PART 1

THEORY

1

BASIC CONCEPTS IN WIRELESS COMMUNICATIONS

1.1 OVERVIEW

In this chapter, we review the important and basic concepts in wireless communications. In Section 1.2, we first review different types of wireless channel models, namely, *time dispersion*, *multipath dispersion*, and *spatial dispersion* in microscopic fading. Concepts of frequency-selective fading, frequency flat fading, fast fading, slow fading, *coherence bandwidth*, *coherence time*, and *coherence distance* will be introduced. In Section 1.3, we establish the equivalence of discrete-time and continuous-time models in wireless communications for both the frequency flat fading and frequency-selective fading channels. In Section 1.4, we review the important and fundamental concepts of entropy, mutual information, and channel capacity, which are critical to the understanding of the materials and approaches in the subsequent chapters. Finally, in Section 1.5, we conclude with a brief summary of main points.

1.2 WIRELESS CHANNEL MODELS

A typical communication system consists of a transmitter, a receiver, and a channel. The *channel* is defined as the physical medium linking the transmitter output and the receiver input. For instance, telephone wire, optical fiber, and the atmosphere are different examples of communication channels. In fact, communication channel plays a very important role in communication system design because the transmitter and receiver designs have to be optimized with respect to the target channel.

Channel-Adaptive Technologies and Cross-Layer Designs for Wireless Systems with Multiple Antennas: Theory and Applications. By V. K. N. Lau and Y.-K. R. Kwok
ISBN 0-471-64865-5

In this book, we focus on the wireless communication channels involving radiofrequencies. In other words, the atmosphere is the medium carrying radiowaves. Please refer to References 15 and 114 for a more detailed introduction to wireless communication channels. Specifically, we briefly review the statistical models of wireless communication channels for single-antenna and multiple-antenna systems, which are frequently used in the analysis and the design of wireless communication systems.

1.2.1 AWGN Channel Model

We consider the simplest wireless channel, the *additive white Gaussian noise* (AWGN) channel. Without loss of generality, we consider single-antenna systems as illustrative in this section. The received signal ($y(t)$) is given by the transmitted signal ($x(t)$) plus a white Gaussian noise ($z(t)$)

$$y(t) = \sqrt{L}x(t) + z(t) \tag{1.1}$$

where L is the power attenuation from the transmitter to the receiver. In free space, L obeys the inverse square law.[1]

The AWGN channel is in fact quite accurate in deep-space communications and the communication links between satellite and Earth station. However, it is far from accurate in most terrestrial wireless communications, due to multipath, reflection, and diffraction. Yet, AWGN channel serves as an important reference on the performance evaluation of communication systems.

In terrestrial wireless communications, signals travel to the receiver via multiple paths, and this creates additional distortion to the transmitted signal on top of the channel noise. In general, the effect of multipath and reflections could be modeled as wireless fading channels or *microscopic fading*. Factors affecting the microscopic fading include multipath propagation, speed of the mobile (unit), speed of the surrounding objects, the transmission symbol duration, and the transmission bandwidth of the signal.

1.2.2 Linear Time-Varying Deterministic Spatial Channel

Consider a general linear channel that can be characterized by a *lowpass equivalent* time-domain impulse response denoted by $h(t; \tau, r)$ (where t is the time-varying parameter, τ is the path delay parameter, and r is the spatial position parameter). The general linear channel is therefore characterized by three independent dimensions: the *time dimension* (characterized by the time parameter t), the *delay dimension* (characterized by the delay parameter τ), and the *spatial dimension* (characterized by the position parameter r). Given a lowpass equivalent input signal $x(t)$, the lowpass equivalent received signal

[1] This is the received power level reduced by 4 times whenever the distance between the transmitter and the receiver increases by 2 times.

$y(t, r)$ through the general linear deterministic channel at time t and position r is given by

$$y(t,r) = \int_{-\infty}^{\infty} h(t;\tau,r)x(t-\tau)d\tau + z(t,r) \tag{1.2}$$

where the input signal (in time domain) is mapped into output signal (in time domain and spatial domain) through the impulse response $h(t; \tau, r)$. For simplicity, we shall discuss the channel characterization based on single-antenna systems. Extension to the MIMO systems will be straightforward. For example, to extend the model to MIMO systems, the transmitted signal $x(t)$ is replaced by the $n_T \times 1$ vector $\mathbf{x}(t)$:

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_{n_T}(t) \end{bmatrix}$$

The received signal $y(t, r)$ and the noise signal $z(t, r)$ are replaced by the $n_R \times 1$ vector $\mathbf{y}(t)$ and $\mathbf{z}(t)$, respectively:

$$\mathbf{y}(t) = \begin{bmatrix} y_1(t, r_1) \\ \vdots \\ y_{n_R}(t, r_{n_R}) \end{bmatrix}$$

$$\mathbf{z}(t) = \begin{bmatrix} z_1(t) \\ \vdots \\ z_{n_R}(t) \end{bmatrix}$$

The time-varying channel impulse response is replaced by the $n_R \times n_T$ matrix $\mathbf{h}(t; \tau, \mathbf{r})$, given by

$$\mathbf{h}(t;\tau,\mathbf{r}) = \begin{bmatrix} h_{1,1}(t;\tau,r_{1,1}) & \cdots & h_{1,n_T}(t;\tau,r_{1,n_T}) \\ \vdots & \ddots & \vdots \\ h_{n_R,1}(t;\tau,r_{n_R,1}) & \cdots & h_{n_R,n_T}(t;\tau,r_{n_R,n_T}) \end{bmatrix}$$

where $\mathbf{h}[i, j]$ is the channel response corresponding to the jth transmit antenna and the ith receive antenna and $\mathbf{r}$ is the corresponding position parameter.

1.2.2.1 Spectral Domain Representations. While Equation (1.2) gives the fundamental input–output relationship of the linear deterministic channels, Fourier transforms are sometimes useful for gaining additional insights in channel analysis. Since the channel impulse responses $h(t; \tau, r)$ are defined

over the time, delay, and position domains, Fourier transforms may be defined for each of these domains, and they are elaborated as follows:

Frequency Domain. The spectral domain of the delay parameter τ is called the *frequency domain* ν. They are related by the Fourier transform relationship $h(t; \tau, r) \leftrightarrow H(t; \nu, r)$. For example, $H(t; \nu, r)$ is given by

$$H(t;\nu,r)=\int_{-\infty}^{\infty} h(t;\tau,r)\exp(-j2\pi\tau\nu)d\tau \tag{1.3}$$

Since $x(t-\tau)=\int_{-\infty}^{\infty}X(\nu)\exp(j2\pi\nu(t-\tau))d\nu$, substituting into Equation (1.2), we have

$$\begin{aligned} y(t,r) &= \int_{-\infty}^{\infty} h(t;\tau,r)\int_{-\infty}^{\infty} X(\nu)\exp(j2\pi\nu(t-\tau))d\nu\, d\tau \\ &= \int_{-\infty}^{\infty} X(\nu)\exp(j2\pi\nu t)\left(\int_{-\infty}^{\infty} h(t;\tau,r)\exp(-j2\pi\nu\tau)d\tau\right)d\nu \\ &= \int_{-\infty}^{\infty} H(t;\nu,r)X(\nu)\exp(j2\pi\nu t)d\nu \end{aligned} \tag{1.4}$$

Hence, the channel response can also be specified by the *time-varying transfer function* $H(t; \nu, r)$. In addition, it can be found from Equation (1.4) that the output signal (in time domain) $y(t, r)$ is mapped from the input signal (in frequency domain) $X(\nu)$ through the time-varying transfer function $H(t; \nu, r)$.

Doppler Domain. The spectral domain of the time parameter t is called the *Doppler domain* f. They are related by the Fourier transform relationship $h(t; \tau, r) \leftrightarrow H(f; \tau, r)$. For example, $H(f; \tau, r)$ is given by

$$H(f;\tau,r)=\int_{-\infty}^{\infty} h(t;\tau,r)\exp(-j2\pi tf)dt \tag{1.5}$$

Similarly, the input signal (in delay domain) $x(\tau)$ can be mapped into the output signal (in frequency domain) $Y(f, r)$ through the transfer function $H(f; \tau, r)$:

$$Y(f,r)=\int_{-\infty}^{\infty} H(f;\tau,r)x(t-\tau)d\tau \tag{1.6}$$

Wavenumber Domain. The spectral domain of the position parameter r is called the *wavenumber domain* k. The wavenumber in three-dimensional space has a physical interpretation of the plane-wave propagation direction. The position and wavenumber domains are related by the Fourier transform relationship $h(t; \tau, r) \leftrightarrow H(t; \tau, k)$. For example, $H(t; \tau, k)$ is given by

$$H(t;\tau,k)=\int_{-\infty}^{\infty} h(t;\tau,r)\exp(-j2\pi rk)dr \tag{1.7}$$

In general, two important concepts are applied to describe these linear deterministic channels: spreading and coherence. The *spreading* concept deals with the physical spreading of the received signal over the parameter space (τ, f, k) when a narrow pulse is transmitted in the corresponding domain. The coherence *concept* deals with the variation of the channel response with respect to another parameter space (v, t, r). These concepts are elaborated in the text below.

1.2.2.2 Channel Spreading. The channel spreading concepts of describing the general linear deterministic channels focus on the spreading of the received signals over the parameter space (τ, f, k) when a narrow pulse in the corresponding parameter is transmitted. We therefore have three types of channel spreading:

Delay Spread. If we transmit a test pulse that is narrow in time, the received signal will have a spread in propagation delay τ due to the sum of different propagation delays of multipaths at the receiver. From Equation (1.2), when the transmit signal is narrow in time, we have $x(\tau) = \delta(\tau)$. Hence, the received signal is given by $y(t, r) = h(t; \tau = t, r)$. The plot of $|h(t; t, r)|^2$ versus time is called the *power-delay profile* as illustrated in Figure 1.1a. The range of delays where we find significant power is called the *delay spread* σ_τ.

Doppler Spread. If we transmit a test pulse narrow in frequency $X(v) = \delta(v)$, the received signal in general will experience a spread in the received spectrum. The range of spectrum spread in the frequency domain of the received signal $Y(f, r)$ refers to Doppler spread. The Doppler spread is given by $f_d = \frac{v}{\lambda}$, where v is the maximum speed between the transmitter and the receiver and λ is the wavelength of the carrier. This is illustrated in Figure 1.2a.

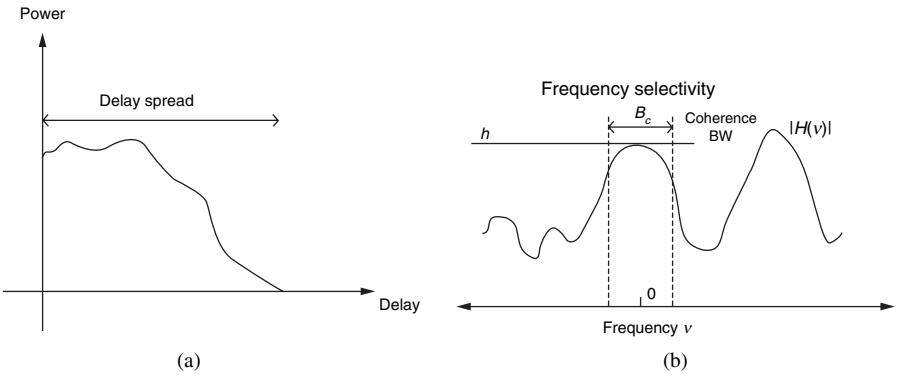


Figure 1.1. Delay spread (a) and coherence bandwidth (b).

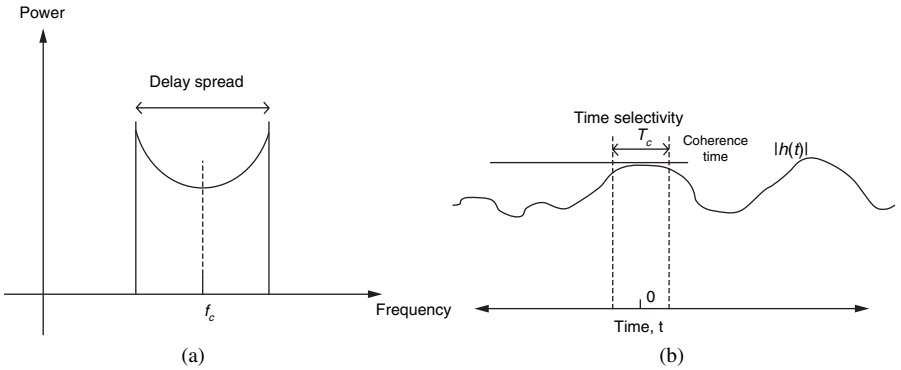


Figure 1.2. Doppler spread (a) and coherence time (b).

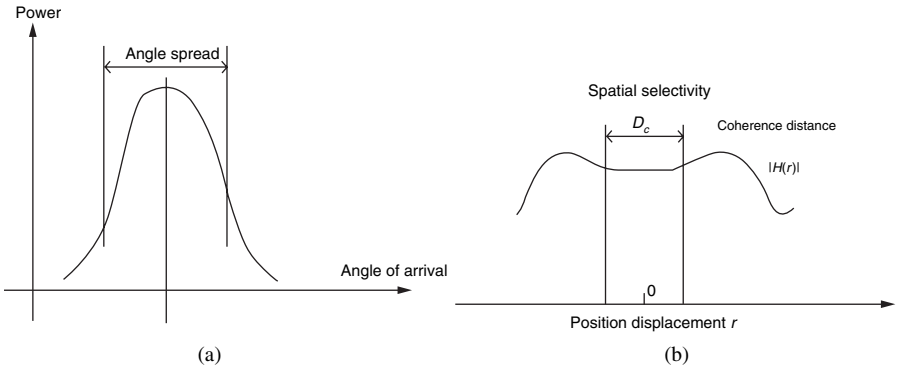


Figure 1.3. Illustration of angle spread and coherence distance.

Angle Spread. Finally, the scattering environment introduces variation in the spatial parameter r, which is equivalent to the spreading in the wavenumber domain k, this is called the *angle spread*. For example, if a test pulse narrow in direction is transmitted, the received signal will experience a spread in the wavenumber domain (angle of arrivals) due to the scattering surroundings; this is called the *angle spread* as illustrated in Figure 1.3a.

1.2.2.3 Channel Coherence. On the other hand, we can describe the linear deterministic channels by looking at the channel coherence or channel selectivity properties over the parameter space (ν, t, r). A channel is said to be selective in the corresponding dimension if the channel response varies as a function of that parameter. The opposite of selectivity is *coherence*. A channel has coherence in the corresponding dimension if it does not change significantly as a function of that parameter. The channel coherence proper-

ties with respect to the frequency, time, and position dimensions are elaborated below.

Frequency Coherence or Frequency Selectivity. A wireless channel has frequency coherence if the magnitude of the carrier wave does not change over a frequency window of interest. This window of interest is usually the bandwidth of the transmitted signal. Hence, mathematically, we can quantify the frequency coherence of the channels by a parameter called the *coherence bandwidth* B_c

$$|H(t; \nu, r)| \approx H_0(t, r) \quad \text{for} \quad |\nu| \leq \frac{B_c}{2} \tag{1.8}$$

where $H_0(t, r)$ is a constant in the frequency domain ν and B_c is the size of the frequency window where we have constant channel response. The largest value of B_c for which Equation (1.8) holds is called the *coherence bandwidth* and can be interpreted as the range of frequencies over which the channel appears static. Figure 1.1b illustrates the concept of coherence bandwidth. In fact, if the bandwidth of the transmitted signal is larger than the coherence bandwidth of the channel, the signal will experience frequency distortion according to Equation (1.4). Such a channel is classified as a *frequency-selective fading channel*. On the other hand, if the transmitted signal has bandwidth smaller than the coherence bandwidth of the channel, frequency distortion will be no introduced to the signal and therefore, the channel will be classified as a *frequency flat fading channel*. Frequency selectivity introduces *intersymbol interference*, and this results in irreducible error floor in the BER (bit error rate) curve. Hence, this is highly undesirable. Whether a signal will experience frequency-selective fading or flat fading depends on both the environment (coherence bandwidth) and the transmitted signal (transmitted bandwidth).

Time Coherence or Time Selectivity. A wireless channel has temporal coherence if the envelope of the unmodulated carrier does not change over a time window of interest. The time coherence of channels can be specified by a parameter called the *coherence time* T_c

$$|H(t; \nu, r)| \approx H_0(\nu, r) \quad \text{for} \quad |t| \leq \frac{T_c}{2} \tag{1.9}$$

where $|H(t; \nu, r)|$ is the envelope of the response at the receiver (at a fixed position r) when a single-tone signal (at a fixed frequency ν) is transmitted, $H_0(\nu, r)$ is a constant in the time domain t and T_c is the size of the time window where we have constant channel response. The largest value of T_c for which Equation (1.9) holds is called the *coherence time* and can be interpreted as the range of time over which the channel

appears static as illustrated in Figure 1.2b. In wireless fading channels, temporal incoherence (or time selectivity) is caused by the motion of the transmitter, the receiver or the scattering objects in the environment. Time selectivity can degrade the performance of wireless communication systems. If the transmit data rate is comparable to the coherence time, it becomes extremely difficult for the receiver to demodulate the transmitted signal reliably because the time selectivity within a symbol duration causes catastrophic distortion on the received pulseshape. Hence, when the transmit symbol duration T_s is longer than the coherence time T_c, we have *fast fading channels*. On the other hand, when the transmit symbol duration is shorter than the coherence time, we have *slow fading channels*. In the extreme case of slow fading the channel remains static for the entire transmit frame.

Spatial Coherence or Spatial Selectivity. A wireless channel has spatial coherence if the magnitude of the carrier wave does not change over a spatial displacement of the receiver. Mathematically, the spatial coherence can be parameterized by the *coherence distance*, D_c

$$|H(t; \nu, r)| \approx H_0(t, \nu) \quad \text{for} \quad |r| \leq \frac{D_c}{2} \tag{1.10}$$

where $|H(t; \nu, r)|$ is the envelope of the response at the receiver when a single-tone signal (at a fixed frequency ν) is transmitted (at a fixed time t), $H_0(t; \nu)$ is a constant with respect to the spatial domain r, and D_c is the size of the spatial displacement where we have constant channel response. The largest value of D_c for which Equation (1.10) holds is called the *coherence distance* and can be interpreted as the range of displacement over which the channel appears static as illustrated in Figure 1.3b. Note that for a wireless receiver moving in three-dimensional space, the coherence distance is a function of the direction that the receiver travels; that is, the position displacement **r** is a vector instead of a scalar. Hence, the study of spatial coherence is much more difficult than the study of the scalar quantities of temporal or frequency coherence. While frequency selectivity is a result of multipath propagation arriving with many different time delays τ, spatial selectivity is caused by the multipath propagation arriving from different directions in space. These multipath waves are superimposed on each other, creating pockets of constructive and destructive interference in the three-dimensional spatial domain so that the received signal power does not appear to be constant over small displacements of receiver position. Hence, if the distance traversed by a receiver is greater than the coherence distance, the channel is said to be *spatially selective* or *small-scale fading*. On the other hand, if the distance traversed by a receiver is smaller than the coherence distance, the channel is said to be *spatially flat*. Spatially selective or spatial flat fading is important when we have to apply spatial diver-

sity (or spatial multiplexing) and beamforming. For instance, in order to produce a beam of energy along the designated direction through antenna array, the dimension of the antenna array must be within the coherence distance of the channels. On the other hand, to effectively exploit the spatial multiplexing or spatial diversity of MIMO systems, the spacing of the antenna array must be larger than coherence distance of the channels.

