Chapter 1

Information Theory

1.1 Issues in Information Theory

The ultimate aim of telecommunications is to communicate information between two geographically separated locations via a communications channel with adequate quality. The theoretical foundations of information theory accrue from Shannon's pioneering work [13–16], and hence most tutorial interpretations of his work over the past fifty years rely fundamentally on [13–16]. This chapter is no exception in this respect. Throughout this chapter we make frequent references to Shannon's seminal papers and to the work of various authors offering further insights into Shannonian information theory. Since this monograph aims to provide an all-encompassing coverage of video compression and communications, we begin by addressing the underlying theoretical principles using a light-hearted approach, often relying on worked examples.

Early forms of human telecommunications were based on smoke, drum or light signals, bonfires, semaphores, and the like. Practical information sources can be classified as analog and digital. The output of an analog source is a continuous function of time, such as, for example, the air pressure variation at the membrane of a microphone due to someone talking. The roots of Nyquist's sampling theorem are based on his observation of the maximum achievable telegraph transmission rate over bandlimited channels [17]. In order to be able to satisfy Nyquist's sampling theorem the analogue source signal has to be bandlimited before sampling. The analog source signal has to be transformed into a digital representation with the aid of time- and amplitude-discretization using sampling and quantization.

The output of a digital source is one of a finite set of ordered, discrete symbols often referred to as an alphabet. Digital sources are usually described by a range of characteristics, such as the source alphabet, the symbol rate, the symbol probabilities, and the probabilistic interdependence of symbols. For example, the probability of $u$ following $q$ in the English language is $p = 1$, as in the word "equation." Similarly, the joint probability of all pairs of consecutive symbols can be evaluated.

In recent years, electronic telecommunications have become prevalent, although
most information sources provide information in other forms. For electronic telecommunications, the source information must be converted to electronic signals by a transducer. For example, a microphone converts the air pressure waveform \( p(t) \) into voltage variation \( v(t) \), where

\[
v(t) = c \cdot p(t - \tau),
\]

and the constant \( c \) represents a scaling factor, while \( \tau \) is a delay parameter. Similarly, a video camera scans the natural three-dimensional scene using optics and converts it into electronic waveforms for transmission.

The electronic signal is then transmitted over the communications channel and converted back to the required form, which may be carried out, for example, by a loudspeaker. It is important to ensure that the channel conveys the transmitted signal with adequate quality to the receiver in order to enable information recovery. Communications channels can be classified according to their ability to support analog or digital transmission of the source signals in a simplex, duplex, or half-duplex fashion over fixed or mobile physical channels constituted by pairs of wires, Time Division Multiple Access (TDMA) time-slots, or a Frequency Division Multiple Access (FDMA) frequency slot.

The channel impairments may include superimposed, unwanted random signals, such as thermal noise, crosstalk via multiplex systems from other users, man-made interference from car ignition, fluorescent lighting, and other natural sources such as lightning. Just as the natural sound pressure wave between two conversing persons will be impaired by the acoustic background noise at a busy railway station, similarly the reception quality of electronic signals will be affected by the above unwanted electronic signals. In contrast, distortion manifests itself differently from additive noise sources, since no impairment is explicitly added. Distortion is more akin to the phenomenon of reverberating loudspeaker announcements in a large, vacant hall, where no noise sources are present.

Some of the channel impairments can be mitigated or counteracted; others cannot. For example, the effects of unpredictable additive random noise cannot be removed or “subtracted” at the receiver. Its effects can be mitigated by increasing the transmitted signal’s power, but the transmitted power cannot be increased without penalties, since the system’s nonlinear distortion rapidly becomes dominant at higher signal levels. This process is similar to the phenomenon of increasing the music volume in a car parked near a busy road to a level where the amplifier’s distortion becomes annoyingly dominant.

In practical systems, the Signal-to-Noise Ratio (SNR) quantifying the wanted and unwanted signal powers at the channel’s output is a prime channel parameter. Other important channel parameters are its amplitude and phase response, determining its usable bandwidth \( (B) \), over which the signal can be transmitted without excessive distortion. Among the most frequently used statistical noise properties are the probability density function (PDF), cumulative density function (CDF), and power spectral density (PSD).

The fundamental communications system design considerations are whether a high-fidelity (HI-FI) or just acceptable video or speech quality is required from a system, which predetermines, among other factors, its cost, bandwidth requirements, as
1.1. ISSUES IN INFORMATION THEORY

well as the number of channels available, and has implementational complexity ramifications. Equally important are the issues of robustness against channel impairments, system delay, and so on. The required transmission range and worldwide roaming capabilities, the maximum available transmission speed in terms of symbols/sec, information confidentiality, reception reliability, convenient lightweight, solar-charged design, are similarly salient characteristics of a communications system.

Information theory deals with a variety of problems associated with the performance limits of the information transmission system, such as that depicted in Figure 1.1. The components of this system constitute the subject of this monograph; hence they will be treated in greater depth later in this volume. Suffice it to say at this stage that the transmitter seen in Figure 1.1 incorporates a source encoder, a channel encoder, an interleaver, and a modulator and their inverse functions at the receiver. The ideal source encoder endeavors to remove as much redundancy as possible from the information source signal without affecting its source representation fidelity (i.e., distortionlessly), and it remains oblivious of such practical constraints as a finite delay and limited signal processing complexity. In contrast, a practical source encoder will have to retain a limited signal processing complexity and delay while attempting to reduce the source representation bit rate to as low a value as possible. This operation seeks to achieve better transmission efficiency, which can be expressed in terms of bit-rate economy or bandwidth conservation.

Figure 1.1: Basic transmission model of information theory.
The channel encoder re-inserts redundancy or parity information but in a controlled manner in order to allow error correction at the receiver. Since this component is designed to ensure the best exploitation of the re-inserted redundancy, it is expected to minimize the error probability over the most common channel, namely, the so-called Additive White Gaussian Noise (AWGN) channel, which is characterized by a memoryless, random distribution of channel errors. However, over wireless channels, which have recently become prevalent, the errors tend to occur in bursts due to the presence of deep received signal fades induced by the distinctively superimposed multipath phenomena. This is why our schematic of Figure 1.1 contains an interleaver block, which is included in order to randomize the bursty channel errors. Finally, the modulator is designed to ensure the most bandwidth-efficient transmission of the source- and channel encoded, interleaved information stream, while maintaining the lowest possible bit error probability. The receiver simply carries out the corresponding inverse functions of the transmitter. Observe in the figure that besides the direct interconnection of the adjacent system components there are a number of additional links in the schematic, which will require further study before their role can be highlighted. Thus, at the end of this chapter we will return to this figure and guide the reader through its further intricate details.

Some fundamental problems transpiring from the schematic of Figure 1.1, which were addressed in depth by a range of references due to Shannon [13-16], Nyquist [17], Hartley [18], Abramson [19], Carlson [20], Raemer [21], and Ferenczy [22] and others are as follows:

- What is the true information generation rate of our information sources? If we know the answer, the efficiency of coding and transmission schemes can be evaluated by comparing the actual transmission rate used with the source’s information emission rate. The actual transmission rate used in practice is typically much higher than the average information delivered by the source, and the closer these rates are, the better is the coding efficiency.

- Given a noisy communications channel, what is the maximum reliable information transmission rate? The thermal noise induced by the random motion of electrons is present in all electronic devices, and if its power is high, it can seriously affect the quality of signal transmission, allowing information transmission only at low-rates.

- Is the information emission rate the only important characteristic of a source, or are other message features, such as the probability of occurrence of a message and the joint probability of occurrence for various messages, also important?

- In a wider context, the topic of this whole monograph is related to the blocks of Figure 1.1 and to their interactions, but in this chapter we lay the theoretical foundations of source and channel coding as well as transmission issues and define the characteristics of an ideal Shannonian communications scheme.

Although numerous excellent treatises are available on these topics, which treat the same subjects with a different flavor [20,22,23], our approach is similar to that of the
above classic sources: since the roots are in Shannon's work, references [13–16, 24, 25] are the most pertinent and authoritative sources.

In this chapter we consider mainly discrete sources, in which each source message is associated with a certain probability of occurrence, which might or might not be dependent on previous source messages. Let us now give a rudimentary introduction to the characteristics of the AWGN channel, which is the predominant channel model in information theory due to its simplicity. The analytically less tractable wireless channels of Chapter 2 will be modeled mainly by simulations in this monograph, although in Chapter 10 some analytical results are also provided in the context of Code Division Multiple Access (CDMA) systems.

1.2 Additive White Gaussian Noise Channel

1.2.1 Background

In this section, we consider the communications channel, which exists between the transmitter and the receiver, as shown in Figure 1.1. Accurate characterization of this channel is essential if we are to remove the impairments imposed by the channel using signal processing at the receiver. Here we initially consider only fixed communications links whereby both terminals are stationary, although mobile radio communications channels, which change significantly with time, are becoming more prevalent.

We define fixed communications channels to be those between a fixed transmitter and a fixed receiver. These channels are exemplified by twisted pairs, cables, wave guides, optical fiber and point-to-point microwave radio channels. Whatever the nature of the channel, its output signal differs from the input signal. The difference might be deterministic or random, but it is typically unknown to the receiver. Examples of channel impairments are dispersion, nonlinear distortions; delay, and random noise.

Fixed communications channels can often be modeled by a linear transfer function, which describes the channel dispersion. The ubiquitous additive Gaussian noise (AWGN) is a fundamental limiting factor in communications via linear time-invariant (LTI) channels. Although the channel characteristics might change due to factors such as aging, temperature changes, and channel switching, these variations will not be apparent over the course of a typical communication session. It is this inherent time invariance that characterizes fixed channels.

An ideal, distortion-free communications channel would have a flat frequency response and linear phase response over the frequency range of $-\infty \ldots +\infty$, although in practice it is sufficient to satisfy this condition over the bandwidth $(B)$ of the signals to be transmitted, as seen in Figure 1.2. In this figure, $A(\omega)$ represents the magnitude of the channel response at frequency $\omega$, and $\phi(\omega) = \omega T$ represents the phase shift at frequency $\omega$ due to the circuit delay $T$.

Practical channels always have some linear distortions due to their bandlimited, nonflat frequency response and nonlinear phase response. In addition, the group-delay response of the channel, which is the derivative of the phase response, is often given.
1.2.2 Practical Gaussian Channels

Conventional telephony uses twisted copper wire pairs to connect subscribers to the local exchange. The bandwidth is approximately 3.4 kHz, and the waveform distortions are relatively benign.

For applications requiring a higher bandwidth, coaxial cables can be used. Their attenuation increases approximately with the square root of the frequency. Hence, for wideband, long-distance operation, they require channel equalization. Typically, coaxial cables can provide a bandwidth of about 50 MHz, and the transmission rate they can support is limited by the so-called skin effect.

Point-to-point microwave radio channels typically utilize high-gain directional transmit and receive antennas in a line-of-sight scenario, where free-space propagation conditions may be applicable.

1.2.3 Gaussian Noise

Regardless of the communications channel used, random noise is always present. Noise can be broadly classified as natural or man-made. Examples of man-made noise are those due to electrical appliances, and fluorescent lighting, and the effects of these sources can usually be mitigated at the source. Natural noise sources affecting radio transmissions include galactic star radiations and atmospheric noise. There exists a low-noise frequency window in the range of 1–10 GHz, where the effects of these sources are minimized.

Natural thermal noise is ubiquitous. This is due to the random motion of electrons, and it can be reduced by reducing the temperature. Since thermal noise contains practically all frequency components up to some $10^{13}$ Hz with equal power, it is often referred to as white noise (WN) in an analogy to white light containing all colors with
1.2. ADDITIVE WHITE GAUSSIAN NOISE CHANNEL

equal intensity. This WN process can be characterized by its uniform power spectral density (PSD) \( N(\omega) = N_0/2 \) shown together with its autocorrelation function (ACF) in Figure 1.3.

The power spectral density of any signal can be conveniently measured by the help of a selective narrowband power meter tuned across the bandwidth of the signal. The power measured at any frequency is then plotted against the measurement frequency. The autocorrelation function \( R(\tau) \) of the signal \( x(t) \) gives an average indication of how predictable the signal \( x(t) \) is after a period of \( \tau \) seconds from its present value. Accordingly, it is defined as follows:

\[
R(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_{-\infty}^{\infty} x(t)x(t + \tau)dt.
\]  

(1.2)

For a periodic signal \( x(t) \), it is sufficient to evaluate the above equation for a single period \( T_0 \), yielding:

\[
R(\tau) = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x(t)x(t + \tau)dt.
\]  

(1.3)

The basic properties of the ACF are:

- The ACF is symmetric: \( R(\tau) = R(-\tau) \).
- The ACF is monotonously decreasing: \( R(\tau) \leq R(0) \).
- For \( \tau = 0 \) we have \( R(0) = \overline{x^2}(t) \), which is the signal’s power.
- The ACF and the PSD form a Fourier transform pair, which is formally stated as the Wiener-Khintchine theorem, as follows:

\[
R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} N(\omega)e^{j\omega\tau}d\omega
\]
CHAPTER 1. INFORMATION THEORY

\[ R(r) = N_0 \text{sinc}(2\pi B r) \]

Figure 1.4: Power spectral density and autocorrelation of bandlimited WN.

where \( \delta(\tau) \) is the Dirac delta function. Clearly, for any timed-domain shift \( \tau > 0 \), the noise is uncorrelated.

Bandlimited communications systems bandlimit not only the signal but the noise as well, and this filtering limits the rate of change of the time-domain noise signal, introducing some correlation over the interval of \( \pm 1/2B \). The stylized PSD and ACF of bandlimited white noise are displayed in Figure 1.4.

After bandlimiting, the autocorrelation function becomes:

\[
R(\tau) = \frac{1}{2\pi} \int_{-B}^{B} \frac{N_0 e^{j\omega \tau}}{2} d\omega = \frac{N_0}{2} \int_{-B}^{B} e^{j2\pi f \tau} df
\]

\[
= \frac{N_0}{2} \left[ \frac{e^{j2\pi f \tau} B}{j2\pi \tau} \right]_{-B}^{B}
\]

\[
= \frac{1}{j2\pi \tau} [\cos 2\pi B \tau + j \sin 2\pi B \tau - \cos 2\pi B \tau - j \sin 2\pi B \tau]
\]

\[
= N_0 B \frac{\sin(2\pi B \tau)}{2\pi B \tau},
\]

which is the well-known sinc-function seen in Figure 1.4.

In the time-domain, the amplitude distribution of the white thermal noise has a normal or Gaussian distribution, and since it is inevitably added to the received signal, it is usually referred to as additive white Gaussian noise (AWGN). Note that AWGN is therefore the noise generated in the receiver. The probability density function (PDF) is the well-known bell-shaped curve of the Gaussian distribution, given by

\[
p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-m)/2\sigma^2},
\]
where \( m \) is the mean and \( \sigma^2 \) is the variance. The effects of AWGN can be mitigated by increasing the transmitted signal power and thereby reducing the relative effects of noise. The signal-to-noise ratio (SNR) at the receiver’s input provides a good measure of the received signal quality. This SNR is often referred to as the channel SNR.

1.3 Information of a Source

Based on Shannon’s work [13–16, 24, 25], let us introduce the basic terms and definitions of information theory by considering a few simple examples. Assume that a simple 8-bit analog-to-digital (ADC) converter emits a sequence of mutually independent source symbols that can take the values \( i = 1, 2, \ldots 256 \) with equal probability. One may wonder, how much information can be inferred upon receiving one of these samples? It is intuitively clear that this inferred information is definitely proportional to the “uncertainty” resolved by the reception of one such symbol, which in turn implies that the information conveyed is related to the number of levels in the ADC. More explicitly, the higher the number of legitimate quantization levels, the lower the relative frequency or probability of receiving any one of them and hence the more “surprising,” when any one of them is received. Therefore, less probable quantized samples carry more information than their more frequent, more likely counterparts.

