1

Introduction

"Information is the resolution of uncertainty."

— Claude Shannon

Reliable communication over noisy channels is a cornerstone of modern information systems, and channel coding serves as a critical mechanism to achieve this reliability. This chapter provides the foundational knowledge necessary to understand the role of channel coding in communication systems, setting the stage for the rest of the book.

We begin by exploring the need for channel coding in ensuring reliable transmission over noisy channels and by reviewing the structure of a generic communication system. This is followed by a discussion of various channel models, including the binary symmetric channel (BSC), binary erasure channel (BEC), additive white Gaussian noise (AWGN) channel, and fading channel. These models illustrate the diverse challenges posed by different communication environments.

The fundamentals of information theory are introduced next, focusing on key concepts such as the meaning of information, discrete entropy, its properties, and mutual information. The Bhattacharyya parameter, a measure of channel reliability, is also explained to provide deeper insights into channel characteristics.

The chapter culminates with an in-depth discussion of channel capacity and the celebrated Shannon Channel Coding Theorem, which defines the theoretical limits of reliable communication. Additional topics include the block error rate for finite-length codes, the Shannon limit on power efficiency, and the computational complexity associated with coding and decoding processes.

By the end of this chapter, readers will have a comprehensive understanding of the role of channel coding in communication systems, the differences between channel models, the fundamentals of information theory, and the principles underpinning the channel coding theorem. This foundational knowledge will

Polar Codes: From Theory to Practice, First Edition. Mohammad Rowshan and Emanuele Viterbo.

enable readers to appreciate the critical role of coding theory in modern communication systems and serve as a basis for exploring advanced topics in the chapters that follow.

1.1 Reliable Transmission over Noisy Channels

Digital communications and storage are essential parts of our modern lives. We rely on them for everything from sending emails to streaming videos. However, these systems are susceptible to errors that can corrupt the data and make them unusable.

Errors in data transmission and storage systems can be caused by a variety of factors, including:

- Random noise: This is caused by the inherent randomness of the physical world, such as the thermal noise of electronic devices.
- Interference: This is caused by other signals that are present in the same transmission medium, such as radio waves or electromagnetic radiation.
- Channel fading: This is caused by changes in the properties of the transmission medium, such as the attenuation of radio waves due to obstacles or the variation of the refractive index of the atmosphere.
- Physical defects: These are caused by imperfections in the physical components of the data transmission or storage system, such as scratches on a disc or defects in a memory chip.

There are two main strategies for combating errors in digital communications and storage: *automatic repeat request* (ARQ) and *forward error correction* (FEC). ARQ systems detect errors in the received data and request that the data be resent. This is a simple and effective way to ensure reliable data transmission, but it can also lead to increased latency and bandwidth usage. FEC systems, on the other hand, not only detect but also correct errors in the received data. FEC systems can achieve higher reliability than ARQ systems with lower latency, but they are also more computationally expensive. The choice of which error control strategy to use depends on the specific application. For example, FEC is often used in real-time applications where retransmission is not possible, such as live streaming. ARQ is often used in applications where latency is less important than channel bandwidth, such as file transfer over a network. In this book, our focus is on FEC, which is the subject of coding theory.

Coding theory originated in the late 1940s with the work of Claude Shannon, Marcel Golay, and Richard Hamming. Shannon's landmark paper *A Mathematical Theory of Communication* in 1948 laid the foundations for both information theory and coding theory. He introduced the concept of channel capacity, which is the maximum rate at which information can be transmitted over a noisy channel

without errors. He also showed that arbitrarily reliable communication is possible at any rate below the channel capacity.

One way to achieve reliable communication is to use error-correcting codes. These codes add redundant bits to the data being transmitted, which can be used to detect and correct errors that occur during transmission. Hamming developed a simple but effective error-correcting code in 1947. His code is now known as the Hamming code. Other examples of communication channels include wireless communication devices and storage systems such as digital video disc (DVD)s and Blu-ray discs. Coding theory is essential for ensuring reliable communication over these channels.

Figure 1.1 illustrates a simple communication channel, ignoring the modulation and demodulation stage for the sake of simplicity. At the source of the communication (transmitter), a message represented as $\mathbf{u} = [u_1 \cdots u_k]$ is intended to be transmitted. However, if this message u is sent directly through the channel without any modifications, the presence of noise could distort u to the extent that its accurate recovery becomes unfeasible. The fundamental concept of coding theory revolves around enhancing the message's resilience by introducing redundancy. This additional information aims to enable the recovery of the original message even in the face of noise-induced corruption during transmission. The process of adding redundancy occurs at the encoder stage, resulting in an augmented message known as a codeword, illustrated as $\mathbf{x} = [x_1 \cdots x_n]$ in the figure. This codeword is then transmitted through the channel, where noise, represented as an error vector $\mathbf{e} = [e_1 \cdots e_n]$, distorts the codeword, leading to the formation of a received vector $\mathbf{y}$.

Subsequently, the received vector undergoes decoding, where the errors are rectified. The surplus redundancy is removed, yielding an estimation $\tilde{\mathbf{u}}$ of the original message. Ideally, $\tilde{\mathbf{u}}$ matches $\mathbf{u}$, ensuring successful recovery. Notably, there exists a direct correspondence between codewords and messages. In many instances, the primary focus shifts from the message $\mathbf{u}$ to the codeword $\mathbf{x}$. From this perspective, the decoder's task involves deriving an estimate $\tilde{\mathbf{y}}$ from $\mathbf{y}$ with the hope that $\tilde{\mathbf{y}}$ aligns with $\mathbf{x}$.

In the case of a DVD or Blu-ray disc, the message source encompasses audio, music, video, or data intended for storage on the disc. The disc itself serves as the communication channel, the DVD or Blu-ray player operates as the decoder, and the recipients consist of listeners or viewers. By refining the communication process through the introduction of redundancy and error correction, coding

Figure 1.1 Error correction coding over a noisy channel.

theory plays a vital role in maintaining the integrity of transmitted information across various contexts and communication channels.

The implementation of error correction comes at a cost. The introduced redundancy functions as an additional layer that consumes transmission resources, such as channel bandwidth or transmission power. Consequently, our objective is to minimize this overhead by minimizing the required redundancy.

