1

Information and Coding Theory

In his classic paper ‘A Mathematical Theory of Communication’, Claude Shannon [1] introduced the main concepts and theorems of what is known as information theory. Definitions and models for two important elements are presented in this theory. These elements are the binary source (BS) and the binary symmetric channel (BSC). A binary source is a device that generates one of the two possible symbols ‘0’ and ‘1’ at a given rate $r$, measured in symbols per second. These symbols are called *bits* (binary digits) and are generated randomly.

The BSC is a medium through which it is possible to transmit one symbol per time unit. However, this channel is not reliable, and is characterized by the error probability $p$ ($0 \leq p \leq 1/2$) that an output bit can be different from the corresponding input. The symmetry of this channel comes from the fact that the error probability $p$ is the same for both of the symbols involved.

Information theory attempts to analyse communication between a transmitter and a receiver through an unreliable channel, and in this approach performs, on the one hand, an analysis of information sources, especially the amount of information produced by a given source, and, on the other hand, states the conditions for performing reliable transmission through an unreliable channel.

There are three main concepts in this theory:

1. The first one is the definition of a quantity that can be a valid measurement of information, which should be consistent with a physical understanding of its properties.
2. The second concept deals with the relationship between the information and the source that generates it. This concept will be referred to as source information. Well-known information theory techniques like compression and encryption are related to this concept.
3. The third concept deals with the relationship between the information and the unreliable channel through which it is going to be transmitted. This concept leads to the definition of a very important parameter called the channel capacity. A well-known information theory technique called error-correction coding is closely related to this concept. This type of coding forms the main subject of this book.

One of the most used techniques in information theory is a procedure called coding, which is intended to optimize transmission and to make efficient use of the capacity of a given channel.
Table 1.1 Coding: a codeword for each message

<table>
<thead>
<tr>
<th>Messages</th>
<th>Codewords</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_1</td>
<td>101</td>
</tr>
<tr>
<td>s_2</td>
<td>01</td>
</tr>
<tr>
<td>s_3</td>
<td>110</td>
</tr>
<tr>
<td>s_4</td>
<td>000</td>
</tr>
</tbody>
</table>

In general terms, coding is a bijective assignment between a set of messages to be transmitted, and a set of codewords that are used for transmitting these messages. Usually this procedure adopts the form of a table in which each message of the transmission is in correspondence with the codeword that represents it (see an example in Table 1.1).

Table 1.1 shows four codewords used for representing four different messages. As seen in this simple example, the length of the codeword is not constant. One important property of a coding table is that it is constructed in such a way that every codeword is uniquely decodable. This means that in the transmission of a sequence composed of these codewords there should be only one possible way of interpreting that sequence. This is necessary when variable-length coding is used.

If the code shown in Table 1.1 is compared with a constant-length code for the same case, constituted from four codewords of two bits, 00, 01, 10, 11, it is seen that the code in Table 1.1 adds redundancy. Assuming equally likely messages, the average number of transmitted bits per symbol is equal to 2.75. However, if for instance symbol s_2 were characterized by a probability of being transmitted of 0.76, and all other symbols in this code were characterized by a probability of being transmitted equal to 0.08, then this source would transmit an average number of bits per symbol of 2.24 bits. As seen in this simple example, a level of compression is possible when the information source is not uniform, that is, when a source generates messages that are not equally likely.

The source information measure, the channel capacity measure and coding are all related by one of the Shannon theorems, the channel coding theorem, which is stated as follows:

*If the information rate of a given source does not exceed the capacity of a given channel, then there exists a coding technique that makes possible transmission through this unreliable channel with an arbitrarily low error rate.*

This important theorem predicts the possibility of error-free transmission through a noisy or unreliable channel. This is obtained by using coding. The above theorem is due to Claude Shannon [1, 2], and states the restrictions on the transmission of information through a noisy channel, stating also that the solution for overcoming those restrictions is the application of a rather sophisticated coding technique. What is not formally stated is how to implement this coding technique.

A block diagram of a communication system as related to information theory is shown in Figure 1.1.

The block diagram seen in Figure 1.1 shows two types of encoders. The channel encoder is designed to perform error correction with the aim of converting an unreliable channel into
a reliable one. On the other hand, there also exists a source encoder that is designed to make the source information rate approach the channel capacity. The destination is also called the information sink.

Some concepts relating to the transmission of discrete information are introduced in the following sections.

1.1 Information

1.1.1 A Measure of Information

From the point of view of information theory, information is not knowledge, as commonly understood, but instead relates to the probabilities of the symbols used to send messages between a source and a destination over an unreliable channel. A quantitative measure of symbol information is related to its probability of occurrence, either as it emerges from a source or when it arrives at its destination. The less likely the event of a symbol occurrence, the higher is the information provided by this event. This suggests that a quantitative measure of symbol information will be inversely proportional to the probability of occurrence.

Assuming an arbitrary message $x_i$ which is one of the possible messages from a set a given discrete source can emit, and $P(x_i) = P_i$ is the probability that this message is emitted, the output of this information source can be modelled as a random variable $X$ that can adopt any of the possible values $x_i$, so that $P(X = x_i) = P_i$. Shannon defined a measure of the information for the event $x_i$ by using a logarithmic measure operating over the base $b$:

$$I_i \equiv - \log_b P_i = \log_b \left( \frac{1}{P_i} \right)$$

(1)

The information of the event depends only on its probability of occurrence, and is not dependent on its content.
The base of the logarithmic measure can be converted by using

\[ \log_a(x) = \log_b(x) \frac{1}{\log_b(a)} \]  

(2)

If this measure is calculated to base 2, the information is said to be measured in \textit{bits}. If the measure is calculated using natural logarithms, the information is said to be measured in \textit{nats}. As an example, if the event is characterized by a probability of \( P_i = 1/2 \), the corresponding information is \( I_i = 1 \) bit. From this point of view, a bit is the amount of information obtained from one of two possible, and equally likely, events. This use of the term \textit{bit} is essentially different from what has been described as the binary digit. In this sense the bit acts as the unit of the measure of information.

Some properties of information are derived from its definition:

\[
I_i \geq 0 \quad 0 \leq P_i \leq 1 \\
I_i \to 0 \quad \text{if} \quad P_i \to 1 \\
I_i > I_j \quad \text{if} \quad P_i < P_j
\]

For any two independent source messages \( x_i \) and \( x_j \) with probabilities \( P_i \) and \( P_j \) respectively, and with joint probability \( P(x_i, x_j) = P_i P_j \), the information of the two messages is the addition of the information in each message:

\[
I_{ij} = \log_b \frac{1}{P_i P_j} = \log_b \frac{1}{P_i} + \log_b \frac{1}{P_j} = I_i + I_j
\]

1.2 Entropy and Information Rate

In general, an information source generates any of a set of \( M \) different symbols, which are considered as representatives of a discrete random variable \( X \) that adopts any value in the range \( A = \{x_1, x_2, \ldots, x_M\} \). Each symbol \( x_i \) has the probability \( P_i \) of being emitted and contains information \( I_i \). The symbol probabilities must be in agreement with the fact that at least one of them will be emitted, so

\[
\sum_{i=1}^{M} P_i = 1
\]

(3)

The source symbol probability distribution is stationary, and the symbols are independent and transmitted at a rate of \( r \) symbols per second. This description corresponds to a discrete memoryless source (DMS), as shown in Figure 1.2.

Each symbol contains the information \( I_i \), so that the set \( \{I_1, I_2, \ldots, I_M\} \) can be seen as a discrete random variable with average information

\[
H_b(X) = \sum_{i=1}^{M} P_i I_i = \sum_{i=1}^{M} P_i \log_b \left( \frac{1}{P_i} \right)
\]

(4)
Discrete memoryless source

![Figure 1.2](image)

Figure 1.2 A discrete memoryless source

The function so defined is called the entropy of the source. When base 2 is used, the entropy is measured in bits per symbol:

$$H(X) = \sum_{i=1}^{M} P_i I_i = \sum_{i=1}^{M} P_i \log_2 \left( \frac{1}{P_i} \right) \text{ bits per symbol} \quad (5)$$

The symbol information value when $P_i = 0$ is mathematically undefined. To solve this situation, the following condition is imposed: $I_i = \infty$ if $P_i = 0$. Therefore $P_i \log_2 \left( \frac{1}{P_i} \right) = 0$ (L’Hôpital’s rule) if $P_i = 0$. On the other hand, $P_i \log \left( \frac{1}{P_i} \right) = 0$ if $P_i = 1$.

**Example 1.1:** Suppose that a DMS is defined over the range of $X$, $A = \{x_1, x_2, x_3, x_4\}$, and the corresponding probability values for each symbol are $P(X = x_1) = \frac{1}{2}$, $P(X = x_2) = P(X = x_3) = \frac{1}{8}$ and $P(X = x_4) = \frac{1}{4}$.

Entropy for this DMS is evaluated as

$$H(X) = \sum_{i=1}^{4} P_i \log_2 \left( \frac{1}{P_i} \right) = \frac{1}{2} \log_2(2) + \frac{1}{8} \log_2(8) + \frac{1}{8} \log_2(8) + \frac{1}{4} \log_2(4)$$

$$= 1.75 \text{ bits per symbol}$$

**Example 1.2:** A source characterized in the frequency domain with a bandwidth of $W = 4000$ Hz is sampled at the Nyquist rate, generating a sequence of values taken from the range $A = \{-2, -1, 0, 1, 2\}$ with the following corresponding set of probabilities $\left\{ \frac{1}{2}, \frac{1}{3}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16} \right\}$.

Calculate the source rate in bits per second.

Entropy is first evaluated as

$$H(X) = \sum_{i=1}^{5} P_i \log_2 \left( \frac{1}{P_i} \right) = \frac{1}{2} \log_2(2) + \frac{1}{4} \log_2(4) + \frac{1}{8} \log_2(8)$$

$$+ 2 \times \frac{1}{16} \log_2(16) = \frac{15}{8} \text{ bits per sample}$$

The minimum sampling frequency is equal to 8000 samples per second, so that the information rate is equal to 15 kbps.