Figure 1.4 summarizes the various behaviors of microscopic fading channels.

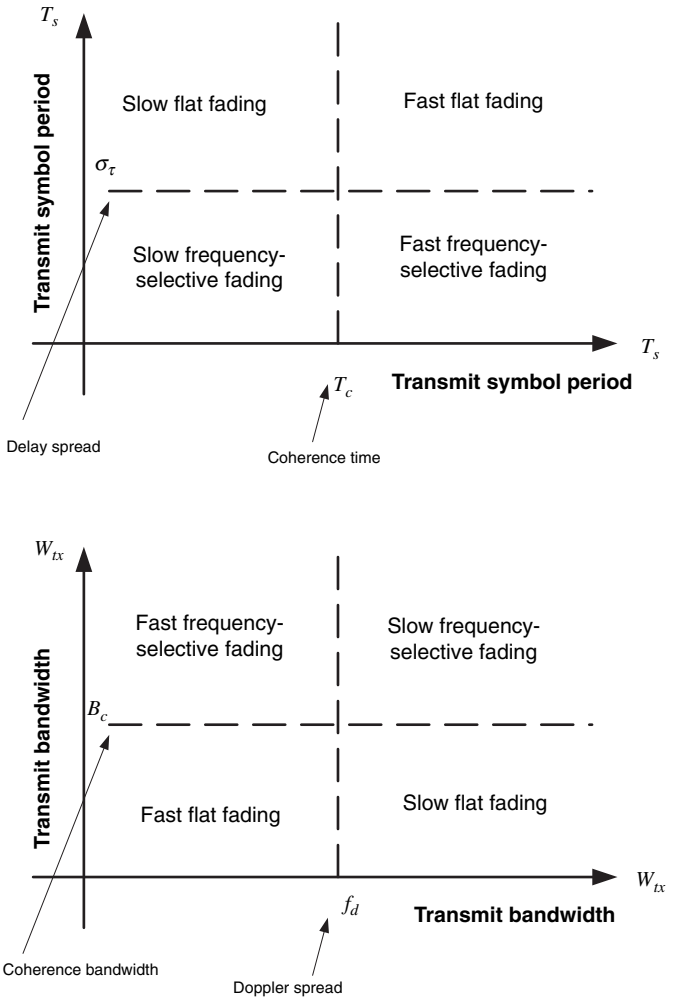


Figure 1.4. Summary of fading channels.

1.2.3 The Random Channels

In Section 1.2.2, we have introduced the general linear deterministic channel where the relationship of output given an input signal is modeled as a general time-varying system. However, in practice, the wireless fading channels we experience are random instead of deterministic; that is, $h(t; \tau, r)$ is a random process instead of a deterministic quantity. Hence, in this section, we shall extend the model of linear deterministic channels to cover the random channels.

For simplicity, let's consider a channel response on the time dimension only. The most common way to characterize the statistical behavior of the random process $h(t)$ is by means of autocorrelation:

$$R_h(t_1, t_2) = \varepsilon[h(t_1)h^*(t_2)] \tag{1.11}$$

The random process is called *wide-sense stationary* (WSS) if the autocorrelation $R_h(t_1, t_2)$ is a function of $|t_1 - t_2|$.

On the other hand, we can also consider the correlation in the spectral domain of t. Specifically, after Fourier transform on $h(t)$, we have a random frequency-varying process $H(f)$. The autocorrelation of the random process $H(f)$ is given by

$$S_H(f_1, f_2) = \varepsilon[H(f_1)H^*(f_2)] \tag{1.12}$$

Lemma 1.1 (Wide-Sense Stationary) A random process is WSS if and only if its spectral components are uncorrelated.

Proof Since $H(f)$ is the Fourier transform of $h(t)$, we have

$$R_h(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varepsilon[H(f_1)H^*(f_2)]\exp(j2\pi[f_1t_1 - f_2t_2])df_1\, df_2$$

Suppose that we have uncorrelated spectral components: $S_H(f_1, f_2) = S_h(f_1)\delta(f_1 - f_2)$. Hence $\varepsilon[H(f_1)H^*(f_2)] = 0$ for $f_1 \neq f_2$. For this case, we can write the complex exponent as $\exp(j2\pi f_1[t_1 - t_2])$ and therefore, $R_h(t_1, t_2)$ is a function of $|t_1 - t_2|$.

On the other hand, if $R_h(t_1, t_2)$ is a function of $|t_1 - t_2|$, the multiplier in the integration must be zero for $f_1 \neq f_2$ because otherwise, there is no way to force $\exp(j2\pi[f_1t_1 - f_2t_2])$ to be a function of $|t_1 - t_2|$ only. Hence, we must have $\varepsilon[H(f_1)H^*(f_2)] = 0$ for $f_1 \neq f_2$.

In fact, the condition of uncorrelated spectral components is often referred to as *uncorrelated scattering* (US).

1.2.3.1 Joint Correlation and Spectrum. Now, let's consider the general random channel response $H(t; \nu, r)$ with respect to the time t, frequency ν, and position r. To accommodate all the random dependencies of such a channel,

it is possible to define a joint correlation of $H(t; \nu, r)$ with respect to (t, ν, r). The joint correlation of the channel response is given by

$$R_H(t_1, \nu_1, r_1; t_2, \nu_2, r_2) = \varepsilon[H(t_1, \nu_1, r_1)H^*(t_2, \nu_2, r_2)] \tag{1.13}$$

For simplicity, we assume the random channel is a *wide-sense stationary, uncorrelated scattering* (WSS-US) random process. Hence, the joint correlation $R_H(t_1, \nu_1, r_1; t_2, \nu_2, r_2)$ is a function of $(\Delta t, \Delta\nu, \Delta r)$ only where $\Delta t = |t_1 - t_2|$, $\Delta\nu = |\nu_1 - \nu_2|$ and $\Delta r = |r_1 - r_2|$. Note that WSS refers to wide-sense stationary with respect to the time parameter t, the frequency parameter ν, and the position parameter r. On the other hand, uncorrelated scattering refers to uncorrelation of the spectral components (as a result of Lemma 1.1) in the *Doppler* parameter f, *delay* parameter τ, and *wavenumber* parameter k

$$\begin{aligned} S_H(f_1, \tau_1, k_1; f_2, \tau_2, k_2) &= \varepsilon[H(f_1, \tau_1, k_1)H^*(f_2, \tau_2, k_2)] \\ &= S_H(f_1, \tau_1, k_1)\delta(f_1 - f_2)\delta(\tau_1 - \tau_2)\delta(k_1 - k_2) \end{aligned} \tag{1.14}$$

where $S_H(f_1, \tau_1, k_1)$ is the *power spectral density* of the random process $H(t, \nu, r)$. The Wiener–Khintchine theorem for WSS-US process leads to the following Fourier transform relationship between the autocorrelation function $R_H(\Delta t, \Delta\nu, \Delta r)$ and the power spectral density $S_H(f, \tau, k)$:

$$R_H(\Delta t, \Delta\nu, \Delta r) \leftrightarrow S_H(f, \tau, k) \tag{1.15}$$

1.2.3.2 Time–Frequency Transform Mapping. Since the joint correlation function $R_H(\Delta t, \Delta\nu, \Delta r)$ is a function of three independent parameters, it is easier for illustration purposes to fix one dimension and focus on the interrelationship between the other two dimensions. For instance, consider single-antenna systems with the receiver at a fixed position r. Thus, this random channel has no dependence on r. Hence, the statistical properties of the random channels can be specified by either the *time–frequency autocorrelation* $R_H(\Delta t, \Delta\nu)$ or the *delay–Doppler spectrum* $S_H(f, \tau)$ as illustrated in Figure 1.5. In a WSS-US channel, knowledge of only one is sufficient as they are two-dimensional Fourier transform pairs.

In Section 1.2.2, we have introduced the concepts of *coherence time* and *coherence bandwidth* or equivalently, *Doppler spread* and *delay spread* for deterministic channels. We will try to extend the definition of these parameters for WSS-US random channels. From the time–frequency autocorrelation function, the correlation in time dimension is given by

$$R_H(\Delta t) = R_H(\Delta t, \Delta\nu)\big|_{\Delta\nu=0} \tag{1.16}$$

The coherence time T_c for the random channel is defined to be value of Δt such that $R_H(\Delta t) < 0.5$.

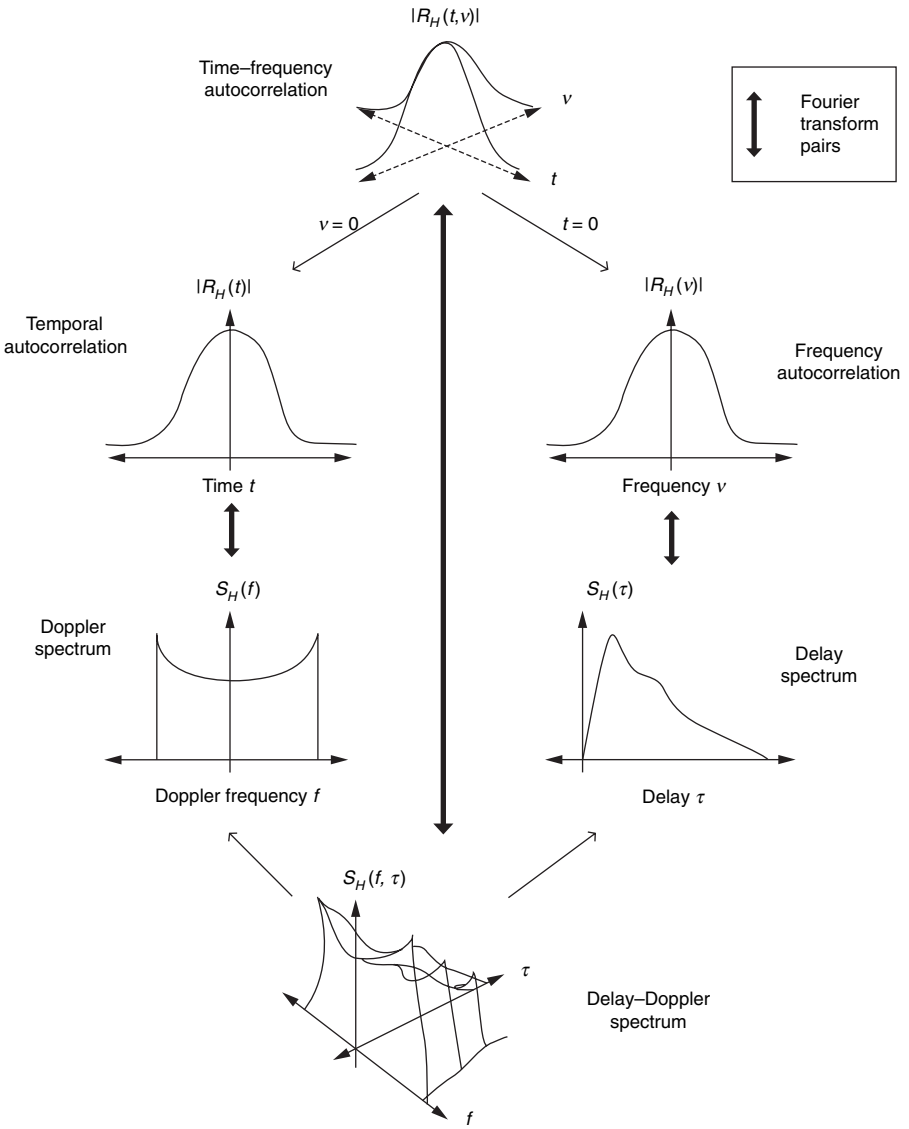


Figure 1.5. Time–frequency autocorrelation and delay–Doppler spectrum.

Similarly, the correlation in the frequency dimension is given by

$$R_H(\Delta\nu) = R_H(\Delta t, \Delta\nu)|_{\Delta t=0} \tag{1.17}$$

The coherence time B_c for the random channel is defined to be value of $\Delta\nu$ such that $R_H(\Delta\nu) < 0.5$.

On the other hand, we can characterize the random channel on the basis of the *delay–Doppler spectrum*. For instance, the *Doppler spectrum* is given by

$$S_H(f) = \int_{-\infty}^{\infty} S_H(f, \tau) d\tau \tag{1.18}$$

The Doppler spread σ_f^2 is defined as the second centered moment of the Doppler spectrum:

$$\sigma_f^2 = \frac{\int_{-\infty}^{\infty} f^2 S_H(f) df}{\int_{-\infty}^{\infty} S_H(f) df} - \left(\frac{\int_{-\infty}^{\infty} f S_H(f) df}{\int_{-\infty}^{\infty} S_H(f) df} \right)^2 \tag{1.19}$$

Similarly, the *power-delay profile* is given by

$$S_H(\tau) = \int_{-\infty}^{\infty} S_H(f, \tau) df \tag{1.20}$$

The delay spread σ_τ^2 is defined as the second centered moment of the power-delay profile:

$$\sigma_\tau^2 = \frac{\int_{-\infty}^{\infty} \tau^2 S_H(\tau) d\tau}{\int_{-\infty}^{\infty} S_H(\tau) d\tau} - \left(\frac{\int_{-\infty}^{\infty} \tau S_H(\tau) d\tau}{\int_{-\infty}^{\infty} S_H(\tau) d\tau} \right)^2 \tag{1.21}$$

Since the Doppler spectrum and the time autocorrelation function are Fourier transform pairs, a large Doppler spread σ_f^2 will result in small coherence time T_c and therefore faster temporal fading and vice versa. Similarly, the power-delay profile and the frequency autocorrelation function are Fourier transform pairs. Hence, a large delay spread σ_τ^2 will result in a small coherence bandwidth B_c and vice versa. In practice, the four parameters are related by

$$B_c \approx \frac{1}{5\sigma_\tau} \tag{1.22}$$

and

$$T_c \approx \frac{1}{5\sigma_f} \tag{1.23}$$

1.2.3.3 Frequency–Space Transform Mapping. For a static channel, we may extend the time–frequency map described in the previous section for the *frequency–space* relationship as illustrated in Figure 1.6.

In this diagram, the joint space–frequency autocorrelation $R_H(\Delta\nu, \Delta r)$ and the joint delay–wavenumber spectrum $S_H(\tau, k)$ are related by Fourier trans-

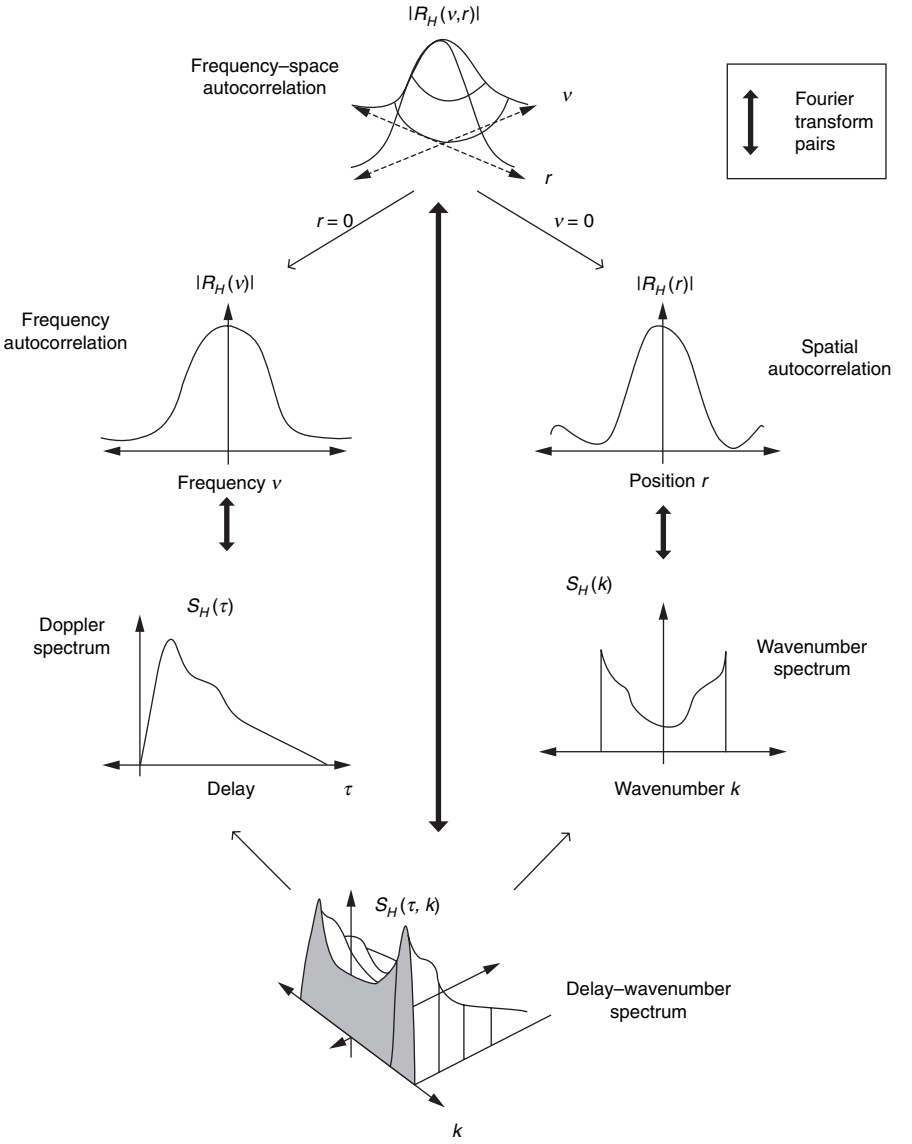


Figure 1.6. Illustration of frequency–space autocorrelation and delay–wavenumber spectrum.

form pairs. In Section 1.2.2, we have introduced the concepts of *coherence distance* and *angle spread* for deterministic channels. We shall try to extend the definition of these parameters for WSS-US random channels. From the frequency–space autocorrelation function, the single dimension spatial autocorrelation of the random channels is given by

$$R_H(\Delta r) = R_H(\Delta v, \Delta r)|_{\Delta v=0} \tag{1.24}$$

The coherence distance D_c is therefore defined as the maximum Δr such that $R_H(\Delta r) < 0.5$.

