Chapter 10

Communication Theory

Previously, we have focused on the electromagnetic properties of a propagation channel, primarily in terms of signal power, noise power, and signal to noise ratio. We will now consider the use of the channel to send information. The goal is to understand the likelihood of bit errors and the maximum rate at which data can be transmitted reliably over a given communications channel in terms of the SNR. As the analysis is quite theoretical, some foundational topics in stochastic theory and random processes will be reviewed first.

10.1 Random Variables and Random Processes

A real-valued random variable is a map X from a sample space Ω into the real line:

$$X : \Omega \to R \tag{10.1}$$

Ω is a set of outcomes for a physical process, such as the roll of a dice or the trajectories of electrons on a warm resistor. X is a measurable parameter associated with the physical process—the number of dots showing on the face up side of a dice or the voltage across a warm resistor at a particular instant in time. Subsets of the space Ω represent combinations of outcomes, or events.

Associated with Ω and X is a measure P which maps well-behaved subsets of Ω to a number between 0 and 1,

$$S \in \Omega, \quad P(S) = \text{Probability of the event } S \tag{10.2}$$

where P satisfies $P(\Omega) = 1$, $P(\emptyset) = 0$, and $P(A \cup B) \leq P(A) + P(B)$. Random variables can be discrete, with a finite or countable number of possible values, or continuous, depending on the properties of the measure P.

The cumulative distribution function (CDF) of a random variable is defined by

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}) \tag{10.3}$$

so that $F(-\infty) = 0$ and $F(\infty) = 1$. The probability distribution function (PDF) is

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{10.4}$$

For a discrete random variable, $f_X(x)$ is a sum of delta functions. For a continuous random variable, X can take on any value within its range, and $F_X(x)$ is a continuous function. A PDF always integrates to one over its domain.

The joint PDF of two random variables X and Y is given by

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} P(X \leq x, Y \leq y) \tag{10.5}$$

The joint PDF expresses the probability of a combination of values for the two random variables. The conditional probability of one random variable given a value for another is

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{10.6}$$

This represents the likelihood of a given value x for the random variable X when the value of Y is known. If the random variables are strongly correlated, then the conditional probability takes on a large value over a small range of possibilities for X. If the random variables are uncorrelated, then the joint PDF is the product of the single variable PDFs and the conditional PDF is equal to the PDF of X.

10.1.1 Expectation and Statistical Moments

Expectation is an idealization of the simple idea of averaging. Intuitively, the expectation of a random process is simply the limit of sample averages of the random variable as the number of samples become large. The problem with this idea is that fairly sophisticated mathematics is required to make it rigorous. Instead of defining expectation in terms of sample averages of one realization of a random process over time, we define it for random variables in terms of the PDF. The sample average idea then arises as an approximation to the expectation that is accurate if certain conditions are met.

The expectation or expected value of a random variable is defined to be

$$\mathrm{E}[X] = \int x f_X(x)\, dx \tag{10.7}$$

This is also called the mean and denoted by $\overline{X}$. The fundamental theorem of expectation is that the expectation of a function of a random variable can be found similarly using

$$\mathrm{E}[g(X)] = \int g(x) f_X(x)\, dx \tag{10.8}$$

where $g(X)$ is a new random variable derived from X according to the function g.

Higher order statistics can also be found using the expectation. The variance of X is

$$\sigma_X^2 = \mathrm{E}[(X - \overline{X})^2] \tag{10.9}$$

The expectation of a power of a random variable is a statistical moment. A normal random variable is completely determined by its first and second order statistics, the mean (first moment) and variance (second central moment).

10.1.2 Correlation

The degree of correlation of two random variables is measured by the expectation of the product,

$$R_{XY} = \mathrm{E}[XY] \tag{10.10}$$

The covariance is $\mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$. If the random variables have zero mean, there is no distinction between correlation and covariance. Correlation can be estimation using sample averages as in (8.18).