To quantitatively assess the amount of redundancy, we introduce the concept of coding rate, denoted as R. This rate is defined as the ratio between the length of the original message and the length of the resulting codeword. For example, if a specific coding scheme produces a codeword of length n from a message of length k, the coding rate can be expressed as:

$$R = k/n$$

In this context, the coding rate reaches its highest value of 1 when no redundancy is incorporated, which basically leaves the message uncoded. The trade-off between coding performance and coding rate is pivotal. With the inclusion of more redundancy, the capability to correct errors becomes more robust, yet this comes at the expense of a reduced coding rate.

An optimal code strives to strike a balance between these factors. It aims to maximize the error correction capability while maintaining the coding rate as close to 1 as possible. This equilibrium ensures that transmission remains efficient in terms of resource utilization while effectively combating the detrimental effects of noise and errors during communication.

So far, we have established some understanding of the advantages offered by channel coding. Next, we discuss some common models for the channels, and in Section 1.4, we delve into the theoretical limit for code rate to have reliable communication.

1.2 Channel Models

There exits various channel models that are used for different applications depending on the specific characteristics of the communication channel and the requirements of the application. In the following, we review the major channel models.

1.2.1 Binary Symmetric Channel (BSC)

The *BSC* is a discrete memoryless channel (DMC) that has two binary symbols as inputs and outputs: 0 and 1. It is symmetric because the probability of receiving a 0 when a 1 is transmitted is the same as the probability of receiving a 1 when a 0 is transmitted. This probability is called the *crossover probability* and is denoted by p.

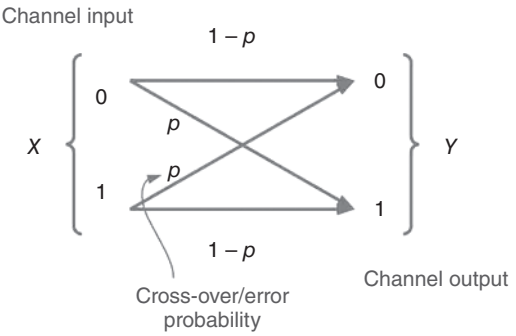

Figure 1.2 Binary symmetric channel.

The probability of no error is $1-p$. Hence, the channel transition probabilities for the BSC are:

$$\begin{cases} P(Y=0|X=0)=1-p, \\ P(Y=0|X=1)=p, \\ P(Y=1|X=0)=p, \\ P(Y=1|X=1)=1-p, \end{cases} \tag{1.1}$$

Figure 1.2 illustrates the BSC model. The *error probability* of a BSC can be computed as

$$\begin{aligned} P_e &= P(X=0,Y=1)+P(X=1,Y=0) \\ &= P(Y=1|X=0)P(X=0)+P(Y=0|X=1)P(X=1)=p \end{aligned}$$

since $P(X=0)+P(X=1)=1$.

1.2.2 Binary Erasure Channel (BEC)

Erasure is a special type of error with known location. The BEC is a communication channel that can either transmit a bit correctly or erase it. If the bit is erased, the receiver will not know what the bit was. The BEC output consists of three symbols: 0, 1, and e, where e represents an erased bit. Figure 1.3 illustrates the BEC model.

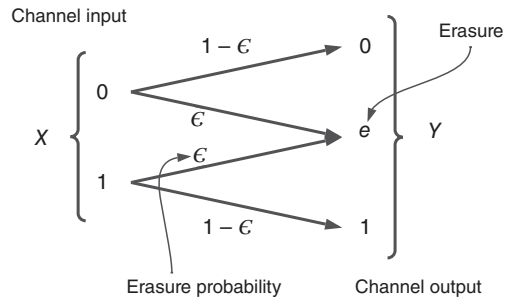

Figure 1.3 Binary erasure channel.

The probability that a bit is erased by the BEC is called the erasure probability and is denoted by ϵ. The channel transition probabilities for the BEC are:

$$\begin{cases} P(Y=0|X=0) = 1-\epsilon, \\ P(Y=e|X=0) = \epsilon, \\ P(Y=1|X=0) = 0 \\ P(Y=0|X=1) = 0 \\ P(Y=e|X=1) = \epsilon \\ P(Y=1|X=1) = 1-\epsilon. \end{cases} \tag{1.2}$$

1.2.3 Additive White Gaussian Noise Channel (AWGN)

The AWGN channel model (as shown in Figure 1.4) is used when the inputs and outputs are real- or complex-valued. In an AWGN channel, the signal is corrupted by white noise z that is characterized by a constant spectral density and a Gaussian distribution. The Gaussian distribution is a probability distribution that is often used to model the thermal noise (also known as Johnson–Nyquist noise) introduced by electronic components of the receiver. The probability density function (PDF) of the Gaussian distribution is given by:

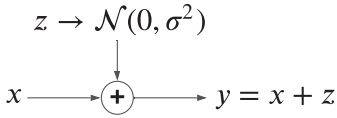

Figure 1.4 Additive white Gaussian noise channel.

$$P(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{z^2}{2\sigma^2}\right), \tag{1.3}$$

where σ^2 is the variance of a Gaussian random process. As a result, the observed signal at the receiver is given by

$$y = x + z, \tag{1.4}$$

where $z \sim \mathcal{N}(0, \sigma^2)$. Given the normalized Gaussian random variable $n \sim \mathcal{N}(0, 1)$, we can alternatively obtain $z = \sigma \cdot n$ and hence y as

$$y = x + \sigma \cdot n. \tag{1.5}$$

Furthermore, knowing $\sigma^2 = N_0/2$ for the AWGN channel, if we have the signal-to-noise ratio (SNR) $= E_s/N_0$ of the channel, assuming E_s is normalized to $E_s = 1$, we have

$$y = x + \sqrt{\frac{1}{2\frac{E_s}{N_0}}} \cdot n, \tag{1.6}$$

where $\frac{E_s}{N_0}$ is in linear form, not in dB. Since SNR [dB] $= 10\log(\text{SNR})$, to convert the SNR from dB to linear, we can use $\text{SNR} = 10^{\frac{\text{SNR [dB]}}{10}}$. If the given SNR is per

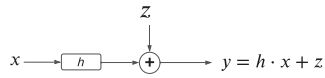

Figure 1.5 Fading channel along with additive Gaussian noise.

information bit that is denoted by E_b/N_0, then we can use the relation $\frac{E_s}{N_0} = R \cdot \frac{E_b}{N_0}$. In the case of higher-order modulation with constellation size $M > 2$, we need to also add the number of bits per symbol, i.e. $m = \log_2 M$. Hence, given SNR per information bit E_b/N_0 and M, then we will have

$$y = x + \sqrt{\frac{1}{2mR\frac{E_b}{N_0}}} \cdot n. \tag{1.7}$$

1.2.4 Fading Channel

The physical wireless channel introduces scaling due to power decay with distance and disturbances via the channel gain coefficient h, as illustrated in Figure 1.5.