Similarly, we can characterize the statistical behavior of the random channels by the delay–wavenumber spectrum $S_H(\tau, k)$. Consider the single-dimension wavenumber spectrum $S_H(k)$:

$$S_H(k) = \int_{-\infty}^{\infty} S_H(\tau, k) d\tau \tag{1.25}$$

The *angle spread* σ_k^2 is defined to be the second centered moment of the wavenumber spectrum:

$$\sigma_k^2 = \frac{\int_{-\infty}^{\infty} k^2 S_H(k) dk}{\int_{-\infty}^{\infty} S_H(k) dk} - \left(\frac{\int_{-\infty}^{\infty} k S_H(k) dk}{\int_{-\infty}^{\infty} S_H(k) dk} \right)^2 \tag{1.26}$$

An important indication of the nature of the channel is called the *spread factor*, given by $B_c T_c$. If $B_c T_c < 1$, the channel is said to be *underspread*; otherwise, it is called *overspread*. In general, if $B_c T_c \ll 1$, the channel impulse response could be easily measured and the measurement could be utilized at the receiver for demodulation and detection or at the transmitter for adaptation. On the other hand, if $B_c T_c \gg 1$, channel measurement would be extremely difficult and unreliable. In this book, we deal mainly with *underspread* fading channels.

1.2.4 Frequency-Flat Fading Channels

We shall consider the effect of fading channels on a transmitted signal. We first look at a simple case called a *flat fading channel*. Let $x(\tau)$ be the lowpass equivalent signal transmitted over the channel and $X(\nu)$ be the corresponding Fourier transform. The lowpass equivalent received signal $y(t, r)$ is given by

$$y(t, r) = \int_{-\infty}^{\infty} h(t; \tau, r) x(t - \tau) d\tau + z(t) = \int_{-\infty}^{\infty} H(t; \nu, r) X(\nu) \exp(j2\pi\nu t) d\nu + z(t) \tag{1.27}$$

where $h(t; \tau, r)$ is the *time-varying impulse response* and $H(t; \nu, r)$ is the *time-varying transfer function* of the channel. Note that both $H(t; \nu, r)$ and $h(t; \tau, r)$ are random processes. Suppose that the two-sided bandwidth W of $x(t)$ is less than the coherence bandwidth B_c. According to the definition of the correlation function $R_H(0; \Delta\nu, 0)$, the random channel fading $H(t, \nu, r)$ is highly correlated within the range of the transmitted bandwidth $\nu \in [-W/2, W/2]$. Hence, all the frequency component of $X(\nu)$ undergoes the same complex fading within the range of frequencies $\nu \in [-W/2, W/2]$. This means that within the bandwidth W of $X(\nu)$, we obtain $H(t, \nu, r) = h(t; \tau, r) = h(t, r)$, where $h(t, r)$ is a complex stationary random process in t and r only. This results in both the

envelope attenuation and phase rotation on the transmitted signal. The received signal simplifies to

$$y(t,r) = h(t,r)\int_{-\infty}^{\infty} X(\nu)\exp(j2\pi\nu t)d\nu + z(t) = h(t,r)x(t) + z(t) \tag{1.28}$$

Hence, the signal $x(t)$ is said to experience the *flat fading channel*. Therefore, the flat fading channel has a time-varying multiplicative effect on the transmitted signal. In this case, the multipath components are not *resolvable* because the signal bandwidth $W \ll B_c = 1/(5T_c)$. In accordance with the central-limit theorem, the random process $h(t, r)$ (as a result of multipath aggregation) can be well approximated by complex stationary zero-mean Gaussian random process. A flat fading channel is also called a *slowly fading channel* if the time duration of a transmitted symbol T_s is much smaller than the coherence time of the channel T_c. Otherwise, it is called a *fast fading channel*. Since in general $W \geq 1/T_s$, a slowly flat fading channel is underspread because $B_cT_c \ll 1$.

1.2.5 Frequency-Selective Fading Channels

When the transmit signal bandwidth $W \gg B_c$, frequency components in $X(\nu)$ have frequency separation greater than B_c. In this case, the random fading components $H(t; \nu_1, r)$ and $H(t; \nu_2, r)$ become uncorrelated whenever $|\nu_1 - \nu_2| \geq B_c$. Hence, some frequency components in $X(\nu)$ (within the transmitted bandwidth W) may experience independent fading. In such case, the channel is said to be *frequency-selective*.

Any lowpass equivalent signal $x(t)$ with two-sided bandwidth W and time duration $t \in \{0, T\}$ may be represented geometrically by a $(N = WT)$-dimensional vector $\mathbf{x} = [x(0), x(1/W), \ldots, x(T - 1/W)]$ in the signal space spanned by the orthonormal basis $\{\psi_0(t), \ldots, \psi_{N-1}(t)\}$. Specifically, by sampling theorem, $\psi_n(t) = \text{sinc}(\pi W(t - n/W))/\pi W(t - n/W)$, and the lowpass equivalent signal is given by

$$x(t) = \sum_n \mathbf{x}[n]\psi_n(t) \tag{1.29}$$

where $\mathbf{x}[n]$ denotes the nth component of the vector $\mathbf{x}$. The corresponding Fourier transform of $x(t)$ is given by

$$x(\nu) = \begin{cases} \dfrac{1}{W}\displaystyle\sum_n \mathbf{x}[n]\exp\frac{-j2\pi\nu n}{W} & |\nu| \leq W/2 \\ 0 & |\nu| > W/2 \end{cases} \tag{1.30}$$

Hence, the received signal as a result of frequency-selective fading is given by

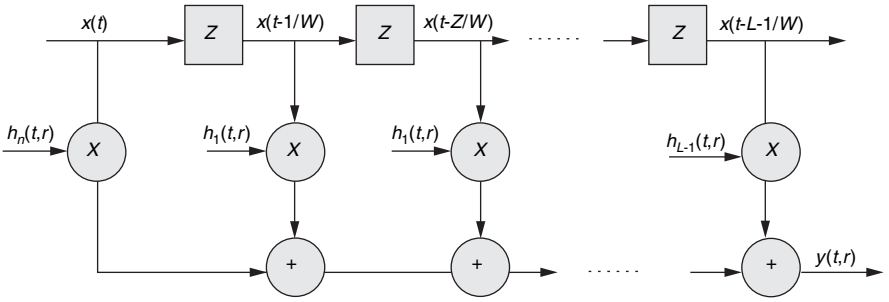


Figure 1.7. Illustration of tapped-delay-line frequency-selective channel model.

$$\begin{aligned} y(t,r) &= \int_{-\infty}^{\infty} H(t;v,r)X(v)\exp(j2\pi vt)dv + z(t) \\ &= \frac{1}{W}\sum_n \mathbf{x}[n]\int_{-\infty}^{\infty} H(t;v,r)\exp\left(j2\pi v\left(t-\frac{n}{W}\right)\right)dv + z(t) \\ &= \frac{1}{W}\sum_n x\left(\frac{n}{W}\right)h\left(t;t-\frac{n}{W},r\right) + z(t) \\ &= \sum_{n=0}^{L-1} x\left(t-\frac{n}{W}\right)h_n(t,r) + z(t) \end{aligned} \tag{1.31}$$

where $h_n(t,r) = \frac{1}{W}h\,(t;n/W,r)$ and $L = [W/B_c]$ is the number of *resolvable multipaths* as seen by the transmitted signal. Hence, the larger the transmitted signal bandwidth W (or the smaller the channel coherence bandwidth B_c), the larger the number of resolvable multipaths that the transmit signal will see. Figure 1.7 illustrates the equivalent view of the frequency-selective channel model in Equation (1.31).

According to the statistical characterization of the channel presented above, the channel taps $h_n(t, r)$ are complex stationary random processes. Specifically, due to the central-limit theorem, the random process can be well approximated by zero-mean complex Gaussian stationary random process. Furthermore, $h_n(t)$ and $h_m(t)$ are statistically independent if $n \neq m$ due to the uncorrelated scattering assumption.

1.3 EQUIVALENCE OF CONTINUOUS-TIME AND DISCRETE-TIME MODELS

While the physical transmitted signals and received signals are all in continuous time domain, it is quite difficult to gain proper design insights by working in the original forms. For example, given the received signal[2] $y(t)$, it is not clear

[2] Without loss of generality, ignore the notation about the position parameter r in the received signal $y(t)$ to simplify notation.

what should be the optimal structure to extract the information bits carried by the signal. Hence, for statistical signal detection and estimation problems, it is usually more convenient to formulate the problems in the *geometric domain* or *discrete-time domain*. In this section, we try to establish the equivalence relationship of the two domains.

1.3.1 Concepts of Signal Space

Theorem 1.1 (Orthonormal Basis for Random Process) For any wide-sense stationary (WSS) complex random process $x(t)$ for $t \in [0,\mathrm{T}]$, there exists a set of orthonormal[3] basis functions $\{\psi_0(t), \ldots, \psi_{N-1}(t)\}$ and complex random variables $\{x_0, \ldots, x_{N-1}\}$ such that $\lim_{N\to\infty}\mathcal{E}[|x(t) - x_N(t)|^2] = 0$ for all t where $x_N(t)$ is given by:

$$x_N(t) = \sum_{n=0}^{N-1} x_n \psi_n(t)$$

and $x_n =< x(t), \psi_n(t) >$. In particular, the basis functions are the solution (called eigenfunctions) of the integral equation:

$$\int_0^T R_x(t-\tau)\psi_n(\tau)d\tau = \lambda_n \psi_n(t)$$

(where $R_x(t)$ is the autocorrelation function of $x(t)$) and $\mathcal{E}[x_n x_m^*] = \lambda_n \delta_{nm}$ (uncorrelated).

In addition, if $x(t)$ is bandlimited to W ($x(t)$ has significant energy only for $f \in [-W, W]$) and $\mathrm{WT} \gg 1$, we have $\lambda_n \approx 0$ for $n > 2WT$ and $\lambda_n \approx 1$ for $n < 2WT$.

Proof Please refer to [122] for the proof.

Since λ_n represents the energy of the basis function $\psi_n(t)$ over $[0, T]$, in other words only about $2WT$ coefficients in $x_N(t)$ have significant energy and we might say that the signal $x(t)$ lies in a signal space of $2WT$ dimensions.

1.3.2 Sufficient Statistics

Before we discuss the equivalence of continuous-time model and discrete-time model, we have to introduce a concept called *sufficient statistics*.

Definition 1.1 (Sufficient Statistics) Suppose that we have a probability density function on the random sample $f(x; \theta)$, where θ is an unknown parameter. Let $T(X_1, \ldots, X_N)$ (a function of N random samples $X_1, \ldots, X_N$ from the population) be a statistical estimate of the unknown parameter θ. The estimate T is called *sufficient statistic* for θ if $\mathbf{X} = (X_1, \ldots, X_N)$ is independent of θ given $T(X_1, \ldots, X_N)$:

$$\Pr[\mathbf{X}|T, \theta] = \Pr[\mathbf{X}|T]$$

[3] Orthonormal basis refers to $\int_0^{\mathrm{T}} \psi_n(t)\psi_m^*(t)dt = \delta(n - m)$.

Example 1.1 (Sufficient Statistics) Given: $X_1, \ldots, X_N, X_n \in \{0, 1\}$, is an independent and identically distributed (i.i.d.) sequence of coin tosses of a coin with unknown parameter $\theta = \Pr[X_n = 1]$. Given N, the number of 1s is a sufficient statistic for θ. Here $T(X_1, \ldots, X_N) = \Sigma_n X_n$. T is a sufficient statistic for θ because

$$\begin{aligned}\Pr\{(X_1, \ldots, X_N) = (x_1, \ldots, x_N) | T = k, \theta\} &= \Pr\{(X_1, \ldots, X_N) = (x_1, \ldots, x_N) | T = k\} \\ &= \begin{cases} \dfrac{1}{\binom{N}{k}}, \text{if } \displaystyle\sum_n x_n = k \\ 0, \text{otherwise} \end{cases}\end{aligned}$$

Example 1.2 (Sufficient Statistics) If X is Gaussian distributed with mean θ and variance 1, and $X_1, \ldots, X_N$ are drawn independently according to this distribution, then $T = \frac{1}{N}\Sigma_n X_n$ is a *sufficient statistic* for θ. This is because $\Pr[(X_1, \ldots, X_N)|T]$ is independent of θ.

In other words, the concept of sufficient statistics allows us to discard data samples that are not related to the parameter in question because the conditional pdf (probability distribution function) of X given T is independent of θ. Note that sufficient statistic is not unique. There could be many sufficient statistics for the same parameter θ. As shown by Examples 1.1 and 1.2, it is relatively easy to *verify* whether a statistic T is sufficient with respect to a parameter θ. However, it is a more difficult problem to *identify* potential sufficient statistics for a parameter θ. Below we summarize the Neyman–Fisher factorization theorem, which can be used to find sufficient statistics in several examples.

Theorem 1.2 (Necessary and Sufficient Condition for Sufficient Statistics) If we can factorize the pdf $p(\mathbf{X}; \theta)$ as $p(\mathbf{X}; \theta) = g(T(\mathbf{X}), \theta)h(\mathbf{X})$, where g is a function depending on $\mathbf{X}$ only through $T(\mathbf{X})$ and h is a function depending only on $\mathbf{X}$, then $T(\mathbf{X})$ is a sufficient statistic for θ. Conversely, if $T(\mathbf{x})$ is a sufficient statistic for θ, then the pdf $p(\mathbf{X}; \theta)$ can be factorized as shown above.

Some examples are given below to illustrate the use of this theorem to find a sufficient statistic.

Example 1.3 (Sufficient Statistics) Let X be Gaussian distributed with unknown mean θ and unit variance, with $X_1, \ldots, X_N$ drawn independently according to this distribution. The pdf of the data is given by

$$f(\mathbf{X}; \theta) = \frac{1}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}\sum_{n=1}^{N}(X_n - \theta)^2\right]$$

Since $\Sigma_{n=1}^{N}(X_n - \theta)^2 = \Sigma_{n=1}^{N} X_n^2 - 2\theta\Sigma_{n=1}^{N} X_n + N\theta^2$, the pdf could be factorized into

$$f(\mathbf{X};\theta)=\underbrace{\frac{1}{(2\pi)^{N/2}}\exp\left[-\frac{1}{2}\left(N\theta^2-2\theta\sum_{n=1}^{N}X_n\right)\right]}_{g(T(\mathbf{X}),\theta)}\underbrace{\exp\left[-\frac{1}{2}\left(\sum_{n=1}^{N}X_n^2\right)\right]}_{h(\mathbf{X})}$$

Hence, $T(\mathbf{X}) = \Sigma_{n=1}^{N} X_n$ is a sufficient statistic for θ because g is a function depending on $\mathbf{X}$ only through T. In fact, any one-to-one mapping of $T(\mathbf{X})$ is also a sufficient statistic for θ.

Example 1.4 (Sufficient Statistics) Let X be Gaussian distributed with zero mean and unknown variance θ, with $X_1, \ldots, X_N$ drawn independently according to this distribution. The pdf of the data is given by

$$f(\mathbf{X};\theta)=\frac{1}{(2\pi\theta)^{N/2}}\exp\left[-\frac{1}{2}\sum_{n=1}^{N}X_n^2\right]=\underbrace{\frac{1}{(2\pi\theta)^{N/2}}\exp\left[-\frac{1}{2}\sum_{n=1}^{N}X_n^2\right]}_{g(T(\mathbf{X}),\theta)}\underbrace{1}_{h(\mathbf{X})}$$

Hence, $T(\mathbf{X}) = \Sigma_{n=1}^{N} X_n^2$ is a sufficient statistic for θ.

1.3.3 Discrete-Time Signal Model—Flat Fading

A digital transmitter can be modeled as a device with input as information bits and continuous-time analog signals matched to the channel or the medium as output. A string of k information bits is passed to a channel encoder (combined with modulator) where redundancy is added to protect the raw information bits. N encoded symbols $\{x_1, \ldots, x_N\}$ (where $x_n \in \mathcal{X}$) are produced at the output of the channel encoder. The N encoded symbols are mapped to the analog signal (modulated) and transmitted out to the channel. The channel input signal $x(t)$ is given by

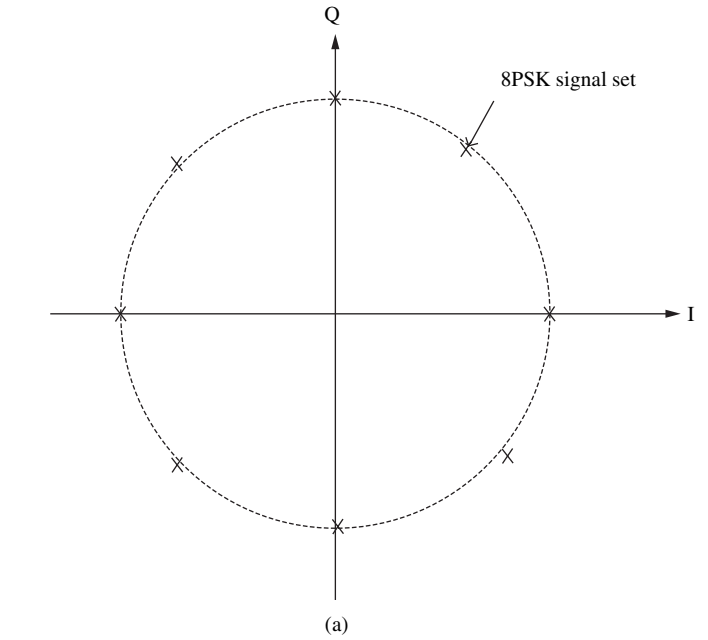
$$x(t)=\sum_{n} x_n g(t-nT_s) \tag{1.32}$$

where T_s is the *symbol duration* and $g(t)$ is the low pass equivalent *transmit pulse* with two-sided bandwidth W and pulse energy $\int_{-\infty}^{\infty}|g(t)|^2dt = 1$. Equation (1.32) is a general model for digitally modulated signals, and the signal set $\mathcal{X}$ is called the *signal constellation*. For example, $\mathcal{X} = \{e^{j2\pi m/M} : m = \{0, 1, \ldots, M-1\}\}$ represents MPSK modulation. $\mathcal{X} = \{x_R + jx_I : x_R, x_I \in \{\pm\frac{1}{2}, \pm\frac{3}{2}, \ldots, \pm\frac{M}{4}\}\}$ represents MQAM modulation. The signal constellations for MPSK and MQAM are illustrated in Figure 1.8.