Not surprisingly, one could resolve this uncertainty by simply asking a maximum of 256 questions, such as “Is the level 1?” “Is it 2? . . .” “Is it 256?” Following Hartley’s approach [18], a more efficient strategy would be to ask eight questions, such as: “Is the level larger than 128?” No. “Is it larger than 64?” No. . . . “Is it larger than 2?” No. “Is it larger than 1?” No. Clearly, the source symbol emitted was of magnitude one, provided that the zero level was not used. We could therefore infer that \( \log_2 256 = 8 \) “Yes/No” binary answers were needed to resolve any uncertainty as regards the source symbol’s level.

In more general terms, the information carried by any one symbol of a \( q \)-level source, where all the levels are equiprobable with probabilities of \( p_i = \frac{1}{q}, i = 1 \ldots q, \) is defined as

\[
I = \log_2 q. \tag{1.7}
\]

Rewriting Equation 1.7 using the message probabilities \( p_i = \frac{1}{q} \) yields a more convenient form:

\[
I = \log_2 \frac{1}{p_i} = -\log_2 p_i, \tag{1.8}
\]

which now is also applicable in case of arbitrary, unequal message probabilities \( p_i \), again, implying the plausible fact that the lower the probability of a certain source symbol, the higher the information conveyed by its occurrence. Observe, however, that for unquantized analog sources, where as regards to the number of possible source symbols we have \( q \to \infty \) and hence the probability of any analog sample becomes infinitesimally low, these definitions become meaningless.

Let us now consider a sequence of \( N \) consecutive \( q \)-ary symbols. This sequence can take \( q^N \) number of different values, delivering \( q^N \) different messages. Therefore, the information carried by one such sequence is:

\[
I_N = \log_2 (q^N) = N \log_2 q, \tag{1.9}
\]
which is in perfect harmony with our expectation, delivering $N$ times the information of a single symbol, which was quantified by Equation 1.7. Doubling the sequence length to $2N$ carries twice the information, as suggested by:

$$I_{2N} = \log_2(q^{2N}) = 2N \cdot \log_2 q.$$  

(1.10)

Before we proceed, let us briefly summarize the basic properties of information following Shannon’s work [13–16,24,25]:

- If for the probability of occurrences of the symbols $j$ and $k$ we have $p_j < p_k$, then as regards the information carried by them we have: $I(k) < I(j)$.

- If in the limit we have $p_k \to 1$, then for the information carried by the symbol $k$ we have $I(k) \to 0$, implying that symbols, whose probability of occurrence tends to unity, carry no information.

- If the symbol probability is in the range of $0 \leq p_k \leq 1$, then as regards the information carried by it we have $I(k) \geq 0$.

- For independent messages $k$ and $j$, their joint information is given by the sum of their information: $I(k,j) = I(k) + I(j)$. For example, the information carried by the statement “My son is 14 years old and my daughter is 12” is equivalent to that of the sum of these statements: “My son is 14 years old” and “My daughter is 12 years old.”

- In harmony with our expectation, if we have two equiprobable messages 0 and 1 with probabilities, $p_1 = p_2 = \frac{1}{2}$, then from Equation 1.8 we have $I(0) = I(1) = 1$ bit.

### 1.4 Average Information of Discrete Memoryless Sources

Following Shannon’s approach [13–16,24,25], let us now consider a source emitting one of $q$ possible symbols from the alphabet $s = s_1, s_2, \ldots, s_i, \ldots, s_q$ having symbol probabilities of $p_i$, $i = 1, 2, \ldots, q$. Suppose that a long message of $N$ symbols constituted by symbols from the alphabet $s = s_1, s_2, \ldots, s_q$ having symbol probabilities of $p_i$ is to be transmitted. Then the symbol $s_i$ appears in every $N$-symbol message on the average $p_i \cdot N$ number of times, provided the message length is sufficiently long. The information carried by symbol $s_i$ is $\log_2 1/p_i$ and its $p_i \cdot N$ occurrences yield an information contribution of

$$I(i) = p_i \cdot N \cdot \log_2 \frac{1}{p_i}.$$  

(1.11)

Upon summing the contributions of all the $q$ symbols, we acquire the total information carried by the $N$-symbol sequence:

$$I = \sum_{i=1}^{q} p_i N \cdot \log_2 \frac{1}{p_i} \text{ [bits].}$$  

(1.12)
1.4. AVERAGE INFORMATION OF DISCRETE MEMORYLESS SOURCES

Averaging this over the \( N \) symbols of the sequence yields the average information per symbol, which is referred to as the source’s entropy [14]:

\[
H = \frac{1}{N} \sum_{i=1}^{q} p_i \cdot \log_2 \frac{1}{p_i} = - \sum_{i=1}^{q} p_i \log_2 p_i \quad \text{[bit/symbol].} \quad (1.13)
\]

Then the average source information rate can be defined as the product of the information carried by a source symbol, given by the entropy \( H \) and the source emission rate \( R_s \):

\[
R = R_s \cdot H \quad \text{[bits/sec].} \quad (1.14)
\]

Observe that Equation 1.13 is analogous to the discrete form of the first moment or in other words the mean of a random process with a probability density function (PDF) of \( p(x) \), as in

\[
\bar{x} = \int_{-\infty}^{\infty} x \cdot p(x) dx, \quad (1.15)
\]

where the averaging corresponds to the integration, and the instantaneous value of the random variable \( x \) represents the information \( \log_2 p_i \) carried by message \( i \), which is weighted by its probability of occurrence \( p_i \) quantified for a continuous variable \( x \) by \( p(x) \).

1.4.1 Maximum Entropy of a Binary Source

Let us assume that a binary source, for which \( q = 2 \), emits two symbols with probabilities \( p_1 = p \) and \( p_2 = (1 - p) \), where the sum of the symbol probabilities must be unity. In order to quantify the maximum average information of a symbol from this source as a function of the symbol probabilities, we note from Equation 1.13 that the entropy is given by:

\[
H(p) = -p \cdot \log_2 p - (1 - p) \cdot \log_2 (1 - p). \quad (1.16)
\]

As in any maximization problem, we set \( \partial H(p) / \partial p = 0 \), and upon using the differentiation chain rule of \( (u \cdot v)' = u' \cdot v + u \cdot v' \) as well as exploiting that \( \log_a x' = \frac{1}{x} \log_a e \)

we arrive at:

\[
\frac{\partial H(p)}{\partial p} = -\log_2 p - \frac{p}{p} \log_2 e + \log_2 (1 - p) + \frac{(1 - p)}{(1 - p)} \log_2 e = 0
\]

\[
\log_2 p = \log_2 (1 - p)
\]

\[
p = (1 - p)
\]

\[
p = 0.5.
\]

This result suggests that the entropy is maximum for equiprobable binary messages. Plotting Equation 1.16 for arbitrary \( p \) values yields Figure 1.5, in which Shannon suggested that the average information carried by a symbol of a binary source is low, if one of the symbols has a high probability, while the other a low probability.
Example: Let us compute the entropy of the binary source having message
probabilities of $p_1 = \frac{1}{8}, p_2 = \frac{7}{8}$.

The entropy is expressed as:

$$H = -\frac{1}{8}\log_2\frac{1}{8} - \frac{7}{8}\log_2\frac{7}{8}.$$ 

Exploiting the following equivalence:

$$\log_2(x) = \log_{10}(x) \cdot \log_2(10) \approx 3.322 \cdot \log_{10}(x)$$  \hspace{1cm} (1.17)

we have:

$$H \approx \frac{3}{8} - \frac{7}{8} \cdot 3.322 \cdot \log_{10}\frac{7}{8} \approx 0.54 \text{ [bit/symbol]},$$

again implying that if the symbol probabilities are rather different, the entropy becomes significantly lower than the achievable 1 bit/symbol. This is because the probability of encountering the more likely symbol is so close to unity that it carries hardly any information, which cannot be compensated by the more “informative” symbol’s reception. For the even more unbalanced situation of $p_1 = 0.1$ and $p_2 = 0.9$ we have:

$$H = -0.1 \log_2 0.1 - 0.9 \cdot \log_2 0.9$$
$$\approx -(0.3322 \cdot \log_{10} 0.1 + 0.9 \cdot 3.322 \cdot \log_{10} 0.9)$$
$$\approx 0.3322 + 0.1368$$
$$\approx 0.47 \text{ [bit/symbol]}.$$ 

In the extreme case of $p_1 = 0$ or $p_2 = 1$ we have $H = 0$. As stated before, the average source information rate is defined as the product of the information
carried by a source symbol, given by the entropy $H$ and the source emission rate $R_s$, yielding $R = R_s \cdot H$ [bits/sec]. Transmitting the source symbols via a perfect noiseless channel yields the same received sequence without loss of information.

### 1.4.2 Maximum Entropy of a $q$-ary Source

For a $q$-ary source the entropy is given by:

$$H = -\sum_{i=1}^{q} p_i \log_2 p_i,$$  \hspace{1cm} (1.18)

where, again, the constraint $\sum p_i = 1$ must be satisfied. When determining the extreme value of the above expression for the entropy $H$ under the constraint of $\sum p_i = 1$, the following term has to be maximized:

$$D = \sum_{i=1}^{q} -p_i \log_2 p_i + \lambda \cdot \left[ 1 - \sum_{i=1}^{q} p_i \right],$$  \hspace{1cm} (1.19)

where $\lambda$ is the so-called Lagrange multiplier. Following the standard procedure of maximizing an expression, we set:

$$\frac{\partial D}{\partial p_i} = -\log_2 p_i - \frac{p_i}{p_i} \cdot \log_2 e - \lambda = 0$$

leading to

$$\log_2 p_i = -(\log_2 e + \lambda) = \text{Constant for } i = 1 \ldots q,$$

which implies that the maximum entropy of a $q$-ary source is maintained, if all message probabilities are identical, although at this stage the value of this constant probability is not explicit. Note, however, that the message probabilities must sum to unity, and hence:

$$\sum_{i=1}^{q} p_i = 1 = q \cdot a,$$  \hspace{1cm} (1.20)

where $a$ is a constant, leading to $a = 1/q = p_i$, implying that the entropy of any $q$-ary source is maximum for equiprobable messages. Furthermore, $H$ is always bounded according to:

$$0 \leq H \leq \log_2 q.$$  \hspace{1cm} (1.21)

### 1.5 Source Coding for a Discrete Memoryless Source

Interpreting Shannon’s work [13–16,24,25] further, we see that source coding is the process by which the output of a $q$-ary information source is converted to a binary
sequence for transmission via binary channels, as seen in Figure 1.1. When a discrete memories source generates $g$-ary equiprobable symbols with an average information rate of $R = R_s \log_2 g$, all symbols convey the same amount of information, and efficient signaling takes the form of binary transmissions at a rate of $R$ bps. When the symbol probabilities are unequal, the minimum required source rate for distortionless transmission is reduced to

$$R = R_s \cdot H < R_s \log_2 q.$$  \hfill (1.22)

Then the transmission of a highly probable symbol carries little information and hence assigning $\log_2 q$ number of bits to it does not use the channel efficiently. What can be done to improve transmission efficiency? Shannon's source coding theorem suggests that by using a source encoder before transmission the efficiency of the system with equiprobable source symbols can be arbitrarily approached.

Coding efficiency can be defined as the ratio of the source information rate and the average output bit rate of the source encoder. If this ratio approaches unity, implying that the source encoder's output rate is close to the source information rate, the source encoder is highly efficient. There are many source encoding algorithms, but the most powerful approach suggested was Shannon's method [13], which is best illustrated by means of the following example, portrayed in Table 1.1, Algorithm 1, and Figure 1.6.

### 1.5.1 Shannon-Fano Coding

The Shannon-Fano coding algorithm is based on the simple concept of encoding frequent messages using short codewords and infrequent ones by long codewords, while reducing the average message length. This algorithm is part of virtually all treatises dealing with information theory, such as, for example, Carlson's work [20]. The formal coding steps listed in Algorithm 1 and in the flowchart of Figure 1.6 can be readily followed in the context of a simple example in Table 1.1. The average codeword length is given by weighting the length of any codeword by its probability, yielding:

$$(0.27 + 0.2) \cdot 2 + (0.17 + 0.16) \cdot 3 + 2 \cdot 0.06 \cdot 4 + 2 \cdot 0.04 \cdot 4 \approx 2.73 \text{ [bit].}$$

The entropy of the source is:

$$H = -\sum_i p_i \log_2 p_i$$

$$= -(\log_2 10) \sum_i p_i \log_{10} p_i$$

$$\approx -3.322 \cdot [0.27 \cdot \log_{10} 0.27 + 0.2 \cdot \log_{10} 0.2$$

$$+0.17 \cdot \log_{10} 0.17 + 0.16 \cdot \log_{10} 0.16$$

$$+2 \cdot 0.06 \cdot \log_{10} 0.06 + 2 \cdot 0.04 \cdot \log_{10} 0.04]$$

$$\approx 2.691 \text{ [bit/symbol].}$$
Algorithm 1 (Shannon-Fano Coding) *This algorithm summarizes the Shannon-Fano coding steps. (See also Figure 1.6 and Table 1.1.)*

1. The source symbols $S_0 \ldots S_7$ are first sorted in descending order of probability of occurrence.

2. Then the symbols are divided into two subgroups so that the subgroup probabilities are as close to each other as possible. This is symbolized by the horizontal divisions in Table 1.1.

3. When allocating codewords to represent the source symbols, we assign a logical zero to the top subgroup and logical one to the bottom subgroup in the appropriate column under ‘coding steps.’

4. If there is more than one symbol in the subgroup, this method is continued until no further divisions are possible.

5. Finally, the variable-length codewords are output to the channel.

<table>
<thead>
<tr>
<th>Symb.</th>
<th>Prob.</th>
<th>Coding Steps</th>
<th>Codeword</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_0$</td>
<td>0.27</td>
<td>0 0</td>
<td>00</td>
</tr>
<tr>
<td>$S_1$</td>
<td>0.20</td>
<td>0 1</td>
<td>01</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.17</td>
<td>1 0 0</td>
<td>100</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.16</td>
<td>1 0 1</td>
<td>101</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0.06</td>
<td>1 1 0 0</td>
<td>1100</td>
</tr>
<tr>
<td>$S_5$</td>
<td>0.06</td>
<td>1 1 0 1</td>
<td>1101</td>
</tr>
<tr>
<td>$S_6$</td>
<td>0.04</td>
<td>1 1 1 0</td>
<td>1110</td>
</tr>
<tr>
<td>$S_7$</td>
<td>0.04</td>
<td>1 1 1 1</td>
<td>1111</td>
</tr>
</tbody>
</table>

Table 1.1: Shannon-Fano Coding Example Based on Algorithm 1 and Figure 1.6

Since the average codeword length of 2.73 bit/symbol is very close to the entropy of 2.691 bit/symbol, a high coding efficiency is predicted, which can be computed as:

$$E \approx \frac{2.691}{2.73} \approx 98.6\%.$$ 

The straightforward 3 bit/symbol binary coded decimal (BCD) assignment gives an efficiency of:

$$E \approx \frac{2.691}{3} \approx 89.69\%.$$ 

In summary, Shannon-Fano coding allowed us to create a set of uniquely invertible mappings to a set of codewords, which facilitate a more efficient transmission of the source symbols, than straightforward BCD representations would. This was possible with no coding impairment (i.e., losslessly). Having highlighted the philosophy of the Shannon-Fano noiseless or distortionless coding technique, let us now concentrate on the closely related Huffman coding principle.
CHAPTER 1. INFORMATION THEORY

Sort source symbols in descending order of probabilities

Derive subgroups of near-equal probabilities

Assign zero & one to top and bottom branches, respectively

Yes

More than one symbols are in subgroups?

No

Stop, output encoded symbols

Figure 1.6: Shannon-Fano Coding Algorithm (see also Table 1.1 and Algorithm 1).

1.5.2 Huffman Coding

The Huffman Coding (HC) algorithm is best understood by referring to the flowchart of Figure 1.7 and to the formal coding description of Algorithm 2, while a simple practical example is portrayed in Table 1.2, which leads to the Huffman codes summarized in Table 1.3. Note that we used the same symbol probabilities as in our Shannon-Fano coding example, but the Huffman algorithm leads to a different codeword assignment. Nonetheless, the code's efficiency is identical to that of the Shannon-Fano algorithm.

The symbol-merging procedure can also be conveniently viewed using the example of Figure 1.8, where the Huffman codewords are derived by reading the associated 1 and 0 symbols from the end of the tree backward, that is, toward the source symbols $S_0 \ldots S_7$. Again, these codewords are summarized in Table 1.3.