The correlation coefficient

$$\rho(X,Y) = \frac{\mathrm{E}[(X - \overline{X})(Y - \overline{Y})]}{\sqrt{\sigma_X^2 \sigma_Y^2}} \tag{10.11}$$

is the covariance normalized by the standard deviations of both random variables, so that $0 \leq |\rho| \leq 1$. Two random variables are uncorrelated if $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$ or $\rho = 0$. Intuitively, if two random variables are uncorrelated they are independent, but it is possible for two random variables to be dependent yet still uncorrelated. Two random variables with correlation coefficient $\rho(X, Y) = 1$ must be related by $Y = aX + b$.

10.1.3 Random Processes

A random process is a parametric family of random variables, representing successive measurements of a random quantity as a function of an independent variable such as position or time. Mathematically, a real-valued random process is a map from a sample space Ω and an interval T into R,

$$X : [\Omega, T] \to R \tag{10.12}$$

so that $X(\cdot, t)$ is a random variable for each fixed time t in the interval T. The CDF of the random process evaluated at time t is

$$F_{X(t)}(x) = P(X(t) \leq x) \tag{10.13}$$

For fixed $\omega \in \Omega$, the time function $X(\omega, \cdot)$ is a realization of the random process. A group of realizations is an ensemble. To simplify the notation, we can suppress the ω dependence and write the random process as $X(t)$ or X_t.

Sample Averages of Random Processes

For a sequence of measurement times $t_1, t_2, \ldots, t_N$, the sample average

$$S_N = \frac{1}{N}\sum_{n=1}^{N} X(t_n) \tag{10.14}$$

is a itself a discrete random process, so that for each N, S_n is a random variable. If the random process X is such that $\mathrm{E}[X(\omega, t_n)] = \overline{X}$ for each n (constant mean) and $\sigma_X^2(t) = \mathrm{E}[(X(t) - \mathrm{E}[X(t)])^2] = \sigma_X^2$ (constant variance), then $\mathrm{E}[S_N] = \overline{X}$ and $\sigma_{S_N}^2 = \sigma_X^2/N$, so that in some sense, the sample average converges to $\overline{X}$ as $N \to \infty$. In simple terms, if the local mean and variance of a random process do not drift over time, then we can find the expected value (ensemble average) from sample or time averages.

Under the above assumptions, rigorously speaking, the sample averages S_N converge to $\overline{X}$ in the sense that $P(|S_N - \overline{X}| > \epsilon)$ converges to zero as N becomes large for any ϵ. This is the weak law of large numbers. There may still be members $X(\omega, t)$ of the ensemble for which sample averages do not converge. In an engineering sense, typically only one realization is available, and we want to be sure that sample averages converge to a measurement of some desired physical quantity. Under an additional assumption, that the fourth moment of the PDF is finite, $P(\{\omega : \lim_{N\to\infty} |S_N(\omega) - \overline{X}| = 0\}) = 1$, which is the strong law of large numbers and provides the desired assurance that sample averages always converge to the mean. A random process with the property that sample averages converge to the mean is ergodic.

10.1.4 Wide-Sense Stationarity

A random process is stationary if the joint distribution of the random variables X_{t_1}, X_{t_2}, $\ldots X_{t_N}$ is independent of a shift $t_1 + t$, $t_2 + t$, $\ldots$, $t_N + t$. In particular, $f_{X(t)}(x) = f_{X(t+s)}(x)$, which implies constant mean and variance. Another implication is that the autocorrelation function

$$R_X(s, t) = \mathrm{E}[X(s)X(t)] \tag{10.15}$$

can be expressed in the form $R_X(\tau)$ where $\tau = t - s$. A weaker property is wide-sense stationarity (WSS), which requires only constant mean, variance, and $R_X(s,t) = R_X(t-s)$. For a wide-sense stationary process, $R_X(\tau)$ has a maximum at $R_X(0)$, and $R_X(\tau)$ is continuous if it is continuous at $\tau = 0$.

10.1.5 Power Spectral Density

The power spectral density (PSD) of a random process is the Fourier transform of the autocorrelation function,

$$S_X(f) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau} d\tau \tag{10.16}$$

For the PSD to be well-defined, X must be WSS.

White noise. White noise is uncorrelated with itself at any offset except zero, so that

$$R_w(\tau) = \frac{N_0}{2}\delta(\tau) \tag{10.17}$$

where $N_0 = k_\text{B}T$ and T is the equivalent temperature of the noise. The units of N_0 are W/Hz = Joules (energy). The PSD is

$$S_w(f) = \frac{N_0}{2} \tag{10.18}$$

If a signal with these properties were applied across a resistor, the time average power dissipated in the resistor would be infinite, so truly white noise cannot exist in practice.