Entropy can be evaluated to a different base by using

$$H_b(X) = \frac{H(X)}{\log_2(b)} \quad (6)$$
Entropy $H(X)$ can be understood as the mean value of the information per symbol provided by the source being measured, or, equivalently, as the mean value experienced by an observer before knowing the source output. In another sense, entropy is a measure of the randomness of the source being analysed. The entropy function provides an adequate quantitative measure of the parameters of a given source and is in agreement with physical understanding of the information emitted by a source.

Another interpretation of the entropy function [5] is seen by assuming that if $n \gg 1$ symbols are emitted, $nH(X)$ bits is the total amount of information emitted. As the source generates $r$ symbols per second, the whole emitted sequence takes $n/r$ seconds. Thus, information will be transmitted at a rate of

$$\frac{nH(X)}{(n/r)} \text{ bps} \quad (7)$$

The information rate is then equal to

$$R = rH(X) \text{ bps} \quad (8)$$

The Shannon theorem states that information provided by a given DMS can be coded using binary digits and transmitted over an equivalent noise-free channel at a rate of

$$r_b \geq R \text{ symbols or binary digits per second}$$

It is again noted here that the bit is the unit of information, whereas the symbol or binary digit is one of the two possible symbols or signals ‘0’ or ‘1’, usually also called bits.

**Theorem 1.1:** Let $X$ be a random variable that adopts values in the range $A = \{x_1, x_2, \ldots, x_M\}$ and represents the output of a given source. Then it is possible to show that

$$0 \leq H(X) \leq \log_2(M) \quad (9)$$

Additionally,

$$H(X) = 0 \quad \text{if and only if } P_i = 1 \text{ for some } i$$

$$H(X) = \log_2(M) \quad \text{if and only if } P_i = 1/M \text{ for every } i \quad (10)$$

The condition $0 \leq H(X)$ can be verified by applying the following:

$$P_i \log_2(1/P_i) \to 0 \quad \text{if } P_i \to 0$$

The condition $H(X) \leq \log_2(M)$ can be verified in the following manner:

Let $Q_1, Q_2, \ldots, Q_M$ be arbitrary probability values that are used to replace terms $1/P_i$ by the terms $Q_i/P_i$ in the expression of the entropy [equation (5)]. Then the following inequality is used:

$$\ln(x) \leq x - 1$$

where equality occurs if $x = 1$ (see Figure 1.3).
After converting entropy to its natural logarithmic form, we obtain

$$\sum_{i=1}^{M} P_i \log_2 \left( \frac{Q_i}{P_i} \right) = \frac{1}{\ln(2)} \sum_{i=1}^{M} P_i \ln \left( \frac{Q_i}{P_i} \right)$$

and if $x = Q_i / P_i$,

$$\sum_{i=1}^{M} P_i \ln \left( \frac{Q_i}{P_i} \right) \leq \sum_{i=1}^{M} P_i \left( \frac{Q_i}{P_i} - 1 \right) = \sum_{i=1}^{M} Q_i - \sum_{i=1}^{M} P_i$$ (11)

As the coefficients $Q_i$ are probability values, they fit the normalizing condition $\sum_{i=1}^{M} Q_i \leq 1$, and it is also true that $\sum_{i=1}^{M} P_i = 1$.

Then

$$\sum_{i=1}^{M} P_i \log_2 \left( \frac{Q_i}{P_i} \right) \leq 0$$ (12)

If now the probabilities $Q_i$ adopt equally likely values $Q_i = 1/M$,

$$\sum_{i=1}^{M} P_i \log_2 \left( \frac{1}{P_i M} \right) = \sum_{i=1}^{M} P_i \log_2 \left( \frac{1}{P_i} \right) - \sum_{i=1}^{M} P_i \log_2(M) = H(X) - \log_2(M) \leq 0$$

$$H(X) \leq \log_2(M)$$ (13)

**Figure 1.3** Inequality $\ln(x) \leq x - 1$
In the above inequality, equality occurs when \(\log_2 \left( \frac{1}{P_i} \right) = \log_2(M)\), which means that \(P_i = \frac{1}{M}\).

The maximum value of the entropy is then \(\log_2(M)\), and occurs when all the symbols transmitted by a given source are equally likely. Uniform distribution corresponds to maximum entropy.

In the case of a binary source \((M = 2)\) and assuming that the probabilities of the symbols are the values

\[
P_0 = \alpha \quad P_1 = 1 - \alpha
\]

(14)

the entropy is equal to

\[
H(X) = \Omega(\alpha) = \alpha \log_2 \left( \frac{1}{\alpha} \right) + (1 - \alpha) \log_2 \left( \frac{1}{1-\alpha} \right)
\]

(15)

This expression is depicted in Figure 1.4.

The maximum value of this function is given when \(\alpha = 1 - \alpha\), that is, \(\alpha = 1/2\), so that the entropy is equal to \(H(X) = \log_2 2 = 1\) bps. (This is the same as saying one bit per binary digit or binary symbol.)

When \(\alpha \to 1\), entropy tends to zero. The function \(\Omega(\alpha)\) will be used to represent the entropy of the binary source, evaluated using logarithms to base 2.

**Example 1.3:** A given source emits \(r = 3000\) symbols per second from a range of four symbols, with the probabilities given in Table 1.2.
The entropy is evaluated as

\[ H(X) = 2 \times \frac{1}{3} \times \log_2(3) + 2 \times \frac{1}{6} \times \log_2(6) = 1.9183 \text{ bits per symbol} \]

And this value is close to the maximum possible value, which is \( \log_2(4) = 2 \text{ bits per symbol} \).

The information rate is equal to

\[ R = rH(X) = (3000)1.9183 = 5754.9 \text{ bps} \]

### 1.3 Extended DMSs

In certain circumstances it is useful to consider information as grouped into blocks of symbols. This is generally done in binary format. For a memoryless source that takes values in the range \( \{x_1, x_2, \ldots, x_M\} \), and where \( P_i \) is the probability that the symbol \( x_i \) is emitted, the order \( n \) extension of the range of a source has \( M^n \) symbols \( \{y_1, y_2, \ldots, y_{M^n}\} \). The symbol \( y_i \) is constituted from a sequence of \( n \) symbols \( x_{ij} \). The probability \( P(Y = y_i) \) is the probability of the corresponding sequence \( x_{i1}, x_{i2}, \ldots, x_{in} \).

\[
P(Y = y_i) = P_{i1}, P_{i2}, \ldots, P_{in}
\]

where \( y_i \) is the symbol of the extended source that corresponds to the sequence \( x_{i1}, x_{i2}, \ldots, x_{in} \). Then

\[
H(X^n) = \sum_{y \in x^n} P(y_i) \log_2 \left( \frac{1}{P(y_i)} \right)
\]  

**Example 1.4:** Construct the order 2 extension of the source of Example 1.1, and calculate its entropy.

Symbols of the original source are characterized by the probabilities \( P(X = x_1) = 1/2, P(X = x_2) = P(X = x_3) = 1/8 \) and \( P(X = x_4) = 1/4 \).

Symbol probabilities for the desired order 2 extended source are given in Table 1.3.

The entropy of this extended source is equal to

\[
H(X^2) = \sum_{i=1}^{M^2} P_i \log_2 \left( \frac{1}{P_i} \right)
\]

\[
= 0.25 \log_2(4) + 2 \times 0.125 \log_2(8) + 5 \times 0.0625 \log_2(16)
\]

\[
+ 4 \times 0.03125 \log_2(32) + 4 \times 0.015625 \log_2(64) = 3.5 \text{ bits per symbol}
\]
Table 1.3  Symbols of the order 2 extended source and their probabilities for Example 1.4

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Probability</th>
<th>Symbol</th>
<th>Probability</th>
<th>Symbol</th>
<th>Probability</th>
<th>Symbol</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₁x₁</td>
<td>0.25</td>
<td>x₂x₁</td>
<td>0.0625</td>
<td>x₃x₁</td>
<td>0.0625</td>
<td>x₄x₁</td>
<td>0.125</td>
</tr>
<tr>
<td>x₁x₂</td>
<td>0.0625</td>
<td>x₂x₂</td>
<td>0.015625</td>
<td>x₃x₂</td>
<td>0.015625</td>
<td>x₄x₂</td>
<td>0.03125</td>
</tr>
<tr>
<td>x₁x₃</td>
<td>0.0625</td>
<td>x₂x₃</td>
<td>0.015625</td>
<td>x₃x₃</td>
<td>0.015625</td>
<td>x₄x₃</td>
<td>0.03125</td>
</tr>
<tr>
<td>x₁x₄</td>
<td>0.125</td>
<td>x₂x₄</td>
<td>0.03125</td>
<td>x₃x₄</td>
<td>0.03125</td>
<td>x₄x₄</td>
<td>0.0625</td>
</tr>
</tbody>
</table>

As seen in this example, the order 2 extended source has an entropy which is twice that of the entropy of the original, non-extended source. It can be shown that the order \( n \) extension of a DMS fits the condition \( H(X^n) = nH(X) \).

1.4 Channels and Mutual Information

1.4.1 Information Transmission over Discrete Channels

A quantitative measure of source information has been introduced in the above sections. Now the transmission of that information through a given channel will be considered. This will provide a quantitative measure of the information received after its transmission through that channel. Here attention is on the transmission of the information, rather than on its generation.

A channel is always a medium through which the information being transmitted can suffer from the effect of noise, which produces errors, that is, changes of the values initially transmitted. In this sense there will be a probability that a given transmitted symbol is converted into another symbol. From this point of view the channel is considered as unreliable. The Shannon channel coding theorem gives the conditions for achieving reliable transmission through an unreliable channel, as stated previously.

1.4.2 Information Channels

**Definition 1.1:** An information channel is characterized by an input range of symbols \( \{x_1, x_2, \ldots, x_U\} \), an output range \( \{y_1, y_2, \ldots, y_V\} \) and a set of conditional probabilities \( P(y_j/x_i) \) that determines the relationship between the input \( x_i \) and the output \( y_j \). This conditional probability corresponds to that of receiving symbol \( y_j \) if symbol \( x_i \) was previously transmitted, as shown in Figure 1.5.