In order to arrive at a fixed average channel bit rate, which is convenient in many communications systems, a long buffer might be needed, causing storage and delay problems. Observe from Table 1.3 that the Huffman coding algorithm gives code-
Algorithm 2 (Huffman Coding)  

This algorithm summarizes the Huffman coding steps.

1. Arrange the symbol probabilities $p_i$ in decreasing order and consider them as "leaf-nodes," as suggested by Table 1.2.

2. While there is more than one node, merge the two nodes having the lowest probability and assign 0/1 to the upper/lower branches, respectively.

3. Read the assigned "transition bits" on the branches from top to bottom in order to derive the codewords.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_0$</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$S_0$</td>
<td>-</td>
</tr>
<tr>
<td>$S_1$</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$S_1$</td>
<td>-</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.17</td>
<td>0</td>
<td>0.33</td>
<td></td>
<td></td>
<td>$S_{23}$</td>
<td>0</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.16</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0.06</td>
<td>0</td>
<td>0.12</td>
<td>0</td>
<td>0.20</td>
<td>$S_{4567}$</td>
<td>0</td>
</tr>
<tr>
<td>$S_5$</td>
<td>0.06</td>
<td>1</td>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
<td>01</td>
</tr>
<tr>
<td>$S_6$</td>
<td>0.04</td>
<td>0</td>
<td>0.08</td>
<td>1</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>$S_7$</td>
<td>0.04</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>11</td>
</tr>
</tbody>
</table>

Table 1.2: Huffman Coding Example Based on Algorithm 2 and Figure 1.7 (for final code assignment see Table 1.3)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{23}$</td>
<td>0.33</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td>1.0</td>
<td>00</td>
</tr>
<tr>
<td>$S_0$</td>
<td>0.27</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>01</td>
</tr>
<tr>
<td>$S_1$</td>
<td>0.20</td>
<td>0</td>
<td>0.4</td>
<td>1</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>$S_{4567}$</td>
<td>0.20</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>11</td>
</tr>
</tbody>
</table>

Table 1.3: Huffman Coding Example Summary of Table 1.2
Figure 1.8: Tree-based Huffman coding example.
words that can be uniquely decoded, which is a crucial prerequisite for its practical employment. This is because no codeword can be a prefix of any longer one. For example, for the following sequence of codewords ..., 00, 10, 010, 110, 1111, ... the source sequence of ...$S_0, S_1, S_2, S_3, S_8$... can be uniquely inferred from Table 1.3.

In our discussions so far, we have assumed that the source symbols were completely independent of each other. Such a source is usually referred to as a memoryless source. By contrast, sources where the probability of a certain symbol also depends on what the previous symbol was are often termed *sources exhibiting memory*. These sources are typically bandlimited sample sequences, such as, for example, a set of correlated or "similar-magnitude" speech samples or adjacent video pixels. Let us now consider sources that exhibit memory.

## 1.6 Average Information of Discrete Sources Exhibiting Memory

Let us invoke Shannon’s approach [13–16,24,25] in order to illustrate sources with and without memory. Let us therefore consider an uncorrelated random white Gaussian noise (WGN) process, which was passed through a low-pass filter. The corresponding autocorrelation functions (ACF) and power spectral density (PSD) functions were portrayed in Figures 1.3 and 1.4. Observe in the figures that through low-pass filtering a WGN process introduces correlation by limiting the rate at which amplitude changes are possible, smoothing the amplitude of abrupt noise peaks. This example suggests that all bandlimited signals are correlated over a finite interval. Most analog source signals, such as speech and video, are inherently correlated, owing to physical restrictions imposed on the analog source. Hence all practical analog sources possess some grade of memory, a property that is also retained after sampling and quantization. An important feature of sources with memory is that they are predictable to a certain extent, hence, they can usually be more efficiently encoded than unpredictable sources having no memory.

### 1.6.1 Two-State Markov Model for Discrete Sources Exhibiting Memory

Let us now introduce a simple analytically tractable model for treating sources that exhibit memory. Predictable sources that have memory can be conveniently modeled by *Markov processes*. A source having a memory of one symbol interval directly “remembers” only the previously emitted source symbol and depending on this previous symbol it emits one of its legitimate symbols with a certain probability, which depends explicitly on the state associated with this previous symbol. A one-symbol-memory model is often referred to as a *first-order model*. For example, if in a first-order model the previous symbol can take only two different values, we have two different states, and this simple two-state first-order Markov model is characterized by the state transition diagram of Figure 1.9. Previously, in the context of Shannon-Fano and Huffman coding of memoryless information sources, we used the notation of $S_i, i = 0,1,...$ for the various symbols to be encoded. In this section, we are dealing with sources
exhibiting memory and hence for the sake of distinction we use the symbol notation of \( X_i, i = 1, 2, \ldots \). If, for the sake of illustration, the previous emitted symbol was \( X_1 \), the state machine of Figure 1.9 is in state \( X_1 \), and in the current signaling interval it can generate one of two symbols, namely, \( X_1 \) and \( X_2 \), whose probability depends explicitly on the previous state \( X_1 \). However, not all two-state Markov models are as simple as that of Figure 1.9, since the transitions from state \( X_1 \) to \( X_2 \) are not necessarily associated with emitting the same symbol as the transitions from state \( X_2 \) to \( X_1 \). Thus more elaborate example will be considered later in this chapter.

Observe in Figure 1.9 that the corresponding transition probabilities from state \( X_1 \) are given by the conditional probabilities \( p_{12} = P(X_2/X_1) \) and \( p_{11} = P(X_1/X_1) = 1 - P(X_2/X_1) \). Similar findings can be observed as regards state \( X_2 \). These dependencies can also be stated from a different point of view as follows. The probability of occurrence of a particular symbol depends not only on the symbol itself, but also on the previous symbol emitted. Thus, the symbol entropy for state \( X_1 \) and \( X_2 \) will now be characterized by means of the conditional probabilities associated with the transitions merging in these states. Explicitly, the symbol entropy for state \( X_i, i = 1, 2 \) is given by:

\[
H_i = \sum_{j=1}^{2} p_{ij} \cdot \log_2 \frac{1}{p_{ij}} \quad i = 1, 2
\]

yielding the symbol entropies, that is, the average information carried by the symbols emitted in states \( X_1 \) and \( X_2 \), respectively, as:

\[
H_1 = p_{11} \cdot \log_2 \frac{1}{p_{11}} + p_{12} \cdot \log_2 \frac{1}{p_{12}}
\]

\[
H_2 = p_{21} \cdot \log_2 \frac{1}{p_{21}} + p_{22} \cdot \log_2 \frac{1}{p_{22}}.
\]

Both symbol entropies, \( H_1 \) and \( H_2 \), are characteristic of the average information conveyed by a symbol emitted in state \( X_1 \) and \( X_2 \), respectively. In order to compute the overall entropy \( H \) of this source, they must be weighted by the probability of occurrence, \( P_1 \) and \( P_2 \), of these states:

\[
H = \sum_{i=1}^{2} P_i H_i
\]

\[
= \sum_{i=1}^{2} P_i \sum_{j=1}^{2} p_{ij} \log_2 \frac{1}{p_{ij}}.
\]

Assuming a highly predictable source having high adjacent sample correlation, it is plausible that once the source is in a given state, it is more likely to remain in that state than to traverse into the other state. For example, assuming that the state machine of Figure 1.9 is in state \( X_1 \) and the source is a highly correlated, predictable source, we are likely to observe long runs of \( X_1 \). Conversely, once in state \( X_2 \), long strings of \( X_2 \) symbols will typically follow.
\begin{align*}
p_{12} &= P(X_2/X_1) \\
p_{21} &= P(X_1/X_2) \\
p_{11} &= P(X_1/X_1) = 1 - P(X_2/X_1) \\
p_{22} &= 1 - P(X_1/X_2) = P(X_2/X_2)
\end{align*}

Figure 1.9: Two-state first-order Markov model.

### 1.6.2 \(N\)-State Markov Model for Discrete Sources Exhibiting Memory

In general, assuming \(N\) legitimate states, (i.e., \(N\) possible source symbols) and following similar arguments, Markov models are characterised by their state probabilities \(P(X_i), i = 1 \ldots N\), where \(N\) is the number of states, as well as by the transition probabilities \(p_{ij} = P(X_i/X_j)\), where \(p_{ij}\) explicitly indicates the probability of traversing from state \(X_j\) to state \(X_i\). Their further basic feature is that they emit a source symbol at every state transition, as will be shown in the context of an example presented in Section 1.7. Similarly to the two-state model, we define the entropy of a source having memory as the weighted average of the entropy of the individual symbols emitted from each state, where weighting is carried out taking into account the probability of occurrence of the individual states, namely \(P_i\). In analytical terms, the symbol entropy for state \(X_i, i = 1 \ldots N\) is given by:

\[
H_i = \sum_{j=1}^{N} p_{ij} \cdot \log_2 \frac{1}{p_{ij}} \quad i = 1 \ldots N. \tag{1.27}
\]

The averaged, weighted symbol entropies give the source entropy:

\[
H = \sum_{i=1}^{N} P_i H_i \\
= \sum_{i=1}^{N} P_i \sum_{j=1}^{N} p_{ij} \log_2 \frac{1}{p_{ij}}. \tag{1.28}
\]

Finally, assuming a source symbol rate of \(v_s\), the average information emission rate \(R\) of the source is given by:

\[
R = v_s \cdot H \quad [\text{bps}]. \tag{1.29}
\]
1.7 Examples

1.7.1 Two-State Markov Model Example

As mentioned in the previous section, we now consider a slightly more sophisticated Markov model, where the symbols emitted upon traversing from state $X_1$ to $X_2$ are different from those when traversing from state $X_2$ to $X_1$. More explicitly:

- Consider a discrete source that was described by the two-state Markov model of Figure 1.9, where the transition probabilities are

  $$
  p_{11} = P(X_1/X_1) = 0.9 \quad p_{22} = P(X_2/X_2) = 0.1
  $$

  $$
  p_{12} = P(X_1/X_2) = 0.1 \quad p_{21} = P(X_2/X_1) = 0.9,
  $$

  while the state probabilities are

  $$
  P(X_1) = 0.8 \quad \text{and} \quad P(X_2) = 0.2.
  $$

  (1.30)

The source emits one of four symbols, namely, $a$, $b$, $c$, and $d$, upon every state transition, as seen in Figure 1.10. Let us find

(a) the source entropy and

(b) the average information content per symbol in messages of one, two, and three symbols.

- **Message Probabilities**

  Let us consider two sample sequences $acb$ and $aab$. As shown in Figure 1.10, the transitions leading to $acb$ are $(1 \sim 1)$, $(1 \sim 2)$, and $(2 \sim 2)$. The probability of encountering this sequence is $0.8 \cdot 0.9 \cdot 0.1 \cdot 0.1 = 0.0072$. The sequence $aab$ has a
probability of zero because the transition from \( a \) to \( b \) is illegal. Further path (i.e., message) probabilities are tabulated in Table 1.4 along with the information of \( H = - \log_2 P \) of all the legitimate messages.

- **Source Entropy**
  - According to Equation 1.27, the entropy of symbols \( X_1 \) and \( X_2 \) is computed as follows:
    \[
    H_1 = -p_{12} \cdot \log_2 p_{12} - p_{11} \cdot \log_2 p_{11} = 0.1 \cdot \log_2 10 + 0.9 \cdot \log_2 \frac{1}{0.9} \approx 0.469 \text{ bit/symbol} \tag{1.31}
    \]
    \[
    H_2 = -p_{21} \cdot \log_2 p_{21} - p_{22} \cdot \log_2 p_{22} \approx 0.469 \text{ bit/symbol} \tag{1.32}
    \]
    - Then their weighted average is calculated using the probability of occurrence of each state in order to derive the average information per message for this source:
      \[
      H \approx 0.8 \cdot 0.469 + 0.2 \cdot 0.469 \approx 0.469 \text{ bit/symbol.}
      \]
    - The average information per symbol \( I_2 \) in two-symbol messages is computed from the entropy \( H_2 \) of the two-symbol messages as follows:
      \[
      h_2 = \sum_{1}^{8} P_{\text{symbol}} \cdot I_{\text{symbol}} = P_{aa} \cdot I_{aa} + P_{ac} \cdot I_{ac} + \ldots + P_{dc} \cdot I_{dc} \approx 1.66 \text{ bits/2 symbols.} \tag{1.33}
      \]
1.7. EXAMPLES

giving $I_2 = h_2/2 \approx 0.83$ bits/symbol information on average upon receiving a two-symbol message.

- There are eight two-symbol messages; hence, the maximum possible information conveyed is $\log_2 8 = 3$ bits/2 symbols, or 1.5 bits/symbol. However, since the symbol probabilities of $P_1 = 0.8$ and $P_2 = 0.2$ are fairly different, this scheme has a significantly lower conveyed information per symbol, namely, $I_2 \approx 0.83$ bits/symbol.

- Similarly, one can find the average information content per symbol for arbitrarily long messages of concatenated source symbols. For one-symbol messages we have:

\[
I_1 = h_1 = \sum_{1}^{4} P_{\text{symbol}} \cdot I_{\text{symbol}}
= P_a \cdot I_a + \ldots + P_d \cdot I_d
\approx 0.72 \times 0.474 + \ldots + 0.18 \times 2.474
\approx 0.341 + 0.113 + 0.292 + 0.445
\approx 1.191 \text{ bit/symbol.}
\]

We note that the maximum possible information carried by one-symbol messages is $h_{1\text{max}} = \log_2 4 = 2$ bit/symbol, since there are four one-symbol messages in Table 1.4.

• Observe the important tendency, in which, when sending longer messages of dependent sources, the average information content per symbol is reduced. This is due to the source’s memory, since consecutive symbol emissions are dependent on previous ones and hence do not carry as much information as independent source symbols. This becomes explicit by comparing $I_1 \approx 1.191$ and $I_2 \approx 0.83$ bits/symbol.

• Therefore, expanding the message length to be encoded yields more efficient coding schemes, requiring a lower number of bits, if the source has a memory. This is the essence of Shannon's source coding theorem.

1.7.2 Four-State Markov Model for a 2-Bit Quantizer

Let us now augment the previously introduced two-state Markov-model concepts with the aid of a four-state example. Let us assume that we have a discrete source constituted by a 2-bit quantizer, which is characterized by Figure 1.11. Assume further that due to bandlimitation only transitions to adjacent quantization intervals are possible, since the bandlimitation restricts the input signal’s rate of change. The probability of the signal samples residing in intervals 1–4 is given by:

\[
P(1) = P(4) = 0.1, \quad P(2) = P(3) = 0.4.
\]

The associated state transition probabilities are shown in Figure 1.11, along with the quantized samples $a, b, c,$ and $d$, which are transmitted when a state transition
takes place, that is, when taking a new sample from the analog source signal at the sampling-rate $f_s$.

Although we have stipulated a number of simplifying assumptions, this example attempts to illustrate the construction of Markov models in the context of a simple practical problem. Next we construct a simpler example for augmenting the underlying concepts and set aside the above four-state Markov-model example as a potential exercise for the reader.

1.8 Generating Model Sources

1.8.1 Autoregressive Model

In evaluating the performance of information processing systems, such as encoders and predictors, it is necessary to have “standardized” or easily described model sources. Although a set of semistandardized speech and images test sequences is widely used by researchers in codec performance testing, in contrast to analytical model sources, real speech or image sources cannot be used in analytical studies. A widely used analytical model source is the Autoregressive (AR) model. A zero mean random sequence $y(n)$
is called an AR process of order $p$, if it is generated as follows:

$$y(n) = \sum_{k=1}^{p} a_k y(n-k) + \epsilon(n), \forall n,$$  (1.35)

where $\epsilon(n)$ is an uncorrelated zero-mean, random input sequence with variance $\sigma^2$; that is,

$$E\{\epsilon(n)\} = 0$$
$$E\{\epsilon^2(n)\} = \sigma^2$$
$$E\{\epsilon(n) \cdot y(m)\} = 0.$$  (1.36)

From Equation 1.35 we surmise that an AR system recursively generates the present output from $p$ number of previous output samples given by $y(n-k)$ and the present random input sample $\epsilon(n)$.