10.1.6 Linear, Time-Invariant Systems

If we apply a linear, time-invariant, casual system or filter with impulse response $h(t)$ to a realization of a stationary random process, the output is

$$y(t) = \int_{-\infty}^{\infty} h(t-\tau)x(\tau)\, d\tau \tag{10.19}$$

The PSD of the output is

$$S_Y(f) = |H(f)|^2 S_X(f) \tag{10.20}$$

where $H(f)$ is the Fourier transform of the impulse response.

10.1.7 Bandlimited Noise

If a white noise signal is passed through an ideal bandpass filter with bandwidth B and center frequency f_c, then from (10.20), the power spectral density becomes

$$S_w(f) = \begin{cases} N_0/2 & |f \pm f_c| \le B/2 \\ 0 & \text{otherwise} \end{cases} \tag{10.21}$$

For a bandlimited white noise signal at baseband ($f_c = 0$), the autocorrelation function is

$$\begin{aligned} R(\tau) &= \int_{-\infty}^{\infty} S_w(f) e^{j2\pi f\tau} df \\ &= \int_{-B/2}^{B/2} N_0 e^{j2\pi f\tau} df \\ &= BN_0 \frac{\sin(\pi B\tau)}{\pi B\tau} \\ &= BN_0 \operatorname{sinc}(B\tau) \end{aligned} \tag{10.22}$$

From this result, we can see that the variance of the bandlimited noise signal is

$$\sigma_\nu^2 = BN_0 \tag{10.23}$$

which is the same value that is obtained if the PSD (10.21) is integrated over all frequencies.

If the white noise random process represents the time-domain voltage $v_{\mathrm{L}}(t)$ across a matched load connected to a warm resistor through a bandpass filter with bandwidth B, then from (2.122) the variance of the noise signal is

$$\sigma_\nu^2 = E[v_{\mathrm{L}}^2] = (\overline{v}_{\mathrm{oc}}/2)^2 = k_{\mathrm{B}}TBR \tag{10.24}$$

By comparing this result with (10.23), we can see that if the load resistance is assumed to be a reference value of 1 Ω, then the power spectral density of $v_L(t)$ is $N_0/2 = k_{\mathrm{B}}T/2$.

10.1.8 Narrowband Random Processes and the Complex Signal Representation

Narrowband signals are fundamental to communications. For such a signal, it is convenient to use a complex baseband representation, defined by

$$x(t) = \mathrm{Re}[\tilde{x}(t)e^{j\omega_c t}] \tag{10.25}$$

The complex quantity $\tilde{x}(t)$ is often written in the form

$$\tilde{x}(t) = x_I(t) + jx_Q(t) \tag{10.26}$$

In this decomposition, I represents in-phase and Q represents the quadrature component, so that this is sometimes called the IQ representation of the signal. If the signal $x(t)$ is a pure tone, then $\tilde{x}(t)$ is a constant and reduces to the standard frequency domain or phasor representation commonly used in electromagnetics and circuit theory. The I and Q signals can also be thought of as a representation in rectangular coordinates of the phase and magnitude of a modulated carrier sinusoid.

10.2 Modulation

Modulation is a way to vary the amplitude and phase of a sinusoidal carrier waveform in order to transmit information. When selecting a modulation scheme, the complexity of implementing the analog or digital processing required to create the modulated waveform is typically balanced with the bandwidth of the resulting signal, the bit rate of the modulation scheme, and the sensitivity to noise. A key parameter of a modulation scheme is the ratio of required bandwidth to bit rate, or spectral efficiency. Older, simpler modulation schemes tend to have poor spectral efficiency, whereas modern digital modulations have spectral efficiencies close to unity.

We will briefly consider a few examples of analog and digital modulation schemes:

AM modulation. For this type of modulation, the carrier amplitude is varied according to

$$s(t) = A_c[1 + \alpha m(t)] \cos(\omega_c t) \tag{10.27}$$

where $m(t)$ is the information-bearing signal, ω_c is the carrier frequency, and α is the modulation strength.