As a result, the observed signal at the receiver is given by

$$y = h \cdot x + z. \tag{1.8}$$

We assume that the scaling factor h varies slowly compared to the duration of the waveform. The channel gain coefficient h is often modeled using a Rayleigh or Rician PDF. The most common fading channel models are

- Flat independent fading channel,
- Block fading channel.

In a flat independent fading channel, the attenuation of the signal remains constant over a single symbol duration but changes from symbol to symbol. This means that the channel is memoryless and that the attenuation of any symbol is independent of the attenuation of any other symbol. In a block fading channel, the attenuation of the signal remains constant over a block of symbols, but changes from block to block. This means that the channel has memory and that the attenuation of any symbol is correlated to the attenuation of the previous symbols in the block. It is important to note that fading channels can be further classified based on the characteristics of the fading, such as slow fading (e.g. Rayleigh fading) and fast fading (e.g. frequency-selective fading).

1.3 Fundamentals of Information Theory

This section reviews the basic concepts of information, discrete entropy, mutual information, their properties, and chain rules. We will use these concepts in Chapter 4. This requires knowledge of probability and random variables.

Continuous entropy and continuous mutual information are very closely related to discrete entropy and discrete mutual information, but they are not in the scope of this section. For the sake of brevity, we skip the proofs; however, interested readers can find some references in the bibliographical notes of this chapter.

1.3.1 The Meaning of Information

Information is a phenomenon that has meaning for living creatures. The human brain has the ability to interpret events, concepts, and objects and to evaluate them based on their information content. This means that information is not just a physical phenomenon but also a mental one. It is something that we perceive and understand and that has meaning for us.

The information content of an event is related to how surprised we are when it occurs. We are more surprised when something unexpected happens, or when something with a low probability of occurrence takes place. This suggests that the information content of an event is inversely related to its probability of occurrence. In other words, the more probable an event is, the less information it conveys. Conversely, the less probable an event is, the more information it conveys. This is because a probable event is something we already expect, so it does not tell us anything new. However, an improbable event is something we did not expect, so it tells us a lot about the world.

In human communication, information is conveyed through speech, writing, and pictures. Speech is a sequence of letters, words, and sentences. Some letters and words are more common than others, so they convey less information. For example, the letter "e" is the most common letter in the English language, so it conveys the least information. Conversely, the letter "q" is the least common letter in the English language, so it conveys the most information. The same principle applies to words and sentences.

We can model an information source as a random variable (RV). Random variables fall into two primary categories: discrete and continuous depending on their range set. Discrete random variables exhibit a countable number of values within their range sets. Conversely, continuous random variables present an uncountable number of values within their range sets.

Let X represent a discrete random variable. Its range set is designated as $\mathcal{X} = \{x_i\}$, with individual symbols x_i. For this discrete RV X, the probability mass function $p(x)$ is introduced, defined as

$$p(x) = \text{Prob}(X = x), \tag{1.9}$$

which quantifies the likelihood of the discrete RV producing the symbol x. The probability associated with the symbol x_i is denoted as $p(x_i)$. Considering that

information and probability are inversely related, the informational significance of symbol x_i can be quantified through a unitless quantity as

$$I\left(x_i\right) \propto \frac{1}{p\left(x_i\right)}. \tag{1.10}$$

Observe that in (1.10) when $p\left(x_i\right)$ becomes very small, $I\left(x_i\right)$ becomes very large, which suggests the use of a logarithmic scale. We use base 2 since the information content of an event is transmitted using bit sequences. This allows us to assign "bits" as a unit of measure as follows:

$$I\left(x_i\right) = \log_2\left(\frac{1}{p\left(x_i\right)}\right) \text{ bits}. \tag{1.11}$$

The logarithmic scale also gives the natural property that the information of a pair of independent symbols is the sum of the information of each symbol. This can be easily verified observing that $p(x_i, x_j) = p(x_i) \cdot p(x_j)$.

1.3.2 Discrete Entropy

Entropy is defined as the average information content of the discrete random variable X, i.e.

$$H(X) = E[I(X)] = \sum_{x_i \in \mathcal{X}} p\left(x_i\right) I\left(x_i\right) = -\sum_{x_i \in \mathcal{X}} p\left(x_i\right) \log\left(p\left(x_i\right)\right),$$

recalling that the expected value of a function of a random variable is $E[g(X)] = \sum_{x_i} p\left(x_i\right) g\left(x_i\right)$.

In the following, we will omit the range set in the sums to simplify notation.

Entropy represents the mean information contained within an information source, specifically the average information content of a random variable representing the source. Imagine receiving symbols from this information source; a higher entropy indicates a substantial inflow of information from the source. On the contrary, a lower entropy suggests a lesser amount of information being received from the same source. For example, if we have a fair coin, the entropy of the coin toss is 1 bit. This is because there are two possible outcomes (heads or tails), and each outcome is equally likely. We have the maximum amount of uncertainty, or least amount of information, about the outcome of the coin toss. If we have a biased coin, where heads are more likely than tails, the entropy of the coin toss will be less than 1 bit. This is because we have less uncertainty, or more information, about the outcome of the coin toss. If we know that the coin is biased toward heads, we can be more confident that the outcome will be heads.

Let X and Y be two discrete random variables with marginal and joint probability mass functions $p_X(x)$, $p_Y(y)$, and $p_{X,Y}(x, y)$, respectively. The *joint entropy* for

the discrete random variables X and Y is defined as

$$H(X,Y) = \sum_{x_i,y_j} p(x_i,y_j) \log\left(\frac{1}{p\left(x_i,y_j\right)}\right) = -\sum_{x,y} p(x,y)\log(p(x,y)).$$

The conditional entropy of the discrete random variable X for a given value y_j of another discrete random variable Y is defined as

$$H\left(X|y_j\right) = -\sum_{x_i} p(x_i|y_j)\log\left(p\left(x_i|y_j\right)\right) = -\sum_{x} p(x|y)\log(p(x|y)),$$

which is interpreted as the amount of average information provided by a single symbol of random variable X if a single symbol of random variable Y is known. If X and Y are not independent random variables, it is obvious that $H(X|y) < H(X)$.