The set of probabilities \( P(y_j/x_i) \) is arranged into a matrix \( P_{ch} \) that characterizes completely the corresponding discrete channel:

\[
P_{ij} = P(y_j/x_i)
\]
Figure 1.5  A discrete transmission channel

\[
P_{\text{ch}} = \begin{bmatrix}
P(y_1/x_1) & P(y_2/x_1) & \cdots & P(y_V/x_1) \\
P(y_1/x_2) & P(y_2/x_2) & \cdots & P(y_V/x_2) \\
\vdots & \vdots & \ddots & \vdots \\
P(y_1/x_U) & P(y_2/x_U) & \cdots & P(y_V/x_U)
\end{bmatrix}
\]  \hspace{1cm} (18)

\[
P_{\text{ch}} = \begin{bmatrix}
P_{11} & P_{12} & \cdots & P_{1V} \\
P_{21} & P_{22} & \cdots & P_{2V} \\
\vdots & \vdots & \ddots & \vdots \\
P_{U1} & P_{U2} & \cdots & P_{UV}
\end{bmatrix}
\]  \hspace{1cm} (19)

Each row in this matrix corresponds to an input, and each column corresponds to an output. Addition of all the values of a row is equal to one. This is because after transmitting a symbol \(x_i\), there must be a received symbol \(y_j\) at the channel output.

Therefore,

\[
\sum_{j=1}^{V} P_{ij} = 1, \quad i = 1, 2, \ldots, U
\]  \hspace{1cm} (20)

**Example 1.5:** The binary symmetric channel (BSC).

The BSC is characterized by a probability \(p\) that one of the binary symbols converts into the other one (see Figure 1.6). Each binary symbol has, on the other hand, a probability of being transmitted. The probabilities of a 0 or a 1 being transmitted are \(\alpha\) and \(1 - \alpha\) respectively.

According to the notation used,

\[x_1 = 0, \quad x_2 = 1 \quad \text{and} \quad y_1 = 0, \quad y_2 = 1\]
The probability matrix for the BSC is equal to

$$P_{ch} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

(21)

**Example 1.6:** The binary erasure channel (BEC).

In its most basic form, the transmission of binary information involves sending two different waveforms to identify the symbols ‘0’ and ‘1’. At the receiver, normally an optimum detection operation is used to decide whether the waveform received, affected by filtering and noise in the channel, corresponds to a ‘0’ or a ‘1’. This operation, often called matched filter detection, can sometimes give an indecisive result. If confidence in the received symbol is not high, it may be preferable to indicate a doubtful result by means of an erasure symbol. Correction of the erasure symbols is then normally carried out by other means in another part of the system.

In other scenarios the transmitted information is coded, which makes it possible to detect if there are errors in a bit or packet of information. In these cases it is also possible to apply the concept of data erasures. This is used, for example, in the concatenated coding system of the compact disc, where on receipt of the information the first decoder detects errors and marks or erases a group of symbols, thus enabling the correction of these symbols in the second decoder. Another example of the erasure channel arises during the transmission of packets over the Internet. If errors are detected in a received packet, then they can be erased, and the erasures corrected by means of retransmission protocols (normally involving the use of a parallel feedback channel).

The use of erasures modifies the BSC model, giving rise to the BEC, as shown in Figure 1.7.

For this channel, $0 \leq p \leq 1/2$, where $p$ is the erasure probability, and the channel model has two inputs and three outputs. When the received values are unreliable, or if blocks are detected.

$$P(0) = \alpha \quad x_1 \quad 0 \quad 1-p \quad 0 \quad y_1$$

$$P(1) = 1 - \alpha \quad x_2 \quad 1 \quad 1-p \quad 1 \quad y_3$$

Figure 1.7 Binary erasure channel
to contain errors, then erasures are declared, indicated by the symbol ‘?’. The probability matrix of the BEC is the following:

\[
P_{\text{ch}} = \begin{bmatrix}
1-p & p & 0 \\
0 & p & 1-p
\end{bmatrix}
\] (22)

### 1.5 Channel Probability Relationships

As stated above, the probability matrix \( P_{\text{ch}} \) characterizes a channel. This matrix is of order \( U \times V \) for a channel with \( U \) input symbols and \( V \) output symbols. Input symbols are characterized by the set of probabilities \( \{ P(x_1), P(x_2), \ldots, P(x_U) \} \), whereas output symbols are characterized by the set of probabilities \( \{ P(y_1), P(y_2), \ldots, P(y_V) \} \).

\[
P_{\text{ch}} = \begin{bmatrix}
P_{11} & P_{12} & \cdots & P_{1V} \\
P_{21} & P_{22} & \cdots & P_{2V} \\
\vdots & \vdots & \ddots & \vdots \\
P_{U1} & P_{U2} & \cdots & P_{UV}
\end{bmatrix}
\]

The relationships between input and output probabilities are the following: The symbol \( y_1 \) can be received in \( U \) different ways. In fact this symbol can be received with probability \( P_{11} \) if symbol \( x_1 \) was actually transmitted, with probability \( P_{21} \) if symbol \( x_2 \) was actually transmitted, and so on.

Any of the \( U \) input symbols can be converted by the channel into the output symbol \( y_1 \). The probability of the reception of symbol \( y_1 \), \( P(y_1) \), is calculated as \( P(y_1) = P_{11}P(x_1) + P_{21}P(x_2) + \cdots + P_{U1}P(x_U) \). Calculation of the probabilities of the output symbols leads to the following system of equations:

\[
\begin{align*}
P_{11}P(x_1) + P_{21}P(x_2) + \cdots + P_{U1}P(x_U) &= P(y_1) \\
P_{12}P(x_1) + P_{22}P(x_2) + \cdots + P_{U2}P(x_U) &= P(y_2) \\
\vdots \\
P_{1V}P(x_1) + P_{2V}P(x_2) + \cdots + P_{UV}P(x_U) &= P(y_V)
\end{align*}
\] (23)

Output symbol probabilities are calculated as a function of the input symbol probabilities \( P(x_i) \) and the conditional probabilities \( P(y_j/x_i) \). It is however to be noted that knowledge of the output probabilities \( P(y_j) \) and the conditional probabilities \( P(y_j/x_i) \) provides solutions for values of \( P(x_i) \) that are not unique. This is because there are many input probability distributions that give the same output distribution.

Application of the Bayes rule to the conditional probabilities \( P(y_j/x_i) \) allows us to determine the conditional probability of a given input \( x_i \) after receiving a given output \( y_j \):

\[
P(x_i/y_j) = \frac{P(y_j/x_i)P(x_i)}{P(y_j)}
\] (24)
By combining this expression with expression (23), equation (24) can be written as

$$P(x_i/y_j) = \frac{P(y_j/x_i)P(x_i)}{\sum_{i=1}^{U} P(y_j/x_i)P(x_i)} \quad (25)$$

Conditional probabilities $P(y_j/x_i)$ are usually called forward probabilities, and conditional probabilities $P(x_i/y_j)$ are known as backward probabilities. The numerator in the above expression describes the probability of the joint event:

$$P(x_i, y_j) = P(y_j/x_i)P(x_i) = P(x_i/y_j)P(y_j) \quad (26)$$

Example 1.7: Consider the binary channel for which the input range and output range are in both cases equal to \{0, 1\}. The corresponding transition probability matrix is in this case equal to

$$P_{ch} = \begin{bmatrix} 3/4 & 1/4 \\ 1/8 & 7/8 \end{bmatrix}$$

Figure 1.8 represents this binary channel.

Source probabilities provide the statistical information about the input symbols. In this case it happens that $P(X = 0) = 4/5$ and $P(X = 1) = 1/5$. According to the transition probability matrix for this case,

- $P(Y = 0/X = 0) = 3/4$
- $P(Y = 1/X = 0) = 1/4$
- $P(Y = 0/X = 1) = 1/8$
- $P(Y = 1/X = 1) = 7/8$

These values can be used to calculate the output symbol probabilities:

- $P(Y = 0) = P(Y = 0/X = 0)P(X = 0) + P(Y = 0/X = 1)P(X = 1)$
  $$= \frac{3}{4} \times \frac{4}{5} + \frac{1}{8} \times \frac{1}{5} = \frac{25}{40}$$

- $P(Y = 1) = P(Y = 1/X = 0)P(X = 0) + P(Y = 1/X = 1)P(X = 1)$
  $$= \frac{1}{4} \times \frac{4}{5} + \frac{7}{8} \times \frac{1}{5} = \frac{15}{40}$$

which confirms that $P(Y = 0) + P(Y = 1) = 1$ is true.
These values can be used to evaluate the backward conditional probabilities:

\[
P(X = 0 / Y = 0) = \frac{P(Y = 0 / X = 0)P(X = 0)}{P(Y = 0)} = \frac{(3/4)(4/5)}{(25/40)} = \frac{24}{25}
\]

\[
P(X = 0 / Y = 1) = \frac{P(Y = 1 / X = 0)P(X = 0)}{P(Y = 1)} = \frac{(1/4)(4/5)}{(15/40)} = \frac{8}{15}
\]

\[
P(X = 1 / Y = 1) = \frac{P(Y = 1 / X = 1)P(X = 1)}{P(Y = 1)} = \frac{(7/8)(1/5)}{(15/40)} = \frac{7}{15}
\]

\[
P(X = 1 / Y = 0) = \frac{P(Y = 0 / X = 1)P(X = 1)}{P(Y = 0)} = \frac{(1/8)(1/5)}{(25/40)} = \frac{1}{25}
\]

### 1.6 The A Priori and A Posteriori Entropies

The probability of occurrence of a given output symbol \(y_j\) is \(P(y_j)\), calculated using expression (23). However, if the actual transmitted symbol \(x_i\) is known, then the related conditional probability of the output symbol becomes \(P(y_j / x_i)\). In the same way, the probability of a given input symbol, initially \(P(x_i)\), can also be refined if the actual output is known. Thus, if the received symbol \(y_j\) appears at the output of the channel, then the related input symbol conditional probability becomes \(P(x_i / y_j)\).

The probability \(P(x_i)\) is known as the a priori probability; that is, it is the probability that characterizes the input symbol before the presence of any output symbol is known. Normally, this probability is equal to the probability that the input symbol has of being emitted by the source (the source symbol probability). The probability \(P(x_i / y_j)\) is an estimate of the symbol \(x_i\) after knowing that a given symbol \(y_j\) appeared at the channel output, and is called the a posteriori probability.