FM modulation. In this case, the carrier phase depends on $m(t)$, so that

$$s(t) = A_c \cos\left[\omega_c t + 2\pi\alpha \int_0^t m(\tau)d\tau\right] \tag{10.28}$$

The derivative of the modulated carrier phase includes a term proportional to the modulating signal $m(t)$, which means that the frequency of the waveform shifts according to the amplitude of the modulating signal.

Binary phase-shift keying (BPSK). This is a digital modulation scheme, meaning that the data source is binary rather than analog. If $p(t)$ is a basic pulse shape and T is the bit duration, then the modulated waveform is

$$m(t) = \sum_k b_k p(t - kT) \tag{10.29}$$

where $b_k = \pm 1$ is the bit sequence to be transmitted. The function $p(t)$ is a pulse shape. For basic PSK, the pulse $p(t)$ is rectangular. Abrupt changes in phase lead to broad bandwidth for the modulated signal for a given bit rate, so the pulse shape can be adjusted by making it a smoother function, leading to a modulation scheme with lower required bandwidth for a given bit rate and higher spectral efficiency.

Quadriphase-shift keying (QPSK). BPSK requires at least twice the bandwidth of the original data stream for transmission, and so is only used in cases where the transmitter or receiver must be simple. The bandwidth can be reduced by modulating both the I and Q components of the signal, so that

$$s(t) = A_c m_1(t) \cos(\omega_c t) + A_c m_2(t) \sin(\omega_c t) \tag{10.30}$$

where m_1 and m_2 are two independent BPSK signals, generated by taking alternate bits from the sequence b_k. Quadriphase-shift keying is used for many different communication systems, from satellite communication to terrestrial wireless.

Offset Quadriphase-shift keying (OQPSK). One problem with QPSK is that the signal often jumps by $\pm 90^\circ$ or $\pm 180^\circ$ in phase, leading to discontinuities in the modulated signal and reduced spectral efficiency. This can be reduced by shifting one of the bit streams in time by half a bit duration.

Pulse shaping. Due to discontinuities in the modulated signal, the rectangular pulse leads to a broadband modulated signal that has poor spectral efficiency and is susceptible to distortion due to frequency dispersion. A smoother pulse with finite bandwidth is used in practice, so that the signal transitions smoothly from one phase state to another. Common choices are the raised cosine or root-raised cosine.

10.2.1 Signal Constellations

One way to represent digital modulation schemes is in terms of the complex baseband representation in the complex plane. If we define the normalized coordinates

$$\begin{aligned}\phi_1(t) &= \sqrt{\frac{2}{T}}\cos(\omega_c t)\\ \phi_2(t) &= \sqrt{\frac{2}{T}}\sin(\omega_c t)\end{aligned}$$

then BPSK can be represented by

$$s(t) = \pm A_c\sqrt{\frac{2}{T}}\phi_1(t) \tag{10.31}$$

In the complex plane, the signal jumps back and forth between two points on the ϕ_1 axis.

The energy contained in one bit is

$$\begin{aligned}E_b &= \int_0^T P(t)\,dt\\ &= \int_0^T \frac{v^2(t)}{R}\,dt\\ &= \frac{1}{R}\int_0^T A_c^2\cos^2(\omega_c t)\,dt\\ &= \frac{A_c^2 T}{2R} \qquad (10.32)\end{aligned}$$

so that we can write

$$s(t) = \pm\sqrt{E_b}\phi_1(t) \tag{10.33}$$

where we have assumed a load resistance of $R = 1\,\Omega$.

For QPSK, the constellation is given by

$$s(t) = \pm\sqrt{E_b}\phi_1(t) \pm \sqrt{E_b}\phi_2(t) \tag{10.34}$$

which corresponds to four points in the complex plane. Each point represents a phase state of the modulated carrier. Other digital modulation schemes have larger constellations. 16-QAM is represented by 16 points on a four by four grid in the complex plane.