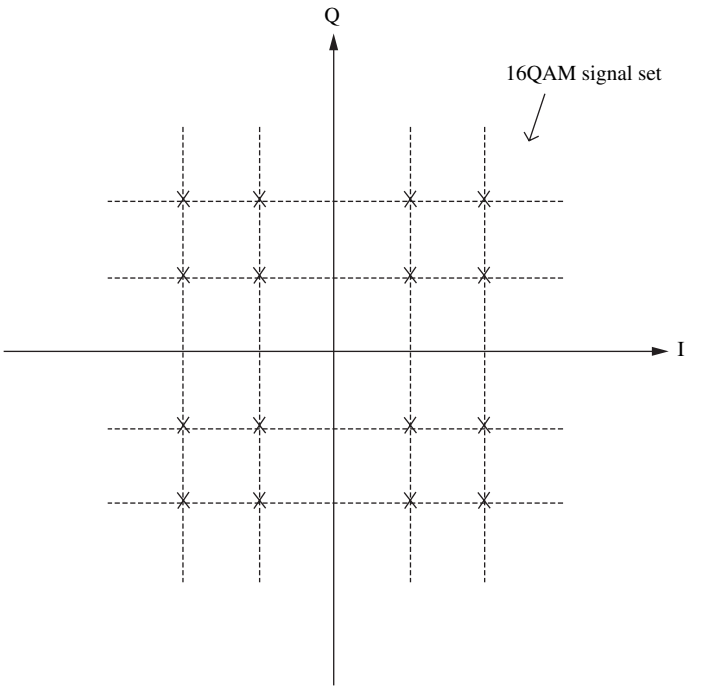
From Equation (1.28), the low pass equivalent received signal through flat fading channel can be expressed as

$$y(t)=h(t)x(t)+z(t)=\sum_{n} x_n h(t)g(t-nT_s)+z(t) \tag{1.33}$$

where $z(t)$ is the low pass equivalent white complex Gaussian noise, $h(t)$ is a zero-mean unit variance complex Gaussian random process, and only $x(t)$ contains information.



(a)



(b)

Figure 1.8. Illustration of signal constellations: (a) MPSK constellation; (b) MQAM constellation.

At the receiver side, the decoder produces the string of k decoded information bits based on the observation $y(t)$ from the channel. Hence, the receiver problem can be modeled as a *detection problem*; that is, given the observation $y(t)$, the receiver has to determine which one of the 2^k *hypothesis* is actually transmitted. It is difficult to gain very useful design insight if we look at the problem from continuous-time domain. In fact, as we have illustrated, the continuous-time signal $y(t)$ can be represent equivalently as vectors in signal space **y**. Hence, the receiver detection problem [assuming knowledge of channel fading $H(t)$] is summarized below.

Problem 1.1 (Detection) Let $\omega \in \{1, 2^k\}$ be the message index at the input of the transmitter. The decoded message $\hat{\omega}$ at the receiver is given by

$$\hat{\omega} = \arg\max_{\omega} p(\mathbf{y}|\omega, h(t))$$

Such a receiver is called the *maximum-likelihood* (ML) receiver, which minimizes the probability of error if all the 2^k messages are equally probable.

Due to the white channel noise term in (1.33), we need infinite dimension signal space in general to represent the received signal $y(t)$ as vector **y**. However, since only $h(t)x(t)$ contains information and $h(t)x(t)$ can be represented as vector in a finite-dimensional signal subspace, not all components in **y** will contain information. Hence, the detection problem in Problem 1.1 could be further simplified, and this is expressed mathematically below.

Let $\ominus_y = \{\Psi_1(t), \ldots, \Psi_n(t), \ldots\}$ denote the infinite-dimensional signal space that contains $y(t)$ and $\ominus_s = \{\Psi_1(t), \ldots, \Psi_{D_s}(t)\}$ denote a D_s-dimensional subspace of $\ominus_y$ that contains $s(t) = h(t)x(t)$; that is

$$y(t) = \sum_{j=1}^{\infty} y_j \Psi_j(t) \tag{1.34}$$

and

$$s(t) = \sum_{j=1}^{D_s} s_j \Psi_j(t) \tag{1.35}$$

where $y_j =< y(t), \Psi_j(t) >= \int_{-\infty}^{\infty} y(t)\Psi_j^*(t)dt$ and $s_j =< s(t), \Psi_j(t) >= \int_{-\infty}^{\infty} s(t)\Psi_j^*(t)dt$.

Let **y** and **s** be the vectors corresponding to $y(t)$ and $s(t) = h(t)x(t)$ with respect to the signal spaces $\ominus_y$ and $\ominus_s$, respectively. Define **v** as the projection of **y** over the signal space $\ominus_s$. Since $p(\mathbf{y}|\omega, h(t)) = p(\mathbf{v}|\omega, h(t))$, **v** forms a *sufficient statistics* for the unknown parameter ω. Therefore, it is sufficient to project the received signal $y(t)$ over the signal space $\ominus_s$, and there is no loss of information with respect to the detection of the message index ω.

The vector representation of $s(t) = h(t)x(t)$ is given by $(s_1, \ldots, s_{D_s})$, where

$$s_j =< h(t)x(t), \Psi_j(t) >= \int_{-\infty}^{\infty} h(t)x(t)\Psi_j^*(t)dt = \sum_n x_n \int_{-\infty}^{\infty} h(t)g(t - nT_s)\Psi_j^*(t)dt$$

In the special case of slow fading where $T_c \gg T_s$, we have $h(t) \approx h_n$ for $t \in [(n-1)T_s, nT_s]$. The vector representation of $s(t)$ becomes

$$\begin{aligned} s_j &= \sum_n x_n \int_{-\infty}^{\infty} h(t) g(t - nT_s) \Psi_j^*(t) dt \\ &= \sum_n x_n h_n \int_{-\infty}^{\infty} g(t - nT_s) \Psi_j^*(t) dt \\ &= \sum_n x_n h_n g_{n,j} \quad \forall j \in [1, D_s] \end{aligned}$$

where $g_{n,j} = \int_{-\infty}^{\infty} g(t - nT_s)\Psi_j^*(t)dt$, $D_s = WNT_s$ and N is the number of transmitted symbols. Therefore, we have $\mathbf{s} = \Sigma_n x_n h_n \mathbf{g}_n$, where $\mathbf{g}_n$ is the vector representation of the time-delayed transmit pulse $g(t - nT_s)$. The equivalent discrete-time flat fading model with channel input $\mathbf{s}$ and channel output $\mathbf{v}$ (the projection of $y(t)$ onto the signal space $\ominus_s$) is therefore given by

$$\mathbf{v} = \mathbf{s} + \mathbf{z} = \sum_n x_n h_n \mathbf{g}_n + \mathbf{z} \tag{1.36}$$

where the vectors are defined with respect to the NWT_s-dimensional signal space $\ominus_x$ that contains $x(t)$. If $z(t)$ is the low pass equivalent white Gaussian noise and the two-sided noise spectral power density of the real and quadrature components are both $\eta_0/2$, the noise covariance of $\mathbf{z}$ is given by

$$\varepsilon[\, z_j z_k^* \,] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varepsilon[z(t)z^*(\tau)] \Psi_j^*(t) \Psi_k(\tau) dt d\tau = \eta_0 \delta(j-k)$$

In other words, we have i.i.d. noise vector components in $\mathbf{z}$, each with variance η_0.

In addition, if the channel noise is a white Gaussian random process, the likelihood function of Problem 1.1 (with knowledge of channel fading at the receiver) can be expressed as

$$p(\mathbf{v}|\omega, \mathbf{h}) = \frac{1}{(\pi\eta_0)^N} \exp\left[-\frac{1}{2\eta_0} |\mathbf{v} - \mathbf{s}(\omega)|^2 \right]$$

Maximizing the likelihood function is equivalent to minimizing the *distance metric* $d(\mathbf{v}, \mathbf{s})$ between the observation $\mathbf{v}$ and the hypothesis $\mathbf{s}(\omega)$. The distance metric is given by: $d(\mathbf{v}, \mathbf{s}) = |\mathbf{v} - \mathbf{s}(\omega)|^2$. This can be further simplified as follows:

$$d(\mathbf{v}, \mathbf{s}) = |\mathbf{v}|^2 - 2\sum_{\mathbf{n}} \mathbf{x_n^*}(\omega)\mathbf{h_n^*}\mathbf{v} \cdot \mathbf{g_n^*} + \left| \sum_{\mathbf{n}} \mathbf{h_n x_n g_n} \right|^2$$

Since $|\mathbf{v}|^2$ is independent of ω, the maximum-likelihood (ML) metric of ω, $\mu(\mathbf{v}, \omega)$, is given by

$$\begin{aligned}\mu(\mathbf{v},\omega) &= 2\sum_n x_n^*(\omega)h_n^*\mathbf{v}\cdot g_n^* - \left|\sum_n h_n x_n g_n\right|^2 \\ &= 2\sum_n x_n^*(\omega)h_n^*\langle y(t), g(t-nT_s)\rangle - \sum_{n,m} h_n h_m^* x_n x_m^* \left\langle g_n, g_m^*\right\rangle \\ &= 2\sum_n x_n^*(\omega)h_n^* \int_{-\infty}^{\infty} y(t)g^*(t-nT_s)dt \\ &\quad - \sum_{n,m} h_n h_m^* x_n x_m^* \int_{-\infty}^{\infty} g(t-nT_s)g^*(t-mT_s)dt \\ &= 2\sum_n x_n^*(\omega)h_n^* q_n - \sum_{n,m} h_n h_m^* x_n x_m^* R_g(n-m) \end{aligned} \tag{1.37}$$

where $R_g(n-m) = \int_{-\infty}^{\infty} g(t-nT_s)g^*(t-mT_s)dt$ and $q_n = \int_{-\infty}^{\infty} y(t)g^*(t-nT_s)dt$ is the matched-filter output (with impulse response $g^*(-t)$) of the received signal $y(t)$ (sampled at $t = nT_s$). Observe that the ML metric $\mu(\mathbf{v}, \omega)$ of ω in Equation (1.37) depends on the received signal $y(t)$ only through $q_n = \int_{-\infty}^{\infty} y(t)g(t-nT_s)dt$. Once $\mathbf{q} = [q_1, \ldots, q_N]$ are known, the ML metric $\mu(\mathbf{v}, \omega)$ can be computed. Hence, the matched-filter outputs, $q_n = \int_{-\infty}^{\infty} y(t)g(t-nT_s)dt$, are also the *sufficient statistics* with respect to ω. Therefore, the ML metric $\mu(\mathbf{v}, \omega)$ in Equation (1.37) can be expressed in terms of the *sufficient statistics* $\mathbf{q} = [q_1, \ldots, q_N]$ directly as $\mu(\mathbf{q}, \omega)$.

If $g(t) = 0$ for all $t > T_s$ or $t < 0$ [*zero intersymbol interference* (ISI) due to the transmit pulseshape $g(t)$], we have $\int_{-\infty}^{\infty} g(t-nT_s)g^*(t-mT_s)dt = \delta(n-m)$ and the ML metric $\mu(\mathbf{q}, \omega)$ can be further simplified as

$$\mu(\mathbf{q},\omega) = \sum_n x_n^*(\omega)h_n^* q_n - \sum_n |h_n|^2 |x_n(\omega)|^2 \tag{1.38}$$

Figure 1.9 illustrates the optimal detector structure based on *matched filtering*.

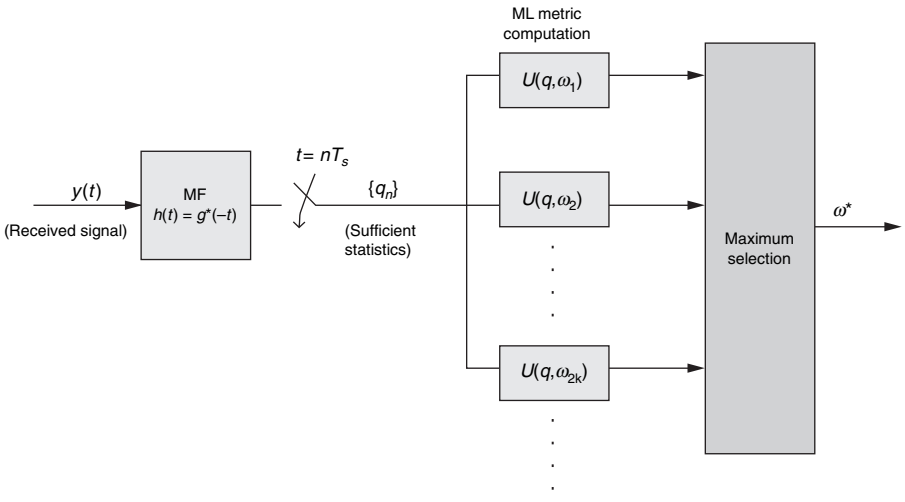


Figure 1.9. Optimal detector structure based on matched filtering for fading channels with zero ISI transmit pulse.

The sufficient statistics $\mathbf{q}$ are first generated from $y(t)$ through matched filtering and periodically sampling at $t = nT_s$. The sequence of sufficient statistics is passed to the ML metric computation, where the ML metrics $\mu(\mathbf{q}, \omega)$ for each hypothesis ω are computed. Hence, the hypothesis that maximizes the metrics is the decoded information $\hat{\omega}$. Finally, the following theorem establish the equivalence of the continuous-time flat fading channel and the discrete-time flat fading channels.

Theorem 1.3 (Equivalence of Continuous-Time and Discrete-Time Models) Given any continuous-time flat fading channels with the transmitted signal $x(t)$ given by Equation (1.32) and the received signal $y(t)$ given by Equation (1.33), there always exists an equivalent discrete-time channel with input x_n and output q_n such that

$$q_n = \sum_k h_k x_k R_g(n-k) + z_n \tag{1.39}$$

where $R_g(n-k) = \int_{-\infty}^{\infty} g(t-kT_s)g^*(t-nT_s)dt$, z_n is a zero-mean Gaussian noise with variance η_0, and η_0 is the power spectral density of $z(t)$. The term *equivalent channels* means that both channels will have the same ML detection results with respect to the message index ω. Furthermore, if $g(t) = 0$ for all $t > T_s$ or $t < 0$ such that

$$q_n = h_n x_n + z_n \tag{1.40}$$

Proof Since q_n is a sufficient statistic with respect to the detection of ω, there is no loss of information if the receiver can only observe q_n instead of the complete received signal $y(t)$. Hence, substituting $y(t)$ in Equation (1.33) into q_n, we have

$$\begin{aligned} q_n &= \sum_k h_k x_k \int_{-\infty}^{\infty} g(t-kT_s)g^*(t-nT_s)dt + \int_{-\infty}^{\infty} z(t)g^*(t-nT_s)dt \\ &= \sum_k h_k x_k R_g(n-k) + z_n \end{aligned}$$

where $R_g(n-k) = \int_{-\infty}^{\infty} g(t-kT_s)g^*(t-nT_s)dt$ and $z_n = \int_{-\infty}^{\infty} z(t)g^*(t-nT_s)dt$. Since the message index ω is contained in the symbols $[x_1, \ldots, x_N]$ only, there is no loss of information if the receiver only observes q_n due to sufficient statistic properties. Hence, the two channels produce the same ML detection results. If the pulse duration of $g(t)$ is T_s, then $g_{n-k} = \delta(n-k)$ and Equation (1.40) follows immediately.

Extending Equation (1.40) to MIMO channels, let $\mathbf{X} \in \mathcal{X}$ be the $(n_T \times 1)$-dimensional input vector, $\mathbf{Y} \in \mathcal{Y}$ be the $n_R \times 1$ output vector, and $\mathbf{H} \in \mathcal{H}$ be the $(n_R \times n_T)$-dimensional channel fading of the discrete-time MIMO channel. The received symbol $\mathbf{y}_n$ is given by

$$\mathbf{y}_n = \mathbf{h}_n \mathbf{x}_n + \mathbf{z}_n \tag{1.41}$$

1.3.4 Discrete-Time Channel Model—Frequency-Selective Fading

The transmitted signal $x(t)$ is given by

$$x(t) = \sum_n x_n g(t - nT_s) \tag{1.42}$$

where T_s is the *symbol duration* and $g(t)$ is the low pass equivalent *transmit pulse* with two-sided bandwidth W and pulse energy $\int_{-\infty}^{\infty} |g(t)|^2 dt = 1$. From Equation (1.31), the received signal after passing through a frequency-selective fading channel is given by

$$y(t) = \sum_{l=0}^{L-1} h_l(t) x(t - lT_s) + z(t) \tag{1.43}$$

where $L = \lfloor \frac{W}{B_c} \rfloor$ is the number of resolvable paths. Letting $s(t) = \Sigma_{l=0}^{L-1} h_l(t) x(t - lT_s)$, using similar arguments, the vector projection of $\mathbf{y}$ onto the signal space containing $s(t)$, $\mathbf{v}$, is a sufficient statistic for the message index ω.

The vector representation of $s(t)$ is given by $(s_1, \ldots, s_{D_s})$, where

$$\begin{aligned} s_j &= < s(t), \Psi_j(t) > = \sum_{l=0}^{L-1} \int_{-\infty}^{\infty} h_l(t) x(t - lT_s) \Psi_j^*(t) dt \\ &= \sum_{l=0}^{L-1} \sum_n x_n \int_{-\infty}^{\infty} h_l(t) g(t - (n+l)T_s) \Psi_j^*(t) dt \end{aligned}$$

In the special case of slow fading where $T_c \gg T_s$, we have $h_l(t) \approx h_{m,l}$ for $t \in [(m-1)T_s, mT_s]$. Hence, the vector representation of $s(t)$ becomes

$$\begin{aligned} s_j &= \sum_{l=0}^{L-1} \sum_n x_n \int_{-\infty}^{\infty} h_l(t) g(t - (n+l)T_s) \Psi_j^*(t) dt \\ &= \sum_{l=0}^{L-1} \sum_n x_n h_{n+l,l} \int_{-\infty}^{\infty} g(t - (n+l)T_s) \Psi_j^*(t) dt \\ &= \sum_{l=0}^{L-1} \sum_n x_n h_{n+l,l} g_{n+l,j} \end{aligned}$$

where $g_{m,j}$ is the component of the projection of $g(t - mT_s)$ onto $\Psi_j(t)$ given by

$$g_{m,j} = \int_{-\infty}^{\infty} g(t - mT_s) \Psi_j^*(t) dt \tag{1.44}$$

Therefore, we have $\mathbf{s} = \Sigma_{l=0}^{L-1} \Sigma_n x_n h_{n+l,l} \mathbf{g}_{n+l}$. The equivalent discrete-time frequency-selective fading model with input vector $\mathbf{s}$ and output vector $\mathbf{v}$ is therefore given by

$$\mathbf{v} = \mathbf{s} + \mathbf{z} = \sum_{l=0}^{L-1} \sum_{n} x_n h_{n+l,l} g_{n+l} + \mathbf{z} \tag{1.45}$$

where the vectors are defined with respect to the signal space that contains $s(t)$. Similarly, since $z(t)$ is the low pass equivalent white Gaussian noise with two-sided power spectral density of the real and quadrature components both given by $\eta_0/2$, we have i.i.d. noise vector components in $\mathbf{z}$ with noise variance η_0.