The conditional entropy of the discrete random variable X given another discrete random variable Y, i.e. all the values (symbols) of Y are known, is defined as

$$H(X|Y) = \sum_{y_j} p(y_j) H\left(X|y_j\right) = \sum_{y} p(y)H(X|y),$$

which can be considered as the total amount of average information per symbol required to know the random variable X assuming that the random variable Y is known.

1.3.3 Properties of the Discrete Entropy

If X and Y are two discrete random variables, and $H(X)$ and $H(Y)$ are the corresponding entropies, then we have the following properties:

1) Discrete entropy is a non-negative quantity, i.e. $H(X) \geq 0$.
2) If the number of elements in the range set of X equals to $|\mathcal{X}|$, the maximum entropy of X equals to

$$H_{\max}(X) = \log\left(|\mathcal{X}|\right).$$

3) For the joint entropy $H(X,Y)$, we have the property

$$H(X,Y) \leq H(X) + H(Y),$$

where equality occurs if X and Y are independent random variables, i.e. $p(x,y) = p(x)p(y)$.

4) For N discrete random variables, property 3 can be generalized as

$$H\left(X_1, X_2, \ldots, X_N\right) \leq H\left(X_1\right) + H\left(X_2\right) + \cdots + H\left(X_N\right),$$

where equality occurs if $X_1, X_2, \ldots, X_N$ are independent random variables.

5) Conditional entropy satisfies

$$H(X|Y) \leq H(X).$$

Chain rule for entropy. Let $X_1, X_2, \ldots, X_n$ be drawn according to a joint pdf $P_{X^n}(x^n) = P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$, where $X^n = (X_1, \ldots, X_n)$ and $x^n = (x_1, \ldots, x_n)$.

Then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1),$$

where $H(X_i | X_{i-1}, \ldots, X_1) := H(X_1)$ for $i = 1$. The above chain rule can also be written as:

$$H(X^n) = \sum_{i=1}^{n} H(X_i | X^{i-1}),$$

where $X^i = (X_1, \ldots, X_i)$.

Chain rule for conditional entropy.

$$H(X_1, X_2, \ldots, X_n | Y) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1, Y)$$

1.3.4 Mutual Information

Mutual information can be used to quantify the relationship between two random variables. It is a measure of the reduction in uncertainty about one random variable that we gain by knowing the value of the other random variable. Mutual information can be used to measure the relationship between any two random variables, regardless of their type or distribution. Let us formally define it.

The mutual information between two random variables X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y) = -\sum_{x} p(x) \log(p(x)) + \sum_{x,y} p(x,y) \log(p(x|y)) \text{bits/symbol.}$$

Recall that $\sum_{x,y} p(x,y) \log(p(x)) = \sum_{x} p(x) \log(p(x))$ and $p(x|y) = \frac{p(x,y)}{p(y)}$. Hence, by substitution and rearrangement, we obtain

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \text{ bits/symbol.}$$

The mutual information can equivalently be written as

$$I(X;Y) = E\left[\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right] \text{ bits/symbol.}$$

The *conditional mutual information* between the discrete random variables X and Y given Z is calculated as:

$$I(X;Y|Z) = E\left[\log \frac{p(X,Y|Z)}{(p(X|Z)p(Y|Z)}\right] = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

Now, let us review the properties of mutual information as follows:

1) Symmetry property states that

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

2) Mutual information is non-negative, i.e.

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

Alternative expressions for mutual information in terms of entropy and conditional entropy are:

$$I(X;Y) = H(Y) - H(Y|X), \tag{1.12}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{1.13}$$

Chain rule for mutual information.

$$I\left(X_1, X_2, \ldots, X_n; Y\right) = \sum_{i=1}^{n} I\left(X_i; Y|X_{i-1}, \ldots, X_1\right),$$

where $I\left(X_i; Y|X_{i-1}, \ldots, X_1\right) := I\left(X_1; Y\right)$ for $i = 1$.

1.3.5 Bhattacharyya Parameter

The Bhattacharyya parameter, named after Anil Kumar Bhattacharyya, is a measure of the similarity between two statistical distributions. Specifically, it quantifies the overlap between PDFs. In the context of communication channels, the Bhattacharyya parameter is related to the probability of error. It is commonly used to evaluate the error performance of a communication system in the presence of noise or other impairments.

The Bhattacharyya parameter (denoted by B) is calculated using the following formula:

$$B = \frac{1}{4} \sum_{i=1}^{M} \sqrt{P_1(i) \cdot P_2(i)}. \tag{1.14}$$

Here, $P_1(i)$ and $P_2(i)$ are the probabilities of the occurrence of the ith symbol in the two distributions being compared and M is the total number of symbols in the common alphabet. The factor 1/4 is often dropped. When comparing two distributions representing different signal constellations or modulation schemes, a lower Bhattacharyya parameter indicates better distinguishability and, consequently, better error performance. It is inversely related to the probability of error, with a smaller B corresponding to a lower probability of error. So, in essence, the Bhattacharyya parameter is related to the probability of error, but it serves as a measure of the distinguishability between probability distributions rather than a direct measure of the error probability.

In Chapter 4, we use the Bhattacharyya parameter as a measure of the reliability of a channel $W : \mathcal{X} \to \mathcal{Y}$, a generic binary discrete memoryless channel (B-DMC) with input alphabet $\mathcal{X} = \{0, 1\}$, arbitrary output alphabet $\mathcal{Y}$, and transition probabilities $W(y|x), x \in \mathcal{X}, y \in \mathcal{Y}$. In this context, the Bhattacharyya parameter as a measure of the reliability of channel W is denoted and defined as

$$Z(W) \triangleq \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}. \tag{1.15}$$

Depending on the type of channel W, the Bhattacharyya parameter can be obtained differently. For a BEC, the Bhattacharyya parameter is

$$Z(W) = \varepsilon, \tag{1.16}$$

where ε is the erasure probability. For a BSC, we have

$$Z(W) = 2\sqrt{p(1-p)}. \tag{1.17}$$

Finally, for an AWGN channel, given the variance of the channel noise σ^2 and the signaling that modulates $\{0, 1\}$ to real values $\{1, -1\}$, the Bhattacharyya parameter is computed as

$$\begin{aligned} Z(W) &= \int_{-\infty}^{\infty} \sqrt{p(y|0)p(y|1)}dy \\ &= \int_{-\infty}^{\infty} \sqrt{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-1)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y+1)^2}{2\sigma^2}}} dy \\ &= e^{-\frac{1}{2\sigma^2}}. \end{aligned} \tag{1.18}$$