10.2.2 Nonlinear Digital Modulation

Other modulation schemes are nonlinear in the sense that the modulated signal cannot be represented as a linear combination of pulses weighted by a bit sequence. These include binary frequency shift keying (BPSK), and variants of BPSK which smooth the phase transitions in the modulated signal and thereby reduce the required channel bandwidth, including minimum shift keying (MSK), continuous phase shift keying (CPSK), and Gaussian filtered minimum phase shift keying (GMSK),

10.3 Bit Error Rate

Channel noise causes a receiver to occasionally detect a 1 when the transmitter sent a 0 and vice versa. One measure of the performance of a communication system is the bit error rate (BER), which is defined to be the average probability of a bit with a error 0 transmitted or with a 1 transmitted, weighted by the probability of a 0 or 1 from the source. With models for channel noise and SNR and the detection scheme used to demodulate the signal, we can estimate the bit error rate analytically.

10.4 Signal Detection

A simple way to detect a digital receiver output is to sample the baseband signal periodically. The problem with this approach is that if the noise is large, on occasion the noise will change the sign of the signal and lead to a bit error. A better detection scheme is to integrate the signal for a bit interval and then sample the output of the integrator at the end of each bit interval. We assume that a clock synchronization step is integrated at the receiver, so that the detector knows when each bit interval ends. We will analyze the performance of this detector for a channel with additive white Gaussian noise (AWGN).

The output of the integrator due to the noise $x_n(t)$ is a random process,

$$y_n = \frac{1}{T}\int_{t_0}^{t_0+T} x_n(t)\,dt \tag{10.35}$$

where t_0 is chosen to be the beginning of a symbol period. The variance of the integrated noise is

$$\begin{aligned}
\sigma^2 &= \mathrm{E}[y_n^2] \\
&= \mathrm{E}\left[\frac{1}{T}\int_{t_0}^{t_0+T} x_n(t)\,dt\,\frac{1}{T}\int_{t_0}^{t_0+T} x_n(s)\,ds\right] \\
&= \frac{1}{T^2}\int_{t_0}^{t_0+T}\int_{t_0}^{t_0+T} \mathrm{E}[x_n(t)x_n(s)]\,dt\,ds \\
&= \frac{1}{T^2}\int_{t_0}^{t_0+T}\int_{t_0}^{t_0+T} R(t-s)\,dt\,ds \\
&= \frac{1}{T^2}\int_{t_0}^{t_0+T}\int_{t_0}^{t_0+T} \frac{N_0}{2}\delta(t-s)\,dt\,ds \\
&= \frac{N_0}{2T}
\end{aligned} \tag{10.36}$$

This result shows that the longer we integrate, the smaller the variance of the integrated noise, which we expect, as the noise waveform is zero mean.

In order to determine the bit error rate (BER) of the channel, we need to know the PDF of the integrated noise. With certain restrictions, the central limit theorem can be applied to the integral of a random variable, so that we can assume y_n is Gaussian distributed. The PDF is

$$f_{y_n}(y_n) = \frac{e^{-y_n^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \tag{10.37}$$

If a 1 was transmitted, then the baseband signal without noise is a rectangular pulse of width T and amplitude A. The output of the integrator at time $t_0 + T$ is

$$y_s = \frac{1}{T}\int_{t_0}^{t_0+T} A\,dt = A \tag{10.38}$$

The symbol energy is

$$\begin{aligned} E_b &= TP_{\text{av}} \\ &= T\frac{1}{T}\int_0^T x_s^2(t)\,dt \\ &= A^2T \end{aligned}$$

so that $y_s = A = \sqrt{E_b/T}$. If y_n is negative and large enough in magnitude that $y_s + y_n < 0$, then a bit error occurs. The probability is

$$\begin{aligned} P(y_n < -y_s) &= P(y_n > y_s) \quad \text{(since the PDF is symmetric)} \\ &= \int_{y_s}^{\infty} \frac{e^{-y_n^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} dy_n \\ &= \int_{y_s/\sqrt{2\sigma^2}}^{\infty} \frac{e^{-z^2}}{\sqrt{2\pi\sigma^2}}\sqrt{2\sigma^2}dz \\ &= \frac{1}{\sqrt{\pi}}\int_{y_s/\sqrt{2\sigma^2}}^{\infty} e^{-z^2}dz \\ &= \frac{1}{2}\text{erfc}(y_s/\sqrt{2\sigma^2}) \\ &= \frac{1}{2}\text{erfc}(\sqrt{E_b/N_0}) \end{aligned} \tag{10.39}$$

Since we get the same result for a zero symbol, this is the BER of the channel.

