Markov chains,". *Ann. Probab.*, Vol. 3, pp. 677–696, 1975.

[57] P. Billingsley. *Probability and Measure.* 2nd Ed. New York: Wiley, 1986.

[58] A.J. Viterbi and J.K. Omura, *Principles of Digital Communications and Coding.* New York: McGraw-Hill, 1979.

[59] E.A. Lee and D.G. Messerschmitt, *Digital Communications.* 2nd Ed. Boston: Kluwer, 1994.

[60] K.J. Raghunath and K.K. Parhi, "Parallel adaptive decision feedback equalizers," *IEEE Trans. Signal Proc.*, Vol. SIG-41, No. 5, pp. 1956–1961, May 1993.

[61] K. Zhou, J.G. Proakis, and F. Ling, "Decision-feedback equalization of time-dispersive channels with coded-modulation," *IEEE Trans. Commun.*, Vol. COM-38, No. 1, pp. 18–24, Jan. 1990.

[62] S.A. Altekar, and N.C. Beaulieu, "Upper bounds to the error probability of decision feedback equalization," *IEEE Trans. Inform. Theory*, Vol IT-39, No. 1, pp. 145–156, Jan. 1993.

[63] D. Blackwell, L. Breiman, and A.J. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Stat. 31*, pp. 558–567, 1960.

[64] R. Ahlswede. Elimination of correlation in random codes for arbitrarily varying channels. *Zeit. Wahrscheinlichkeitstheorie verw. Gebiete 44*, pp. 159–175, 1978.

[65] R.V. Cox, J. Hagenauer, N. Seshadri, and C.-E. W. Sundberg, "Variable rate sub-band speech coding and matched convolutional channel coding for mobile radio channels,". *IEEE Trans. Signal Proc.*, Vol. SP-39, No. 8, pp. 1717–1731, Aug. 1991.

[66] C.-E.W. Sundberg and N. Seshadri, "Multi-level block coded modulations with unequal error protection for the Rayleigh fading channel," *Europ. Trans. Telecomm. and Related Technol.* Vol. 4, No. 3, pp. 325–334, May-June 1993.

[67] H.S. Wang and N. Moayeri, "Modeling, capacity, and joint source/channel coding for Rayleigh fading channels," Technical Report WINLAB-TR-32, Wireless Information Network Laboratory, Rutgers University, May 1992.

[68] P.P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inform. Theory*, Vol IT-19, No. 2, pp. 197–207, March 1973.

[69] P.P. Bergmans and T.M. Cover, "Cooperative broadcasting," *IEEE Trans. Inform. Theory*, Vol IT-20, No. 3, pp. 317–324, May 1974.

[70] H. Imai and S. Hirakawa, "A new multilevel coding method using error correcting codes," *IEEE Trans. Inform. Theory*, Vol IT-23, No. 3, pp. 371–377, May 1977.

[71] J.W. Modestino and D.G. Daut, "Combined source-channel coding of images," *IEEE Trans. Commun.*, Vol. COM-27, No. 11, pp. 1644–1659, Nov. 1979.

[72] A.R. Calderbank and N. Seshadri, "Multilevel codes for unequal error protection," *IEEE Trans. Inform. Theory*, Vol IT-39, No. 4, pp. 1234–1248, July 1993.

[73] L.-F. Wei, "Coded modulation with unequal error protection," *IEEE Trans. Commun.*, Vol. COM-41, No. 10, pp. 1439–1449, Oct. 1993.

[74] N. Seshadri and C.-E.W. Sundberg, "Multilevel trellis coded modulations for the Rayleigh fading channel," *IEEE Trans. Commun.*, Vol. COM-41, No. 9, pp. 1300–1310, Sept. 1993.

[75] C.-E.W. Sundberg and N. Seshadri, "Coded modulations for fading channels: An overview," *Europ. Trans. Telecomm. and Related Technol.* Vol. 4, No. 3, pp. 309–323, May-June 1993.

[76] A.J. Viterbi, "The orthogonal-random waveform dichotomy for digital mobile personal communications," *IEEE Personal Commun. Mag.*, Vol PC-1, No. 1, pp. 18–24, First Quarter 1994.

[77] R.S. Cheng and S. Verdú, "Gaussian multiaccess channels with ISI: Capacity region and multiuser water-filling," *IEEE Trans. Inform. Theory*, Vol IT-39, No. 3, pp. 773–785, May 1993.

[78] P.P. Bergmans, "Degraded broadcast channels," Ph.D. dissertation, Dept. Elec. Engin., Stanford University, Stanford, Calif., 1972.

[79] P.P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. Inform. Theory*, Vol IT-20, No. 2, pp. 279–280, March 1974.

[80] S. Verdú, "The capacity region of the symbol-asynchronous Gaussian multiple-access channel," *IEEE Trans. Inform. Theory*, Vol IT-35, No. 4, pp. 733–751, July 1989.