Similar to Section 1.2.3, knowing $\sigma^2 = N_0/2$ and assuming $E_s = 1$, we can rewrite $Z(W)$ in terms of $\text{SNR} = E_s/N_0$ as $Z(W) = e^{-\frac{1}{2\sigma^2}} = e^{-SNR}$. Note that SNR in this equation is not in dB, but it is in linear form. Since SNR is commonly used in dB, we can use the following equation instead:

$$Z(W) = e^{-\frac{1}{2\sigma^2}} = e^{-\text{SNR}} = \exp\left(-10^{\frac{\text{SNR [dB]}}{10}}\right). \tag{1.19}$$

1.4 Channel Capacity and Channel Coding Theorem

Shannon's noisy channel coding theorem states that it is possible to transmit data over a noisy channel at any desired rate, as long as the coding scheme is designed correctly. The theorem is based on the concept of channel capacity, which is the maximum rate at which data can be transmitted over a noisy channel without errors. The channel capacity is determined by the properties of the channel, such as noise level and bandwidth. Shannon defined the discrete channel capacity as the

maximum of the average mutual information, which is computed over all possible probability distributions associated with the input symbol alphabet as

$$C = \max_{p(X)} I(X;Y) = \max_{p(X)} [H(X) - H(X|Y)]. \tag{1.20}$$

From (1.20), since $I(X;Y) \geq 0$, we can conclude that $C \geq 0$. Furthermore, since $\max H(X) = \log |\mathcal{X}|$, where $\mathcal{X}$ is the channel input alphabet, then $C \leq \log |\mathcal{X}|$ with equality holding when the channel is error-free.

Formally, the *channel coding theorem* states that for any DMC with input $\mathcal{X}$, output $\mathcal{Y}$ and capacity C, there exists a code with a sufficiently long block length n such that the transmission rate R can be made arbitrarily close to the channel capacity C, and the probability of error ϵ in decoding can be made arbitrarily small. In practical terms, this means that by using longer block codes, the rate can approach channel capacity ($R \to C$), and with sufficiently long codes ($n \to \infty$), the probability of error can be made as small as desired ($\epsilon \to 0$). The channel coding theorem provides a theoretical foundation for designing codes that achieve reliable communication in the presence of noise and disturbances. The goal of coding theory is to find codes that meet the conditions of Shannon's theorem. Many different codes have been proposed, and the best code to use depends on the specific characteristics of the channel.

Now, let us review the capacity of the channels discussed in Section 1.2. The capacity of a BSC with crossover probability p is given by

$$C = \max_{p(X)} I(X;Y) = 1 - H(p) = 1 - p\log p + (1-p)\log(1-p) \text{ bits/channel use}, \tag{1.21}$$

where p indicates the probability of error or the probability of transmitting a symbol and receiving another. The capacity of the BEC (in bits per channel use), in which the parameter e represents the erasure and the probability of erasure $0 \leq \epsilon \leq 1$, is as

$$C = \max_{p(X)} I(X;Y) = 1 - \epsilon \text{ bits/channel use}. \tag{1.22}$$

This means that on average, the $1-\epsilon$ fraction of symbols can be transmitted successfully without being erased.

The capacity of AWGN channels (in bits per second per Hertz) is given by

$$C = \log_2\left(1 + \frac{P}{N_0}\right), \tag{1.23}$$

where P is the average transmitted power, and N_0 is the power of the AWGN. The term $\frac{P}{N_0}$ is called the signal-to-noise ratio (SNR) of the channel.

Finally, the capacity of a fading channel is more complex than that of other channels due to the time-varying nature of the channel caused by fading.

The capacity of a fading channel is usually considered in terms of ergodic capacity, which represents the average capacity over different fading realizations. The capacity of a fading channel (in bits per second per Hertz) is given by

$$C = \mathbf{E}_h \left[\log_2 \left(1 + \frac{P|h|^2}{N_0} \right) \right], \tag{1.24}$$

where E_h denotes the expectation over all possible channel fading states, P is the average transmitted power, $|h|^2$ is the squared magnitude of the fading coefficient, and N_0 is the power of the AWGN. The term $\frac{P|h|^2}{N_0}$ represents the instantaneous SNR) in the fading channel. The expectation over the fading states captures the average behavior of the channel.

1.5 Block Error Rate for Finite Length Codes

In this section, we discuss a non-asymptotic approximation for the maximal sustainable rate of an AWGN channel as a function of block length N and error probability ϵ, i.e. $R^*(N, \epsilon)$. The discussion will be limited to introducing an approximate expression rather than the full information-theoretic derivation. The interested readers are referred to the suggested references in the bibliographical notes.

As an asymptotic limit, it is known from classic texts that for a general class of channels, the channel capacity C is the highest rate R at which information can be transmitted reliably, given that the block length N is allowed to grow without bound. That is,

$$\lim_{N\to\infty} R^*(N, \epsilon) = C. \tag{1.25}$$

However, at a fixed finite block length, we have to backoff from capacity and choose a rate smaller than capacity $R < C$ to maintain the desired error probability. Figure 1.6 illustrates the achievable rate for a desired block error rate $R^*(N, \epsilon)$ at different finite block lengths N.

The gap to the channel capacity can be characterized by a coefficient called channel dispersion, which measures the stochastic variability of the channel relative to a deterministic channel with the same capacity. The channel dispersion coefficient is defined as

$$V = \lim_{\epsilon\to 0} \lim_{N\to\infty} \sup \frac{(NC - \log M^*(N, \epsilon))^2}{N} \rho. \tag{1.26}$$

where $\rho = \frac{1}{2\ln\frac{1}{\epsilon}}$ and $M^*(N, \epsilon)$ is the maximum cardinality of a codebook of block length N which can be decoded with block error probability $P_e \leq \epsilon$. As can be seen in (1.26), the channel dispersion is measured in squared information units per

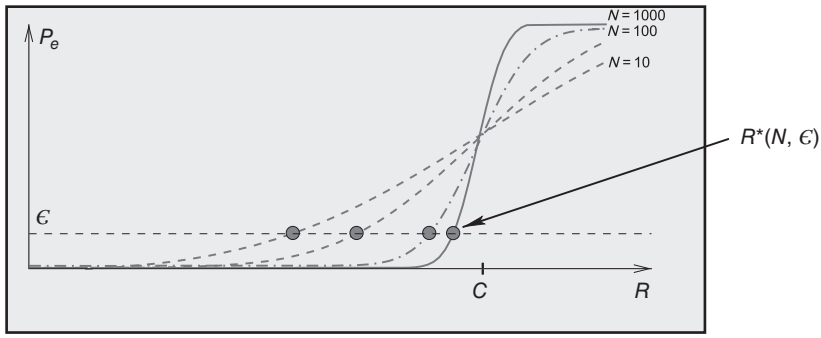

Figure 1.6 Trade-off between the reliability and the code rate: the maximum code rate for a desired ϵ at different blocklength N.