When the channel noise is white Gaussian, the likelihood function is given by

$$p(\mathbf{v}|\omega, \mathbf{h}) = \frac{1}{(\pi\eta_0)^N} \exp\left[-\frac{1}{2\eta_0} |\mathbf{v} - \mathbf{s}|^2 \right]$$

Factorizing the likelihood function according to Theorem 1.2, we have the ML metric $\mu(\mathbf{v}, \omega)$ given by

$$\begin{aligned}
\mu(\mathbf{r}, \omega) &= 2 \sum_n \sum_{l=0}^{L-1} x_n^*(\omega) h_{n+l,l}^* \mathbf{v} \cdot \mathbf{g}_{n+l}^* - \left| \sum_{l=0}^{L-1} \sum_n x_n h_{n+l,l} \mathbf{g}_{n+l} \right|^2 \\
&= 2 \sum_n \sum_{l=0}^{L-1} h_{n+l,l}^* x_n^* \langle y(t), g(t-(n+l)T_s) \rangle \\
&\quad - \sum_{n,n'} \sum_{l,l'} h_{n+l,l} h_{n'+l',l'}^* x_n x_{n'}^* \langle g(t-(n+l)T_s), g(t-(n'+l')T_s) \rangle \\
&= 2 \sum_n \sum_{l=0}^{L-1} h_{n+l,l}^* x_n^* \int_{-\infty}^{\infty} y(t) g^*(t-(n+l)T_s) dt \\
&\quad - \sum_{n,n'} \sum_{l,l'} h_{n,l} h_{n',l'}^* x_n x_{n'}^* \int_{-\infty}^{\infty} g(t-(n+l)T_s) g^*(t-(n'+l')T_s) dt \\
&= 2 \sum_n \sum_{l=0}^{L-1} h_{n+l,l}^* x_n^* q_{n+l} - \sum_{n,n'} \sum_{l,l'} h_{n,l} h_{n',l'}^* x_n x_{n'}^* R_g(n+l-n'-l')
\end{aligned}$$

where

$$q_n = \int_{-\infty}^{\infty} y(t) g^*(t - nT_s) dt \tag{1.46}$$

is the matched-filter output [with impulse response $g^*(-t)$] of the received signal $y(t)$ sampled at $t = nT_s$. Observe that the ML metric depends on the received signal $y(t)$ only through the matched-filter outputs $\mathbf{q} = [q_1, \ldots, q_N]$. Once $\mathbf{q}$ is known, the ML metric $\mu(\mathbf{v}, \omega)$ can be computed and hence, the ML detection of $\hat{\omega}$ can be obtained directly from $\mathbf{q}$. In other words, there is no loss of information if the receiver can only observe $\mathbf{q}$ instead of the complete received signal $y(t)$. $\mathbf{q}$ is therefore another sufficient statistics with respect to the detection of the message ω. If $g(t) = 0$ for all $t > T_s$ or $t < 0$ [*zero ISI* due to the transmit pulse $g(t)$], the ML metric can be further simplified as

$$\mu(\mathbf{v}, \omega) = \sum_n \sum_{l=0}^{L-1} x_n^*(\omega) h_{n+l,l}^* q_{n+l} - \sum_{n,n'} \sum_{l,l',n+l=n'+l'} h_{n+l,l} h_{n'+l',l'}^* x_n x_{n'}^*. \tag{1.47}$$

Figure 1.9 illustrates the optimal detector structure for frequency-selective fading channel with zero ISI pulse based on matched filtering.

Similar to the flat fading case, we can establish a discrete-time equivalent channel with any given continuous-time frequency-selective fading channels and summarize the results by the following theorem.

Theorem 1.4 (Equivalence of Continuous-Time and Discrete-Time Models) For any continuous-time frequency-selective fading channels with input $x(t)$ given by Equation (1.42) and received signal $y(t)$ given by Equation (1.43), there always exists an equivalent discrete-time frequency-selective fading channel with discrete-time inputs $\{x_n\}$ and discrete-time outputs $\{q_n\}$ such that

$$q_n = \sum_k \sum_{l=0}^{L-1} h_{k+l,l} x_k R_g(n-k-l) + z_n \tag{1.48}$$

where $R_g(m) = \int_{-\infty}^{\infty} g(t)g^*(\mathrm{t} - mT_s)dt$, $h_{m,l} = h_l(t)$ for $t \in [mT_s, (m+1)T_s]$, z_n is a zero-mean Gaussian noise with variance η_0, and η_0 is the power spectral density of $z(t)$. If $g(t) = 0$ for all $t > T_s$ or $t < 0$ (zero ISI), we have $R_g(m) = \delta(mT_s)$. Hence

$$q_n = \sum_{l=0}^{L-1} h_{n,n-l} x_{n-l} + z_n \tag{1.49}$$

Proof Similar to the flat fading case in Theorem 1.3, $\mathbf{q}$ is a sufficient statistic with respect to the detection of ω and there is no loss of information if the receiver can observe $\mathbf{q}$ only. Substituting $y(t)$ from Equation (1.43) into q_n, the results follow.

Extending Equation (1.49) to MIMO channels, the received symbol $\mathbf{y}_n$ is given by

$$\mathbf{y}_n = \sum_{l=0}^{L-1} \mathbf{h}_{n,n-l} \mathbf{x}_{n-1} + \mathbf{z}_n \tag{1.50}$$

1.4 FUNDAMENTALS OF INFORMATION THEORY

In this section, we give an overview on the background and mathematical concepts in information theory in order to establish Shannon's channel coding theorem. Concepts of ergodic and outage capacities will be followed. Finally, we give several examples of channel capacity in various channels. This will form the basis for the remaining chapters in the book.

1.4.1 Entropy and Mutual Information

The first important concept in information theory is *entropy*, which is a measure of uncertainty in a random variable.

Definition 1.2 (Entropy of Discrete Random Variable) The entropy of a discrete random variable X with probability mass function $p(X)$ is given by

$$H(X) = -\sum_{x} p(x)\log_2(p(x)) = -\varepsilon[\log_2(p(X))] \tag{1.51}$$

where the expectation is taken over X.

Definition 1.3 (Entropy of Continuous Random Variable) On the other hand, the entropy of a continuous random variable X with probability density function $f(X)$ is given by

$$H(X) = -\int_{x} f(x)\log_2(f(x))dx \tag{1.52}$$

For simplicity, we shall assume discrete random variable unless otherwise specified.

Definition 1.4 (Joint Entropy) The joint entropy of two random variables X_1, X_2 is defined as

$$H(X_1, X_2) = -\sum_{x_1, x_2} p(x_1, x_2)\log_2(p(x_1, x_2)) = -\varepsilon[\log_2(p(X_1, X_2))] \tag{1.53}$$

where the expectation is taken over (X_1, X_2).

Definition 1.5 (Conditional Entropy) The conditional entropy of a random variable X_2 given X_1 is defined as

$$H(X_2|X_1) = -\sum_{x_1, x_2} p(x_1, x_2)\log_2(p(x_2|x_1)) = -\varepsilon[\log_2(p(X_2|X_1))] \tag{1.54}$$

where the expectation is taken over (X_1, X_2).

After introducing the definitions of entropy, let's look at various properties of entropy. They are summarized as lemmas below. Please refer to the text by Cover and Thomas [30] for the proof.

The first lemma gives a lower bound on entropy.

Lemma 1.2 (Lower Bound of Entropy)

$$H(X) \geq 0 \tag{1.55}$$

Equality holds if and only if there exists $x_0 \in \mathcal{X}$ such that $p(x_0) = 1$.

Proof Directly obtained from Equation (1.51).

The following lemma gives an upper bound on entropy for discrete and continuous random variable X.

Lemma 1.3 (Upper Bound of Entropy) If X is a discrete random variable, then

$$H(X) \leq \log_2(|X|) \tag{1.56}$$

Equality holds if and only if $p(X) = 1/|X|$. On the other hand, if X is a continuous random variable, then

$$H(X) \leq \frac{1}{2}\log_2(2\pi e\sigma_X^2) \tag{1.57}$$

where $\sigma_X^2 = \varepsilon[|X|^2]$ and equality holds if and only if X is Gaussian distributed with an arbitrary mean μ and variance σ_X^2.

From the lemmas presented-above, entropy can be interpreted as a measure of *information* because $H(X) = 0$ if there is no uncertainty about X. On the other hand, $H(X)$ is maximized if X is equiprobable or X is Gaussian distributed.

The chain rule of entropy is given by the following lemma.

Lemma 1.4 (Chain Rule of Entropy)

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1) \tag{1.58}$$

On the other hand, as summarized in the lemma below, conditioning reduces entropy.

Lemma 1.5 (Conditioning Reduces Entropy)

$$H(X|Y) \leq H(X) \tag{1.59}$$

Equality holds if and only if $p(XY) = p(X)p(Y)$ (X and Y are independent).

Lemma 1.6 (Concavity of Entropy) $H(X)$ is a concave function of $p(X)$.

Lemma 1.7 If X and Y are independent, then $H(X + Y) \geq H(X)$.

Lemma 1.8 (Fano's Inequality) Given two random variables X and Y, let $\hat{X} = g(Y)$ be an estimate of X given Y. Define the probability of error as

$$P_e = \Pr\left[\hat{X} \neq X\right]$$

we have

$$H_2(P_e) + P_e \log_2(|X| - 1) \geq H(X|Y) \tag{1.60}$$

where $H_2(p) = -p\log_2(p) - (1-p)\log_2(1-p)$.

After we have reviewed the definitions and properties of entropy, we shall discuss the concept of *mutual information*.

Definition 1.6 (Mutual Information) The mutual information of two random variables X and Y is given by

$$I(X;Y) = \sum_{x,y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) = \varepsilon\left[\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)\right] \tag{1.61}$$

From Equation (1.53), we have

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \tag{1.62}$$

Figure 1.10 illustrates the mutual information.

Hence, mutual information is the reduction in uncertainty of X, due to knowledge of Y. Therefore, if X is the transmitted symbol and Y is the received

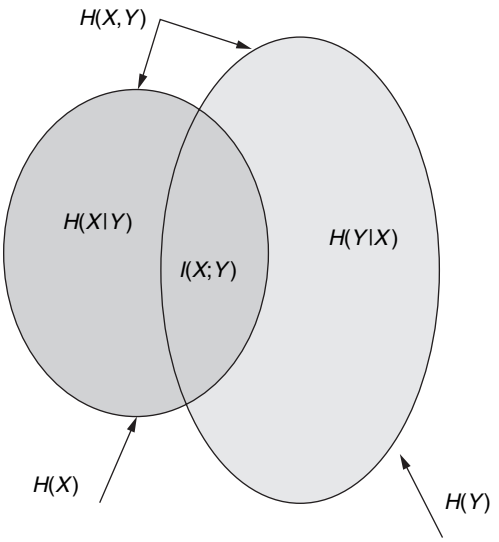


Figure 1.10. Illustration of mutual information.

symbol, mutual information could be interpreted as a measure of the amount of *information* communicated to the receiver. We shall formally establish this interpretation in the next section when we discuss Shannon's channel coding theorem.

Definition 1.7 (Conditional Mutual Information) The mutual information of 2 random variables X and Y conditioned on Z is given by

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z) \tag{1.63}$$

We shall summarize the properties of mutual information below.

Lemma 1.9 (Symmetry)

$$I(X;Y) = I(Y;X)$$

The proof can be obtained directly from definition.

Lemma 1.10 (Self-Information)

$$I(X;X) = H(X)$$

The proof can be obtained directly from the definition.

Lemma 1.11 (Lower Bound)

$$I(X;Y) \geq 0$$

Equality holds if and only if X and Y are independent.

Lemma 1.12 (Chain Rule of Mutual Information)

$$I(X_1,\ldots,X_N;Y) = \sum_{n=1}^{N} I(X_n;Y|X_1,\ldots,X_{n-1})$$

Lemma 1.13 (Convexity of Mutual Information) Let (X, Y) be the random variable with the joint pdf given by $p(X)p(Y|X)$. $I(X; Y)$ is a concave function of $p(X)$ for a given $p(Y|X)$. On the other hand, $I(X; Y)$ is a convex function of $p(Y|X)$ for a given $p(X)$.

Lemma 1.14 (Data Processing Inequality) X, Y, Z are said to form a Markov chain $X \to Y \to Z$ if $p(x, y, z) = p(x)p(y|x)p(z|y)$. If $X \to Y \to Z$, we have

$$I(X;Y) \geq I(X;Z)$$

Equality holds if and only if $X \to Z \to Y$.

Corollary 1.1 (Preprocessing at the Receiver) If $Z = g(Y)$, then $X \to Y \to Z$ and we have $I(X; Y|Z) \leq I(X; Y)$ and $I(X; Z) \leq I(X; Y)$. Equality holds in both cases if and only if Y, X are independently conditioned on $g(Y)$.

If we assume that X is the channel input and Y is the channel output, then the function of received data Y cannot increase the mutual information about X. On the other hand, the dependence of X and Y is decreased by the observation of a downstream variable $Z = g(Y)$. Since $I(X; Y)$ gives the channel capacity as we shall introduced in the next section, Corollary 1.1 states that the channel capacity will not be increased by any *preprocessing* at the receiver input.

Corollary 1.2 (Postprocessing at the Transmitter) If $Y = g(X)$, then $X \to Y \to Z$ and we have $I(X; Z) \leq I(Y; Z)$. Equality holds in both cases if and only if X, Z are independently conditioned on $g(X)$.

Similarly, any *postprocessing* at the transmitter cannot increase the channel capacity as illustrated by Corollary 1.2. These points are illustrated in Figure 1.11.

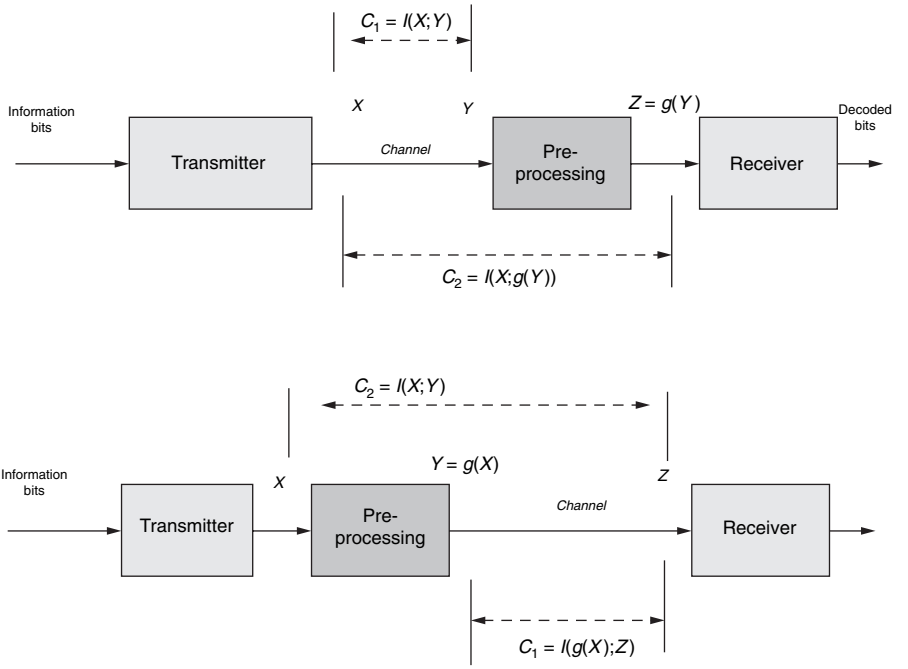


Figure 1.11. Illustration of data processing inequality. Preprocessing at the receiver and postprocessing at the transmitter cannot increase mutual information; $C_1 \geq C_2$.

1.4.2 Shannon's Channel Coding Theorem

The Shannon's channel coding theorem is in fact based on the *law of large numbers*. Before we proceed to establish the coding theorem, we shall discuss several important mathematical concepts and tools.

Convergence of Random Sequence

Definition 1.8 (Random Sequence) A *random sequence* or a *discrete-time random process* is a sequence of random variables $X_1(\xi), \ldots, X_n(\xi), \ldots$, where ξ denotes an outcome of the underlying sample space of the random experiment.

For a specific event ξ, $X_n(\xi)$ is a sequence of of numbers that might or might not converge. Hence, the notion of convergence of a random sequence might have several interpretations:

Convergence Everywhere. A random sequence $\{X_n\}$ converges everywhere if the sequence of numbers $X_n(\xi)$ converges in the traditional sense[4] for every event ξ. In other words, the limit of the random sequence is a random variable $X(\xi)$, which is $X_n(\xi) \to X(\xi)$ as $n \to \infty$.

Convergence Almost Everywhere. Define $\mathcal{V} = \{\xi : \lim_{n\to\infty} x_n(\xi) = x(\xi)\}$ be the set of events such that the number sequence $x_n(\xi)$ converges to $x(\xi)$ as $n \to \infty$. The random sequence x_n is said to converge almost everywhere (or with probability 1) to x if there exists such a set $\mathcal{V}$ such that $\Pr[\mathcal{V}] = 1$ as $n \to \infty$.

Convergence in the Mean-Square Sense. The random sequence x_n tends to the random variable x in the mean-square sense if $\mathcal{E}[|x_n - x|^2] \to 0$ as $n \to \infty$.

Convergence in Probability. The random sequence x_n converges to the random variable x in probability if for any $\varepsilon > 0$, the probability $\Pr[|x_n - x| > \varepsilon] \to 0$ as $n \to \infty$.

Convergence in Distribution. The random sequence x_n is said to converge to a random variable x in distribution if the cumulative distribution function (cdf) of $x_n(F_n(\alpha))$ converges to the cdf of $x(F(\alpha))$ as $n \to \infty$

$$F_n(\alpha) \to F(\alpha) \quad n \to \infty$$

for every point α of the range of x. Note that in this case, the sequence $x_n(\xi)$ needs not converge for any event ξ.

Figure 1.12 illustrates the relationship between various convergence modes. For instance, "almost-everywhere convergence" implies convergence in prob-

[4] A sequence of number x_n tends to a limit x if, given any $\varepsilon > 0$, there exists a number n_0 such that $|x_n - x| < \varepsilon$ for every $n > n_0$.