[81] W. Hirt and J.L. Massey, "Capacity of the discrete-time Gaussian channel with intersymbol interference," *IEEE Trans. Inform. Theory*, Vol IT-34, No. 3, pp. 380–388, May 1988.

[82] S. Verdú, "Multiple-access channels with memory with and without frame synchronism," *IEEE Trans. Inform. Theory*, Vol IT-35, No. 3, pp. 605–619, May 1989.

[83] J.C.-I. Chuang, "Performance issues and algorithms for dynamic channel assignment," *IEEE J. Selected Areas Commun.*, Vol. SAC-11, No. 6, pp. 955–963, Aug. 1993.

[84] S. Ihara, "On the capacity of channels with additive non-Gaussian noise," *Information and Control*, Vol. 37, pp. 34-39, 1978.

[85] P. Jung, P.W. Baier, and A. Steil, "Advantages of CDMA and spread spectrum techniques over FDMA and TDMA in cellular mobile radio applications," *IEEE Trans. Vehic. Technol.*, Vol. VT-42, No. 3, pp. 357–364, Aug. 1993.

[86] J.-P. Linnartz, *Narrowband Land-Mobile Radio Networks.* Norwood, MA: Artech House, 1993.

[87] T.S. Rappaport and L.B. Milstein, "Effects of radio propagation path loss on DS-CDMA cellular frequency reuse efficiency for the reverse channel," *IEEE Trans. Vehic. Technol.*, Vol. VT-41, No. 3, pp. 231–242, Aug. 1992.

[88] B. Gundmundson, J. Sköld, and J.K. Ugland, "A comparison of CDMA and TDMA systems," *IEEE Vehic. Technol. Conf. Rec.*, pp. 732–735, May 1992.

[89] P.T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *Bell System Tech. J.*, Vol 47, pp. 73–91, Jan. 1968.

[90] S. Verdú, "Minimum probability of error for asynchronous Gaussian multiple-access channels," *IEEE Trans. Inform. Theory*, Vol IT-32, No. 1, pp. 85–96, Jan. 1986.

[91] R. Lupas and S. Verdú, "Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Trans. Inform. Theory*, Vol. IT-35, No. 1, pp. 123–136, Jan. 1989.

[92] R. Lupas and S. Verdú, "Near-far resistance of multiuser detectors in asynchronous channels," *IEEE Trans. Commun.*, Vol. COM-38, No. 4, pp. 496–508, April 1990.

[93] M.K. Varanasi and B. Aazhang, "Multistage detection in asynchronous code-division multiple-access communications," *IEEE Trans. Commun.*, Vol. COM-38, No. 4, pp. 509–519, April 1990.

[94] M.K. Varanasi and B. Aazhang, "Near-optimum detection in synchronous code-division multiple-access systems," *IEEE Trans. Commun.*, Vol. COM-39, No. 5, pp. 725–736, May 1991.

[95] A. Duel-Hallen, "Decorrelating decision-feedback multiuser detector for synchronous code-division multiple-access channel," *IEEE Trans. Commun.*, Vol. COM-41, No. 2, pp. 285–290, Feb. 1993.

[96] S. Vasudevan and M.K. Varanasi, "Optimum diversity combiner based multiuser detection for time-dispersive Rician fading CDMA channels," *IEEE J. Selected Areas Commun.*, Vol. SAC-12, No. 4, pp. 580–592, May 1994.

[97] Z. Zvonar and D. Brady, "Multiuser detection in single-path fading channels," *IEEE Trans. Commun.*, Vol. COM-42, No. 2-4, pp. 1729–1739, Feb.-April 1994.

[98] N. Bambos and G.J. Pottie, "Power control based admission policies in cellular radio networks," *Globecom Conf. Rec.*, pp. 863–867, Dec. 1992.

[99] B. Jabbarin, "Intelligent network concepts in mobile communications," *IEEE Commun. Mag.*, Vol. 30, No. 2, pp. 64–69, Feb. 1992.

[100] D.C. Cox. "Wireless network access for personal communications," *IEEE Commun. Mag.*, Vol. 30, No. 12, pp. 96–115, Dec. 1992.

[101] J.H. Lodge. "Mobile satellite communications systems: Toward global personal communications," *IEEE Commun. Mag.*, Vol. 29, No. 11, pp. 24–30, Nov. 1991.

[102] D.J. Goodman. "Trends in cellular and cordless communications," *IEEE Commun. Mag.*, Vol. 29, No. 6, pp. 31–40, June 1991.

[103] J. Walrand, *Communication Networks: A First Course.* Homewood, IL: Aksen Associates, 1991.

[104] A.S. Tanenbaum, *Computer Networks.* 2nd Ed. Englewood Cliffs, N.J.: Prentice-Hall, 1988.

[105] S. Chia, "The universal mobile telecommunication system," *IEEE Commun. Mag.*, Vol. 30, No. 12, pp. 54–62, Dec. 1992.

[106] R.H. Katz, "Adaptation and mobility in wireless information systems," *IEEE Personal Commun. Mag.*, pp. 6–17, Vol. 1, No. 1, Jan.-March 1994.