channel use. From the definition of dispersion and its unit, one can picture it as the "variance of the channel capacity." Hence, for finite block-length N, we can consider the maximum code rate to approximately follow a normal distribution of $R \sim \mathcal{N}(C, \frac{V}{N})$ as demonstrated in Figure 1.7 that results in

$$R + \text{constant} \cdot \sqrt{\frac{V}{N}} \lesssim C. \tag{1.27}$$

To be more precise, the maximum code rate via the so-called normal approximation (NA) based on the gap to capacity C is

$$R = C - \sqrt{\frac{V}{N}} \log_2(e)\, Q^{-1}(\epsilon) + \frac{\log_2(N)}{2N}. \tag{1.28}$$

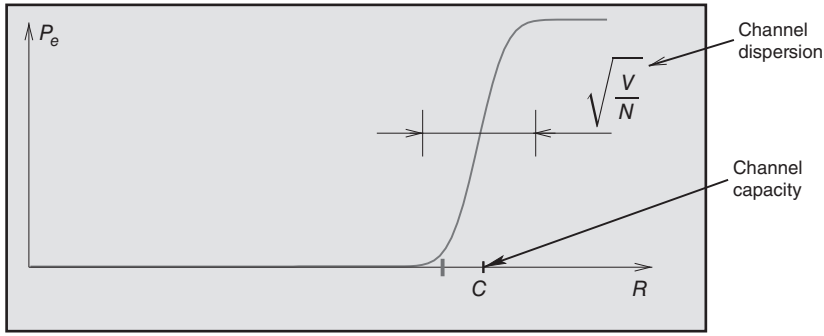

Figure 1.7 A sketch of channel dispersion relative to channel capacity.

This approximation is known as the Polyanskiy–Poor–Verdù (PPV) meta-converse approximation. By rearranging this approximation, we can get a tight approximation for the block error rate ϵ (roughly for block-lengths larger than 100 bits) as

$$\epsilon^*(N,R) = Q\left(\frac{C - R + \frac{\log_2(N)}{2N}}{\sqrt{\frac{V}{N}} \cdot \log_2(e)}\right). \tag{1.29}$$

In the case of the binary-input AWGN channel the capacity and the channel dispersion coefficient can be obtained by using the Laplace transform as

$$C = 1 + \frac{H'(0)}{\ln(2)}, \tag{1.30}$$

and

$$V = H''(0) - (H'(0))^2, \tag{1.31}$$

respectively, and the derivatives of $H(\cdot)$ for every SNR Ω can be numerically evaluated by

$$H^{(\ell)}(s) = \frac{1}{\sqrt{2\pi\Omega}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\Omega}(x-\Omega)^2} e^{-s \cdot h(x)} (-\ln(1 + e^{-2x}))^{\ell} \, dx. \tag{1.32}$$

where ℓ denotes the order of derivative and $h(x) = \ln\left(1 + e^{-2x}\right)$.

The MATLAB script C.1 in Appendix C provides the NA (1.29) for the binary AWGN channel.

1.6 Shannon Limit on Power Efficiency

In this section, we look at the minimum energy per bit required by a transmitter to achieve reliable communication in a band-limited communication system, regardless of modulation or coding scheme. By considering a particular modulation or encoding scheme, the limit is called the constrained Shannon limit.

We are interested in designing a communication system that requires the smallest possible power for reliable transmission of information, that is, with an acceptable bit-error rate (BER). This performance measure is usually illustrated in terms of BER versus E_b/N_0, where E_b/N_0 is the ratio of signal energy per information bit E_b to noise power spectral density per bit N_0. The *power efficiency* η_P refers to the E_b/N_0 that is required at the receiver input to achieve a certain BER.

We learned earlier that the Shannon channel capacity of an AWGN channel with a bandwidth B [Hz], received power P [W], and a constant noise power spectral density $\sigma_n^2 = \frac{N_0}{2}$ [W/Hz] is given by

$$C_{\text{awgn}} = B \log_2\left(1 + \frac{P}{BN_0}\right) \quad \text{bits/s}, \tag{1.33}$$

Here, $P = RE_b$ is the average power, where E_b [J] is the average signal energy per information bit, and R [bits/s] is the data transmission rate. Hence, we have

$$\frac{P}{BN_0} = \frac{RE_b}{BN_0}. \tag{1.34}$$

Keep in mind that this is a simplified relationship and that actual system behavior may vary depending on the modulation scheme, the coding scheme, and the specific characteristics of the channel. Furthermore, the equation assumes that the SNR–block error rate (BLER) relationship is linear over the range of interest, which may not be the case in all situations.

The channel capacity restricts the data rate R to be smaller than the capacity C to achieve reliable transmission. Here, we define the ratio of the data rate R to the bandwidth B as *spectral efficiency*, $\eta_B = \frac{R}{B}$ [bits/s/Hz]. Now, given the requirement $R < C$, by substituting (1.34) into (1.33) and rearranging it, we have the following:

$$\frac{R}{B} < \frac{C}{B} = \log_2\left(1 + \frac{RE_b}{BN_0}\right). \tag{1.35}$$

Then, rewriting (1.35) gives a constraint on the SNR per bit E_b/N_0 for reliable data transmission as

$$\frac{E_b}{N_0} > \frac{2^{\eta_B} - 1}{\eta_B}, \quad \eta_B = \frac{R}{B}. \tag{1.36}$$

Observe that in the case of infinite bandwidth, the spectral efficiency η_B approaches zero. As $\lim_{\eta_B \to 0} \frac{2^{\eta_B}-1}{\eta_B}$ gives the indeterminate form of $0/0$, using the L'Hopital rule and simplification, we get the minimum required E_b/N_0 for infinite bandwidth as

$$\frac{E_b}{N_0} > \lim_{\eta_B \to 0} 2^{\eta_B} \ln 2 = \ln 2 = 0.693, \quad \text{or} \quad \frac{E_b}{N_0} > -1.59 \text{ dB}. \tag{1.37}$$

Note that the power efficiency limit is independent of the BER. In fact, this limit determines the minimum possible E_b/N_0 to achieve an arbitrarily small BER.