As has been defined, the source entropy is an average calculated over the information of a set of symbols for a given source:

\[
H(X) = \sum_i P(x_i) \log_2 \left[ \frac{1}{P(x_i)} \right]
\]

This definition corresponds to the a priori entropy. The a posteriori entropy is given by the following expression:

\[
H(X / y_j) = \sum_i P(x_i / y_j) \log_2 \left[ \frac{1}{P(x_i / y_j)} \right] \quad i = 1, 2, \ldots, U
\]

### Example 1.8: Determine the a priori and a posteriori entropies for the channel of Example 1.7.

The a priori entropy is equal to

\[
H(X) = \frac{4}{5} \log_2 \left( \frac{5}{4} \right) + \frac{1}{5} \log_2(5) = 0.7219 \text{ bits}
\]
Assuming that a ‘0’ is present at the channel output,
\[ H(X/0) = \frac{24}{25} \log_2 \left( \frac{25}{24} \right) + \frac{1}{25} \log_2 (25) = 0.2423 \text{ bits} \]
and in the case of a ‘1’ present at the channel output,
\[ H(X/1) = \frac{8}{15} \log_2 \left( \frac{15}{8} \right) + \frac{7}{15} \log_2 \left( \frac{15}{7} \right) = 0.9968 \text{ bits} \]
Thus, entropy decreases after receiving a ‘0’ and increases after receiving a ‘1’.

1.7 Mutual Information

According to the description of a channel depicted in Figure 1.5, \( P(x_i) \) is the probability that a given input symbol is emitted by the source, \( P(y_j) \) determines the probability that a given output symbol \( y_j \) is present at the channel output, \( P(x_i, y_j) \) is the joint probability of having symbol \( x_i \) at the input and symbol \( y_j \) at the output, \( P(y_j/x_i) \) is the probability that the channel converts the input symbol \( x_i \) into the output symbol \( y_j \) and \( P(x_i/y_j) \) is the probability that \( x_i \) has been transmitted if \( y_j \) is received.

1.7.1 Mutual Information: Definition

Mutual information measures the information transferred when \( x_i \) is sent and \( y_j \) is received, and is defined as
\[
I(x_i, y_j) = \log_2 \frac{P(x_i/y_j)}{P(x_i)} \text{ bits} \tag{28}
\]
In a noise-free channel, each \( y_j \) is uniquely connected to the corresponding \( x_i \), and so they constitute an input–output pair \((x_i, y_j)\) for which \( P(x_i/y_j) = 1 \) and \( I(x_i, y_j) = \log_2 \frac{1}{P(x_i)} \) bits; that is, the transferred information is equal to the self-information that corresponds to the input \( x_i \).
In a very noisy channel, the output \( y_j \) and the input \( x_i \) would be completely uncorrelated, and so \( P(x_i/y_j) = P(x_i) \) and also \( I(x_i, y_j) = 0 \); that is, there is no transference of information. In general, a given channel will operate between these two extremes.

The mutual information is defined between the input and the output of a given channel. An average of the calculation of the mutual information for all input–output pairs of a given channel is the average mutual information:
\[
I(X, Y) = \sum_{i,j} P(x_i, y_j) I(x_i, y_j) = \sum_{i,j} P(x_i, y_j) \log_2 \left[ \frac{P(x_i/y_j)}{P(x_i)} \right] \text{ bits per symbol} \tag{29}
\]
This calculation is done over the input and output alphabets. The average mutual information measures the average amount of source information obtained from each output symbol.
The following expressions are useful for modifying the mutual information expression:

\[ P(x_i, y_j) = P(x_i/y_j)P(y_j) = P(y_j/x_i)P(x_i) \]
\[ P(y_j) = \sum_i P(y_j/x_i)P(x_i) \]
\[ P(x_i) = \sum_j P(x_i/y_j)P(y_j) \]

Then

\[ I(X, Y) = \sum_{i,j} P(x_i, y_j)I(x_i, y_j) = \sum_{i,j} P(x_i, y_j) \log_2 \left[ \frac{1}{P(x_i)} \right] - \sum_{i,j} P(x_i, y_j) \log_2 \left[ \frac{1}{P(x_i/y_j)} \right] \]  

\[ \sum_{i,j} P(x_i, y_j) \log_2 \left[ \frac{1}{P(x_i)} \right] = \sum_i \left[ \sum_j P(x_i/y_j)P(y_j) \right] \log_2 \frac{1}{P(x_i)} \]
\[ I(X, Y) = H(X) \]

\[ I(X, Y) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{1}{P(x_i/y_j)} \]  

where \( H(X/Y) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{1}{P(x_i/y_j)} \) is usually called the **equivocation**.

In a sense, the equivocation can be seen as the information lost in the noisy channel, and is a function of the backward conditional probability. The observation of an output symbol \( y_j \) provides \( H(X) - H(X/Y) \) bits of information. This difference is the mutual information of the channel.

### 1.7.2 Mutual Information: Properties

Since

\[ P(x_i/y_j)P(y_j) = P(y_j/x_i)P(x_i) \]

the mutual information fits the condition

\[ I(X, Y) = I(Y, X) \]

And by interchanging input and output it is also true that

\[ I(X, Y) = H(Y) - H(Y/X) \]  

where

\[ H(Y) = \sum_j P(y_j) \log_2 \frac{1}{P(y_j)} \]
which is the destination entropy or output channel entropy:

$$H(Y/X) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{1}{P(y_j/x_i)}$$  \hspace{1cm} (33)

This last entropy is usually called the noise entropy.

Thus, the information transferred through the channel is the difference between the output entropy and the noise entropy. Alternatively, it can be said that the channel mutual information is the difference between the number of bits needed for determining a given input symbol before knowing the corresponding output symbol, and the number of bits needed for determining a given input symbol after knowing the corresponding output symbol, $I(X, Y) = H(X) - H(X/Y)$.

As the channel mutual information expression is a difference between two quantities, it seems that this parameter can adopt negative values. However, and in spite of the fact that for some $y_j$, $H(X/y_j)$ can be larger than $H(X)$, this is not possible for the average value calculated over all the outputs:

$$\sum_{i,j} P(x_i, y_j) \log_2 \frac{P(x_i/y_j)}{P(x_i)} = \sum_{i,j} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$$

then

$$-I(X, Y) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{P(x_i)P(y_j)}{P(x_i, y_j)} \leq 0$$

because this expression is of the form

$$\sum_{i=1}^{M} P_i \log_2 \left( \frac{Q_i}{P_i} \right) \leq 0$$  \hspace{1cm} (34)

which is the expression (12) used for demonstrating Theorem 1.1. The above expression can be applied due to the factor $P(x_i)P(y_j)$, which is the product of two probabilities, so that it behaves as the quantity $Q_i$, which in this expression is a dummy variable that fits the condition $\sum_i Q_i \leq 1$.

It can be concluded that the average mutual information is a non-negative number. It can also be equal to zero, when the input and the output are independent of each other.

A related entropy called the joint entropy is defined as

$$H(X, Y) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{1}{P(x_i, y_j)}$$

$$= \sum_{i,j} P(x_i) \log_2 \frac{P(x_i)P(y_j)}{P(x_i, y_j)} + \sum_{i,j} \log_2 \frac{P(y_j)}{P(x_i)P(y_j)}$$  \hspace{1cm} (35)

Then the set of all the entropies defined so far can be represented in Figure 1.9. The circles define regions for entropies $H(X)$ and $H(Y)$, the intersection between these two entropies is the mutual information $I(X, Y)$, while the differences between the input and output entropies are $H(X/Y)$ and $H(Y/X)$ respectively (Figure 1.9). The union of these entropies constitutes the joint entropy $H(X, Y)$. 
Example 1.9: Entropies of the binary symmetric channel (BSC).

The BSC is constructed with two inputs \((x_1, x_2)\) and two outputs \((y_1, y_2)\), with alphabets over the range \(A = \{0, 1\}\). The symbol probabilities are \(P(x_1) = \alpha\) and \(P(x_2) = 1 - \alpha\), and the transition probabilities are \(P(y_1/x_2) = P(y_2/x_1) = p\) and \(P(y_1/x_1) = P(y_2/x_2) = 1 - p\) (see Figure 1.10). This means that the error probability \(p\) is equal for the two possible symbols. The average error probability is equal to

\[
P = P(x_1)P(y_2/x_1) + P(x_2)P(y_1/x_2) = \alpha p + (1 - \alpha)p = p
\]

The mutual information can be calculated as

\[
I(X, Y) = H(Y) - H(Y/X)
\]

The output \(Y\) has two symbols \(y_1\) and \(y_2\), such that \(P(y_2) = 1 - P(y_1)\). Since

\[
P(y_1) = P(y_1/x_1)P(x_1) + P(y_1/x_2)P(x_2) = (1 - p)\alpha + p(1 - \alpha)
\]

\[= \alpha - p\alpha + p - p\alpha = \alpha + p - 2p\alpha\] \hspace{1cm} (36)

the destination or sink entropy is equal to

\[
H(Y) = P(y_1)\log_2 \frac{1}{P(y_1)} + [1 - P(y_1)]\log_2 \frac{1}{[1 - P(y_1)]} = \Omega[P(y_1)]
\]

\[= \Omega(\alpha + p - 2p\alpha)\] \hspace{1cm} (37)

Figure 1.10 BSC of Example 1.9
The noise entropy $H(Y/X)$ can be calculated as

$$H(Y/X) = \sum_{i,j} P(x_i, y_j) \log_2 \frac{1}{P(y_j/x_i)}$$

$$= \sum_{i,j} P(y_j/x_i) P(x_i) \log_2 \frac{1}{P(y_j/x_i)}$$

$$= \sum_i P(x_i) \left[ \sum_j P(y_j/x_i) \log_2 \frac{1}{P(y_j/x_i)} \right]$$

$$= P(x_1) \left[ P(y_2/x_1) \log_2 \frac{1}{P(y_2/x_1)} + P(y_1/x_1) \log_2 \frac{1}{P(y_1/x_1)} \right]$$

$$+ P(x_2) \left[ P(y_2/x_2) \log_2 \frac{1}{P(y_2/x_2)} + P(y_1/x_2) \log_2 \frac{1}{P(y_1/x_2)} \right]$$

$$= \alpha \left[ p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p} \right] + (1 - \alpha) \left[ (1 - p) \log_2 \frac{1}{1 - p} + p \log_2 \frac{1}{p} \right]$$

$$= p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p} = \Omega(\alpha)$$  \hspace{1cm} (38)

Note that the noise entropy of the BSC is determined only by the forward conditional probabilities of the channel, being independent of the source probabilities. This facilitates the calculation of the channel capacity for this channel, as explained in the following section.