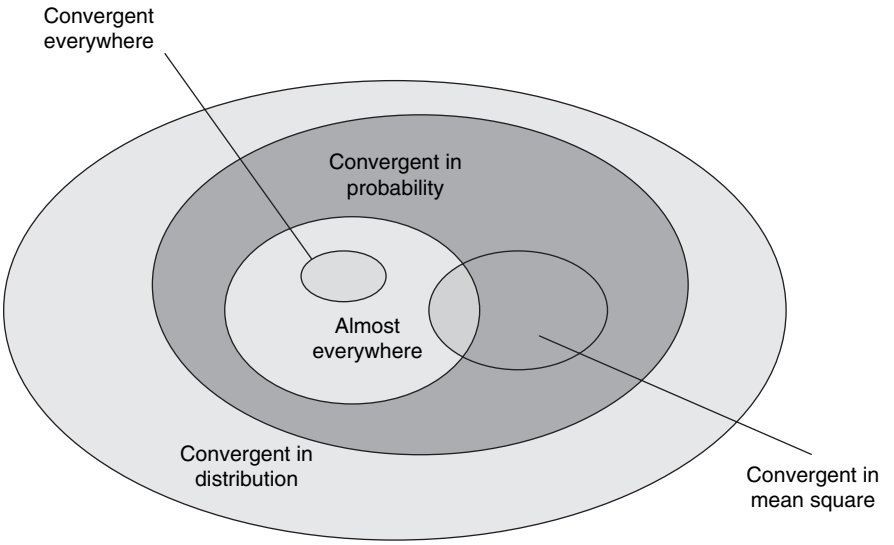


Figure 1.12. Various modes of random sequence convergence.

ability but not the converse. Hence, almost-everywhere convergence is a stronger condition than convergence in probability. Obviously, convergence everywhere implies convergence almost everywhere but not the converse. Hence, convergence everywhere is an even stronger condition compared with almost everywhere convergence. Another example is convergence in mean-square sense implies convergence in probability because

$$\Pr[|x_n - x| > \varepsilon] \le \frac{\varepsilon\left[|x_n - x|^2\right]}{\varepsilon^2}$$

as a result of Chebyshev's inequality.[5] Hence, if $x_n \to x$ in mean-square sense, then for any fixed $\varepsilon > 0$, the right side tends to 0 as $n \to \infty$. Again, the converse is not always true. Note that convergence in mean-square sense does not always imply convergence almost everywhere and vice versa.

Law of Large Numbers. The law of large numbers is the fundamental magic behind information theory.

Weak Law of Large Numbers. If $X_1, \ldots, X_n$ are i.i.d. random variables, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_n X_n = \varepsilon|X| \text{ in probability}$$

[5] For a random variable X with mean μ_X and variance σ_X^2, the probability $\Pr[|X - \mu_X| > x_0]$ is upper-bounded by $\frac{\sigma_X^2}{\sigma_0^2}$ for all $x_0 > 0$.

The expression on the left side is also called the *sample mean*.

Strong Law of Large Numbers. If $X_1, \ldots, X_n$ are i.i.d. random variables, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_n X_n = \varepsilon|X|$$

with probability 1 (almost everywhere).

We shall explain the difference between the two versions of law of large numbers by an example below.

Suppose that the random variable X_n has two possible outcomes {0, 1} and the probability of the event $X_n = 1$ in a given experiment equals p and we wish to estimate p to within an error $\varepsilon = 0.1$ using the sample mean $\bar{x}(n) = (1/n)\Sigma_n x_n$ as its estimate. If $n \geq 1000$, then $\Pr[|\bar{x}(n) - p| < 0.1] \geq 1 - 1/(4n\varepsilon^2) \geq \frac{39}{40}$. Hence, if we repeat the experiment 1000 times, then in 39 out of 40 such runs, our error $|\bar{x}(n) - p|$ will be less than 0.1.

Suppose now that we perform the experiment 2000 times and we determine the sample mean $\bar{x}(n)$ not for one n but for every n between 1000 and 2000. The weak law of large numbers leads to the following conclusion. If our experiment is repeated a large number of times, then for a specific n larger than 1000, the error $|\bar{x}(n) - p|$ will exceed 0.1 only in one run out of 40. Hence, 97.5% of the runs will be "good." We cannot draw the conclusion that in the good runs, the error will be less than 0.1 for every n between 1000 and 2000. This conclusion, however, is correct but can be deduced only from the strong law of large numbers.

Typical Sequence. As a result of the weak law of large numbers, we have the following corollary.

Corollary 1.3 (Typical Sequence) If $X_1, \ldots, X_n$ are i.i.d. random variables with distribution $p(X)$, then

$$\frac{1}{n}\log_2(p(X_1, \ldots, X_n)) \to H(X)$$

in probability.

The proof follows directly by recognizing $(1/n)\log_2(p(X_1, \ldots, X_n)) = (1/n)\Sigma_n \log_2 p(X_n)$, which converges to $\varepsilon[\log_2(p(X))]$ in probability. This is called *asymptotic equipartition* (AEP) *theorem*. As a result of the AEP theorem, we define a typical sequence as follows.

Definition 1.9 (Typical Sequence) A random sequence $(x_1, \ldots, x_N)$ is called a *typical sequence* if $2^{-N(H(X)+\varepsilon)} \leq p(x_1, \ldots, x_N) \leq 2^{-N(H(X)-\varepsilon)}$. The set of typical sequences is called the typical set and is denoted by

$$\mathcal{A}_\varepsilon = \{(x_1, \ldots, x_N) : 2^{-N(H(X)+\varepsilon)} \le p(x_1, \ldots, x_N) \le 2^{-N(H(X)-\varepsilon)}\}$$

Together with the AEP theorem, we have the following properties concerning typical set $\mathcal{A}_\varepsilon$. Please refer to the text by Cover and Thomas [30] for the proof of the lemmas below.

Lemma 1.15 If $(x_1, \ldots, x_N) \in \mathcal{A}_\varepsilon$, then

$$H(X) - \varepsilon \le -\frac{1}{N}\log_2 p(x_1, \ldots, x_N) \le H(X) + \varepsilon$$

The proof follows directly from the definition of typical set $\mathcal{A}_\varepsilon$.

Lemma 1.16

$$\Pr[(x_1, \ldots, x_N) \in \mathcal{A}_\varepsilon] > 1 - \varepsilon$$

for sufficiently large N.

Lemma 1.17

$$|\mathcal{A}_\varepsilon| \le 2^{N(H(X)+\varepsilon)}$$

for sufficiently large N.

Lemma 1.18

$$|\mathcal{A}_\varepsilon| \ge (1-\varepsilon)2^{N(H(X)-\varepsilon)}$$

for sufficiently large N.

Figure 1.13 illustrates the concepts of typical set. As a result of AEP and Lemma 1.16, the probability of a sequence $(x_1, \ldots, x_N)$ being a member of the typical set $\mathcal{A}_\varepsilon$ is arbitrarily high for sufficiently large N. In other words, the typical set contains all *highly probable sequences*. Suppose that X_n is a binary random variable; then the total number of possible combinations in a sequence $(x_1, \ldots, x_N)$ is 2^N. However, as a result of Lemmas 1.17 and 1.18, the cardinality of $\mathcal{A}_\varepsilon$ is around $2^{NH(X)}$. Since $H(X) \le 1$, we have $|\mathcal{A}_\varepsilon| \le 2^N$. This means that the typical set is in general smaller than the set of all possible sequences, resulting in possible *data compression*. The key to achieve effective compression is to group the symbols into a long sequence to exploit the law of large numbers.

Jointly Typical Sequence. We can extend the concept of typical sequence to *jointly typical sequence*.

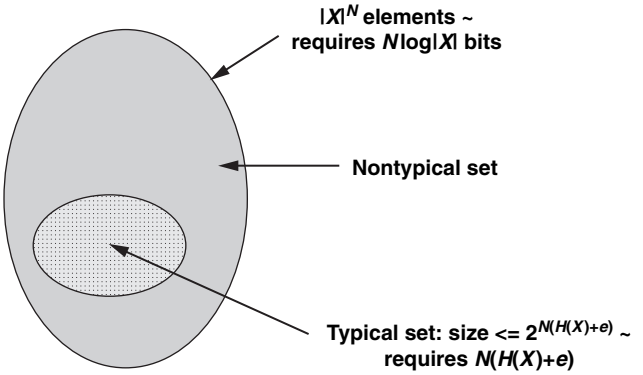


Figure 1.13. Illustration of a typical set.

Definition 1.10 (Jointly Typical Sequence) Let $\mathbf{x}^N$ and $\mathbf{y}^N$ be two sequences of length N generated according to $p(\mathbf{x}^N, \mathbf{y}^N)$. $(\mathbf{x}^N, \mathbf{y}^N)$ is called a *jointly typical sequence* if

$$\left| -\frac{1}{N}\log_2(p(\mathbf{x}^N)) - H(X) \right| < \varepsilon$$
$$\left| -\frac{1}{N}\log_2(p(\mathbf{y}^N)) - H(Y) \right| < \varepsilon$$

and

$$\left| -\frac{1}{N}\log_2(p(\mathbf{x}^N, \mathbf{y}^N)) - H(X,Y) \right| < \varepsilon$$

The jointly typical set $\mathcal{A}_\varepsilon$ is defined as

$$\mathcal{A}_\varepsilon = \{(\mathbf{x}^N, \mathbf{y}^N): \text{the preceding three conditions are satisfied.}\}$$

Similarly, we have the following interesting properties about jointly typical sequences.

Lemma 1.19 Let $\mathbf{x}^N$ and $\mathbf{y}^N$ be two sequences of length N drawn i.i.d. according to $p(\mathbf{x}^N, \mathbf{y}^N) = \Pi_{n=1}^N p(x_n, y_n)$. For any $\varepsilon > 0$, we have

$$\Pr\{(\mathbf{x}^N, \mathbf{y}^N) \in \mathcal{A}_\varepsilon\} \to 1 \text{ as } N \to \infty \tag{1.64}$$

and

$$|\mathcal{A}_\varepsilon| \le 2^{NH(X,Y)+\varepsilon} \text{ for some sufficiently large } N \tag{1.65}$$

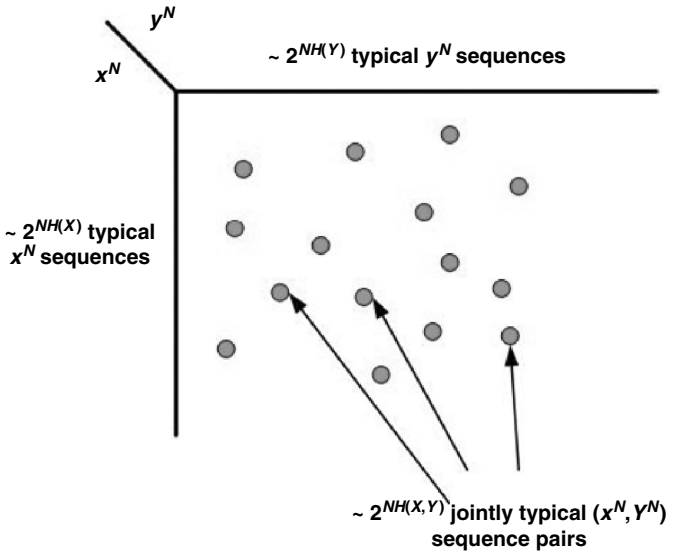


Figure 1.14. Illustration of a jointly typical set.

Furthermore, if $(\tilde{\mathbf{x}}^N, \tilde{\mathbf{y}}^N)$ are two sequences drawn i.i.d. according to $p(\tilde{\mathbf{x}}^N, \tilde{\mathbf{y}}^N) = p(\tilde{\mathbf{x}}^N)p(\tilde{\mathbf{y}}^N)$ (i.e., the distributions of $\tilde{\mathbf{X}}^N$ and $\tilde{\mathbf{Y}}^N$ have the same marginal distribution as $\mathbf{X}^N$ and $\mathbf{Y}^N$ but $\tilde{\mathbf{X}}^N$ and $\tilde{\mathbf{Y}}^N$ are independent), then

$$(1-\varepsilon)2^{-N(I(X;Y)+3\varepsilon)} \le \Pr\{(\tilde{\mathbf{x}}^N, \tilde{\mathbf{y}}^N) \in \mathcal{A}_\varepsilon\} \le 2^{-N(I(X;Y)-3\varepsilon)} \tag{1.66}$$

for sufficiently large N.

Hence, from these properties regarding typical sequences, we can see that when N is sufficiently large, there are about $2^{NH(X)}$ typical X sequences and about $2^{NH(Y)}$ typical Y sequences in the typical sets. However, not all pairs of typical X sequence and typical Y sequence are *jointly typical* as illustrated in Figure 1.14. There are about $2^{NH(X,Y)}$ such *jointly typical sequences* between X and Y. If $\mathbf{x}^N$ and $\mathbf{y}^N$ are naturally related according to their joint density function $p(X, Y)$, then there is a high chance that $(\mathbf{x}^N, \mathbf{y}^N)$ is in the *jointly typical set* for sufficiently large N. However, if $\mathbf{x}^N$ and $\mathbf{y}^N$ are randomly picked from their respective typical sets, the chance for $(\mathbf{x}^N, \mathbf{y}^N)$ in the *jointly typical set* is very small. This serves as an important property that we shall utilize to prove the channel coding theorem in the next section.

Channel Coding Theorem. We have now introduced the essential mathematical tools to establish the channel coding theorem. Before we proceed, here is a summary of the definitions of *channel encoder*, *generic channel*, *channel decoder*, and *error probability* in discrete-time domain.

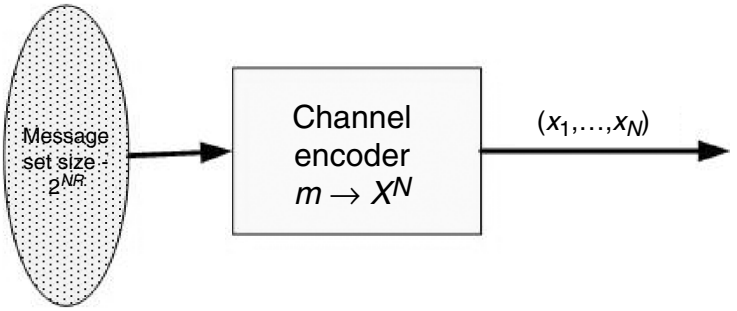


Figure 1.15. Channel encoder.

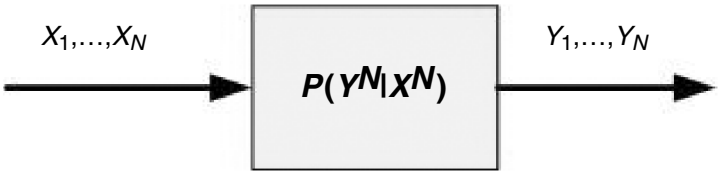


Figure 1.16. Generic channel.

Definition 1.11 (Channel Encoder) The *channel encoder* is a device that maps a message index $m \in \{1, 2, \ldots, M = 2^{NR}\}$ to the channel input $\mathbf{X}^N = [X_1, \ldots, X_N]$ *as* illustrated in Figure 1.15.

Hence, the transmitter has to use the channel N times to deliver the message m through the transmit symbols $X_1, \ldots, X_N$ and is specified by a function mapping $f: f: m \in \{1, 2, \ldots, M = 2^{NR}\} \rightarrow \mathcal{X}^N$. Alternatively, the channel encoder could be characterized by a codebook $\{\mathbf{X}^N(1), \mathbf{X}^N(2), \ldots, \mathbf{X}^N(M)\}$ with M elements. When the current message index is m, the channel encoder output will be $f(m) = \mathbf{X}^N(m)$. The encoding rate of the channel encoder is the number of information bits delivered per channel use and is given by

$$R = \frac{\log_2(M)}{N}$$

Definition 1.12 (Generic Channel) A *generic channel* in discrete time is a probabilistic mapping between the channel input $\mathbf{X}^N$ and the channel output $\mathbf{Y}^N$ as illustrated in Figure 1.16, namely, $p(\mathbf{Y}^N|\mathbf{X}^N)$. The channel is called *memoryless* if $p(\mathbf{Y}^N|\mathbf{X}^N) = \Pi_{n=1}^N p(Y_n|X_n)$, meaning that the current output symbol Y_n is independent of the past transmitted symbols.[6]

[6] Strictly speaking, this definition of memoryless channel is valid only if the encoder has no knowledge of the past channel outputs. i.e. $p(X_n|\mathbf{X}^{n-1}, \mathbf{Y}^{n-1}) = p(X_n|\mathbf{X}^{n-1})$. This condition holds in this book as the feedback channel is used to carry channel state only. Otherwise, a more general definition of memoryless channel should be: $p(Y_n|\mathbf{X}^n\mathbf{Y}^{n-1}) = p(Y_n|X_n)$ for all $n \geq 1$[100]. This more general definition reduces to the original definition if the encoder has no knowledge of past channel outputs. This is because $p(\mathbf{Y}^N|\mathbf{X}^N) = \Pi_n p(Y_n|\mathbf{X}^N\mathbf{Y}^{n-1}) = \Pi_n p(Y_n|\mathbf{X}^n\mathbf{Y}^{n-1}) = \Pi_n p(Y_n|X_n)$.

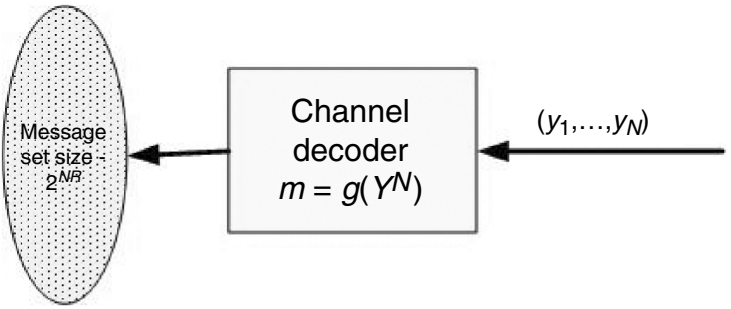


Figure 1.17. Illustration of channel decoder.

Definition 1.13 (Channel Decoder) A *channel decoder* is a deterministic mapping from the received symbols $\mathbf{Y}^N = [Y_1, \ldots, Y_N]$ to a decoded message index $\hat{m} \in \{1, 2, \ldots, 2^{NR}\}$ as illustrated in Figure 1.17

$$\hat{m} = g(\mathbf{Y}^N)$$

for some decoding functions g. The error probability associated with a decoding function is defined as

$$P_e = \Pr[g(Y^N) \neq m | m \text{ is transmitted}]$$

For simplicity and illustration purposes, we shall focus on the memoryless channel unless otherwise specified. Shannon's coding theorem is summarized below.