The MATLAB script in C.2 in Appendix C gives the Shannon limit. Also, for more information on modulation schemes, we invite the readers to see Chapter 12.

Before closing this section, let us find the relationship between SNR or $E_s/N_0 = P/N_0$ and R. E_s/N_0 and E_b/N_0 are both SNR metrics used in digital communications. E_s/N_0 is the ratio of the signal energy per symbol to the noise power spectral density. E_b/N_0 is the ratio of energy per bit to the noise power spectral density.

The relationship between E_s/N_0 and E_b/N_0 depends on the modulation scheme used. For example, for binary phase-shift keying (BPSK), E_s/N_0 is equal to E_b/N_0. For quadrature phase-shift keying (QPSK), E_s/N_0 is equal to $2E_b/N_0$. For the 16-quadrature amplitude modulation (16-QAM), E_s/N_0 is equal to $4E_b/N_0$.

In general, the relationship between E_s/N_0 and E_b/N_0 can be expressed as follows:

$$\frac{E_s}{N_0} = \frac{E_b \cdot \log_2(M) \cdot R}{N_0}, \tag{1.38}$$

where R is the code rate, M is the constellation size, and $m = \log_2 M$ is the transmission rate in bits per symbol (or bits per channel use).

Denote by SNR_1 the initial SNR corresponding to a specific BLER and by R_1 the initial code rate. Now, if you change the code rate to R_2, the relationship between the new SNR (SNR_2) and the initial SNR $\left(\mathrm{SNR}_1\right)$ can be expressed as follows:

$$\mathrm{SNR}_2 = \mathrm{SNR}_1 + 10 \cdot \log_{10}\left(\frac{R_2}{R_1}\right). \tag{1.39}$$

This formula is derived from the fact that increasing the code rate generally requires a higher SNR to maintain the same level of performance (in terms of BLER).

1.7 Time and Space Complexity

Anyone can implement the same algorithm to solve an a encoding or decoding problem in various ways. In addition, there may be multiple algorithms available to solve a single problem. Therefore, we require a metric to compare different algorithms and their implementations. The big O notation denoted as $O(f(n))$ is used to describe how the run-time in terms of the number of operations (steps), or the space/memory requirements of algorithms, represented by the function $f(n)$, grows as the input size n increases. Here, the capital letter O stands for the *order of approximation*.

This notation focuses on the dominant factor that influences an algorithm's complexity while disregarding constant factors and lower-order terms. Hence, the big O notation only describes the asymptotic behavior of complexity, not the exact complexity. This attribute makes it an effective tool for comparing and analyzing algorithms, free from the complexities of implementation details.

Example 1.1 Let us consider $O(n^2)$, pronounced as "Big O squared", which is the complexity of the selection sort algorithm. Selection sort is a sorting method that iterates through the list of n elements, ensuring that each element at a given index i is the ith least/greatest element in the list. This sorting algorithm can be implemented by two nested loops where the outer loop iterates through the list (n iterations in total), while for the ith element in the list, the inner loop finds the smallest/largest element in the remaining part of the list (by $n - i$ iterations). Observe that this actually gives a geometric sum, meaning the inner loop repeats

Table 1.1 Example time complexities.

Complexity	Name	Example
$O(1)$	Constant	Array access
$O(\log n)$	Logarithmic	Binary search
$O(n)$	Linear	Linear search
$O(n \log n)$	Linearithmic	Merge sort
$O\left(n^2\right)$	Quadratic	Bubble sort
$O\left(n^2\right)$	Cubic	Matrix–vector multiplication
$O\left(n^3\right)$	Cubic	Matrix–matrix multiplication
$O\left(n^c\right)$	Polynomial	
$O\left(c^n\right), c > 1$	Exponential	Recursive Fibonacci
$O(n!)$	Factorial	Permutations of a set

$n + (n-1) + \cdots + 2 + 1$ times, which equals $n(n-1)/2 = n^2/2 - n/2$ times. This is the time complexity of the selection sorting algorithm. Now, since the dominant term is $n^2/2$, ignoring the coefficient, we get $O(n^2)$.

Table 1.1 lists common time complexities and the corresponding examples, from the best to the worst in terms of complexity:

The formal definition of Big O is as follows:

Definition 1.1 Given two functions $f(n)$ and $g(n)$, then $f(n) = O(g(n))$ if there exist constants $c > 0$ and $n_0 \leq 0$ such that $f(n) \leq c \cdot g(n)$ for all $n \leq n_0$. Observe that in this case, $f(n)$ does not grow faster than $c \cdot g(n)$ for all $n \leq n_0$.

1.8 Historical Milestones in Channel Coding

The field of error correction codes and channel coding has been pivotal in enabling reliable communication over noisy channels, transforming theoretical insights into practical engineering marvels. Rooted in Claude Shannon's 1948 groundbreaking work on information theory, this field has evolved through decades of innovation, driven by the quest to approach the channel capacity—the theoretical limit of reliable communication. Early milestones, such as Hamming codes and Reed–Solomon codes, provided foundational methods for error detection and correction. The introduction of convolutional codes and the Viterbi algorithm revolutionized real-time communications, while advances like turbo

codes and low-density parity-check (LDPC) codes approached channel capacity with iterative decoding. More recently, the advent of polar codes has offered a capacity-achieving solution with efficient implementations, finding applications in modern standards like 5G. Each milestone reflects the ingenuity of researchers in tackling the challenges of noise, bandwidth, and complexity, shaping the technologies that underpin modern communication and data storage systems. Here is a timeline of major historical milestones in channel coding, highlighting key contributions, authors, and publications:

1948: Foundations of Information Theory

- **Author**: Claude E. Shannon
- **Publication**: *A Mathematical Theory of Communication* (Bell System Technical Journal)
- **Contribution**: Shannon's seminal work established the theoretical foundations of information theory, including the concept of channel capacity and the idea that error-free communication is possible with proper coding.

1950: Hamming Codes

- **Author**: Richard W. Hamming
- **Publication**: *Error Detecting and Error Correcting Codes* (Bell System Technical Journal)
- **Contribution**: Hamming introduced the first practical error-correcting code, known as Hamming codes, which could detect and correct single-bit errors in data.

1954: Reed–Muller Codes

- **Authors**: Irving S. Reed and David E. Muller
- **Publication**: *A Class of Multiple-Error-Correcting Codes and Their Decoding Scheme* (IRE Transactions on Information Theory)
- **Contribution**: Reed–Muller codes provided a flexible class of error-correcting codes that could correct multiple errors and were used in early space communication systems.