Finally,

$$I(X, Y) = H(Y) - H(Y/X) = \Omega(\alpha + p - 2\alpha p) - \Omega(p)$$  \hspace{1cm} (39)

The average mutual information of the BSC depends on the source probability $\alpha$ and on the channel error probability $p$.

When the channel error probability $p$ is very small, then

$$I(X, Y) \approx \Omega(\alpha) - \Omega(0) \approx \Omega(\alpha) = H(X)$$

This means that the average mutual information, which represents the amount of information transferred through the channel, is equal to the source entropy. On the other hand, when the channel error probability approaches its maximum value $p \approx 1/2$, then

$$I(X, Y) = \Omega(\alpha + 1/2 - \alpha) - \Omega(1/2) = 0$$

and the average mutual information tends to zero, showing that there is no transference of information between the input and the output.

**Example 1.10**: Entropies of the binary erasure channel (BEC).

The BEC is defined with an alphabet of two inputs and three outputs, with symbol probabilities $P(x_1) = \alpha$ and $P(x_2) = 1 - \alpha$, and transition probabilities $P(y_1/x_1) = 1 - p$ and $P(y_2/x_1) = p$, $P(y_3/x_1) = 0$ and $P(y_1/x_2) = 0$, and $P(y_2/x_2) = p$ and $P(y_3/x_2) = 1 - p$. 
Now to calculate the mutual information as \( I(X, Y) = H(Y) - H(Y/X) \), the following values are determined:

\[
P(y_1) = P(y_1/x_1)P(x_1) + P(y_1/x_2)P(x_2) = \alpha(1 - p) \\
P(y_2) = P(y_2/x_1)P(x_1) + P(y_2/x_2)P(x_2) = p \\
P(y_3) = P(y_3/x_1)P(x_1) + P(y_3/x_2)P(x_2) = (1 - \alpha)(1 - p)
\]

In this way the output or sink entropy is equal to

\[
H(Y) = P(y_1) \log_2 \frac{1}{P(y_1)} + P(y_2) \log_2 \frac{1}{P(y_2)} + P(y_3) \log_2 \frac{1}{P(y_3)} \\
= \alpha(1 - p) \log_2 \frac{1}{\alpha(1 - p)} + p \log_2 \frac{1}{p} + (1 - \alpha)(1 - p) \log_2 \frac{1}{(1 - \alpha)(1 - p)} \\
= (1 - p)\Omega(\alpha) + \Omega(p)
\]

The noise entropy \( H(Y/X) \) remains to be calculated:

\[
H(Y/X) = \sum_{i,j} P(y_j/x_i)P(x_i) \log_2 \frac{1}{P(y_j/x_i)} = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{(1 - p)} = \Omega(p)
\]

after which the mutual information is finally given by

\[
I(X, Y) = H(Y) - H(Y/X) = (1 - p)\Omega(\alpha)
\]

### 1.8 Capacity of a Discrete Channel

The definition of the average mutual information allows us to introduce the concept of channel capacity. This parameter characterizes the channel and is basically defined as the maximum possible value that the average mutual information can adopt for a given channel:

\[
C_s = \max_{P(x)} I(X, Y) \text{ bits per symbol}
\]  \hspace{1cm} (40)

It is noted that the definition of the channel capacity involves not only the channel itself but also the source and its statistical properties. However the channel capacity depends only on the conditional probabilities of the channel, and not on the probabilities of the source symbols, since the capacity is a value of the average mutual information given for particular values of the source symbols.

Channel capacity represents the maximum amount of information per symbol transferred through that channel.

In the case of the BSC, maximization of the average mutual information is obtained by maximizing the expression

\[
C_s = \max_{P(x)} I(X, Y) = \max_{P(x)} \{H(Y) - H(Y/X)\} \\
= \max_{P(x)} \{\Omega(\alpha + p - 2\alpha p) - \Omega(p)\} = 1 - \Omega(p) = 1 - H(p)
\]  \hspace{1cm} (41)

which is obtained when \( \alpha = 1 - \alpha = 1/2 \).
If the maximum rate of symbols per second, $s$, allowed in the channel is known, then the capacity of the channel per time unit is equal to

$$C = sC_s \text{ bps} \quad (42)$$

which, as will be seen, represents the maximum rate of information transference in the channel.

### 1.9 The Shannon Theorems

#### 1.9.1 Source Coding Theorem

The source coding theorem and the channel coding (channel capacity) theorem are the two main theorems stated by Shannon [1, 2]. The source coding theorem determines a bound on the level of compression of a given information source. The definitions for the different classes of entropies presented in previous sections, and particularly the definition of the source entropy, are applied to the analysis of this theorem.

Information entropy has an intuitive interpretation [1, 6]. If the DMS emits a large number of symbols $n_f$ taken from an alphabet $A = \{x_1, x_2, \ldots, x_M\}$ in the form of a sequence of symbols, symbol $x_1$ will appear $n_fP(x_1)$ times, symbol $x_2$, $n_fP(x_2)$ times, and symbol $x_M$, $n_fP(x_M)$ times. These sequences are known as typical sequences and are characterized by the probability

$$P \approx \prod_{i=1}^{M} \left[ P(x_i) \right]^{n_i P(x_i)} \quad (43)$$

since

$$P(x_i) = 2 \log_2[P(x_i)]$$

$$P \approx \prod_{i=1}^{M} \left[ P(x_i) \right]^{n_i P(x_i)} = \prod_{i=1}^{M} 2^{n_i \log_2[P(x_i)]} = \prod_{i=1}^{M} 2^{n_i \log_2[P(x_i)]} P(x_i)$$

$$= 2^{\sum_{i=1}^{M} n_i \log_2[P(x_i)]} = 2^{-n(H(X))} \quad (44)$$

Typical sequences are those with the maximum probability of being emitted by the information source. Non-typical sequences are those with very low probability of occurrence. This means that of the total of $M^{n_f}$ possible sequences that can be emitted from the information source with alphabet $A = \{x_1, x_2, \ldots, x_M\}$, only $2^{n(H(X))}$ sequences have a significant probability of occurring. An error of magnitude $\varepsilon$ is made by assuming that only $2^{n(H(X))}$ sequences are transmitted instead of the total possible number of them. This error can be arbitrarily small if $n_f \to \infty$. This is the essence of the data compression theorem.

This means that the source information can be transmitted using a significantly lower number of sequences than the total possible number of them.
If only $2^{n_1 H(X)}$ sequences are to be transmitted, and using a binary format of representing information, there will be $n_1 H(X)$ bits needed for representing this information. Since sequences are constituted of symbols, there will be $H(X)$ bits per symbol needed for a suitable representation of this information. This means that the source entropy is the amount of information per symbol of the source.

For a DMS with independent symbols, it can be said that compression of the information provided by this source is possible only if the probability density function of this source is not uniform, that is, if the symbols of this source are not equally likely. As seen in previous sections, a source with $M$ equally likely symbols fits the following conditions:

$$H(X) = \log_2 M, \quad 2^{n_1 H(X)} = 2^{n_1 \log_2 M} = M^n$$  \hspace{1cm} (45)

The number of typical sequences for a DMS with equally likely symbols is equal to the maximum possible number of sequences that this source can emit.

This has been a short introduction to the concept of data and information compression. However, the aim of this chapter is to introduce the main concepts of a technique called error-control coding, closely related to the Shannon channel coding (capacity) theorem.

### 1.9.2 Channel Capacity and Coding

Communication between a source and a destination happens by the sending of information from the former to the latter, through a medium called the communication channel. Communication channels are properly modelled by using the conditional probability matrix defined between the input and the output, which allows us to determine the reliability of the information arriving at the receiver. The important result provided by the Shannon capacity theorem is that it is possible to have an error-free (reliable) transmission through a noisy (unreliable) channel, by means of the use of a rather sophisticated coding technique, as long as the transmission rate is kept to a value less than or equal to the channel capacity. The bound imposed by this theorem is over the transmission rate of the communication, but not over the reliability of the communication.

In the following, transmission of sequences or blocks of $n$ bits over a BSC is considered. In this case the input and the output are $n$-tuples or vectors defined over the extensions $X^n$ and $Y^n$ respectively. The conditional probabilities will be used:

$$P(X/Y) = \prod_{i=1}^{n} P(x_i/y_j)$$

Input and output vectors $X$ and $Y$ are words of $n$ bits. By transmitting a given input vector $X$, and making the assumption that the number of bits $n$ is relatively large, the error probability $p$ of the BSC determines that the output vector $Y$ of this channel will differ in $np$ positions with respect to the input vector $X$.

On the other hand, the number of sequences of $n$ bits with differences in $np$ positions is equal to

$$\binom{n}{np}$$  \hspace{1cm} (46)
By using the Stirling approximation [6]

\[ n! \approx n^n e^{-n} \sqrt{2\pi n} \tag{47} \]

it can be shown that

\[ \left( \frac{n}{np} \right) \approx 2^{n\Omega(p)} \tag{48} \]

This result indicates that for each input block of \( n \) bits, there exists \( 2^{n\Omega(p)} \) possible output sequences as a result of the errors introduced by the channel.

On the other hand, the output of the channel can be considered as a discrete source from which \( 2^{nH(Y)} \) typical sequences can be emitted. Then the amount

\[ M = \frac{2^{nH(Y)}}{2^{n\Omega(p)}} = 2^{n[H(Y) - \Omega(p)]} \tag{49} \]

represents the maximum number of possible inputs able to be transmitted and to be converted by the distortion of the channel into non-overlapping sequences.