Theorem 1.5 (Shannon's Channel Coding Theorem) Letting $C = \max_{p(X)} I(X; Y)$. For rate $R < C$, there exists at least one channel encoder f and one channel decoder g for the memoryless channel $p(y|x)$ such that the error probability $P_e \to 0$ as $N \to \infty$. On the other hand, when $R > C$, the error probability is bounded away from zero for any channel encoder and channel decoder.

In other words, the channel coding theorem establishes a rigid bound on the maximal supportable data rate over a communication channel. From the definition of C above, the channel capacity [aftermaximization over $p(x)$] is a function of the channel $p(y|x)$ only. Note that $I(X; Y)$ is a function of both transmitter design ($p(x)$) and the channel ($p(y|x)$). We shall exploit the typical sequence to prove this powerful coding theorem on the basis of the *random codebook* argument. It is also important to go over the proof as it helps us understand the meaning of channel capacity and decoding mechanism.

Proof (Achievability) Let's first consider the proof for the achievability part, which is, given $R < C$, there exist a channel encoder f and a channel decoder g with arbitrarily low error probability. Consider a fixed distribution on x ($p(x)$). We generate 2^{NR} independent codewords at random according to the distribution $p(x)$. Hence, we formed a *random codebook* given by the matrix

$$\Omega = \begin{bmatrix} x_1(1) & \cdots & x_N(1) \\ \vdots & \vdots & \vdots \\ x_1(2^{NR}) & \cdots & x_N(2^{NR}) \end{bmatrix}$$

where the mth row represents the mth codeword that will be transmitted when the message index is m. The codebook Ω is then revealed to both the transmitter and the receiver. A message $m \in [1, 2^{NR}]$ is chosen according to a uniform distribution. The mth codeword $\Omega(m)$ {which has N symbols $\mathbf{x}(m) = [x_1(m), \ldots, x_N(m)]$} is sent out of the transmitter. The receiver receives a sequence $\mathbf{y} = [y_1, \ldots, y_N]$ according to the distribution $p(\mathbf{y}|\mathbf{x}) = \Pi_{n=1}^{N} p(y_n|x_n)$.

At the receiver, the decoder guesses which message was transmitted according to typical set decoding. The receiver declares that $\hat{m}$ was transmitted if there is one and only one index $\hat{m}$ such that $(\mathbf{x}(\hat{m}), \mathbf{y})$ is jointly typical. Hence, if there does not exist such index $\hat{m}$ or there is more than one such index, the receiver declares error. There will also be decoding error if $\hat{m} \neq m$.

To prove the achievability part, we shall show that the error probability of such encoder and decoder can achieve an arbitrarily low level. The error probability averaged over all possible random codebook realizations Ω is given by

$$\begin{aligned} \overline{P}_e &= \sum_{\Omega} P(\Omega) P_e^{(N)}(\Omega) = \sum_{\Omega} P(\Omega) \frac{1}{2^{NR}} \sum_{m} \lambda_m(\Omega) \\ &= \frac{1}{2^{NR}} \sum_{m} \left[\sum_{\Omega} P(\Omega) \lambda_m(\Omega) \right] \\ &= \sum_{\Omega} P(\Omega) \lambda_1(\Omega) \\ &= \Pr(E|W = 1) \end{aligned}$$

where

$$\lambda_m(\Omega) = \Pr\{\hat{m} \neq m | \mathbf{X}(m) \text{ is transmitted}\}$$

By symmetry of code construction, the average probability of error averaged over all codes does not depend on the particular index m being sent. Without loss of generality, assume that message $m = 1$ is transmitted. Define the events E_i as

$$E_i = \left\{(\mathbf{X}(i), \mathbf{Y}) \in \mathcal{A}_{\varepsilon}^{(N)}\right\} \forall i \in [1, \ldots, 2^{NR}]$$

Thus E_i is the event that the ith codeword $\mathbf{x}(i)$ and $\mathbf{y}$ are jointly typical. Error occurs when the transmitted codeword and the received sequence are not jointly typical or a wrong codeword is jointly typical with the received sequence. On the basis of the union bound, we have

$$\overline{P}_e = \Pr\{E_1^c \cup E_2 \cup \ldots \cup E_{2^{NR}}\} \le P(E_1^c) + \sum_{n=2}^{2^{NR}} P(E_n)$$

and by the jointly AEP theorem, we have $P(E_1^c) \le \varepsilon$ for sufficiently large N. Since by random code generation, $\mathbf{x}(1)$ and $\mathbf{x}(i)$ are independent, so are $\mathbf{y}$ and $\mathbf{x}(i)$. By the jointly AEP theorem, we have

$$P(E_i) \le 2^{-N(I(X;Y)-3\varepsilon)}$$

Hence, the average error probability is bounded by

$$\overline{P}_e \le \varepsilon + \sum_{i=2}^{2^{NR}} 2^{-N(I(X;Y)-3\varepsilon)} = \varepsilon + (2^{NR}-1)2^{-N(I(X;Y)-3\varepsilon)} \le \varepsilon + 2^{3N\varepsilon}2^{-N(I(X;Y)-R)}$$

If $R < I(X; Y) - 3\varepsilon$, we can choose ε and N such that $\overline{P}_e \le \varepsilon$. Since the average error probability is averaged over all codebook realizations $\{\Omega\}$, there exists at least one codebook Ω^* with a small average probability of error $P_e^*(\Omega^*) \le \varepsilon$. This proved that there exists at least one achievable code with rate $R < I(X; Y) - 3\varepsilon$ for arbitrarily small error probability ε.

Converse In the converse, we have to prove that any $(N, 2^{NR})$ codes with $\lim_{N\to\infty} P_e^{(N)} = 0$ must have $R \le C$. From Fano's inequality, let $\mathbf{Y}$ be the channel output from a discrete memoryless channel when $\mathbf{X}$ is the channel input. We have $I(\mathbf{X}; \mathbf{Y}) \le NC$ for all $p(\mathbf{X})$ because

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= \text{“}H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})\text{”} \\ &= H(\mathbf{Y}) - \sum_{n=1}^{N} H(Y_n|X_n) \\ &\le \sum_n H(Y_n) - \sum_n H(Y_n|X_n) \\ &= \sum_n I(X_n;Y_n) \le NC \end{aligned}$$

where the first equality is due to the fact that Y_n depends on X_n only and the last inequality is obtained directly from the definition of C. Let W be the message index drawn according to a uniform distribution over $[1, 2^{NR}]$. We have

$$\begin{aligned} NR = H(W) &= H(W|\mathbf{Y}) + I(W;\mathbf{Y}) \le H(W|\mathbf{Y}) + I(\mathbf{X}(W);\mathbf{Y}) \\ &\le 1 + P_e^{(N)}NR + I(\mathbf{X}(W);\mathbf{Y}) \le 1 + P_e^{(N)}NR + NC \end{aligned}$$

Thus

$$R \le P_e^{(N)} R + \frac{1}{N} + C \Rightarrow P_e^{(N)} \ge 1 - \frac{C}{R} - \frac{1}{NR}$$

Hence, if $R > C$, the error probability is bounded away from 0.

As a result of the *random coding proof*, we can see that good codes are actually very easy to find. For instance, we can randomly generate the M codewords to form a random codebook. We reveal this randomly generated codebook realization to the receiver so that the receiver can employ typical sequence decoding as illustrated in the proof. The random codebook argument presented above shows that there is a high chance that this randomly generated codebook is indeed a good code (or capacity achieving code). However, in practice, the problem lies in the receiving side. If our receiver has infinite processing power enabling us to exhaustively search in the decoding process, then there is almost no need for any code design at the transmitter. Because of limited processing power at the receiver, we have to create some *artificial structures* in the channel encoder to facilitate simple decoding at the receiver. For example, we have various types of codes such as *block codes*, *trellis codes*, and *turbo codes* in the literature. Once these artificial structures are added, good codes that can achieve the Shannon capacity are much harder to find. Yet, the channel coding theorem offers a tight upper bound on the maximum achievable rate so that we know how far we are from the best possible performance when we design practical codes.

1.4.3 Examples of Channel Capacity

In this section, we try to illustrate the concept of channel capacity by looking at several examples in wireless channels. We shall focus on some simple channels here. More sophisticated channels (such as the finite-state channels) will be discussed in detail in Chapter 2. In general, there is no closed-form expression for channel capacity except for a few special cases.

Discrete Memoryless Channels. Real-world channels are continuous-time with continuous-time input and continuous output. From Section 1.3, we established the equivalence between continuous-time and discrete-time channels. Hence, without loss of generality, we assume a discrete-time channel model with channel input $X \in \mathcal{X}$ and channel output $Y \in \mathcal{Y}$.

In general, the channel input $\mathcal{X}$ and the channel output $\mathcal{Y}$ are continuous complex numbers. However, if we consider a superchannel including a M-ary digital modulator, a memoryless physical channel, and a digital demodulator with Q quantization levels, we have a M-input, Q-output discrete memoryless channel as illustrated in Figure 1.18.

The $M \times Q$ discrete memoryless channel is specified by the *channel transition matrix* $p(y|x)$:

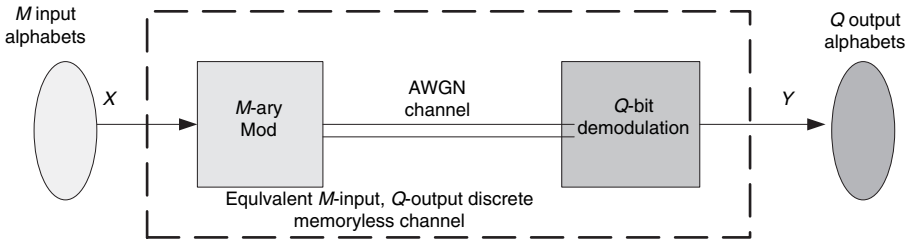


Figure 1.18. An *M*-input, *Q*-output discrete memoryless channel.

$$p(y|x) = \begin{bmatrix} p_{11} & \cdots & p_{1Q} \\ \vdots & \ddots & \vdots \\ p_{M1} & \cdots & p_{MQ} \end{bmatrix}$$

In addition, if the channel transition matrix is *symmetric* (i.e., all the rows are permutation of each other and so are the columns), the channel is called a *symmetric channel*.

Example 1.5 (Binary Symmetric Channel) A *binary symmetric channel* is a discrete channel with $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$. Furthermore, the probability of error $p(y = 0|x = 1) = p(y = 1|x = 0) = p$ and the probability of correct transmission is $p(y = 0|x = 0) = p(y = 1|x = 1) = 1 - p$ as illustrated in Figure 1.19.

The mutual information is given by

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X = x) \\ &= H(Y) - \sum_x p(x)H_2(p) \\ &= H(Y) - H_2(p) \\ &\le 1 - H_2(p) \end{aligned}$$

where $H_2(p) = -p\log_2 p - (1 - p)\log_2(1 - p)$ and the last inequality follows because Y is a binary random variable. Equality holds if and only if $p(Y = 1) = p(Y = 0) = 0.5$. This is equivalent to uniform input distribution $p(X = 0) = p(X = 1) = 0.5$. Hence, we have $C = 1 - H_2(p) = 1 + p\log_2 p + (1 - p)\log_2(1 - p)$.

Example 1.6 (Binary Erasure Channel) A *binary erasure channel* is a discrete channel with $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, e\}$, where e denotes erasure. Furthermore, the probability of erasure is $p(y = e|x = 1) = p(y = e|x = 0) = p$, and the probability of correct transmission is $p(y = 0|x = 0) = p(y = 1|x = 1) = 1 - p$ as illustrated in Figure 1.20.

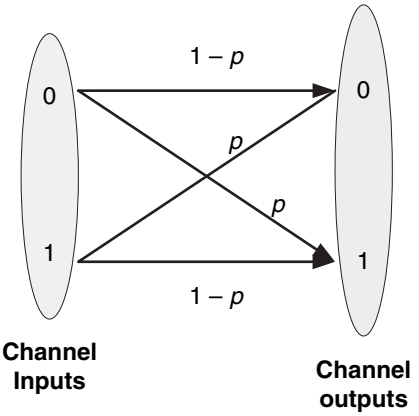


Figure 1.19. Binary symmetric channel.

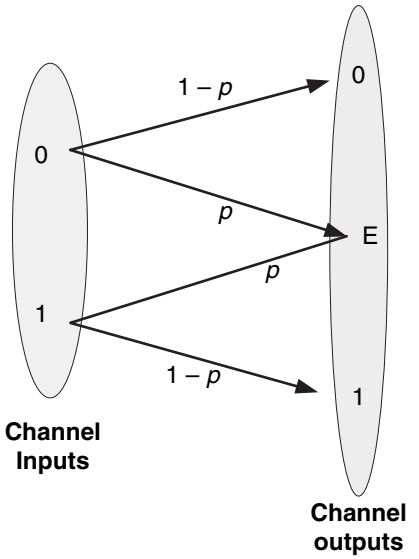


Figure 1.20. Binary erasure channel.

The channel capacity of the binary erasure channel is calculated as follows:

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} (H(Y) - H_2(p)) \end{aligned}$$

While the first guess for the maximum of $H(Y)$ is $\log_2(3)$, this is not achievable by any choice of input distribution $p(X)$. Let E be the event $\{Y = e\}$,

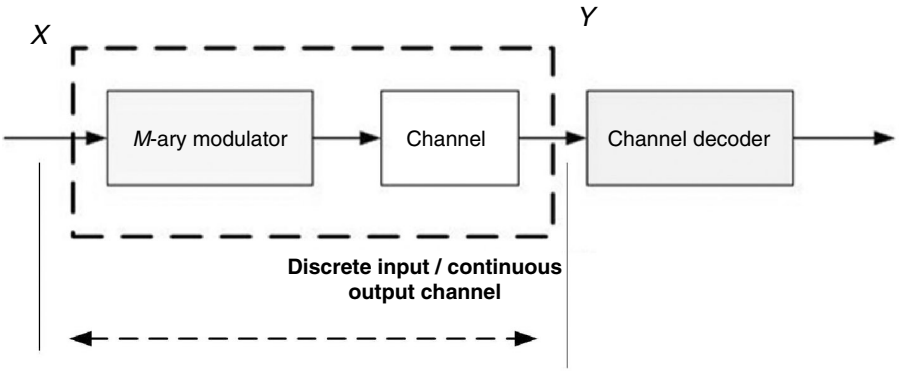


Figure 1.21. Discrete-input continuous-output channel.

(erasure). We have $H(Y) = H(Y, E) = H(E) + H(Y|E)$. Letting $\pi = p(X = 1)$, we have $H(Y) = H_2(p) + (1 - p)H_2(\pi)$. Hence

$$\begin{aligned} C &= \max_{p(x)}(H(Y) - H_2(p)) \\ &= \max_{p(x)}((1-p)H_2(\pi) + H_2(p) - H_2(p)) \\ &= \max_{\pi}((1-p)H_2(\pi)) \\ &= 1-p \end{aligned}$$

where the capacity achieving input distribution is $\pi = 0.5$. The expression for capacity has some intuitive meaning. Since a proportion p of the bits are erased in the channel, we can recover at most a proportion $(1 - p)$ of the bits. Hence, the channel capacity is at most $(1 - p)$.

Discrete-Input Continuous-Output Channels. A channel is called *discrete-input continuous-output* if the channel input alphabet $\mathcal{X}$ is discrete and the channel output alphabet $\mathcal{Y}$ is continuous. For example, the *superchannel* including a digital modulator and a physical channel as illustrated in Figure 1.21 belongs to this type of channel. The channel is characterized by a transition density $f(\mathbf{y}^N|\mathbf{x}^N)$. Similarly, the channel is memoryless if $f(\mathbf{y}^N|\mathbf{x}^N) = \Pi_n f(y_n|x_n)$.

Example 1.7 (Binary Input Continuous-output AWGN Channel) Consider a superchannel with a binary modulator and an AWGN physical channel. The discrete-time received signal Y_n is given by:

$$Y_n = X_n + Z_n$$

where X_n is the binary input and Z_n is a zero-mean white Gaussian noise with variance σ_z^2. Hence,we have $f(\mathbf{y}^N|\mathbf{x}^N) = \Pi_n f(y_n|x_n)$, where $f(y_n|x_n = -A)\mathcal{N}(-A, \sigma_z^2)$ and $f(y_n|x_n = A)\mathcal{N}(A, \sigma_z^2)$. The mutual information is given by

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) = H(Y) - (H(Y|X=A)\pi + 1(1-\pi)H(Y|X=-A)) \\
&= -\int_{-\infty}^{\infty} f(y)\log_2(f(y))dy + \pi\int_{-\infty}^{\infty} f(y|x=A)\log_2(f(y|x=A))dy \\
&\quad + (1-\pi)\int_{-\infty}^{\infty} f(y|x=-A)\log_2(f(y|x=-A))dy \\
&= \pi\int_{-\infty}^{\infty} f(y|x=A)\log_2\left(\frac{f(y|x=A)}{f(y)}\right)dy \\
&\quad + (1-\pi)\int_{-\infty}^{\infty} f(y|x=-A)\log_2\left(\frac{f(y|x=-A)}{f(y)}\right)dy
\end{aligned}
$$

where $\pi = p(X = A)$. Hence, the mutual information is maximized when $\pi = 0.5$. Note that $C \leq 1$ because the digital modulator in the "superchannel" is a binary modulator delivering at most 1 bit per modulation symbol. At finite channel noise, some of the bits transmitted will be in error. However, we observe that C approaches 1 bit per channel use as $\sigma_z^2 \to 0$. This means that at high SNR, nearly all of the transmitted bits could be decoded at the receiver and therefore, achieving a capacity of 1 bit per channel use.

In general, for M-ary modulator, the channel capacity is given by

$$
\begin{aligned}
C &= \max_{\pi}[H(Y) - H(Y|X)] \\
&= \max_{\pi}\left[\sum_{\pi}\pi(x)\int_{y}\log_2\left(\frac{f(y|x)}{\sum_x \pi(x)f(y|x)}\right)\right]
\end{aligned}
$$

where $\pi = [\pi_1, \ldots, \pi_M]$ and $\pi_q = p(X = q)$. In general, the capacity achieving distribution π has to be evaluated numerically.