### 1.8.2 AR Model Properties

AR models are very useful in studying information processing systems, such as speech and image codecs, predictors, and quantizers. They have the following basic properties:

1. The first term of Equation 1.35, which is repeated here for convenience,

$$\hat{y}(n) = \sum_{k=1}^{p} a_k y(n-k)$$

defines a predictor, giving an estimate $\hat{y}(n)$ of $y(n)$, which is associated with the minimum mean squared error between the two quantities.

2. Although $\hat{y}(n)$ and $y(n)$ depend explicitly only on the past $p$ number of samples of $y(n)$, through the recursive relationship of Equation 1.35 this entails the entire past of $y(n)$. This is because each of the previous $p$ samples depends on their predecessors.

3. Then Equation 1.35 can be written in the form of:

$$y(n) = \hat{y}(n) + \epsilon(n),$$  (1.37)

where $\epsilon(n)$ is the prediction error and $\hat{y}(n)$ is the minimum variance prediction estimate of $y(n)$.

4. Without proof, we state that for a random Gaussian distributed prediction error sequence $\epsilon(n)$ these properties are characteristic of a $p^{th}$ order Markov process portrayed in Figure 1.12. When this model is simplified for the case of $p = 1$, we arrive at the schematic diagram shown in Figure 1.13.
5. The power spectral density (PSD) of the prediction error sequence $e(n)$ is that of a random "white-noise" sequence, containing all possible frequency components with the same energy. Hence, its autocorrelation function (ACF) is the Kronecker delta function, given by the Wiener-Khintchine theorem:

$$E\{e(n) \cdot e(m)\} = \sigma^2 \delta(n - m).$$  \hspace{1cm} (1.38)

### 1.8.3 First-Order Markov Model

A variety of practical information sources are adequately modeled by the analytically tractable first-order Markov model depicted in Figure 1.13, where the prediction order is $p = 1$. With the aid of Equation 1.35 we have

$$y(n) = e(n) + ay(n-1),$$
where \( a \) is the adjacent sample correlation of the process \( y(n) \). Using the following recursion:

\[
y(n - 1) = \varepsilon(n - 1) + a_1 y(n - 2)
\]

\[
\vdots \quad \vdots \quad \vdots
\]

\[
y(n - k) = \varepsilon(n - k) + a_1 y(n - k - 1)
\]  

(1.39)

we arrive at:

\[
y(n) = \varepsilon(n) + a_1 \varepsilon(n - 1) + a_1^2 y(n - 2)
\]

which can be generalized to:

\[
y(n) = \sum_{j=0}^{\infty} a^j \varepsilon(n - j).
\]  

(1.40)

Clearly, Equation 1.40 describes the first-order Markov process by the help of the adjacent sample correlation \( a_1 \) and the uncorrelated zero-mean random Gaussian process \( \varepsilon(n) \).

1.9 Run-Length Coding for Discrete Sources Exhibiting Memory

1.9.1 Run-Length Coding Principle [20]

For discrete sources having memory, (i.e., possessing intersample correlation), the coding efficiency can be significantly improved by predictive coding, allowing the required transmission rate and hence the channel bandwidth to be reduced. Particularly amenable to run-length coding are binary sources with inherent memory, such as black and white documents, where the predominance of white pixels suggests that a Run-Length-Coding (RLC) scheme, which encodes the length of zero runs, rather than repeating long strings of zeros, provides high coding efficiency.

Following Carlson’s interpretation [20], a predictive RLC scheme can be constructed according to Figure 1.14. The q-ary source messages are first converted to binary bit format. For example, if an 8-bit analog-digital converter (ADC) is used, the 8-bit digital samples are converted to binary format. This bit-stream, \( x(i) \), is then compared with the output signal of the predictor, \( \hat{x}(i) \), which is fed with the prediction error signal \( e(i) \). The comparator is a simple mod-2 gate, outputting a logical 1, whenever the prediction fails; that is, the predictor’s output is different from the incoming bit \( x(i) \). If, however, \( x(i) = \hat{x}(i) \), the comparator indicates this by outputting a logical 0. For highly correlated signals from sources with significant memory the predictions are usually correct, and hence long strings of 0 runs are emitted, interspersed with an occasional 1. Thus, the prediction error signal \( e(i) \) can be efficiently run-length encoded by noting and transmitting the length of zero runs.

The corresponding binary run-length coding principle becomes explicit from Table 1.5 and from our forthcoming coding efficiency analysis.
Following Jain’s interpretation [26], let us now investigate the RLC efficiency by assuming that a run of \( r \) successive logical 0s is followed by a 1. Instead of directly transmitting these strings, we represent such a string as an \( n \)-bit word giving the length of the 0-run between successive logical ones. When a 0-run longer than \( N = 2^n - 1 \) bits occurs, this is signaled as the all 1 codeword, informing the decoder to wait for the next RLC codeword before releasing the decoded sequence. Again, the scheme’s operation is characterized by Table 1.5. Clearly, data compression is achieved if the average number of 0 data bits per run \( d \) is higher than the number of bits, \( n \), required to encode the 0-run length. Let us therefore compute the average number of bits per run without RLC. If a run of \( r \) logical zeros are followed by a 1, the run-length is \( (r + 1) \). The expected or mean value of \( (r + 1) \), namely, \( d = (r + 1) \), is calculated by weighting each specific \( (r + 1) \) with its probability of occurrence that
1.9. RUN-LENGTH CODING

Figure 1.15: CDF and PDF of the geometric distribution of run-length \( l \).

is, with its discrete PDF \( c(r) \) and then averaging the weighted components, in:

\[
d = (r + 1) = \sum_{r=0}^{N-1} (r + 1) \cdot c(r) + Nc(N). \tag{1.41}
\]

The PDF of a run of \( r \) zeros followed by a 1 is given by:

\[
c(r) = \begin{cases} 
p^r(1-p) & 0 \leq r \leq N-1 \\
p^N & r = N,
\end{cases} \tag{1.42}
\]

since the probability of \( N \) consecutive zeros is \( p^N \) if \( r = N \), while for shorter runs the joint probability of \( r \) zeros followed by a 1 is given by \( p^r \cdot (1-p) \). The PDF and CDF of this distribution are shown in Figure 1.15 for \( p = 0.9 \) and \( p = 0.1 \), where \( p \) represents the probability of a logical zero bit. Substituting Equation 1.42 in Equation 1.41 gives:

\[
d = N \cdot p^N + \sum_{r=0}^{N-1} (r + 1) \cdot p^r \cdot (1-p)
\]

\[
= N \cdot p^N + 1 \cdot p^0 \cdot (1-p) + 2 \cdot p \cdot (1-p) + \ldots + N \cdot p^{N-1} \cdot (1-p)
\]

\[
= N \cdot p^N + 1 + 2p + 3p^2 + \ldots + N \cdot p^{N-1} - p - 2p^2 \ldots - N \cdot p^N
\]

\[
= 1 + p + p^2 + \ldots + p^{N-1}. \tag{1.43}
\]

Equation 1.43 is a simple geometric progression, given in closed form as:

\[
d = \frac{1 - p^N}{1 - p}. \tag{1.44}
\]

**RLC Example:** Using a run-length coding memory of \( M = 31 \) and a zero symbol probability of \( p = 0.95 \), characterize the RLC efficiency.

Substituting \( N \) and \( p \) into Equation 1.44 for the average run-length we have:

\[
d = \frac{1 - 0.95^{31}}{1 - 0.95} \approx \frac{1 - 0.204}{0.05} \approx 15.92. \tag{1.45}
\]
The compression ratio $C$ achieved by RLC is given by:

$$C = \frac{d}{n} = \frac{1 - p^N}{n(1 - p)} \approx \frac{15.92}{5} \approx 3.18.$$  

(1.46)

The achieved average bit rate is

$$B = \frac{n}{d} \approx 0.314 \text{ bit/pixel},$$

and the coding efficiency is computed as the ratio of the entropy (i.e., the lowest possible bit rate and the actual bit rate). The source entropy is given by:

$$H \approx -0.95 \cdot 3.322 \cdot \log_{10} 0.95 - 0.05 \cdot 3.322 \cdot \log_{10} 0.05$$

$$\approx 0.286 \text{ bit/symbol},$$  

(1.47)

giving a coding efficiency of:

$$E = \frac{H}{B} \approx \frac{0.286}{0.314} \approx 91\%.$$  

This concludes our RLC example.

### 1.10 Information Transmission via Discrete Channels

Let us now return to Shannon's classic references [13-16,24,25] and assume that both the channel and the source are discrete, and let us evaluate the amount of information transmitted via the channel. We define the channel capacity characterizing the channel and show that according to Shannon nearly error-free information transmission is possible at rates below the channel capacity via the binary symmetric channel (BSC). Let us begin our discourse with a simple introductory example.

#### 1.10.1 Binary Symmetric Channel Example

Let us assume that a binary source is emitting a logical 1 with a probability of $P(1) = 0.7$ and a logical 0 with a probability of $P(0) = 0.3$. The channel’s error probability is $p_e = 0.02$. This scenario is characterized by the binary symmetric channel (BSC) model of Figure 1.16. The probability of error-free reception is given by that of receiving 1, when a logical 1 is transmitted plus the probability of receiving a 0 when 0 is transmitted, which is also plausible from Figure 1.16. For example, the first of these two component probabilities can be computed with the aid of Figure 1.16 as the product of the probability $P(1)$ of a logical 1 being transmitted and the conditional probability $P(1|1)$ of receiving a 1, given the condition that a 1 was transmitted:

$$P(Y_1, X_1) = P(X_1) \cdot P(Y_1|X_1)$$  

(1.48)

$$P(1, 1) = P(1) \cdot P(1|1) = 0.7 \cdot 0.98 = 0.686.$$
1.10. INFORMATION TRANSMISSION VIA DISCRETE CHANNELS

Similarly, the probability of the error-free reception of a logical 0 is given by:

\[ P(Y_0, X_0) = P(X_0) \cdot P(Y_0/X_0) \]
\[ P(0, 0) = P(0) \cdot P(0/0) = 0.3 \cdot 0.98 = 0.294, \]

giving the total probability of error-free reception as:

\[ P_{\text{correct}} = P(1, 1) + P(0, 0) = 0.98. \]

Following similar arguments, the probability of erroneous reception is also given by two components. For example, using Figure 1.16, the probability of receiving a 1 when a 0 was transmitted is computed by multiplying the probability \( P(0) \) of a logical 0 being transmitted by the conditional probability \( P(1/0) \) of receiving a logical 1, given the fact that a 0 is known to have been transmitted:

\[ P(Y_1, X_0) = P(X_0) \cdot P(Y_1/X_0) \]
\[ P(1, 0) = P(0) \cdot P(1/0) = 0.3 \cdot 0.02 = 0.006. \]

Conversely,

\[ P(Y_0, X_1) = P(X_1) \cdot P(Y_0/X_1) \]
\[ P(0, 1) = P(1) \cdot P(0/1) = 0.7 \cdot 0.02 = 0.014, \]

yielding a total error probability of:

\[ P_{\text{error}} = P(1, 0) + P(0, 1) = 0.02, \]

which is constituted by the above two possible error events.

Viewing events from a different angle, we observe that the total probability of receiving 1 is that of receiving a transmitted 1 correctly plus a transmitted 0 incorrectly:

\[ P_1 = P(1) \cdot (1 - p_e) + P(0) \cdot p_e \]
\[ = 0.7 \cdot 0.98 + 0.3 \cdot 0.02 = 0.686 + 0.006 = 0.692. \]
Figure 1.17: BSC performance for $p_e = 0, 0.125, 0.25, 0.375,$ and 0.5.

On the same note, the probability of receiving 0 is that of receiving a transmitted 0 correctly plus a transmitted 1 incorrectly:

$$P_0 = P(0) \cdot (1 - p_e) + P(1) \cdot p_e$$

$$= 0.3 \cdot 0.98 + 0.7 \cdot 0.02 = 0.294 + 0.014 = 0.308.$$  

(1.50)

In the next example, we further study the performance of the BSC for a range of different parameters in order to gain a deeper insight into its behavior.

**Example:** Repeat the above calculations for $P(1) = 1, 0.9, 0.5$, and $p_e = 0, 0.1, 0.2, 0.5$ using the BSC model of Figure 1.16. Compute and tabulate the probabilities $P(1,1)$, $P(1,0)$, $P(0,1)$, $P_{\text{correct}}$, $P_{\text{error}}$, $P_1$, and $P_0$ for these parameter combinations, including also their values for the previous example, namely, for $P(1) = 0.7$, $P(0) = 0.3$ and $p_e = 0.02$. Here we neglected the details of the calculations and summarized the results in Table 1.6. Some of the above quantities are plotted for further study in Figure 1.17, which reveals the interdependency of the various probabilities for the interested reader.

Having studied the performance of the BSC, the next question that arises is, how much information can be inferred upon reception of a 1 and a 0 over an imperfect
<table>
<thead>
<tr>
<th>$p_e$</th>
<th>$P(1)$</th>
<th>$P(0)$</th>
<th>$(1 - p_e) = P(1/1)$</th>
<th>$P(1, 1)$</th>
<th>$P(0, 0)$</th>
<th>$(1 - p_e) = P(0/0)$</th>
<th>$P(1, 0)$</th>
<th>$P(0, 0)$</th>
<th>$P_1$</th>
<th>$P_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.1</td>
<td>1</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.3</td>
<td>1</td>
<td>0.7</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.02</td>
<td>1</td>
<td>0</td>
<td>0.98</td>
<td>0.98</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0.02</td>
<td>0.98</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.1</td>
<td>0.98</td>
<td>0.882</td>
<td>0.098</td>
<td>0.02</td>
<td>0.002</td>
<td>0.018</td>
<td>0.884</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.3</td>
<td>0.98</td>
<td>0.686</td>
<td>0.294</td>
<td>0.02</td>
<td>0.006</td>
<td>0.014</td>
<td>0.692</td>
<td>0.308</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
<td>0.98</td>
<td>0.49</td>
<td>0.49</td>
<td>0.02</td>
<td>0.01</td>
<td>0.491</td>
<td>0.509</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>0</td>
<td>0.9</td>
<td>0.9</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.1</td>
<td>0.9</td>
<td>0.81</td>
<td>0.09</td>
<td>0.1</td>
<td>0.01</td>
<td>0.09</td>
<td>0.81</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.3</td>
<td>0.9</td>
<td>0.63</td>
<td>0.27</td>
<td>0.1</td>
<td>0.03</td>
<td>0.07</td>
<td>0.723</td>
<td>0.277</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
<td>0.9</td>
<td>0.45</td>
<td>0.45</td>
<td>0.1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.455</td>
<td>0.545</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0</td>
<td>0.8</td>
<td>0.8</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td>0.2</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.1</td>
<td>0.8</td>
<td>0.72</td>
<td>0.08</td>
<td>0.2</td>
<td>0.02</td>
<td>0.18</td>
<td>0.722</td>
<td>0.278</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.3</td>
<td>0.8</td>
<td>0.56</td>
<td>0.24</td>
<td>0.2</td>
<td>0.06</td>
<td>0.14</td>
<td>0.566</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
<td>0.8</td>
<td>0.40</td>
<td>0.40</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.1</td>
<td>0.5</td>
<td>0.45</td>
<td>0.05</td>
<td>0.5</td>
<td>0.05</td>
<td>0.45</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.3</td>
<td>0.5</td>
<td>0.35</td>
<td>0.15</td>
<td>0.5</td>
<td>0.15</td>
<td>0.35</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.25</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 1.6: BSC Performance Table
Figure 1.18: Forward transition probabilities of the nonideal binary symmetric channel.

(i.e., error-prone) channel. In order to answer this question, let us first generalize the above intuitive findings in the form of Bayes’ rule.

1.10.2 Bayes’ Rule

Let \( Y_j \) represent the received symbols and \( X_i \) the transmitted symbols having probabilities of \( P(Y_j) \) and \( P(X_i) \), respectively. Let us also characterize the forward transition probabilities of the binary symmetric channel as suggested by Figure 1.18.