The smaller the error probability of the channel, the larger is the number of non-overlapping sequences. By applying the base 2 logarithmic function,

\[ \log_2 M = n [H(Y) - \Omega(p)] \]

and then

\[ R_s = \frac{\log_2 M}{n} = H(Y) - \Omega(p) \tag{50} \]

The probability density function of the random variable \( Y \) depends on the probability density function of the message and on the statistical properties of the channel. There is in general terms a probability density function of the message \( X \) that can maximize the entropy \( H(Y) \). If the input is characterized by a uniform probability density function and the channel is a BSC, the output has a maximum entropy, \( H(Y) = 1 \). This makes the expression (50) adopt its maximum value

\[ R_s = 1 - \Omega(p) \tag{51} \]

which is valid for the BSC.

This is indeed the parameter that has been defined as the channel capacity. This will therefore be the maximum possible transmission rate for the BSC if error-free transmission is desired over that channel. This could be obtained by the use of a rather sophisticated error coding technique.

Equation (51) for the BSC is depicted in Figure 1.11.

The channel capacity is the maximum transmission rate over that channel for reliable transmission. The worst case for the BSC is given when \( p = 1/2 \) because the extreme value \( p = 1 \) corresponds after all to a transmission where the roles of the transmitted symbols are interchanged (binary transmission).

So far, a description of the channel coding theorem has been developed by analysing the communication channel as a medium that distorts the sequences being transmitted.

The channel coding theorem is stated in the following section.
1.9.3 Channel Coding Theorem

The channel capacity of a discrete memoryless channel is equal to

\[ C_s = \max_{P(x)} I(X, Y) \text{ bits per symbol} \]  

(52)

The channel capacity per unit time \( C \) is related to the channel capacity \( C_s \) by the expression \( C = sC_s \). If the transmission rate \( R \) fits the condition \( R < C \), then for an arbitrary value \( \varepsilon > 0 \), there exists a code with block length \( n \) that makes the error probability of the transmission be less than \( \varepsilon \). If \( R > C \) then there is no guarantee of reliable transmission; that is, there is no guarantee that the arbitrary value of \( \varepsilon \) is a bound for the error probability, as it may be exceeded. The limiting value of this arbitrary constant \( \varepsilon \) is zero.

**Example 1.11**: Determine the channel capacity of the channel of Figure 1.12 if all the input symbols are equally likely, and

\[
\begin{align*}
P(y_1/x_1) &= P(y_2/x_2) = P(y_3/x_3) = 0.5 \\
P(y_1/x_2) &= P(y_1/x_3) = 0.25 \\
P(y_2/x_1) &= P(y_2/x_3) = 0.25 \\
P(y_3/x_1) &= P(y_3/x_2) = 0.25
\end{align*}
\]
The channel capacity can be calculated by first determining the mutual information and then maximizing this parameter. This maximization consists of looking for the input probability density function that makes the output entropy be maximal.

In this case the input probability density function is uniform and this makes the output probability density function be maximum. However this is not always the case. In a general case, the probability density function should be selected to maximize the mutual information.

For this example,

$$H(Y/X) = P(x_1)H(Y/X = x_1) + P(x_2)H(Y/X = x_2) + P(x_3)H(Y/X = x_3)$$

and

$$H(Y/X = x_1) = H(Y/X = x_2) = H(Y/X = x_3) = \frac{1}{4} \log_2(4) + \frac{1}{4} \log_2(4) + \frac{1}{2} \log_2(2)$$

$$= 0.5 + 0.5 + 0.5 = 1.5$$

$$H(Y/X) = 1.5$$

Therefore,

$$I(X, Y) = H(Y) - 1.5$$

The output entropy is maximal for an output alphabet with equally likely symbols, so that

$$H(Y) = \frac{1}{3} \log_2(3) + \frac{1}{3} \log_2(3) + \frac{1}{3} \log_2(3) = \log_2(3) = 1.585$$

Then

$$C_s = 1.585 - 1.5 = 0.085 \text{ bits per symbol}$$

This rather small channel capacity is a consequence of the fact that each input symbol has a probability of 1/2 of emerging from the channel in error.
1.10 Signal Spaces and the Channel Coding Theorem

The theory of vector spaces can be applied to the field of the communication signals and is a very useful tool for understanding the Shannon channel coding theorem [2, 5, 6].

For a given signal \( x(t) \) that is transmitted through a continuous channel with a bandwidth \( B \), there is an equivalent representation that is based on the sampling theorem:

\[
x(t) = \sum_{k} x_k \text{sinc}(2Bt - k) \tag{53}
\]

where

\[
x_k = x(kT_s) \quad \text{and} \quad T_s = 1/2B \tag{54}
\]

with \( x_k = x(kT_s) \) being the samples of the signal obtained at a sampling rate \( 1/T_s \).

Signals are in general power limited, and this power limit can be expressed as a function of the samples \( x_k \) as

\[
P = x_k^2 \tag{55}
\]

Assuming that the signal has duration \( T \), this signal can be represented by a discrete number of samples \( n = T/T_s = 2BT \). This means that the \( n \) numbers \( x_1, x_2, \ldots, x_n \) represent this signal. This is true because of the sampling theorem, which states that the signal can be perfectly reconstructed if this set of \( n \) samples is known. This set of numbers can be thought of as a vector, which becomes a vectorial representation of the signal, with the property of allowing us a perfect reconstruction of this signal by calculating

\[
x(t) = \sum_{k} x(kT_s) \text{sinc}(f_s t - k)
\]

\[
f_s = \frac{1}{T_s} \geq 2W \tag{56}
\]

where \( W \) is the bandwidth of the signal that has to fit the condition

\[
W \leq B \leq f_s - W \tag{57}
\]

This vector represents the signal \( x(t) \) and is denoted as \( X = (x_1, x_2, \ldots, x_n) \) with \( n = 2BT = 2WT \). The reconstruction of the signal \( x(t) \), based on this vector representation [expression (53)], is given in terms of a signal representation over a set of orthogonal functions like the sinc functions. The representative vector is \( n \) dimensional. Its norm can be calculated as

\[
||X||^2 = x_1^2 + x_2^2 + \cdots + x_n^2 = \sum_{i=1}^{n} x_i^2 \tag{58}
\]

If the number of samples is large, \( n \gg 1 \), and

\[
\frac{1}{n} ||X||^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \overline{x_k^2} = P \tag{59}
\]

\[
||X|| = \sqrt{nP} = \sqrt{2BTp} \tag{60}
\]
so the norm of the vector is proportional to its power. By allowing the components of the vector \( X \) vary through all their possible values, a hypersphere will appear in the corresponding vector space. This hypersphere is of radius \( ||X|| \), and all the possible vectors will be enclosed by this sphere. The volume of this hypersphere is equal to \( \text{Vol}_n = K_n ||X||^n \).

As noise is also a signal, it can adopt a vector representation. This signal is usually passed through a filter of bandwidth \( B \) and then sampled, and so this set of samples constitutes a vector that represents the filtered noise. This vector will be denoted as \( N = (N_1, N_2, \ldots, N_n) \), and if \( P_N \) is the noise power, then this vector has a norm equal to \( ||N|| = \sqrt{n}P \). Thus, signals and noise have vector representation as shown in Figure 1.13.

Noise in this model is additive and independent of (uncorrelated with) the transmitted signals. During transmission, channel distortion transforms the input vector \( X \) into an output vector \( Y \) whose norm will be equal to \( ||Y|| = ||X + N|| = \sqrt{n}(P + P_N) \) \[2\] (see Figure 1.14). (Signal and noise powers are added as they are uncorrelated.)

1.10.1 Capacity of the Gaussian Channel

The Gaussian channel resulting from the sampling of the signals is a discrete channel, which is described in Figure 1.15.

The variable \( N \) represents the samples of a Gaussian variable and is in turn a Gaussian random variable with squared variance \( P_N \). The signal has a power \( P \). If all the variables are represented by vectors of length \( n \), they are related by

\[
Y = X + N
\]

(61)
Gaussian channel, signals represented by real numbers

\[ Y = X + N \]
\[ \frac{1}{n} \sum_{i=1}^{n} X_i^2 \leq P \]

**Figure 1.15**  Gaussian channel

If the number of samples, \( n \), is large, the noise power can be calculated as the average of the noise samples:

\[ \frac{1}{n} \sum_{i=1}^{n} N_i^2 = \frac{1}{n} \sum_{i=1}^{n} |Y_i - X_i|^2 \leq P_N \]

which means that

\[ |Y - X|^2 \leq nP_N \]  \hspace{1cm} (62)

This can be seen as the noise sphere representing the tip of the output vector \( Y \) around the transmitted (or true) vector \( X \), whose radius is \( \sqrt{nP_N} \), which is proportional to the noise power at the input. Since noise and transmitted signals are uncorrelated,

\[ \frac{1}{n} \sum_{i=1}^{n} X_i^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 + \frac{1}{n} \sum_{i=1}^{n} N_i^2 \leq P + P_N \]  \hspace{1cm} (64)

Then

\[ |Y|^2 \leq n(P + P_N) \]  \hspace{1cm} (65)

and the output sequences are inside \( n \)-dimensional spheres of radius \( \sqrt{n(P + P_N)} \) centred at the origin, as shown in Figure 1.16.

**Figure 1.16**  A representation of the output vector space
Figure 1.16 can be understood as follows. The transmission of the input vector $X$ generates an associated sphere whose radius is proportional to the noise power of the channel, $\sqrt{nP_N}$. In addition, output vectors generate the output vector space, a hypersphere with radius $\sqrt{n(P + P_N)}$. The question is how many spheres of radius $\sqrt{nP_N}$ can be placed, avoiding overlapping, inside a hypersphere of radius $\sqrt{n(P + P_N)}$?

For a given $n$-dimensional hypersphere of radius $R_c$, the volume is equal to

$$V_{ol,n} = K_n R_c^n$$

where $K_n$ is a constant and $R_c$ is the radius of the sphere. The number of non-overlapped messages that are able to be transmitted reliably in this channel is [2, 6]

$$M = \frac{K_n [n(P + P_N)]^{n/2}}{K_n (nP_N)^{n/2}} = \left( \frac{P + P_N}{P_N} \right)^{n/2}$$

The channel capacity, the number of possible signals that can be transmitted reliably, and the length of the transmitted vectors are related as follows:

$$C_s = \frac{1}{n} \log_2(M) = \frac{1}{n} \log_2 \left( 1 + \frac{P}{P_N} \right) = \frac{1}{2} \log_2 \left( 1 + \frac{P}{P_N} \right)$$

A continuous channel with power spectral density $N_0/2$, bandwidth $B$ and signal power $P$ can be converted into a discrete channel by sampling it at the Nyquist rate. The noise sample power is equal to

$$P_N = \int_{-B}^{B} \frac{N_0}{2} \, df = N_0 B$$

Then

$$C_s = \frac{1}{2} \log_2 \left( 1 + \frac{P}{N_0 B} \right)$$

A given signal with bandwidth $W$ transmitted through this channel and sampled at the Nyquist rate will fulfil the condition $W = B$, and will be represented by $2W = 2B$ samples per second.