Continuous-Input Continuous-Output Channels. In this case, both the channel input $\mathcal{X}$ and the channel output $\mathcal{Y}$ are continuous values (or complex values in general). The channel is specified by the *transition density* $f(y|x)$. In fact, we can consider a continuous-input channel as a limiting case of a discrete-input channel when we have infinitely dense constellation at the digital modulator. For any input distribution $p(X = x_1), \ldots, p(X = x_M)$ of an M-ary digital modulator, we could find an equivalent input distribution $f(X)$ in the continuous input case. In other words, we expect the channel capacity of continuous-input channels to be greater than or equal to the channel capacity of discrete input channels because the input distribution of the latter case is a subset of that in the former case.

Example 1.8 (Discrete Time AWGN Channel) Consider a discrete-time AWGN channel where $X \in \mathcal{X}$ is the channel input and $Y \in \mathcal{Y}$ is the channel output. Both the channel input and channel output are complex numbers. The channel output is given by:

$$Y_n = X_n + Z_n$$

where Z_n is the white Gaussian channel noise with variance σ_z^2. Therefore, the channel transition probability is given by $f(y|x) = \frac{1}{2\pi\sigma_z}\exp\left[-\frac{1}{2\sigma_z^2}|y-x|^2\right]$. The channel capacity is given by

$$\begin{aligned} C &= \max_{f(x)}[H(Y) - H(Y|X)] \\ &= \max_{f(x)} H(Y) - H(Z) \\ &= \max_{f(y)} H(Y) - H(Z) \\ &= \log_2\left(1 + \frac{\sigma_x^2}{\sigma_z^2}\right) \end{aligned}$$

where $\sigma_x^2 = \varepsilon[|X|^2]$. Note that the second equality is due to the fact that $H(Z)$ is independent of $f(x)$ and the final equality is due to the fact that $H(Y)$ is maximized if and only if Y is complex Gaussian, which is equivalent to X being complex Gaussian.

Continuous-Time Channels. In this section, we try to connect the channel capacity of discrete-time channels to continuous-time channels. The channel capacity of discrete-time channels is expressed in units of bits per channel use. On the other hand, the unit of channel capacity of continuous-time channels is bits per second. To facilitate the discussion, let's consider the following example.

Example 1.9 (Continuous-Time AWGN Channel) Consider a continuous-time AWGN channel with channel input $X(t)$ (with two-sided bandwidth W) and channel output $Y(t)$. Both the channel input and channel output are complex-valued random processes. The channel output is given by

$$Y(t) = X(t) + Z(t)$$

where $Z(t)$ is the white Gaussian channel noise with two-sided power spectral density η_0.

From the discussions in previous section, a bandlimited random process $X(t)$ could be represented in geometric domain as vector $\mathbf{X}$ in WT dimension signal space over the complex field. Hence, the equivalent discrete time channel is given by

$$\mathbf{y} = \mathbf{x} + \mathbf{z}$$

where $\mathbf{z}$ is a zero-mean Gaussian i.i.d. sequence with variance σ_z^2. The total noise power is given by

$$P_z = \frac{\varepsilon\left[\|\mathbf{z}\|^2\right]}{T} = \frac{\sum_{i=1}^{WT}\sigma_z^2}{T} = W\eta_0$$

Hence, we have $\sigma_z^2 = \eta_0$. The channel transition density is given by $f(\mathbf{y}|\mathbf{x}) = \Pi_{n=1}^{WT} f(y_n|x_n)$, where

$$f(y_n|x_n) = \frac{1}{2\pi\eta_0}\exp\left[-\frac{1}{\eta_0}|y_n - x_n|^2\right]$$

Treating $\mathbf{X}$ as a supersymbol, the asymptotic channel capacity [in bits per second (bps)] is given by

$$C = \lim_{T\to\infty}\max_{p(\mathbf{x})}\frac{1}{T}I(\mathbf{X};\mathbf{Y})$$

Because of the i.i.d. nature of the noise sequence $\mathbf{Z}$, we have

$$\max_{p(\mathbf{X})} I(\mathbf{X};\mathbf{Y}) = \sum_{n=1}^{WT}\max_{p(X_n)} I(X_n;Y_n) = WT\log_2\left(1+\frac{\sigma_x^2}{\eta_0}\right)$$

where $\sigma_x^2 = \varepsilon[|X_n|^2]$. On the other hand, the transmit power is given by

$$P_{tx} = \frac{1}{T}\int_0^T |x(t)|^2\,dt = \frac{1}{T}\|\mathbf{x}\|^2 = W\left(\frac{1}{WT}\sum_n |x_n|^2\right) \to W\varepsilon\left[|x_n|^2\right] = W\sigma_x^2 \quad (1.67)$$

as $T \to \infty$. Substituting into the capacity equation, the channel capacity (in bps) is given by

$$C = W\log_2\left(1+\frac{\sigma_x^2}{\eta_0}\right) = W\log_2\left(1+\frac{P_{tx}}{W\eta_0}\right) \quad (1.68)$$

Note that $\eta_0 W$ equals to the total noise power and hence, the channel capacity could also be written as $C = \log_2(1 + SNR)$.

Define bandwidth efficiency as $\eta = R/W$, where R denotes the transmission bit rate and W denotes the bandwidth. From Shannon's capacity, the bandwidth efficiency could be expressed as

$$\eta_{\max} = \log_2\left(1+\frac{P_{tx}}{W\eta_0}\right) = \log_2\left(1+\frac{RE_b}{W\eta_0}\right)$$

where R is the bit rate and E_b is the bit energy. Since $R/W = \eta$, the equation above has a solution as follows:

$$\frac{E_b}{\eta_0} \geq \frac{2^{\eta_{\max}}-1}{\eta_{\max}} \quad (1.69)$$

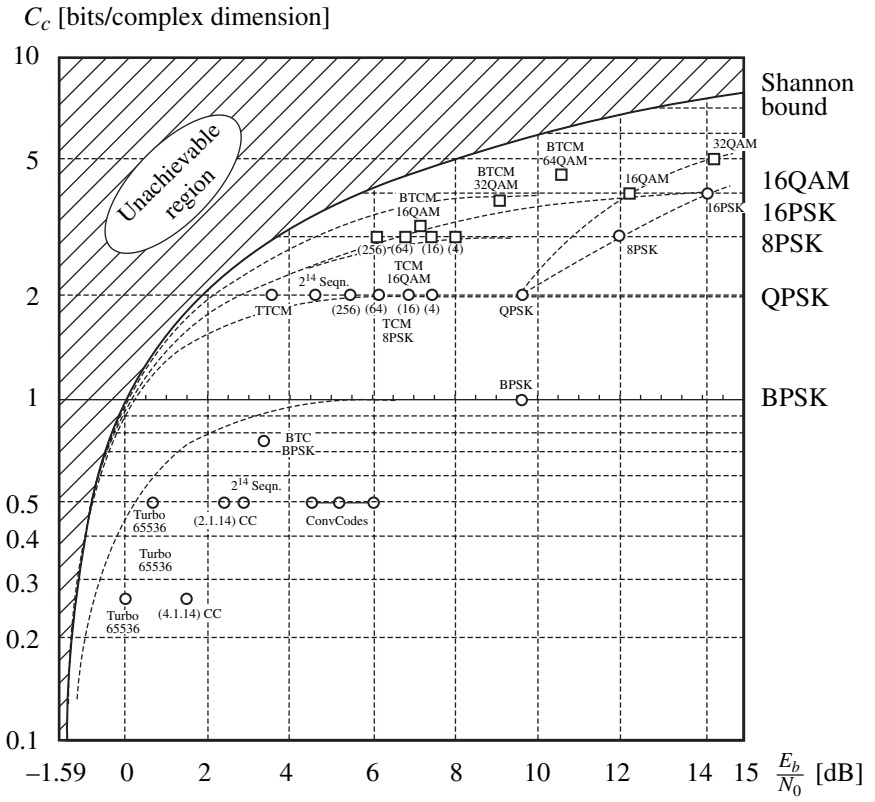


Figure 1.22. A plot of bandwidth efficiency versus E_b/η_0 for AWGN channel.

In the limiting case of very large bandwidth, we have $\eta_{\max} \to 0$ and $E_b/\eta_0 \geq \log(2) = -1.59$ dB. This represents the absolute minimum bit energy : noise ratio that is required to reliably transmit one bit of information over a very large bandwidth.

Figure 1.22 illustrates the bandwidth efficiency versus E_b/η_0 for various modulation schemes over the AWGN channel. As mentioned, the capacity of a *continuous modulator* is an upper bound of all other discrete modulators such as BPSK, QPSK, and 16QAM. At a very low E_b/η_0 region (such as in deep-space communications), we have to sacrifice bandwidth to trade for *power efficiency* because power is the limiting factor. Hence, modulation schemes operating in the region are called *bandwidth expansion modes*. Examples are M-ary orthogonal modulations. On the other hand, when bandwidth is a limited resource (such as in terrestrial communications), we have to sacrifice power for bandwidth efficiency. Examples are 16QAM and 64QAM modulations. This region of operation is therefore referred to as consisting of *high-bandwidth-efficiency modes*.

1.5 SUMMARY

In this chapter, we have reviewed the fundamental concepts in wireless communications. We have discussed the wireless fading channels in continuous-time domain and established the equivalence between continuous-time and discrete-time models based on the concept of signal space and sufficient statistics. The wireless fading channel is modeled as a random process $H(t, \nu, r)$ in the time domain t, frequency domain ν, and the position domain r. Equivalently,we can also characterize the random fading channels from the spectral domain $H(f, \tau, k)$, which involves the *Doppler domain* f, the *delay domain* τ, and the *wavenumber domain* k. In fact, the two random processes are related by three-dimensional Fourier transform pairs ($H(t, \nu, r) \leftrightarrow H(f, \tau, k)$).

In most cases, we are concerned with the types of random channels that are wide-sense stationary (WSS) uncorrelated scattering (US). Hence, the autocorrelation of $H(t, \nu, r)$ is a function of $(\Delta t, \Delta \nu, \Delta r)$ only. To characterize the statistical properties of the WSS-US random channels, we can define *coherence time* T_c, *coherence bandwidth* B_c, and *coherence distance* D_c on the basis of the one-dimensional autocorrelation $R_H(\Delta t)$, $R_H(\Delta \nu)$ and $R_H(\Delta r)$.

On the other hand, from the spectral domains (f, τ, k), the uncorrelated scattering property implies that $\varepsilon[H(f, \tau, k)H^*(f', \tau', k')] = S_H(f, \tau, k)\delta(f - f')\delta(\tau - \tau')\delta(k - k')$. Hence, we can also define the corresponding dual parameters on the basis of the scattering function $S_H(f, \tau, k)$: the *Doppler spread* σ_f^2, the *delay spread* σ_τ^2, and the *angle spread* σ_k^2. With respect to the coherence bandwidth (or delay spread) and the transmitted bandwidth, we can deduce whether flat fading channels ($W < B_c$) or frequency-selective fading channels ($W > B_c$) will be experienced by the signal. On the other hand, on the basis of the coherence time (or Doppler spread) and the transmitted symbol duration, we can deduce whether we have fast fading channels ($T_c < T_s$) or slow fading channels ($T_c > T_s$).

Finally, we reviewed the fundamentals of information theory including entropy, mutual information, and channel capacity. We illustrate the application of information theory by evaluating the channel capacity of several simple channels.

From now on, unless specified otherwise, we shall further develop the theories and concepts based on the discrete-time model.

EXERCISES

1. *Spreading and Coherence in different environment* Recall that delay, doppler and angle spread are defined by the environment and so as the corresponding frequency, time and spatial coherence. Using the knowledge of them, determine the order of the following environment in terms of the magnitudes of spreads and coherence with 1 being the smallest and 6 being the largest.

(a) Office
(b) Elevator
(c) Residential Building
(d) Underground Trains
(e) Bullet Trains
(f) Pedestrian Pathway

2. *Wide Sense Stationary* A random processes $X(t, \nu, r)$ is regarded as a wide sense stationary process if the correlation with itself only depends on the difference of its parameters. Therefore, samples taking at an interval of time, frequency and space are no different from samples taken at the shifts of the interval. In other words, the randomness of the process does not depends on its parameters. $X(t, \nu, r)$ is the input random process of the system characterized by a system function $H(t, \nu, r)$ which is also random in nature and $Y(t, \nu, r)$ is the output random process. Suppose there is noise $N(t, \nu, r)$ in the system, the input-output equation is therefore

$$Y(t, \nu, r) = H(t, \nu, r)X(t, \nu, r) + N(t, \nu, r)$$

(a) Construct a random process $X(t, \nu, r)$ which is WSS in time, frequency and position.
(b) Under which condition(s) is output $Y(t, \nu, r)$ WSS?
(c) Under which condition(s) are $X(t, \nu, r)$ and $Y(t, \nu, r)$ jointly WSS?

3. *Upper Bound of differential entropy* Differential entropy is the entropy of a continuous random variable. It can be shown that the entropy is upper bounded by the following.

$$H(X) \leq \frac{1}{2}\log_2(2\pi e\sigma_x^2)$$

where σ_x^2 is the variance of the random variable X.
(a) verify that the upper bound is achieved if X is of guassian distributed.

4. *Entropy of morphism* Compare the entropy of 2 random variables X and Y if
(a) the mapping from X to Y is isomorphic
(b) the mapping from X to Y is endomorphic or subjective

5. *Entropy of markov chains* Suppose there are 2 states A and B as shown in the diagram 1. At each state, one can choose to either stay in the same state or jump to the other state. p_{AA}, p_{AB}, p_{BA}, p_{BB} are the corresponding transition probabilities. Define X_n to be the state at time instant n and X_0 is the initial state. Assume that it is equally probable for X_0 to be A or B.

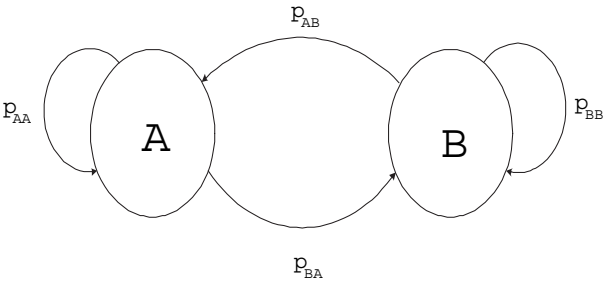


Figure 1. Markov chains with 2 states A and B.

(a) Find the probability if X_3 equals A.

(b) What is the entropy of X_n, $H(X_n)$ as $n \to \infty$?

(c) verify the result obtained in (b) if the transition probabilities are symmetric. That is $p_{AA} = p_{BB}$ and $p_{AB} = p_{BA}$.

(d) verify the result obtained in (b) if the transition probabilities are biased, say to A. That is $p_{AA} \gg p_{BA}$ and $p_{AB} \gg p_{BB}$.

6. *Fano's inequality* Suppose Y is an estimate of the random variable X. If there is no error, we can say that the conditional entropy $H(X|Y) = 0$. Thus, the estimate error is expected to be small when the conditional entropy is small. Recall the Fano's inequality,

$$H(P_e) + P_e \log(|x| - 1) \geq H(X|Y)$$

where P_e is the probability of error and $|x|$ denote the size of the support set of x. Therefore, given a target or required probability of error, it gives an upper bound of the conditional entropy. To have more insight into this inequality, we can reorganize the terms to be the following

$$P_e \geq \frac{H(X|Y) - 1}{\log(|x|)}$$

Given the conditional entropy, Fano's inequality provides a lower bound of the probability of error. Define an indicator variable E such that

$$E = \begin{cases} 1, & \text{if } \hat{X} \neq X; \\ 0, & \text{if } \hat{X} = X. \end{cases}$$

(a) Use chain rule of entropy to expand $H(E, X|Y)$

(b) Prove that $H(E|Y) \leq H(P_e)$.

(c) Prove that $H(X|E, Y) \leq P_e \log(|x| - 1)$

(d) Combining parts (a), (b) and (c) prove the Fano's inequality.

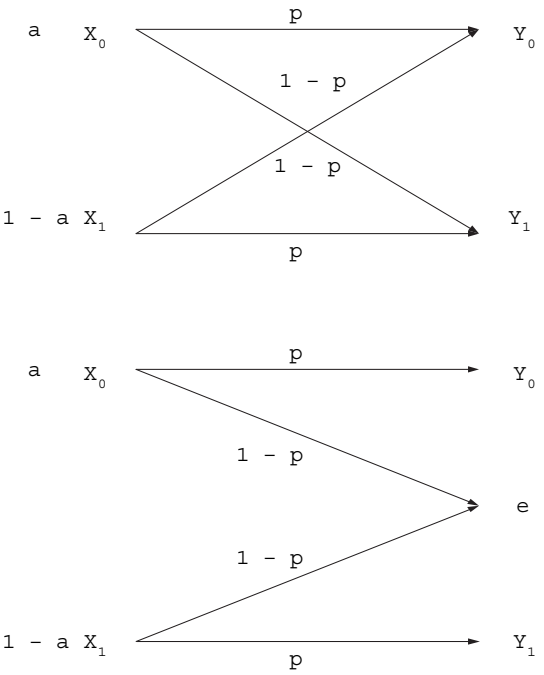


Figure 2. Above: Binary symmetric channel Below Binary erasure channel.

7. *Binary symmetric channel v.s. Binary Erasure channel* Suppose the input X of the channel is equally probable to take values X_0 and X_1, ($a = 0.5$), whereas the output can take values of Y_0 and Y_1. The probability of receiving the right symbol is denoted by p. In other words, $p(Y_0|X_0) = p(Y_1|X_1) = p$. In the binary erasure channel, the output Y can take one more value e which is called the erasure. The probability of getting the erasure is $1 - p$ and $p(Y_0|X_1) = p(Y_1|X_0) = 0$. Both channels are shown as reference in figure 2.

(a) Compute $H(X|Y)$ in the binary symmetric channel

(b) Compute $H(X|Y)$ in the binary erasure channel

(c) Compute $H(Y)$ in both cases.

(d) What does $H(Y)$ tend to if $p \to 1$? How about $p \to 0$? Explain.

8. *Venn Diagram* Use Venn Diagram to illustrate the following

(a) $I(X; Y|Z)$

(b) $I(X_1, X_2, X_3; Y)$

9. *Channel Capacity of AWGN Channels*

(a) Given an AWGN channel with channel output given by $Y_n = X_n + Z_n$, show that the capacity is

$$C = \log_2\left(1 + \frac{\sigma_x^2}{\sigma_z^2}\right)$$

where σ_x^2 is the variance of X and σ_z^2 is the variance of Z. Assume everything is complex. What is the unit of the capacity obtained.

(b) In practice, we transmit continuous waveforms $x(t)$ instead of discrete-time symbols, X_n. Using (a) and Nyquist sampling theorem, show that the channel capacity of the continuous waveform AWGN channel with bandwidth W is given by

$$C = W\log_2\left(1 + \frac{P}{W\eta_0}\right)$$

where η_0 is the single-sided noise power spectral density. What is the unit of the capacity obtained.