Then in general, following from the previous introductory example, the joint probability \( P(Y_j, X_i) \) of receiving \( Y_j \), when the transmitted source symbol was \( X_i \), is computed as the probability \( P(X_i) \) of transmitting \( X_i \), multiplied by the conditional probability \( P(Y_j/X_i) \) of receiving \( Y_j \), when \( X_i \) is known to have been transmitted:

\[
P(Y_j, X_i) = P(X_i) \cdot P(Y_j/X_i),
\]

a result that we have already intuitively exploited in the previous example. Since for the joint probabilities \( P(Y_j, X_i) = P(X_i, Y_j) \) holds, we have:

\[
P(X_i, Y_j) = P(Y_j) \cdot P(X_i/Y_j) = P(X_i) \cdot P(Y_j/X_i).
\]

Equation 1.52 is often presented in the form:

\[
P(X_i/Y_j) = \frac{P(X_i, Y_j)}{P(Y_j)} = \frac{P(Y_j) \cdot P(X_i/Y_j)}{P(Y_j)},
\]

which is referred to as Bayes’ rule.

Logically, the probability of receiving a particular \( Y_j = Y_{i_0} \) is the sum of all joint probabilities \( P(X_i, Y_{i_0}) \) over the range of \( X_i \). This corresponds to the probability of
receiving the transmitted $X_i$ correctly, giving rise to the channel output $Y_{i_0}$ plus the sum of the probabilities of all other possible transmitted symbols giving rise to $Y_{i_0}$:

$$P(Y_j) = \sum_X P(X_i, Y_j) = \sum_X P(X_i)P(Y_j/X_i). \quad (1.54)$$

Similarly:

$$P(X_i) = \sum_Y P(X_i, Y_j) = \sum_Y P(Y_j)P(X_i/Y_j). \quad (1.55)$$

### 1.10.3 Mutual Information

In this section, we elaborate further on the ramifications of Shannon's information theory [13-16,24,25]. Over nonideal channels impairments are introduced, and the received information might be different from the transmitted information. In this section, we quantify the amount of information that can be inferred from the received symbols over noisy channels. In the spirit of Shannon's fundamental work [13] and Carlson's classic reference [20], let us continue our discourse with the definition of mutual information. We have already used the notation $P(X_i)$ to denote the probability that the source symbol $X_i$ was transmitted and $P(Y_i)$ to denote the probability that the symbol $Y_j$ was received. The joint probability that $X_i$ was transmitted and $Y_j$ was received had been quantified by $P(X_i, Y_j)$, and $P(X_i/Y_j)$ indicated the conditional probability that $X_i$ was transmitted, given that $Y_j$ was received, while $P(Y_j/X_i)$ was used for the conditional probability that $Y_j$ was received given that $X_i$ was transmitted.

In case of $i = j$, the conditional probabilities $P(Y_j/X_j)j = 1 \cdots q$ represent the error-free transmission probabilities of the source symbols $j = 1 \cdots q$. For example, in Figure 1.18 the probabilities $P(Y_0/X_0)$ and $P(Y_1/X_1)$ are the probabilities of the error-free reception of a transmitted $X_0$ and $X_1$ source symbol, respectively. The probabilities $P(Y_j/X_i)j \neq i$, on the other hand, give the individual error probabilities, which are characteristic of error events that corrupted a transmitted symbol $X_i$ to a received symbol of $Y_j$. The corresponding error probabilities in Figure 1.18 are $P(Y_0/X_1)$ and $P(Y_1/X_0)$.

Let us define the mutual information of $X_i$ and $Y_j$ as:

$$I(X_i, Y_j) = \log_2 \frac{P(X_i/Y_j)}{P(X_i)} = \log_2 \frac{P(X_i, Y_j)}{P(X_i) \cdot P(Y_j)} = \log_2 \frac{P(Y_j/X_i)}{P(Y_j)} \text{ bits}, \quad (1.56)$$

which quantifies the amount of information conveyed, when $X_i$ is transmitted and $Y_j$ is received. Over a perfect, noiseless channel, each received symbol $Y_j$ uniquely identifies a transmitted symbol $X_i$ with a probability of $P(X_i/Y_j) = 1$. Substituting this probability in Equation 1.56 yields a mutual information of:

$$I(X_i, Y_j) = \log_2 \frac{1}{P(X_i)}, \quad (1.57)$$

which is identical to the self-information of $X_i$ and hence no information is lost over the channel. If the channel is very noisy and the error probability becomes 0.5, then
the received symbol $Y_j$ becomes unrelated to the transmitted symbol $X_i$, since for a binary system upon its reception there is a probability of 0.5 that $X_0$ was transmitted and the probability of $X_1$ is also 0.5. Then formally $X_i$ and $Y_j$ are independent and hence

$$P(X_i/Y_j) = \frac{P(X_i,Y_j)}{P(Y_j)} = \frac{P(X_i) \cdot P(Y_j)}{P(Y_j)} = P(X_i),$$

(1.58)
giving a mutual information of:

$$I(X_i,Y_j) = \log_2 \frac{P(X_i)}{P(X_i,Y_j)} = \log_2 1 = 0,$$

(1.59)
implying that no information is conveyed via the channel. Practical communications channels perform between these extreme values and are usually characterized by the average mutual information defined as:

$$I(X,Y) = \sum_{x,y} P(X_i,Y_j) \cdot I(X_i,Y_j)$$

$$= \sum_{x,y} P(X_i,Y_j) \cdot \log_2 \frac{P(X_i,Y_j)}{P(X_i)} \quad \text{[bit/symbol].}$$

(1.60)

Clearly, the average mutual information in Equation 1.60 is computed by weighting each component $I(X_i,Y_j)$ by its probability of occurrence $P(X_i,Y_j)$ and summing these contributions for all combinations of $X_i$ and $Y_j$. The average mutual information $I(X,Y)$ defined above gives the average amount of source information acquired per received symbol, as distinguished from that per source symbol, which was given by the entropy $H(X)$. Let us now consolidate these definitions by working through the following numerical example.

### 1.10.4 Mutual Information Example

Using the same numeric values as in our introductory example as regards to the binary symmetric channel in Section 1.10.1, and exploiting that from Bayes’ rule in Equation 1.53, we have:

$$P(X_i/Y_j) = \frac{P(X_i,Y_j)}{P(Y_j)}.$$

The following probabilities can be derived, which will be used at a later stage, in order to determine the mutual information:

$$P(X_1/Y_1) = P(1/1) = \frac{P(1,1)}{P_1} = \frac{0.686}{0.692} \approx 0.9913$$

and

$$P(X_0/Y_0) = P(0/0) = \frac{P(0,0)}{P_0} = \frac{0.294}{0.3080} \approx 0.9545,$$

where $P_1 = 0.692$ and $P_0 = 0.3080$ represent the total probability of receiving 1 and 0, respectively, which is the union of the respective events of error-free and erroneous
1.10. INFORMATION TRANSMISSION VIA DISCRETE CHANNELS

receptions yielding the specific logical value concerned. The *mutual information* from Equation 1.56 is computed as:

\[
I(X_1, Y_1) = \log_2 \frac{P(X_1 | Y_1)}{P(X_1)}
\]

\[
\approx \log_2 \frac{0.9913}{0.7} \approx 0.502 \text{ bit} \quad (1.61)
\]

\[
I(X_0, Y_0) \approx \log_2 \frac{0.9545}{0.3} \approx 1.67 \text{ bit.} \quad (1.62)
\]

These figures must be contrasted with the amount of source information conveyed by the source symbols \(X_0, X_1\):

\[
I(0) = \log_2 \frac{1}{0.3} \approx \log_2 3.33 \approx 1.737 \text{ bit/symbol} \quad (1.63)
\]

and

\[
I(1) = \log_2 \frac{1}{0.7} \approx \log_2 1.43 \approx 0.5146 \text{ bit/symbol.} \quad (1.64)
\]

The amount of information "lost" in the noisy channel is given by the difference between the amount of information carried by the source symbols and the mutual information gained upon inferring a particular symbol at the noisy channel's output. Hence, the lost information can be computed from Equations 1.61, 1.62, 1.63, and 1.64, yielding \((1.737 - 1.67) \approx 0.067 \text{ bit}\) and \((0.5146 - 0.502) \approx 0.013 \text{ bit}\), respectively. These values may not seem catastrophic, but in relative terms they are quite substantial and their values rapidly escalate, as the channel error probability is increased. For the sake of completeness and for future use, let us compute the remaining mutual information terms, namely, \(I(X_0, Y_1)\) and \(I(X_1, Y_0)\), which necessitate the computation of:

\[
P(X_0 | Y_1) = \frac{P(X_0, Y_1)}{P(Y_1)}
\]

\[
P(0/1) = \frac{P(0, 1)}{P_1} = \frac{0.3 \cdot 0.02}{0.692} \approx 0.00867
\]

\[
P(X_1 | Y_0) = \frac{P(X_1, Y_0)}{P(Y_0)}
\]

\[
P(1/0) = \frac{P(1, 0)}{P_0} = \frac{0.7 \cdot 0.02}{0.308} \approx 0.04545
\]

\[
I(X_0, Y_1) = \log_2 \frac{P(X_0 | Y_1)}{P(X_0)} \approx \log_2 \frac{0.00867}{0.3} \approx -5.11 \text{ bit} \quad (1.65)
\]

\[
I(X_1, Y_0) = \log_2 \frac{P(X_1 | Y_0)}{P(X_1)} \approx \log_2 \frac{0.04545}{0.7} \approx -3.945 \text{ bit,} \quad (1.66)
\]

where the negative sign reflects the amount of "misinformation" as regards, for example, \(X_0\) upon receiving \(Y_1\). In this example we informally introduced the definition of mutual information. Let us now set out to formally exploit the benefits of our deeper insight into the effects of the noisy channel.
1.10.5 Information Loss via Imperfect Channels

Upon rewriting the definition of mutual information in Equation 1.56, we have:

\[
I(X_i, Y_j) = \log_2 \frac{P(X_i/Y_j)}{P(X_i)}
\]

\[
= \log_2 \frac{1}{P(X_i)} - \log_2 \frac{1}{P(X_i/Y_j)}
\]

\[
= I(X_i) - I(X_i/Y_j). \quad (1.67)
\]

Following Shannon’s [13–16, 24, 25] and Ferenczy’s [22] approach and rearranging Equation 1.67 yields:

\[
I(X_i) - I(X_i, Y_j) = I(X_i/Y_j). \quad (1.68)
\]

Briefly returning to figure 1.18 assists the interpretation of \( P(X_i/Y_j) \) as the probability or certainty/uncertainty that \( X_i \) was transmitted, given that \( Y_j \) was received, which justifies the above definition of the information loss. It is useful to observe from this figure that, as it was stated before, \( P(Y_j/X_i) \) represents the probability of erroneous or error-free reception. Explicitly, if \( j = i \), then \( P(Y_j/X_i) = P(Y_j/X_j) \) is the probability of error-free reception, while if \( j \neq i \), then \( P(Y_j/X_i) \) is the probability of erroneous reception.

With the probability \( P(Y_j/X_i) \) of erroneous reception in mind, we can actually associate an error information term with it:

\[
I(Y_j/X_i) = \log_2 \frac{1}{P(Y_j/X_i)}. \quad (1.69)
\]

Let us now concentrate on the average mutual information’s expression in Equation 1.60 and expand it as follows:

\[
I(X, Y) = \sum_{X,Y} P(X_i, Y_j) \cdot \log_2 \frac{1}{P(X_i)} - \sum_{X,Y} P(X_i, Y_j) \log_2 \frac{1}{P(X_i/Y_j)}. \quad (1.70)
\]

Considering the first term at the right-hand side (rhs) of the above equation and invoking Equation 1.55, we have:

\[
\sum_X \left[ \sum_Y P(X_i, Y_j) \right] \log_2 \frac{1}{P(X_i)} = \sum_X P(X_i) \log_2 \frac{1}{P(X_i)} = H(X). \quad (1.71)
\]

Then rearranging Equation 1.70 gives:

\[
H(X) - I(X, Y) = \sum_{X,Y} P(X_i, Y_j) \log_2 \frac{1}{P(X_i/Y_j)}, \quad (1.72)
\]
where $H(X)$ is the average source information per symbol and $I(X, Y)$ is the average conveyed information per received symbol.

Consequently, the rhs term must be the average information per symbol lost in the noisy channel. As we have seen in Equation 1.67 and Equation 1.68, the information loss is given by:

$$I(X_i/Y_j) = \log_2 \frac{1}{P(X_i/Y_j)}.$$  \hfill (1.73)

The average information loss $H(X/Y)$ _equivocation_, which Shannon [15] terms is computed as the weighted sum of these components:

$$H(X/Y) = \sum_X \sum_Y P(X_i, Y_j) \cdot \log_2 \frac{1}{P(X_i/Y_j)}.$$  \hfill (1.74)

Following Shannon, this definition allowed us to express Equation 1.72 as:

$$H(X) - I(X, Y) = H(X/Y).$$  \hfill (1.75)

1.10.6 Error Entropy via Imperfect Channels

Similarly to our previous approach and using the probability $P(Y_j/X_i)$ of erroneous reception associated with the information term of:

$$I(Y_j/X_i) = \log_2 \frac{1}{P(Y_j/X_i)}$$  \hfill (1.76)

we can define the average “error information” or error entropy. Hence, the above error information terms in Equation 1.76 are weighted using the probabilities $P(X_i, Y_j)$ and averaged for all $X$ and $Y$ values, defining the _error entropy_:

$$H(Y/X) = \sum_X \sum_Y P(X_i, Y_j) \cdot \log_2 \frac{1}{P(Y_j/X_i)}.$$  \hfill (1.77)

Using Bayes’ rule from Equation 1.52, we have

$$P(X_i/Y_j) \cdot P(Y_j) = P(Y_j/X_i) \cdot P(X_i)$$  \hfill (1.78)

Following from this, for the average mutual information in Equation 1.56 we have:

$$I(X, Y) = I(Y, X),$$  \hfill (1.79)

which, after interchanging $X$ and $Y$ in Equation 1.75, gives:

$$H(Y) - I(Y, X) = H(Y/X).$$  \hfill (1.80)
### Table 1.7: Summary of Definitions ©Ferenczy [22]

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source inf.</td>
<td>$I(X_i) = -\log_2 P(X_i)$</td>
</tr>
<tr>
<td>Received inf.</td>
<td>$I(Y_j) = -\log_2 P(Y_j)$</td>
</tr>
<tr>
<td>Joint inf.</td>
<td>$I_{X_i,Y_j} = -\log_2 P(X_i,Y_j)$</td>
</tr>
<tr>
<td>Mutual inf.</td>
<td>$I(X_i,Y_j) = \log_2 \frac{P(X_i,Y_j)}{P(X_i)}$</td>
</tr>
<tr>
<td>Av. Mut. inf.</td>
<td>$I(X,Y) = \sum_X \sum_Y P(X_i,Y_j) \log_2 \frac{P(X_i,Y_j)}{P(X_i)}$</td>
</tr>
<tr>
<td>Source entropy</td>
<td>$H(X) = -\sum_X P(X_i) \cdot \log_2 P(X_i)$</td>
</tr>
<tr>
<td>Destination entr.</td>
<td>$H(Y) = -\sum_Y P(Y_j) \log_2 P(Y_j)$</td>
</tr>
<tr>
<td>Equivocation</td>
<td>$H(X/Y) = -\sum_X \sum_Y P(X_i,Y_j) \log_2 P(X_i/Y_j)$</td>
</tr>
<tr>
<td>Error entropy</td>
<td>$H(Y/X) = -\sum_X \sum_Y P(X_i,Y_j) \log_2 P(Y_j/X_i)$</td>
</tr>
</tbody>
</table>

Subtracting the conveyed information from the destination entropy gives the error entropy, which is nonzero, if the destination entropy and conveyed information are not equal due to channel errors. Let us now proceed following Ferenczy’s approach [22] and summarize the most important definitions for future reference in Table 1.7 before we attempt to augment their physical interpretations using the forthcoming numerical example.

**Example** Using the BSC model of Figure 1.16, as an extension of the worked examples of Subsections 1.10.1 and 1.10.4 and following Ferenczy’s interpretation [22] of Shannon’s elaborations [13–16, 24, 25], let us compute the following range of system characteristics:

(a) The **joint information**, as distinct from the mutual information introduced earlier, for all possible channel input/output combinations.