The channel capacity per second is calculated by multiplying $C_s$, the capacity per symbol or sample, by the number of samples per second of the signal:

$$C = 2BC_s = B \log_2 \left( 1 + \frac{P}{N_0 B} \right) \text{ bps}$$

This Shannon equation states that in order to reliably transmit signals through a given channel they should be selected by taking into account that, after being affected by noise, at the channel output the noise spheres must remain non-overlapped, so that each signal can be properly distinguished.

There will be therefore a number $M$ of coded messages of length $T$, that is, of $M$ coded vectors of $n$ components each, resulting from evaluating how many spheres of radius $\sqrt{nP_N}$
can be placed in a hypersphere (output vector space) of radius \( \sqrt{n(P + P_N)} \), where

\[
M = \left[ \frac{\sqrt{n(P + P_N)}}{\sqrt{n} P_N} \right]^n = \left( 1 + \frac{P}{P_N} \right)^{n/2}
\]

Assuming now that during time \( T \), one of \( \mu \) possible symbols is transmitted at \( r \) symbols per second, the number of different signals that can be constructed is

\[
M = \mu^r T
\]  

(72)

In the particular case of binary signals, that is for \( \mu = 2 \), the number of possible non-overlapped signals is \( M = 2^r T \).

The channel capacity, as defined by Shannon, is the maximum amount of information that can be transmitted per unit time. The Shannon theorem determines the amount of information that can be reliably transmitted through a given channel. The number of possible messages of length \( T \) that can be reliably transmitted is \( M \). From this point of view, the combination of source and channel can be seen as a discrete output source \( Y \) with an alphabet of size \( M \). The maximum entropy of this destination source (or information sink) is achieved when all the output symbols are equally likely, and this entropy is equal to \( \log_2 M \), which is in turn the amount of information provided at this sink. The maximum rate measured in symbols per second is then equal to \( (1/T) \log_2 M \). The channel capacity is measured as the limiting value of this maximum rate when the length of the message tends to infinity:

\[
C = \lim_{T \to \infty} \frac{1}{T} \log_2 M \text{ bps}
\]  

(73)

Then, taking into account previous expressions,

\[
C = \lim_{T \to \infty} \frac{1}{T} \log_2 M = \lim_{T \to \infty} \frac{1}{T} \log_2 (\mu^r T) = \lim_{T \to \infty} \frac{r T}{T} \log_2 \mu = r \log_2 \mu \text{ bps}
\]  

(74)

This calculation considers the channel to be noise free. For instance, in the case of the binary alphabet, \( \mu = 2 \) and \( C = r \). The number of distinguishable signals for reliable transmission is equal to

\[
M \leq \left( 1 + \frac{P}{P_N} \right)^{n/2} \text{ with } n = 2BT
\]  

(75)

\[
C = \lim_{T \to \infty} \frac{1}{T} \log_2 M = \lim_{T \to \infty} \frac{1}{T} \log_2 \left( 1 + \frac{P}{P_N} \right)^{n/2} = \frac{2BT}{2T} \log_2 \left( 1 + \frac{P}{P_N} \right)
\]

\[
= B \log_2 \left( 1 + \frac{P}{P_N} \right) \text{ bps}
\]  

(76)

**Example 1.12:** Find the channel capacity of the telephony channel, assuming that the minimum signal-to-noise ratio of this system is \( (P/P_N)_{\text{dB}} = 30 \text{ dB} \) and the signal and channel transmission bandwidths are both equal to \( W = 3 \text{ kHz} \).

\[
\text{As } (P/P_N)_{\text{dB}} = 30 \text{ dB and } (P/P_N) = 1000
\]

\[
C = 3000 \log_2 (1 + 1000) \approx 30 \text{ kbps}
\]
1.11 Error-Control Coding

Figure 1.17 shows the block diagram of an encoding scheme. The source generates information symbols at a rate $R$. The channel is a BSC, characterized by a capacity $C_s = 1 - \Omega(p)$ and a symbol rate $s$. Mutual information is maximized by previously coding the source in order to make it have equally likely binary symbols or bits. Then, an encoder takes blocks of bits and converts them into $M$-ary equally likely symbols [1, 2] that carry $\log_2 M$ bits per symbol each. In this way, information adopts a format suitable for its transmission through the channel. The BSC encoder represents each symbol by a randomly selected vector or word of $n$ binary symbols.

Each binary symbol of the vector of length $n$ carries an amount of information equal to $\log_2 M / n$. Since $s$ symbols per second are transmitted, the encoded source information rate is

$$R = \frac{s \log_2 M}{n}$$  \hfill (77)

The Shannon theorem requires that

$$R \leq C = sC_s$$

which in this case means that

$$\frac{\log_2 M}{n} \leq C_s$$

$$\log_2 M \leq nC_s$$  \hfill (78)

$$M = 2^{n(C_s-\delta)}$$  \hfill (79)

$$0 \leq \delta < C_s$$  \hfill (80)

$\delta$ can be arbitrarily small, and in this case $R \to C$.

Assume now that the coded vectors of length $n$ bits are in an $n$-dimensional vector space. If the vector components are taken from the binary field, the coordinates of this vector representation adopt one of the two possible values, one or zero. In this case the distance between any two vectors can be evaluated using the number of different components they have. Thus, if $c$ is a given vector of the code, or codeword, and $c'$ is a vector that differs in $l$ positions with respect to $c$, the distance between $c$ and $c'$ is $l$, a random variable with values between 0 and $n$. If the value of $n$ is very large, vector $c'$ is always within a sphere of radius $d < n$. The decoder will decide that $c$ has been the transmitted vector when receiving $c'$ if this received vector is inside the sphere of radius $d$ and none of the remaining $M - 1$ codewords are inside that.
sphere. Incorrect decoding happens when the number of errors produced during transmission is such that the received vector is outside the sphere of radius $d$ and lies in the sphere of another codeword different from $c$. Incorrect decoding also occurs even if the received vector lies in the sphere of radius $d$, but another codeword is also inside that sphere. Then, the total error probability is equal to

$$P_e = P_{le} + P_{ce}$$  (81)

where $P_{le}$ is the probability of the fact that the received vector is outside the sphere of radius $d$ and $P_{ce}$ is the probability that two or more codewords are inside the same sphere. It is noted that this event is possible because of the random encoding process, and so two or more codewords can be within the same sphere of radius $d$.

$P_{le}$ is the probability of the error event $l \geq d$. Transmission errors are statistically independent and happen with a probability $p < 1/2$, and the number of errors $l$ is a random variable governed by the binomial distribution

$$I = np, \quad \sigma^2 = n(1 - p)p$$  (82)

If the sphere radius is adopted as

$$d = n\beta, \quad p < \beta < 1/2$$  (83)

it is taken as slightly larger than the number of expected errors per word. The error probability $P_{le}$ is then equal to

$$P_{le} = P(l \geq d) \leq \left(\frac{\sigma}{d - l}\right)^2 = \frac{p(1 - p)}{n(\beta - p)^2}$$  (84)

and if the conditions expressed in (83) are true, then $P_{le} \to 0$ as $n \to \infty$.

On the other hand, in order to estimate the error probability, a number $m$ is defined to describe the number of words or vectors contained within the sphere of radius $d$ surrounding a particular one of the $M$ codewords. As before, Shannon assumed a random encoding technique for solving this problem. From this point of view, the remaining $M - 1$ codewords are inside the $n$-dimensional vector space, and the probability that a randomly encoded vector or codeword is inside the sphere containing $m$ vectors is

$$m \choose 2^n$$  (85)

Apart from the particular codeword selected, there exist $M - 1$ other code vectors, so that using equation (79)

$$P_{ce} = (M - 1)m 2^{-n} < Mm 2^{-n} < m 2^{-n} 2^{n(C,\beta)}$$

$$= m 2^{-n} 2^{n[1-Omega(p)\beta]} = m 2^{-n[Omega(p)+\delta]}$$  (86)

All the $m$ vectors that are inside the sphere of radius $d$, defined around the codeword $c$, have $d$ different positions with respect to the codeword, or less. The number of possible codewords with $d$ different positions with respect to the codeword is equal to $\binom{n}{d}$. In general, the number
m of codewords inside the sphere of radius \( d \) is

\[
m = \sum_{i=0}^{d} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}, \quad d = n\beta
\]  

(87)

Among all the terms in the above expression, the term \( \binom{n}{d} \) is the largest, and it can be considered as a bound on the sum of the \( d + 1 \) other terms as follows:

\[
m \leq (d + 1) \binom{n}{d} = \frac{n!}{(n-d)!d!}(d+1)
\]  

(88)

Since \( d = n\beta \), and by using the Stirling approximation for the factorial number \( n! \approx n^n e^{-n} \sqrt{2\pi n} \) if \( n \gg 1 \),

\[
m \leq (d + 1) \binom{n}{d} = \frac{n^\beta + 1}{\sqrt{2\pi n\beta(1-\beta)}}
\]

and by combining it with expression (86),

\[
P_{ce} \leq \frac{n^\beta + 1}{\sqrt{2\pi n\beta(1-\beta)}} 2^{-n[\delta + \Omega(\beta) - \Omega(p)]} = \frac{n^\beta + 1}{\sqrt{2\pi n\beta(1-\beta)}} 2^{-n[\delta + \Omega(p)]} 2^{-n[\delta - \Omega(\beta) - \Omega(p)]}
\]  

(89)

The above expression says that the random coding error probability \( P_{ce} \) tends to zero if \( \delta \geq \Omega(\beta) - \Omega(p) \), which is a function of the parameter \( p \), the error probability of the BSC, and if \( n \to \infty \). Once again the error probability tends to zero if the length of the codeword tends to infinity. The value of the parameter \( \delta \) is the degree of sacrifice of the channel capacity, and it should fit the condition \( 0 \leq \delta < C_s \).