The quantity E_b/N_0 that controls the BER can be related to the SNR using

$$\begin{aligned} \frac{E_b}{N_0} &= \frac{A^2T}{N_0} \\ &= \frac{A^2T}{BN_0T}BT \\ &= \frac{P_s}{P_{\text{n}}}BT \\ &= (BT)\,\text{SNR} \end{aligned} \tag{10.40}$$

where B is the bandwidth of the noise. The frequency response of the integrator is

$$\begin{aligned} H(f) &= \frac{1}{T}\int_{T/2}^{T/2} e^{j2\pi ft}dt \\ &= \frac{\sin(\pi fT)}{\pi fT} \\ &= \text{sinc}(fT) \end{aligned}$$

The sinc function has an approximate bandwidth $B_p \simeq 1/T$, which is called the bit-rate bandwidth. So, $BT \simeq 1$, and

$$\text{SNR} \simeq \frac{E_b}{N_0} \tag{10.41}$$

If the noise at the receiver is bandlimited white noise instead of ideal white noise, then the integrated noise variance becomes

$$\begin{aligned}\sigma^2 &= \frac{1}{T^2}\int_{t_0}^{t_0+T}\int_{t_0}^{t_0+T} R(t-s)\,dt\,ds\\ &= \frac{2}{T}\int_0^T R(\tau)(1-\tau/T)\,d\tau\\ &= \frac{2}{T}\int_0^T N_0 B\frac{\sin(2\pi B\tau)}{2\pi B\tau}(1-\tau/T)\,d\tau\\ &= \frac{N_0}{2T}\left[\frac{\cos(2\pi BT)-1}{\pi^2B^2T}+\frac{2\,\mathrm{si}(2\pi BT)}{\pi}\right] \qquad (10.42)\end{aligned}$$

If $BT \simeq 1$, then

$$\sigma^2 \simeq 0.9\frac{N_0}{2T} \qquad (10.43)$$

which is close to the previous result. The effect is to cut down the sidelobes of the sinc transfer function of the integrator, which reduces the noise at the output slightly. The variance of the bandlimited noise after integration can also be written as

$$\sigma^2 \simeq 0.9BN_0\frac{1}{2BT} = 0.9\frac{\sigma_\nu^2}{2BT} \qquad (10.44)$$

so that the noise is reduced by a factor of $2BT$ by the integrator.

10.4.1 Matched Filter

The detector described above can be viewed as a convolution of the received signal with the rectangular pulse $p(t)$. This can be used to generalize the detector to other symbol pulses which are not rectangular, by convolving the received signal with the symbol pulse shape $p(t)$. This is a matched filter.

10.4.2 Rayleigh Fading Channel

To analyze the bit error rate for a Rayleigh fading channel, we can view the additive noise of the previous section as receiver noise and let the signal power be governed by the Rayleigh fading model. To find the bit error rate, we weight the probability of error (10.39) by the PDF of the local SNR for the channel given by (9.18) and integrate over local SNR. This leads to

$$\mathrm{BER} = \frac{1}{2}\left(1-\sqrt{\frac{\Gamma}{1+\Gamma}}\right) \qquad (10.45)$$

where the mean SNR is $\Gamma = \mathrm{E}[\gamma] \simeq \mathrm{E}[E_b/N_0]$ and $\gamma \simeq E_b/N_0$ is the local SNR.

10.5 Information Theory

Claude Shannon revolutionized communication theory when he laid out the foundations of information theory in 1948 in terms of three basic theorems, which we will briefly survey in this section. In order to state these theorems, we need to define some measures of the uncertainty and information content of random variables that represent information sources.

10.5.1 Definitions

Source. Information bearing signal generated by some physical signal (digitized speech, data, video). We assume that the possible values of the signal are discrete, so that an analog signal is digitized before transmission. When designing a communication system, we do not care about the details of a specific instance of a signal to be transmitted, so we will model the source stochastically as a random process.