(b) The entropy, i.e., the average information of both the source and the sink.

(c) The **average joint information** $H(X,Y)$.

(d) The **average mutual information per symbol conveyed**.

(e) The average information loss and average error entropy.

With reference to Figure 1.16 and to our introductory example from Section 1.10.1 we commence by computing further parameters of the BSC. Recall that the source information was:

$$I(X_0) = \log_2 \frac{1}{0.3} \approx 3.322 \log_{10} 3.333 \approx 1.737 \text{ bit}$$

$$I(X_1) = \log_2 \frac{1}{0.7} \approx 0.515 \text{ bit}.$$  

The probability of receiving a logical 0 was 0.308 and that of logical 1 was 0.692, of whether 0 or 1 was transmitted. Hence, the information inferred upon the reception of 0 and 1, respectively, is given by:

$$I(Y_0) = \log_2 \frac{1}{0.308} \approx 3.322 \log_{10} 3.247 \approx 1.699 \text{ bit}$$
Observe that because of the reduced probability of receiving a logical 1 from 0.7 → 0.692 as a consequence of channel-induced corruption, the probability of receiving a logical 0 is increased from 0.3 → 0.308. This is expected to increase the average destination entropy, since the entropy maximum of unity is achieved, when the symbols are equiprobable. We note, however, that this does not give more information about the source symbols, which must be maximized in an efficient communications system. In our example, the information conveyed increases for the reduced probability logical 1 from 0.515 bit → 0.531 bit and decreases for the increased probability 0 from 1.737 bit → 1.699 bit. Furthermore, the average information conveyed is reduced, since the reduction from 1.737 to 1.699 bit is more than the increment from 0.515 to 0.531. In the extreme case of an error probability of 0.5 we would have \( P(0) = P(1) = 0.5 \), and \( I(1) = I(0) = 1 \) bit, associated with receiving equiprobable random bits, which again would have a maximal destination entropy, but a minimal information concerning the source symbols transmitted. Following the above interesting introductory calculations, let us now turn our attention to the computation of the joint information.

**a)** The *joint information*, as distinct from the mutual information introduced earlier in Equation 1.56, of all possible channel input/output combinations is computed from Figure 1.16 as follows:

\[
I_{X_iY_j} = - \log_2 P(X_i, Y_j)
\]

\[
I_{00} = - \log_2(0.3 \cdot 0.98) \approx -3.322 \cdot \log_{10} 0.294 \approx 1.766 \text{ bit}
\]

\[
I_{01} = - \log_2(0.3 \cdot 0.02) \approx 7.381 \text{ bit}
\]

\[
I_{10} = - \log_2(0.7 \cdot 0.02) \approx 6.159 \text{ bit}
\]

\[
I_{11} = - \log_2(0.7 \cdot 0.98) \approx 0.544 \text{ bit}.
\]

These information terms can be individually interpreted formally as the information carried by the simultaneous occurrence of the given symbol combinations. For example, as it accrues from their computation, \( I_{00} \) and \( I_{11} \) correspond to the favorable event of error-free reception of a transmitted 0 and 1, respectively, which hence were simply computed by formally evaluating the information terms. By the same token, in the computation of \( I_{01} \) and \( I_{10} \), the corresponding source probabilities were weighted by the channel error probability rather than the error-free transmission probability, leading to the corresponding information terms. The latter terms, namely, \( I_{01} \) and \( I_{10} \), represent low-probability, high-information events due to the low channel error probability of 0.02.

Lastly, a perfect channel with zero error probability would render the probability of the error-events zero, which in turn would assign infinite information contents to the corresponding terms of \( I_{01} \) and \( I_{10} \), while \( I_{00} \) and \( I_{11} \) would be identical to the self-information of the 0 and 1 symbols. Then, if under zero error probability we evaluate the effect of the individual symbol probabilities on the remaining
joint information terms, the less frequently a symbol is emitted by the source, the higher its associated joint information term becomes and vice versa, which is seen by comparing $I_{00}$ and $I_{11}$. Their difference can be equalized by assuming an identical probability of 0.5 for both, which would yield $I_{00} = I_{11} = 1$-bit. The unweighted average of $I_{00}$ and $I_{11}$ would then be lower than in case of the previously used probabilities of 0.3 and 0.7, respectively, since the maximum average would be associated with the case of 0 and 1, where the associated $\log_2$ terms would be 0 and $-\infty$, respectively. The appropriately weighted average joint information terms will be evaluated under paragraph c/ during our later calculations. Let us now move on to evaluate the average information of the source and sink.

b/ Calculating the entropy, that is, the average information for both the source and the sink, is quite straightforward and ensues as follows:

$$H(X) = \sum_{i=1}^{2} P(X_i) \cdot \log_2 \frac{1}{P(X_i)}$$

$$\approx 0.3 \cdot \log_2 3.333 + 0.7 \cdot \log_2 1.429$$

$$\approx 0.5211 + 0.3605$$

$$\approx 0.8816 \text{ bit/symbol.} \quad (1.82)$$

For the computation of the sink’s entropy, we invoke Equations 1.49 and 1.50, yielding:

$$H(Y) = 0.308 \cdot \log_2 \frac{1}{0.308} + 0.692 \cdot \log_2 \frac{1}{0.692}$$

$$\approx 0.5233 + 0.3676$$

$$\approx 0.8909 \text{ bit/symbol.} \quad (1.83)$$

Again, the destination entropy $H(Y)$ is higher than the source entropy $H(X)$ due to the more random reception caused by channel errors, approaching $H(Y) = 1 \text{ bit/symbol}$ for a channel bit error rate of 0.5. Note, however, that unfortunately this increased destination entropy does not convey more information about the source itself.

c/ Computing the average joint information $H(X,Y)$ gives:

$$H(X,Y) = - \sum_{i=1}^{2} \sum_{j=1}^{2} P(X_i, Y_j) \log_2 P(X_i, Y_j)$$

$$= - \sum_{i=1}^{2} \sum_{j=1}^{2} P(X_i, Y_j) I_{X_i,Y_j}.$$. \quad (1.84)

Upon substituting the $I_{X_i,Y_j}$ values calculated in Equation 1.81 into Equation 1.84, we have:

$$H(X,Y) \approx 0.3 \cdot 0.98 \cdot 1.766 + 0.3 \cdot 0.02 \cdot 7.381$$
In order to interpret $H(X, Y)$, let us again scrutinize the definition given in Equation 1.84, which weights the joint information terms of Equation 1.81 by their probability of occurrence. We have argued before that the joint information terms corresponding to erroneous events are high due to the low error probability of 0.02. Observe, therefore, that these high-information symbol combinations are weighted by their low-probability of occurrence, causing $H(X, Y)$ to become relatively low. It is also instructive to consider the above terms in Equation 1.84 for the extreme cases of zero and 0.5 error probabilities and for different source emission probabilities, which are left for the reader to explore. Here we proceed considering the average conveyed mutual information per symbol.

**d/ The average conveyed mutual information per symbol** was defined in Equation 1.60 in order to quantify the average source information acquired per received symbol, which is repeated here for convenience as follows:

$$I(X, Y) = \sum_{X} \sum_{Y} P(X_i, Y_j) \log_2 \frac{P(X_i/Y_j)}{P(X_i)}$$

$$= \sum_{X} \sum_{Y} P(X_i, Y_j) \cdot I(X_i, Y_j).$$

Using the individual mutual information terms from Equations 1.61–1.66 in Section 1.10.4, we get the average mutual information representing the average amount of source information acquired from the received symbols, as follows:

$$I(X, Y) \approx 0.3 \cdot 0.98 \cdot 1.67 + 0.3 \cdot 0.02 \cdot (-5.11) + 0.7 \cdot 0.02 \cdot (-3.945) + 0.7 \cdot 0.98 \cdot 0.502$$

$$\approx 0.491 - 0.03066 - 0.05523 + 0.3444$$

$$\approx 0.7495 \text{ bit/symbol. (1.85)}$$

In order to interpret the concept of mutual information, in Section 1.10.4 we noted that the amount of information “lost” owing to channel errors was given by the difference between the amount of information carried by the source symbols and the mutual information gained upon inferring a particular symbol at the noisy channel’s output. These were given in Equations 1.61–1.64, yielding $(1.737 - 1.67) \approx 0.067$ bit and $(0.5146 - 0.502) \approx 0.013$ bit, for the transmission of a 0 and 1, respectively. We also noted that the negative sign of the terms corresponding to the error-events reflected the amount of misinformation as regards, for example, $X_0$ upon receiving $Y_1$. Over a perfect channel, the cross-coupling transitions of Figure 1.16 are eliminated, since the associated error probabilities are 0, and hence there is no information loss over the channel. Consequently, the error-free mutual information terms become identical to the self-information of the source symbols, since exactly the same amount of information can be inferred upon reception of a symbol, as much is carried by its appearance at the output of the source.
It is also instructive to study the effect of different error probabilities and source symbol probabilities in the average mutual information definition of Equation 1.84 in order to acquire a better understanding of its physical interpretation and quantitative power as regards the system’s performance. It is interesting to note, for example, that assuming an error probability of zero will therefore result in average mutual information, which is identical to the source and destination entropy computed above under paragraph b. It is also plausible that $I(X, Y)$ will be higher than the previously computed 0.7495 bits/symbol, if the symbol probabilities are closer to 0.5, or in general in case of $q$-ary sources closer to $1/q$. As expected, for a binary symbol probability of 0.5 and error probability of 0, we have $I(X, Y)=1$ bit/symbol.

e/ Lastly, let us determine the average information loss and average error entropy, which were defined in Equations 1.74 and 1.80 and are repeated here for convenience. Again, we will be using some of the previously computed probabilities from Sections 1.10.1 and 1.10.4, beginning with computation of the average information loss of Equation 1.74:

$$H(X/Y) = - \sum_{X} \sum_{Y} P(X_i, Y_j) \log_2 P(X_i/Y_j)$$

$$= -P(X_0, Y_0) \log_2 P(X_0/Y_0) - P(X_0, Y_1) \log_2 P(X_0/Y_1)$$
$$- P(X_1, Y_0) \log_2 P(X_1/Y_0) - P(X_1, Y_1) \log_2 P(X_1/Y_1)$$
$$= P(0,0) \cdot \log_2 P(0/0) + P(0,1) \cdot \log_2 P(0/1)$$
$$P(1,0) \cdot \log_2 P(1/0) + P(1,1) \cdot \log_2 P(1/1)$$
$$\approx -0.3 \cdot 0.98 \cdot \log_2 0.9545 - 0.3 \cdot 0.02 \cdot \log_2 0.0086$$
$$-0.7 \cdot 0.02 \cdot \log_2 0.04545 - 0.7 \cdot 0.98 \cdot \log_2 0.9913$$
$$\approx 0.0198 + 0.0411 + 0.0624 + 0.0086$$
$$\approx 0.132 \text{ bit/symbol.}$$

In order to augment the physical interpretation of the above-average information loss expression, let us examine the main contributing factors in it. It is expected to decrease as the error probability decreases. Although it is not straightforward to infer the clear effect of any individual parameter in the equation, experience shows that as the error probability increases, the two middle terms corresponding to the error events become more dominant. Again, the reader may find it instructive to alter some of the parameters on a one-by-one basis and study the way its influence manifests itself in terms of the overall information loss.

Moving on to the computation of the average error entropy, we find its definition equation is repeated below, and on inspecting Figure 1.16 we have:

$$H(Y/X) = - \sum_{X} \sum_{Y} P(X_i, Y_j) \cdot \log_2 P(Y_j/X_i)$$

$$= -P(X_0, Y_0) \log_2 P(Y_0/X_0) - P(X_0, Y_1) \log_2 P(Y_1/X_0)$$
$$- P(X_1, Y_0) \log_2 P(Y_0/X_1) - P(X_1, Y_1) \log_2 P(Y_1/X_1)$$

$$P(Y_0/X_0) = 0.98$$
\begin{align*}
P(Y_0/X_1) &= 0.02 \\
P(Y_1/X_0) &= 0.02 \\
P(Y_1/X_1) &= 0.98 \\
H(Y/X) &= P(0,0) \cdot \log_2 P(0/0) + P(0,1) \cdot \log_2 P(0/1) \\
&\quad + P(1,0) \cdot \log_2 P(1/0) + P(1,1) \cdot \log_2 P(1/1) \\
&= -0.294 \cdot \log_2 0.98 - 0.014 \cdot \log_2 0.02 \\
&\quad -0.006 \cdot \log_2 0.02 - 0.686 \cdot \log_2 0.98 \\
&\approx 0.0086 + 0.079 + 0.034 + 0.02 \\
&\approx 0.141 \text{ bit/symbol.}
\end{align*}

The average error entropy in the above expression is expected to fall as the error probability is reduced and vice versa. Substituting different values into its definition equation further augments its practical interpretation. Using our previous results in this section, we see that the \textit{average loss of information per symbol or equivocation} denoted by $H(X/Y)$ is given by the difference between the source entropy of Equation 1.82 and the average mutual information of Equation 1.85, yielding:

$$H(X/Y) = H(X) - I(X,Y) \approx 0.8816 - 0.7495 \approx 0.132 \text{ bit/symbol},$$

which according to Equation 1.75, is identical to the value of $H(X/Y)$ computed earlier. In harmony with Equation 1.80, the error entropy can also be computed as the difference of the average entropy $H(Y)$ in Equation 1.83 of the received symbols and the mutual information $I(X,Y)$ of Equation 1.85, yielding:

$$H(Y) - I(X,Y) \approx 0.8909 - 0.7495 \approx 0.141 \text{ bit/symbol},$$

as seen above for $H(Y/X)$.

Having defined the fundamental parameters summarized in Table 1.7 and used in the information-theoretical characterization of communications systems, let us now embark on the definition of channel capacity. Initially, we consider discrete noiseless channels, leading to a brief discussion of noisy discrete channels, and then we proceed to analog channels, before exploring the fundamental message of the Shannon-Hartley law.

### 1.11 Capacity of Discrete Channels [15, 22]

Shannon [15] defined the \textit{channel capacity} $C$ of a channel as the maximum achievable information transmission rate at which error-free transmission can be maintained over the channel.