Finally, replacing the two terms of the error probability corresponding, respectively, to the effect of the noise and to the random coding [5], we obtain

\[
P_e = P_e + P_{ce} \leq \frac{p(1-p)}{n(\beta - p)^2} + \frac{n^\beta + 1}{\sqrt{2\pi n\beta(1-\beta)}} 2^{-n[\delta + \Omega(\beta) - \Omega(p)]} \]

(90)

For a given value of \( p \), if \( \beta \) is taken according to expression (83), and fitting also the condition \( \delta \geq \Omega(\beta) - \Omega(p) \), then \( \delta' = \delta - [\Omega(\beta) - \Omega(p)] > 0 \) and the error probability is

\[
P_e = P_e + P_{ce} \leq \frac{K_1}{n} + \sqrt{n} K_2 2^{-a K_3} + \frac{K_4}{\sqrt{n}} 2^{-a K_3}
\]  

(91)

where \( K_1, K_2, K_3 \) and \( K_4 \) are positive constants. The first and third terms clearly tend to zero as \( n \to \infty \), and the same happens with the term \( \sqrt{n} K_2 2^{-a K_3} = \frac{\sqrt{n} K_2}{2^a K_3} \) if it is analysed using the L'Hopital rule. Hence, \( P_e \to 0 \) as long as \( n \to \infty \), and so error-free transmission is possible when \( R < C \).

1.12 Limits to Communication and their Consequences

In a communication system operating over the additive white Gaussian noise (AWGN) channel for which there exists a restriction on the bandwidth, the Nyquist and Shannon theorems are enough to provide a design framework for such a system [5, 7].
An ideal communication system characterized by a given signal-to-noise ratio $P_s/N_0 = S/N$ and a given bandwidth $B$ is able to perform error-free transmission at a rate $R = B \log_2(1 + S/N)$. The ideal system as defined by Shannon is one as seen in Figure 1.18 [5].

The source information is provided in blocks of duration $T$ and encoded as one of the $M$ possible signals such that $R = \log_2 M/T$. There is a set of $M = 2^{RT}$ possible signals. The signal $y(t) = x(t) + n(t)$ is the noisy version of the transmitted signal $x(t)$, which is obtained after passing through the band-limited AWGN channel. The Shannon theorem states that

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

The transmission rate of the communication system tends to the channel capacity, $R \to C$, if the coding block length, and hence the decoding delay, tends to infinity, $T \to \infty$. Then, from this point of view, this is a non-practical system.

An inspection of the expression $C = B \log_2(1 + S/N)$ leads to the conclusion that both the bandwidth and the signal-to-noise ratio contribute to the performance of the system, as their increase provides a higher capacity, and their product is constant for a given capacity, and so they can be interchanged to improve the system performance. This expression is depicted in Figure 1.19.

For a band-limited communication system of bandwidth $B$ and in the presence of white noise, the noise power is equal to $N = N_0B$, where $N_0$ is the power spectral density of the noise in that channel. Then

$$C/B = \log_2 \left(1 + \frac{S}{N_0B}\right)$$

There is an equivalent expression for the signal-to-noise ratio described in terms of the average bit energy $E_b$ and the transmission rate $R$.

If $R = C$ then

$$\frac{E_b}{N_0} = \frac{S}{N_0R} = \frac{S}{N_0C}$$

$$\frac{C}{B} = \log_2 \left(1 + \frac{E_bC}{N_0B}\right), \quad 2^{C/B} = 1 + \frac{E_b}{N_0} \left(\frac{C}{B}\right)$$

$$\frac{E_b}{N_0} = \frac{B}{C} \left(2^{C/B} - 1\right)$$
The Shannon limit can be now analysed from equation

\[
\frac{C}{B} = \log_2 \left( 1 + \frac{E_b C}{N_0 B} \right)
\]  
(96)

making use of the expression

\[
\lim_{x \to 0} (1 + x)^{1/x} = e
\]

where \(x = \frac{E_b C}{N_0 B} \).

Since \(\log_2 (1 + x) = \frac{1}{x} \log_2 (1 + x) = x \log_2 [(1 + x)^{1/x}] \) [7],

\[
\frac{C}{B} = \frac{C}{B} \frac{E_b}{N_0} \log_2 \left( 1 + \frac{E_b C}{N_0 B} \right)^{N_0 B / C E_b}
\]

\[
\Rightarrow 1 = \frac{E_b}{N_0} \log_2 \left( 1 + \frac{E_b C}{N_0 B} \right)^{N_0 B / C E_b}
\]  
(97)

If \(\frac{C}{B} \to 0\), obtained by letting \(B \to \infty\),

\[
\frac{E_b}{N_0} = \frac{1}{\log_2 (e)} = 0.693
\]
or

\[
\left( \frac{E_b}{N_0} \right)_{\text{dB}} = -1.59 \text{ dB}
\]

(98)

This value is usually called the Shannon limit. This is a performance bound on the value of the ratio \( E_b/N_0 \), using a rather sophisticated coding technique, and for which the channel bandwidth and the code length \( n \) are very large. This means that if the ratio \( E_b/N_0 \) is kept slightly higher than this value, it is possible to have error-free transmission by means of the use of such a sophisticated coding technique.

From the equation

\[
2^{C/B} = 1 + \frac{E_b}{N_0} \left( \frac{C}{B} \right)
\]

(99)

a curve can be obtained relating the normalized bandwidth \( B/C \) (Hz/bps) and the ratio \( E_b/N_0 \).

For a particular transmission rate \( R \),

\[
2^{R/B} \leq 1 + \frac{E_b}{N_0} \left( \frac{R}{B} \right)
\]

(100)

\[
\frac{E_b}{N_0} \geq \left( \frac{B}{R} \right) \left( 2^{R/B} - 1 \right)
\]

(101)

Expression (100) can be also depicted, and it defines two operating regions, one of practical use and another one of impractical use [1, 2, 5, 7]. This curve is seen in Figure 1.20, which represents the quotient \( R/B \) as a function of the ratio \( E_b/N_0 \). The two regions are separated by the curve that corresponds to the case \( R = C \) [equation (99)]. This curve shows the Shannon limit when \( R/B \to 0 \). However, for each value of \( R/B \), there exists a different bound, which can be obtained by using this curve.

![Figure 1.20 Practical and non-practical operation regions. The Shannon limit](image-url)
Bibliography and References


Problems

1.1 A DMS produces symbols with the probabilities as given in Table P.1.1.

<table>
<thead>
<tr>
<th>A</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
</tr>
<tr>
<td>D</td>
<td>0.1</td>
</tr>
<tr>
<td>E</td>
<td>0.05</td>
</tr>
<tr>
<td>F</td>
<td>0.05</td>
</tr>
</tbody>
</table>

(a) Find the self-information associated with each symbol, and the entropy of the source.
(b) Calculate the maximum possible source entropy, and hence determine the source efficiency.

1.2 (a) Calculate the entropy of a DMS that generates five symbols \{A, B, C, D, E\} with probabilities \(P_A = 1/2, P_B = 1/4, P_C = 1/8, P_D = 1/16\) and \(P_E = 1/16\).
(b) Determine the information contained in the emitted sequence DADED.
1.3 Calculate the source entropy, the transinformation $I(X, Y)$ and the capacity of the BSC defined in Figure P.1.1.

$$P(0) = \alpha = 0.2$$

$\rho = 0.25$

$1 - \rho = 0.75$

$P(1) = 1 - \alpha = 0.8$

$1 - \rho = 0.75$

Figure P.1.1  A binary symmetric channel

1.4 Show that for the BSC, the entropy is maximum when all the symbols of the discrete source are equally likely.

1.5 An independent-symbol binary source with probabilities 0.25 and 0.75 is transmitted over a BSC with transition (error) probability $p = 0.01$. Calculate the equivocation $H(X/Y)$ and the transinformation $I(X, Y)$.

1.6 What is the capacity of the cascade of BSCs as given in Figure P.1.2?

Figure P.1.2  A cascade of BSCs

1.7 Consider a binary channel with input and output alphabets $\{0, 1\}$ and the transition probability matrix:

$$P_{ch} = \begin{bmatrix} 3/5 & 2/5 \\ 1/5 & 4/5 \end{bmatrix}$$

Determine the a priori and the two a posteriori entropies of this channel.

1.8 Find the conditional probabilities $P(x_i/y_j)$ of the BEC with an erasure probability of 0.469, when the source probabilities are 0.25 and 0.75. Hence find the equivocation, transinformation and capacity of the channel.

1.9 Calculate the transinformation and estimate the capacity of the non-symmetric erasure channel given in Figure P.1.3.
1.10 Figure P.1.4 shows a non-symmetric binary channel. Show that in this case
\[ I(X, Y) = \Omega [q + (1 - p - q)\alpha] - \alpha \Omega (\rho) - (1 - \alpha) \Omega (q). \]

\[ P(0) = \alpha \quad 0 \quad 1 \quad 0 \]
\[ P(1) = 1 - \alpha \quad 0 \quad 1 \quad 0 \]

**Figure P.1.4** A non-symmetric binary channel

1.11 Find the transinformation, the capacity and the channel efficiency of the symmetric erasure and error channel given in Figure P.1.5.

\[ \alpha = 0.25 \quad 0 \quad 0.9 \quad 0 \]
\[ 1 \quad 0.08 \quad 0.02 \quad 0.02 \]
\[ 0 \quad 0.08 \quad 0.02 \]
\[ 0 \quad 0 \]

**Figure P.1.5** A symmetric erasure and error channel

1.12 Consider transmission over a telephone line with a bandwidth \( B = 3 \) kHz. This is an analogue channel which can be considered as perturbed by AWGN, and for which the power signal-to-noise ratio is at least 30 dB.

(a) What is the capacity of this channel, in the above conditions?

(b) What is the required signal-to-noise ratio to transmit an \( M \)-ary signal able to carry 19,200 bps?

1.13 An analogue channel perturbed by AWGN has a bandwidth \( B = 25 \) kHz and a power signal-to-noise ratio SNR of 18 dB. What is the capacity of this channel in bits per second?