Entropy. A measure of the information content of a signal. For a source which emits symbols chosen from an alphabet of K symbols, the entropy

$$H = \sum_{k=1}^{K} p_k \log_2(1/p_k) \tag{10.46}$$

where p_k is the probability that the symbol s_k is emitted by the source. With the base 2 logarithm, H is in units of bits. A source which emits only one symbol has zero entropy, and a source has maximum entropy of $\log_2 K$ when the probabilities of each symbol are equal. $\log_2 K$ is the number of bits required to assign the symbols to K equal-length bit sequences.

The greater the redundancy in the sequence of symbols, the lower the entropy, so entropy can be viewed as a measure of uncertainty, which agrees with the thermodynamics concept of entropy, since a fluid or gas with greater randomness in the distribution of molecules has higher thermodynamic entropy. As we will see shortly, the source coding theorem implies that the higher the entropy of a source, the greater the number of bits that are required to transmit the source data over a communication channel.

Mutual information. The mutual information of random variables X and Y is defined to be

$$I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{10.47}$$

where X represents the source symbols to be transmitted across the channel and Y represents the received signal. The more reliable the channel, the larger the mutual information $I(X;Y)$. If the channel is noiseless, then X and Y are completely correlated and the mutual information takes on the maximum value $I(X;Y) = H(X)$.

Mutual information can be computed in terms of the entropy of Y reduced by the entropy of Y given X, so that

$$I(X;Y) = H(Y) - H(Y|X) \tag{10.48}$$

where $H(X)$ and $H(Y)$ are the entropies of X and Y and $H(Y|X)$ is the conditional entropy of Y given X, which measures the uncertainty of Y that remains if X is known. This relationship helps to explain mutual information, because $I(X;Y)$ is the uncertainty in Y reduced by the uncertainty that remains if X is known. For a noiseless channel, $H(Y) = H(X)$ and $H(Y|X) = 0$.

For continuous random variables, the mutual information is

$$I(X;Y) = \int\int p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, dx\, dy \tag{10.49}$$

Capacity. Maximum error-free information rate of a channel in bits/sec or bits/use for a discrete channel. The capacity is defined to be the maximum value of the mutual information $I(X;Y)$ for a source X and received symbol Y, so that

$$C = \max_{p(x)} I(X;Y) \tag{10.50}$$

In words, the capacity of a channel is the maximum of the mutual information over all possible sources. Since we can encode the source data before transmission, we can design the probability distribution $p(x)$ of the source symbols in such a way that the mutual information and channel capacity are maximized.

Probability of error. Probability that noise introduced by a channel will cause the receiver to detect a different bit than was transmitted. Using codes that represent a particular sequence of bits with a longer string of bits, errors can be detected and corrected.

10.5.2 Source Coding Theorem

Given a discrete, memoryless source characterized by a certain entropy, the average code-word length for a distortionless source encoding scheme is bounded above by the entropy. Source coding is used to compress the signal so that redundancy is removed and a channel can be used more efficiently. The degree of compression that can be obtained is bounded by

$$L \geq H \quad \text{(bits/symbol)} \tag{10.51}$$

so that it takes at least H transmitted bits per source symbol on average to encode the source signal.

The source coding theorem is easy to understand intuitively. If the source consists of a typical English language text, the letter e appears much more often than other letters like q or z. Therefore, we can assign the letter e a short bit sequence as a codeword, and less common letters longer codewords. In this way, the average number of bits per text symbol can be smaller than $\log_2 27 = 4.8$ bits/symbol (including 27 letters and the space as symbols). This is reflected in the entropy, since in (10.46) the contribution to the entropy of symbols with very low probability is small. The entropy of English text is close to four, meaning that by assigning short bit sequences to common letters like e and r and longer bit sequences to uncommon letters like q and z, on average four bits per letter are required to transmit the text. The entropy does not tell the full story, however. Because certain letter pairs and triplets occur much more frequently than others, the entropy rate taking into account joint probabilities for letter combinations is even lower: 0.6 to 1.3, depending on the measurement approach. This surprisingly low value implies that highly efficient compression algorithms can be designed for English text. The best known algorithms require about 1.5 bits per character.

10.5.3 Channel Coding Theorem

If a memoryless channel has capacity C and a source generates information at a rate less than C, then there exists a coding scheme such that the signal can be transmitted with arbitrarily low probability of error.