Every practical channel is noisy, but transmitting at a sufficiently high power the channel error probability $p_e$ can be kept arbitrarily low, providing us with a simple initial channel model for our further elaborations. Following Ferenczy's approach [22],
assume that the transmission of symbol $X_i$ requires a time interval of $t_i$, during which an average of

$$H(X) = \sum_{i=1}^{q} P(X_i) \log_2 \frac{1}{P(X_i)} \text{ bit symbol}$$

(1.86)

information is transmitted, where $q$ is the size of the source alphabet used. This approach assumes that a variable-length coding algorithm, such as the previously described Shannon-Fano or the Huffman coding algorithm may be used in order to reduce the transmission rate to as low as the source entropy. Then the average time required for the transmission of a source symbol is computed by weighting $t_i$ with the probability of occurrence of symbol $X_i$, $i = 1 \ldots q$:

$$t_{av} = \sum_{i=1}^{q} P(X_i) t_i \frac{\text{sec symbol}}{\text{symbol}}.$$ 

(1.87)

Now we can compute the average information transmission rate $v$ by dividing the average information content of a symbol by the average time required for its transmission:

$$v = \frac{H(X)}{t_{av}} \frac{\text{bit sec}}{\text{sec}}.$$ 

(1.88)

The maximum transmission rate $v$ as a function of the symbol probability $P(X_i)$ must be found. This is not always an easy task, but a simple case occurs when the symbol duration is constant; that is, we have $t_i = t_0$ for all symbols. Then the maximum of $v$ is a function of $P(X_i)$ only and we have shown earlier that the entropy $H(X)$ is maximized by equiprobable source symbols, where $P(X_i) = \frac{1}{q}$. Then from Equations 1.86 and 1.87 we have an expression for the channel’s maximum capacity:

$$C = v_{max} = \frac{H(X)}{t_{av}} = \frac{\log_2 q}{t_0} \text{ bit sec}.$$ 

(1.89)

Shannon [15] characterized the capacity of discrete noisy channels using the previously defined mutual information describing the amount of average conveyed information, given by:

$$I(X,Y) = H(Y) - H(Y/X),$$

(1.90)

where $H(Y)$ is the average amount of information per symbol at the channel’s output, while $H(Y/X)$ is the error entropy. Here a unity symbol-rate was assumed for the sake of simplicity. Hence, useful information is transmitted only via the channel if $H(Y) > H(Y/X)$. Via a channel with $p_e = 0.5$, where communication breaks down, we have $H(Y) = H(Y/X)$, and the information conveyed becomes $I(X,Y) = 0$. The amount of information conveyed is maximum if the error entropy $H(Y/X) = 0$. Therefore, Shannon [15] defined the noisy channel’s capacity as the maximum value of the conveyed information $I(X,Y)$:

$$C = I(X,Y)_{MAX} = [H(Y) - H(Y/X)]_{MAX},$$

(1.91)

where the maximization of Equation 1.91 is achieved by maximizing the first term and minimizing the second term.
In general, the maximization of Equation 1.91 is an arduous task, but for the BSC seen in Figure 1.19 it becomes fairly simple. Let us consider this simple case and assume that the source probabilities of 1 and 0 are \( P(0) = P(1) = 0.5 \) and the error probability is \( p_e \). The entropy at the destination is computed as:

\[
H(Y) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \text{ bit/symbol},
\]

while the error entropy is given by:

\[
H(Y/X) = - \sum_X \sum_Y P(X_i, Y_j) \cdot \log_2 P(Y_j/X_i). \tag{1.92}
\]

In order to be able to compute the capacity of the BSC as a function of the channel's error probability, let us substitute the required joint probabilities of:

\[
P(0,0) = P(0)(1 - p_e)
P(0,1) = P(0)p_e
P(1,0) = P(1)p_e
P(1,1) = P(1)(1 - p_e). \tag{1.93}
\]

and the conditional probabilities of:

\[
P(0/0) = (1 - p_e)
P(0/1) = p_e
P(1/0) = p_e
P(1/1) = (1 - p_e). \tag{1.94}
\]

into Equation 1.92, yielding:

\[
H(Y/X) = -[P(0)(1 - p_e) \cdot \log_2 (1 - p_e) + P(0) \cdot p_e \log_2 p_e
+P(1) \cdot p_e \log_2 p_e + P(1)(1 - p_e) \log_2 (1 - p_e)]
\]
$0.8$

$0.6$

$0.4$

$0.2$

$0.0$

Figure 1.20: Channel capacity versus $p_e$ for the BSC.

$$C = 1 + (1 - p_e) \log_2 (1 - p_e) + p_e \log_2 p_e.$$  \hspace{1cm} (1.95)

Finally, upon substituting $H(Y)$ and $H(Y/X)$ from above into Equation 1.91, the BSC’s channel capacity becomes:

$$C = 1 + (1 - p_e) \log_2 (1 - p_e) + p_e \log_2 p_e.$$  \hspace{1cm} (1.96)

Following Ferenczy’s [22] interpretation of Shannon’s lessons [13–16,24,25], the graphic representation of the BSC’s capacity is depicted in Figure 1.20 using various $p_e$ error probabilities.

Observe, for example, that for $p_e = 10^{-2}$ the channel capacity is $C \approx 0.9$ bit/symbol, that is, close to its maximum of $C = 1$ bit/symbol, but for higher $p_e$ values it rapidly decays, falling to $C \approx 0.5$ bit/symbol around $p_e = 10^{-1}$. If $p_e = 50\%$, we have $C = 0$ bit/symbol; since no useful information transmission takes place, the channel delivers random bits. Notice also that if $P(0) \neq P(1) \neq 0.5$, then $H(Y) < 1$ bit/symbol and hence $C < C_{\text{max}} = 1$ bit/symbol, even if $p_e = 0$. 
1.12 Shannon’s Channel Coding Theorem [19, 27]

In the previous section, we derived a simple expression for the capacity of the noisy BSC in Equation 1.96, which was depicted in Figure 1.20 as a function of the channel’s error probability $p_e$. In this section, we focus on Shannon’s channel coding theorem, which states that as long as the information transmission rate does not exceed the channel’s capacity, the bit error rate can be kept arbitrarily low [24, 25]. In the context of the BSC channel capacity curve of Figure 1.20, this theorem implies that noise over the channel does not preclude the reliable transmission of information; it only limits the rate at which transmission can take place. Implicitly, this theorem prophesies the existence of an appropriate error correction code, which adds redundancy to the original information symbols. This reduces the system’s useful information throughput but simultaneously allows error correction coding. Instead of providing a rigorous proof of this theorem, following the approach suggested by Abramson [19], which was also used by Hey and Allen [27] in their compilation of Feynman’s lectures, we will make it plausible.

The theorem is stated more formally as follows. Let us assume that a message of $K$ useful information symbols is transmitted by assigning it to an $N$-symbol so-called block code, where the symbols are binary and the error probability is $p_e$. Then, according to Shannon, upon reducing the coding rate $R = \frac{K}{N}$ beyond every limit, the error probability obeys the following relationship:

$$R = \frac{K}{N} \leq C = 1 + (1 - p_e) \log_2(1 - p_e) + p_e \cdot \log_2 p_e.$$  

(1.97)

As Figure 1.20 shows upon increasing the bit error rate $p_e$, the channel capacity reduces gradually toward zero, which forces the channel coding rate $R = \frac{K}{N}$ to zero in the limit. This inequality therefore implies that an arbitrarily low BER is possible only when the coding rate $R$ tends to zero, which assumes an infinite-length block code and an infinite coding delay. By scrutinizing Figure 1.20, we can infer that, for example, for a BER of $10^{-1}$ an approximately $R = \frac{K}{N} \approx \frac{1}{2}$ so-called half-rate code is required in order to achieve asymptotically perfect communications, while for $BER = 10^{-2}$ an approximately $R \approx 0.9$ code is required.

Shannon’s channel coding theorem does not specify how to create error correction codes, which can achieve this predicted performance; it merely states their existence. Hence, the error correction coding community has endeavored over the years to create such good codes but until 1993 had only limited success. Then in that year Berrou et al. [28] invented the family of iteratively decoded turbo-codes, which are capable of approaching the Shannonian predictions within a fraction of a dB.

Returning to the channel coding theorem, Hey and Feynman [27] offered a witty approach to deepening the physical interpretation of this theorem, which we briefly highlight below. Assuming that the block-coded sequences are long, in each block on the average there are $t = p_e \cdot N$ number of errors. In general, $t$ number of errors can be allocated over the block of $N$ positions in

$$C_N^t = \binom{N}{t} = \frac{N!}{t!(N-t)!}.$$
different ways, which are associated with the same number of error patterns. The number of additional parity bits added during the coding process is \( P = (N - K) \), which must be sufficiently high for identifying all the \( C \) number of error patterns, in order to allow inverting (i.e., correcting) the corrupted bits in the required positions. Hence, we have [27]:

\[
\frac{N!}{t!(N-t)!} \leq 2^{(N-K)}.
\]

Upon exploiting the Stirling formula of

\[
N! \approx \sqrt{2\pi N} \cdot \left(\frac{N}{e}\right)^N = \sqrt{2\pi} \cdot \sqrt{N} \cdot N^N \cdot e^{-N}
\]

and taking the logarithm of both sides, we have:

\[
\log_e N! \approx \log_e \sqrt{2\pi} + \frac{1}{2} \log_e N + N \log_e N - N.
\]

Furthermore, when \( N \) is large, the first and second terms are diminisingly small in comparison to the last two terms. Thus, we have:

\[
\log_e N! \approx N \log_e N - N.
\]

Then, after taking the logarithm, the factorial expression on the left-hand side (L) of Equation 1.98 can be written as:

\[
L \approx [N \log_e N - N] - [t \log_e t - t] - [(N - t) \log_e (N - t) - (N - t)].
\]

Now taking into account that \( t \approx p_e \cdot N \), we have [27]:

\[
L \approx [N \log_e N - N] - [p_e N \log_e (p_e N) - p_e N] \\
- [(N - p_e N) \log_e N - p_e N] - (N - p_e N) \\
\approx [N \log_e N - N] - [p_e N \log_e p_e + p_e N \log_e N - p_e N] \\
- [N \log_e (N(1 - p_e)) - p_e N \log_e (N(1 - p_e)) - (N - p_e N)] \\
\approx [N \log_e N - N] - [p_e N \log_e p_e + p_e N \log_e N - p_e N] \\
- [N \log_e N + N \log_e (1 - p_e) - p_e N \log_e N] \\
- p_e N \log_e (1 - p_e) - (N - p_e N) \\
\approx N[\log_e N - 1 - p_e \log_e p_e - p_e \log_e N + p_e] \\
- \log_e N - \log_e (1 - p_e) + p_e \log_e N \\
+ p_e \log_e (1 - p_e) + 1 - p_e \\
\approx N[-p_e \log_e p_e - \log_e (1 - p_e) + p_e \log_e (1 - p_e)] \\
\approx N[-p_e \log_e p_e - (1 - p_e) \log_e (1 - p_e)].
\]

If we consider that \( \log_e a = \log_2 a \cdot \log_2 2 \), then we can convert the \( \log_e \) terms to \( \log_2 \) as follows [27]:

\[
L \approx N \log_2 2[-p_e \log_2 p_e - (1 - p_e) \log_2 (1 - p_e)].
\]
Finally, upon equating this term with the logarithm of the right-hand side expression of Equation 1.98, we arrive at:

\[ N \log_e 2[-p_e \log_2 p_e - (1 - p_e) \log_2 (1 - p_e)] \leq (N - K) \log_e 2, \]

which can be simplified to:

\[-p_e \log_2 p_e - (1 - p_e) \log_2 (1 - p_e) \leq 1 - \frac{K}{N} \]

or to a form, identical to Equation 1.97:

\[ \frac{K}{N} \leq 1 + (1 - p_e) \log_2 (1 - p_e) + p_e \log_2 p_e. \]

### 1.13 Capacity of Continuous Channels [16, 22]

During our previous discussions, it was assumed that the source emitted discrete messages with certain finite probabilities, which would be exemplified by an 8-bit analog-to-digital converter emitting one of 256 discrete values with a certain probability. However, after digital source encoding and channel encoding according to the basic schematic of Figure 1.1 the modulator typically converts the digital messages to a finite set of bandlimited analog waveforms, which are chosen for maximum “transmission convenience.” In this context, transmission convenience can imply a range of issues, depending on the communications channel. Two typical constraints are predominantly power-limited or bandwidth-limited channels, although in many practical scenarios both of these constraints become important. Because of their limited solar power supply, satellite channels tend to be more severely power-limited than bandlimited, while typically the reverse situation is experienced in mobile radio systems.

The third part of Shannon’s pioneering paper [16] considers many of these issues. Thus, in what follows we define the measure of information for continuous signals, introduce a concept for the continuous channel capacity, and reveal the relationships among channel bandwidth, channel capacity, and channel signal-to-noise ratio, as stated by the Shannon-Hartley theorem. Finally, the ideal communications system transpiring from Shannon’s pioneering work is characterized, before concluding with a brief discussion of the ramifications of wireless channels as regards the applicability of Shannon’s results.

Let us now assume that the channel’s analog input signal \( x(t) \) is bandlimited and hence that it is fully characterized by its Nyquist samples and by its probability density function (PDF) \( p(x) \). The analogy of this continuous PDF and that of a discrete source are characterized by \( P(X_i) \approx p(x_i) \Delta X \), which reflects the practical way of experimentally determining the histogram of a bandlimited analog signal by observing the relative frequency of events, when its amplitude resides in a \( \Delta X \) wide amplitude bin-centered around \( X_i \). As an analogy to the discrete average information or entropy expression of:

\[ H(X) = - \sum_i P(X_i) \cdot \log_2 P(X_i), \] (1.99)
Shannon [16] introduced the entropy of analog sources, as it was also noted and exploited, for example, by Ferenczy [22], as follows:

\[ H(x) = - \int_{-\infty}^{\infty} p(x) \log_2 p(x) dx. \]  

(1.100)

For our previously used discrete sources, we have shown that the source entropy is maximized for equiprobable messages. The question that arises is whether this is also true for continuous PDF's. Shannon [16] derived the maximum of the analog signal's entropy under the constraints of:

\[ \int_{-\infty}^{\infty} p(x) dx = 1 \]  

(1.101)

\[ \sigma_x^2 = \int_{-\infty}^{\infty} x^2 \cdot p(x) dx = \text{Constant} \]

(1.102)

based on the calculus of variations. He showed that the entropy of a signal \( x(t) \) having a constant variance of \( \sigma_x^2 \) is maximum, if \( x(t) \) has a Gaussian distribution given by:

\[ p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{x^2}{2\sigma^2}\right)}. \]

(1.103)

Then the maximum of the entropy can be derived upon substituting this PDF into the expression of the entropy. Let us first take the natural logarithm of both sides of the PDF and convert it to base two logarithm by taking into account that \( \log_e 2 = \frac{1}{\log_2 e} \), in order to be able to use it in the entropy's \( \log_2 \) expression. Then the PDF of Equation 1.103 can be written as:

\[ -\log_2 p(x) = + \log_2 \sqrt{2\pi}\sigma + \left(\frac{x^2}{2\sigma^2}\right) \cdot \frac{1}{\log_e 2}, \]

(1.104)

and upon exploiting that \( \log_e 2 = 1/\log_2 e \), the entropy is expressed according to Shannon [16] and Ferenczy [22] as:

\[ H_{\text{max}}(x) = - \int p(x) \cdot \log_2 p(x) dx \]

\[ = \int p(x) \cdot \log_2 \sqrt{2\pi}\sigma dx + \int p(x) \frac{x^2}{2\sigma^2} \cdot \log_2 e \cdot dx \]

\[ = \log_2 \sqrt{2\pi}\sigma \int p(x) dx + \frac{\log_2 e}{2\sigma^2} \int x^2 p(x) dx \]

\[ = \log_2 \sqrt{2\pi}\sigma + \frac{\sigma^2}{2\sigma^2} \log_2 e \]

\[ = \log_2 \sqrt{2\pi}\sigma + \frac{\log_2 e}{2} \]

\[ = \log_2 \sqrt{2\pi e}\sigma. \]

(1.105)
Since the maximum of the entropy is proportional to the logarithm of the signal’s average power $S_x = \sigma^2_x$, upon quadrupling the signal’s power the entropy is increased by one bit because the range of uncertainty as regards where the signal samples can reside is expanded.

We are now ready to formulate the channel capacity versus channel bandwidth and versus channel SNR relationship of analog channels. Let us assume white, additive, signal-independent noise with a power of $N$ via the channel. Then the received (signal+noise) power is given by:

$$\sigma^2_y = S + N. \quad (1.106)$$

By the same argument, the channel’s output entropy is maximum if its output signal $y(t)$ has a Gaussian PDF and its value is computed from Equation 1.105 as:

$$H_{max}(y) = \frac{1}{2} \log_2(2\pi e\sigma^2_y) = \frac{1}{2} \log_2 2\pi e(S + N). \quad (1.107)$$

We proceed by taking into account the channel impairments, reducing the amount of information conveyed by the amount of the error entropy $H(y/x)$ giving:

$$I(x,y) = H(y) - H(y/x), \quad (1.108)$$

where again the noise is assumed to be Gaussian and hence:

$$H(y/x) = \frac{1}{2} \log_2(2\pi e N). \quad (1.109)$$

Upon substituting Equation 1.107 and Equation 1.109 in Equation 1.108, we have:

$$I(x,y) = \frac{1}{2} \log_2 \left( \frac{2\pi e(S + N)}{2\pi eN} \right)$$

$$= \frac{1}{2} \log_2 \left( 1 + \frac{S}{N} \right), \quad (1.110)$$

where, again, both the channel’s output signal and the noise are assumed to have Gaussian distribution.

The analog channel’s capacity is then calculated upon multiplying the information conveyed per source sample by the Nyquist sampling rate of $f_s = 2 \cdot f_B$, yielding [24]:

$$C = f_B \cdot \log_2 \left( 1 + \frac{S}{N} \right) \text{ bit/sec}. \quad (1.111)$$

Equation 1.111 is the well-known Shannon-Hartley law,\(^1\) establishing the relationship among the channel capacity $C$, channel bandwidth $f_B$, and channel signal-to-noise ratio (SNR).

Before analyzing the consequences of the Shannon-Hartley law following Shannon’s deliberations [24], we make it plausible from a simple practical point of view. As we

\(^1\)Comment by the Authors: Although the loose definition of capacity is due to Hartley, the underlying relationship is entirely due to Shannon.
have seen, the root mean squared (RMS) value of the noise is $\sqrt{N}$, and that of the signal plus noise at the channel’s output is $\sqrt{S+N}$. The receiver has to decide from the noisy channel’s output signal what signal has been input to the channel, although this has been corrupted by an additive Gaussian noise sample. Over an ideal noiseless channel, the receiver would be able to identify what signal sample was input to the receiver. However, over noisy channels it is of no practical benefit to identify the corrupted received message exactly. It is more beneficial to quantify a discretized version of it using a set of decision threshold values, where the resolution is dependent on how corrupted the samples are. In order to quantify this SNR-dependent receiver dynamic range resolution, let us consider the following argument.

Having very densely spaced receiver detection levels would often yield noise-induced decision errors, while a decision-level spacing of $\sqrt{N}$ according to the RMS noise-amplitude intuitively seems a good compromise between high information resolution and low decision error rate. Then assuming a transmitted sample, which resides at the center of a $\sqrt{N}$ wide decision interval, noise samples larger than $\sqrt{N}/2$ will carry samples across the adjacent decision boundaries. According to this spacing, the number of receiver reconstruction levels is given by:

$$q = \frac{\sqrt{S+N}}{\sqrt{N}} = \left(1 + \frac{S}{N}\right)^{\frac{1}{2}},$$

(1.112)

which creates a scenario similar to the transmission of equiprobable $q$-ary discrete symbols via a discrete noisy channel, each conveying $\log_2 q$ amount of information at the Nyquist sampling rate of $f_s = 2 \cdot f_B$. Therefore, the channel capacity becomes [24]:

$$C = 2 \cdot f_B \cdot \log_2 q = f_B \cdot \log_2 \left(1 + \frac{S}{N}\right),$$

(1.113)

as seen earlier in Equation 1.111.

### 1.13.1 Practical Evaluation of the Shannon-Hartley Law

The Shannon-Hartley law of Equation 1.111 and Equation 1.113 reveals the fundamental relationship of the SNR, bandwidth, and channel capacity. This relationship can be further studied following Ferenczy’s interpretation [22] upon referring to Figure 1.21.

Observe from the figure that a constant channel capacity can be maintained, even when the bandwidth is reduced, if a sufficiently high SNR can be guaranteed. For example, from Figure 1.21 we infer that at $f_B = 10$ KHz and SNR = 30 dB the channel capacity is as high as $C = 100$ kbps. Surprisingly, $C \approx 100$ kbps can be achieved even for $f_B = 5$ KHz, if SNR = 60 dB is guaranteed.

Figure 1.22 provides an alternative way of viewing the Shannon-Hartley law, where the SNR is plotted as a function of $f_B$, parameterized with the channel capacity $C$. It is important to notice how dramatically the SNR must be increased in order to maintain a constant channel capacity $C$, as the bandwidth $f_B$ is reduced below $0.1 \cdot C$, where $C$ and $f_B$ are expressed in kbit/s and Hz, respectively. This is due to the $\log_2(1 + \text{SNR})$ function in Equation 1.111, where a logarithmically increasing SNR value is necessitated to compensate for the linear reduction in terms of $f_B$. 
From our previous discourse, the relationship between the relative channel capacity $C/f_B$ expressed from Equation 1.113, and the channel SNR now becomes plausible. This relationship is quantified in Table 1.8 and Figure 1.23 for convenience. Notice that due to the logarithmic SNR scale expressed in dBs, the $C/f_B$ curve becomes near-linear, allowing a near-linearly proportional relative channel capacity improvement upon increasing the channel SNR. A very important consequence of this relationship is that if the channel SNR is sufficiently high to support communications using a high number of modulation levels, the channel is not exploited to its full capacity upon using $C/f_B$ values lower than is afforded by the prevailing SNR. Proposing various techniques in order to exploit this philosophy was the motivation of reference [29].

The capacity $C$ of a noiseless channel with $SNR = \infty$ is $C = \infty$, although noiseless channels do not exist. In contrast, the capacity of an ideal system with $f_B = \infty$ is finite [20,23]. Assuming additive white Gaussian noise (AWGN) with a double-sided power spectral density (PSD) of $\eta/2$, we have $N = \frac{\eta}{2} \cdot 2 \cdot f_B = \eta \cdot f_B$, and applying the Shannon-Hartley law gives [20]:

$$C = f_B \cdot \log_2 \left(1 + \frac{S}{\eta f_B}\right)$$
$C = f_B \log_2(1+S/N)$ \text{[bit/s]} for $f_B$ in Hz

$C = f_B \log_2(1+S/N)$ \text{[Kbit/s]} for $f_B$ in KHz

$C = f_B \log_2(1+S/N)$ \text{[Mbit/s]} for $f_B$ in MHz

Figure 1.22: SNR versus $f_B$ relations according to the Shannon-Hartley law. ©Ferenczy [22].

<table>
<thead>
<tr>
<th>SNR Ratio</th>
<th>dB</th>
<th>$C/f_B$ [bit/sec/Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4.8</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>8.5</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>11.8</td>
<td>4</td>
</tr>
<tr>
<td>31</td>
<td>14.9</td>
<td>5</td>
</tr>
<tr>
<td>63</td>
<td>18.0</td>
<td>6</td>
</tr>
<tr>
<td>127</td>
<td>21.0</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1.8: Relative Channel Capacity versus SNR
Our aim is now to determine $C_{\infty} = \lim_{f_B \to \infty} C$. Upon exploiting that:

$$
\lim_{x \to 0} (1 + x)^{\frac{1}{x}} = e
$$

(1.115)

where $x = S/ (\eta \cdot f_B)$, we have

$$
C_{\infty} = \lim_{f_B \to \infty} C = \frac{S}{\eta} \log_2 e = 1.45 \cdot \left( \frac{S}{\eta} \right)
$$

(1.116)

which is the capacity of the channel with $f_B = \infty$. The practically achievable transmission rate $R$ is typically less than the channel capacity $C$, although complex turbo-coded modems [28] can approach its value. For example, for a telephone channel with a signal-to-noise ratio of $S/N = 10^3 = 30 dB$ and a bandwidth of $B = 3.4$ kHz from Equation 1.113, we have $C = 3.4 \cdot \log_2 (1 + 10^3) \cdot \frac{kbit}{sec} \approx 3.4 \cdot 10 = 34 kbit/s$, which is fairly close to the rate of the V.34 CCITT standard 28.8 kbit/s telephone-channel modem that was recently standardized.
CHAPTER 1. INFORMATION THEORY

Figure 1.24: Shannon’s ideal communications system for AWGN channels.

In this chapter, we have been concerned with various individual aspects of Shannon’s information theory [13–16, 24, 25]. Drawing nearer to concluding our discourse on the foundations of information theory, let us now outline in broad terms the main ramifications of Shannon’s work [13–16].

1.13.2 Shannon’s Ideal Communications System for Gaussian Channels

The ideal Shannonian communications system shown in Figure 1.24 has the following characteristics. The system’s information-carrying capacity is given by the information rate

\[ C = B \log_2 (1 + S/N) \]

while as regards its error rate we have \( p_e \rightarrow 0 \). The transmitted and received signals are bandlimited Gaussian random variables, which facilitate communicating at the highest possible rate over the channel.

Information from the source is observed for \( T \) seconds, where \( T \) is the symbol duration and encoded as equiprobable \( M \)-ary symbols with a rate of

\[ R = \frac{\log_2 M}{T} \]

Accordingly, the signaling waveform generator of Figure 1.24 assigns a bandlimited AWGN representation having a maximum frequency of \( f_B \) from the set of \( M = 2^{RT} \) possible waveforms to the source message, uniquely representing the signal \( x(t) \) to be transmitted for a duration of \( T \). The noisy received signal \( y(t) = x(t) + n(t) \) is compared to all \( M = 2^{RT} \) prestored waveforms at the receiver, and the most “similar” is chosen to identify the most likely transmitted source message. The observation intervals at both the encoder and decoder amount to \( T \), yielding an overall coding delay of \( 2T \). Signaling at a rate equal to the channel capacity is only possible, if the source signal’s observation interval is infinitely long, that is, \( T \rightarrow \infty \).

Before concluding this chapter, we offer a brief discussion of the system-architectural ramifications of transmitting over wireless channels rather than over AWGN channels.

1.14 Shannon’s Message and Its Implications for Wireless Channels

In wireless communications over power- and bandlimited channels it is always of prime concern to maintain an optimum compromise in terms of the contradictory
requirements of low bit rate, high robustness against channel errors, low delay, and low complexity. The minimum bit rate at which distortionless communications is possible is determined by the entropy of the speech source message. Note, however, that in practical terms the source rate corresponding to the entropy is only asymptotically achievable as the encoding memory length or delay tends to infinity. Any further compression is associated with information loss or coding distortion. Note that the optimum source encoder generates a perfectly uncorrelated source-coded stream, where all the source redundancy has been removed; therefore, the encoded symbols are independent, and each one has the same significance. Having the same significance implies that the corruption of any of the source-encoded symbols results in identical source signal distortion over imperfect channels.

Under these conditions, according to Shannon's pioneering work [13], which was expanded, for example, by Hagenauer [30] and Viterbi [31], the best protection against transmission errors is achieved if source and channel coding are treated as separate entities. When using a block code of length $N$ channel coded symbols in order to encode $K$ source symbols with a coding rate of $R = K/N$, the symbol error rate can be rendered arbitrarily low, if $N$ tends to infinity and the coding rate to zero. This condition also implies an infinite coding delay. Based on the above considerations and on the assumption of additive white Gaussian noise (AWGN) channels, source and channel coding have historically been separately optimized.

Mobile radio channels are subjected to multipath propagation and so constitute a more hostile transmission medium than AWGN channels, typically exhibiting path-loss, log-normal slow fading and Rayleigh fast-fading [32]. Furthermore, if the signaling rate used is higher than the channel’s coherence bandwidth, over which no spectral-domain linear distortion is experienced, then additional impairments are inflicted by dispersion, which is associated with frequency-domain linear distortions. Under these circumstances the channel’s error distribution versus time becomes bursty, and an infinite-memory symbol interleaver is required in Figure 1.1 in order to disperse the bursty errors and hence to render the error distribution random Gaussian-like, such as over AWGN channels. For mobile channels, many of the above mentioned, asymptotically valid ramifications of Shannon’s theorems have limited applicability.

A range of practical limitations must be observed when designing mobile radio speech or video links. Although it is often possible to further reduce the prevailing typical bit rate of state-of-art speech or video codecs, in practical terms this is possible only after a concomitant increase of the implementational complexity and encoding delay. A good example of these limitations is the half-rate GSM speech codec, which was required to approximately halve the encoding rate of the 13 kbps full-rate codec, while maintaining less than quadrupled complexity, similar robustness against channel errors, and less than doubled encoding delay. Naturally, the increased algorithmic complexity is typically associated with higher power consumption, while the reduced number of bits used to represent a certain speech segment intuitively implies that each bit will have an increased relative significance. Accordingly, their corruption may inflict increasingly objectionable speech degradations, unless special attention is devoted to this problem.

In a somewhat simplistic approach, one could argue that because of the reduced
source rate we could accommodate an increased number of parity symbols using a more powerful, implementationally more complex and lower rate channel codec, while maintaining the same transmission bandwidth. However, the complexity, quality, and robustness trade-off of such a scheme may not always be attractive.

A more intelligent approach is required to design better speech or video transceivers for mobile radio channels [30]. Such an intelligent transceiver is portrayed in Figure 1.1. Perfect source encoders operating close to the information-theoretical limits of Shannon's predictions can only be designed for stationary source signals, a condition not satisfied by most source signals. Further previously mentioned limitations are the encoding complexity and delay. As a consequence of these limitations the source-coded stream will inherently contain residual redundancy, and the correlated source symbols will exhibit unequal error sensitivity, requiring unequal error protection. Following Hagenauer [30], we will refer to the additional knowledge as regards the different importance or vulnerability of various speech-coded bits as source significance information (SSI). Furthermore, Hagenauer termed the confidence associated with the channel decoder's decisions as decoder reliability information (DRI). These additional links between the source and channel codecs are also indicated in Figure 1.1. A variety of such techniques have successfully been used in robust source-matched source and channel coding [33,34].

The role of the interleaver and de-interleaver seen in Figure 1.1 is to rearrange the channel coded bits before transmission. The mobile radio channel typically inflicts bursts of errors during deep channel fades, which often overload the channel decoder's error correction capability in certain speech or video segments. In contrast other segments are not benefiting from the channel codec at all, because they may have been transmitted between fades and hence are error-free even without channel coding. This problem can be circumvented by dispersing the bursts of errors more randomly between fades so that the channel codec is always faced with an "average-quality" channel, rather than the bimodal faded/nonfaded condition. In other words, channel codecs are most efficient if the channel errors are near-uniformly dispersed over consecutive received segments.

In its simplest manifestation, an interleaver is a memory matrix filled with channel coded symbols on a row-by-row basis, which are then passed on to the modulator on a column-by-column basis. If the transmitted sequence is corrupted by a burst of errors, the de-interleaver maps the received symbols back to their original positions, thereby dispersing the bursty channel errors. An infinite memory channel interleaver is required in order to perfectly randomize the bursty errors and therefore to transform the Rayleigh-fading channel's error statistics to that of a AWGN channel, for which Shannon's information theoretical predictions apply. Since in interactive video or speech communications the tolerable delay is strictly limited, the interleaver's memory length and efficiency are also limited.

A specific deficiency of these rectangular interleavers is that in case of a constant vehicular speed the Rayleigh-fading mobile channel typically produces periodic fades and error bursts at traveled distances of $\lambda/2$, where $\lambda$ is the carrier's wavelength, which may be mapped by the rectangular interleaver to another set of periodic bursts of errors. A range of more random rearrangement or interleaving algorithms exhibiting a higher performance than rectangular interleavers have been proposed for mobile
channels in [35], where a variety of practical channel coding schemes have also been portrayed.

Returning to Figure 1.1, the soft-decision information (SDI) or channel state information (CSI) link provides a measure of confidence with regard to the likelihood that a specific symbol was transmitted. Then the channel decoder often uses this information in order to invoke maximum likelihood sequence estimation (MLSE) based on the Viterbi algorithm [35] and thereby improve the system's performance with respect to conventional hard-decision decoding. Following this rudimentary review of Shannon's information theory, let us now turn our attention to the characterization of wireless communications channels.

1.15 Summary and Conclusions

An overview of Shannonian information theory has been given, in order to establish a firm basis for our further discussions throughout the book. Initially we focussed our attention on the basic Shannonian information transmission scheme and highlighted the differences between Shannon’s theory valid for ideal source and channel codecs as well as for Gaussian channels and its ramifications for Rayleigh channels. We also argued that practical finite-delay source codecs cannot operate at transmission rates as low as the entropy of the source. However, these codecs do not have to operate losslessly, since perceptually unobjectionable distortions can be tolerated. This allows us to reduce the associated bit rate.

Since wireless channels exhibit bursty error statistics, the error bursts can only be randomized with the aid of infinite-length channel interleavers, which are not amenable to real-time communications. Although with the advent of high-delay turbo channel codecs it is possible to operate near the Shannonian performance limits over Gaussian channels, over bursty and dispersive channels different information-theoretical channel capacity limits apply.

We considered the entropy of information sources both with and without memory and highlighted a number of algorithms, such as the Shannon-Fano, the Huffman and run-length coding algorithms, designed for the efficient encoding of sources exhibiting memory. This was followed by considering the transmission of information over noise-contaminated channels leading to Shannon’s channel coding theorem. Our discussions continued by considering the capacity of communications channels in the context of the Shannon-Hartley law. The chapter was concluded by considering the ramifications of Shannon’s messages for wireless channels.