This is a remarkable result! As long as we do not send bits too fast over a channel, even in the presence of noise it is possible to achieve essentially zero transmission errors. The role of the code is to reintroduce controlled redundancy into the symbol stream so that errors caused by noise can be corrected. The basic idea is that very long code words can still be recognized even if many bits are flipped by noise.

The catch to this result is that the proof of the theorem guarantees that a code exists but does not construct such a code. Coding theory has been concerned for many years with developing every better codes that allow error detection and correction and achieve performance close to Shannon's limit without requiring too much processing capability to encode or decode the information. Turbo codes, for example, are a recent advance based on feedback between multiple decoders that have moved the state-of-the-art closer to the theoretical limit.

10.5.4 Channel Capacity Theorem

The capacity of an AWGN channel with bandwidth B (Hz) and SNR P/σ^2, where P is the time average received power and σ^2 is the variance of Gaussian noise introduced by the channel, is bounded by

$$\begin{aligned} C &= B\log_2(1+P/\sigma^2) \\ &= B\log_2(1+\text{SNR}) \quad \text{(Bits/sec)} \end{aligned} \tag{10.52}$$

This is the Shannon-Hartley capacity bound for an AWGN channel. If we send information below this rate, using coding we can achieve an arbitrarily low bit error rate. Above this rate, errors are unavoidable. This expression shows the importance of bandwidth, because C is linear in B but only increases logarithmically with transmitted power. It is common to express the capacity relative to a bandwidth of $B = 1\,\text{Hz}$ (i.e., bits/sec/Hz rather than bits/sec) to separate the dependence of capacity on the frequency allocation for the channel from the propagation environment which determines SNR.

Another way to interpret capacity in bits/sec/Hz is in terms of the capacity of a temporally discrete channel expressed in bits per channel use or bits per transmission. By Shannon's sampling theorem, a real signal of bandwidth B corresponds to a bit rate of $2B$ bits/sec. Dividing the capacity in bits/sec by the channel use rate leads to

$$C = \frac{1}{2}\log_2(1+\text{SNR}) \quad \text{(bits/use, real channel)} \tag{10.53}$$

For a complex channel, we can view the real and imaginary parts as independent signals, so that the capacity doubles:

$$C = \log_2(1+\text{SNR}) \quad \text{(bits/use, complex channel)} \tag{10.54}$$

10.5.5 Binary Symmetric Channel

The proof of the capacity bound (10.52) for the AWGN channel is beyond the scope of this treatment, but we can illustrate some of the concepts by considering a simpler discrete channel. The binary symmetric channel takes an input bit from a source and transmits that bit to a receiver, with a probability p that a given bit will be flipped due to noise. If we use the channel N times, how many bits can we transmit without error? Even though each bit is unreliable, Shannon proved that by using long codewords we can still send information error free as long as we use the channel enough times. In this way, we can send

$$M = CN < N \tag{10.55}$$

bits without error, where C is the channel capacity in bits/use.

The binary channel is discrete, whereas the analog AWGN channel is continuous, so that here, $C < 1$, whereas $C > 1$ is possible for the analog channel if the SNR is not too small. We can convert the continuous AWGN channel to a binary symmetric channel by sending two possible signal levels (e.g., BPSK). The capacity in this case is necessarily less than one, since we can only transmit one bit per channel use, and the possibility of a bit error requires additional coding to eliminate errors. The two-level quantization does not make optimal use of the channel, since with an analog channel we can send symbols from a large constellation such that each symbol represents many bits.

For the given channel, the mutual information of the transmitted and received bits is

$$\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - \sum_{x=0,1} p(x) H(y|X=x) \\
&= H(Y) - \sum_{x=0,1} p(x) H(p) \\
&\leq 1 - H(p)
\end{aligned}$$

so that the capacity is bounded by

$$\begin{aligned}
C &= 1 - H(p) \\
&= 1 + p \log_2 p + (1-p) \log_2 (1-p)
\end{aligned}$$

If the channel is completely reliable, so that p is 0 or 1, then $C = 1$ and no coding is required to correct errors. If $p = 1/2$, $C = 0$, so that no information can be reliably transmitted.