1954: Golay Codes

- **Author**: Marcel Golay
- **Publication**: *Notes on Digital Coding*
- **Contribution**: Golay introduced the Golay code, a perfect error-correcting code that has been used in various applications, including space exploration.

1955: Convolutional Codes

- **Author**: Peter Elias
- **Publication**: *Coding for Noisy Channels* (IRE Convention Record)
- **Contribution**: Elias introduced convolutional codes, which encode data using a sliding window and are widely used in wireless communication systems.

1957: Cyclic Codes

- **Author**: W. Wesley Peterson
- **Publication**: *Encoding and Error-Correction Procedures for the Bose–Chaudhuri Codes* (IRE Transactions on Information Theory)
- **Contribution**: Peterson developed cyclic codes, including BCH codes, which are efficient for correcting multiple errors and are widely used in digital communication and storage systems.

1960: Bose–Chaudhuri–Hocquenghem (BCH) Codes

- **Authors**: Raj Chandra Bose, Dwijendra K. Ray-Chaudhuri, and Alexis Hocquenghem
- **Publication**: Independent papers in 1959 and 1960
- **Contribution**: BCH codes are a class of cyclic codes that can correct multiple errors and are widely used in applications like satellite communications and QR codes.

1962: Low-Density Parity-Check (LDPC) Codes

- **Author**: Robert G. Gallager
- **Publication**: Low-Density Parity-Check Codes (*IRE Transactions on Information Theory*)
- **Contribution**: Gallager introduced LDPC codes in his doctoral dissertation at MIT, which are capacity-approaching codes. Although initially overlooked, they became widely used in modern communication systems (e.g. Wi-Fi, 5G).

1963: Fano Algorithm

- **Author**: Robert Fano
- **Publication**: A Heuristic Discussion of Probabilistic Decoding (*IEEE Transactions on Information Theory*)
- **Contribution**: The Fano algorithm is a sequential decoding algorithm for convolutional codes, providing an efficient way to decode long constraint-length codes.

1967: Reed-Solomon Codes

- **Authors**: Irving S. Reed and Gustave Solomon
- **Publication**: Polynomial Codes over Certain Finite Fields (*Journal of the Society for Industrial and Applied Mathematics*)
- **Contribution**: Reed–Solomon codes are highly effective for correcting burst errors and are used in CDs, DVDs, and deep-space communication.

1969: Berlekamp–Massey Algorithm

- **Authors**: Elwyn Berlekamp and James Massey
- **Publication**: On the Solution of Algebraic Equations over Finite Fields (*Information and Control*)
- **Contribution**: The Berlekamp–Massey algorithm is an efficient method for decoding BCH and Reed–Solomon codes by solving the key equation in error correction.

1970: Goppa Codes

- **Author**: Valerii D. Goppa
- **Publication**: A New Class of Linear Error-Correcting Codes (*Problemy Peredachi Informatsii*)
- **Contribution**: Goppa codes are used in algebraic coding theory and cryptography, particularly in the McEliece cryptosystem.

1972: Viterbi Algorithm

- **Author**: Andrew J. Viterbi
- **Publication**: *Principles of Coherent Communication* (McGraw-Hill)
- **Contribution**: The Viterbi algorithm provides an efficient method for decoding convolutional codes and is widely used in digital communication systems.

1974: BCJR Algorithm

- **Authors**: Bahl, Cocke, Jelinek, and Raviv
- **Publication**: Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate (*IEEE Transactions on Information Theory*)
- **Contribution**: The BCJR algorithm is a forward–backward algorithm for maximum a posteriori (MAP) decoding of convolutional codes, later used in turbo codes.

1976: Concatenated Codes

- **Author**: David Forney
- **Publication**: *Concatenated Codes* (MIT Press)
- **Contribution**: Forney introduced concatenated codes, which combine two or more error-correcting codes to achieve high performance. These were used in deep-space communication.

1982: Trellis-Coded Modulation (TCM)

- **Author**: Gottfried Ungerboeck
- **Publication**: *Channel Coding with Multilevel/Phase Signals* (IEEE Transactions on Information Theory)
- **Contribution**: TCM integrates coding and modulation to improve error performance without increasing bandwidth, revolutionizing digital communication.

1993: Turbo Codes

- **Authors**: Claude Berrou, Alain Glavieux, and Punya Thitimajshima
- **Publication**: *Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes* (Proceedings of ICC'93)
- **Contribution**: Turbo codes introduced iterative decoding and achieved near-Shannon-limit performance, revolutionizing wireless communication (e.g. 3G and 4G).

1996: Rediscovery of LDPC Codes

- **Authors**: David MacKay and Radford Neal

- **Publication**: *Near Shannon Limit Performance of Low Density Parity Check Codes* (IEEE Transactions on Information Theory)
- **Contribution**: MacKay and Neal rediscovered and modernized LDPC codes, demonstrating their near-Shannon-limit performance and leading to practical implementations in modern communication systems.

2002: Fountain Codes

- **Author**: Michael Luby
- **Publication**: *LT Codes* (Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science)
- **Contribution**: Fountain codes, such as LT (Luby Transform) codes and Raptor codes, introduced rateless coding, which is essential for efficient data streaming and file delivery systems.

2009: Polar Codes

- **Author**: Erdal Arıkan
- **Publication**: *Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels* (IEEE Transactions on Information Theory)
- **Contribution**: Polar codes are the first class of codes proven to achieve channel capacity and are used in 5G networks.

1.9 Why Study Polar Codes?

Polar codes, introduced by Arıkan in 2009, represent a landmark in coding theory as the first class of codes proven to achieve the capacity of symmetric B-DMCs with an explicit construction. This breakthrough has sparked immense interest in both academia and industry due to their theoretical elegance and practical potential.

The primary reasons to study polar codes include:

- Capacity achievement: Polar codes achieve channel capacity asymptotically under successive cancellation (SC) decoding, making them a foundational element in information theory.
- Superior performance for short code-lengths: The error correction performance of short and medium-length polar codes concatenated with cyclic codes or rate-1 convolutional codes under SC list decoding is superior to that of other modern codes.
- Adoption in standards: Polar codes are used in 5G New Radio (NR) standards for control channel communications, underscoring their relevance in practical applications.
- Flexibility: Polar codes support rate adaptation, which makes them suitable for a wide range of communication scenarios, from ultra-reliable low-latency communications (URLLCs) to massive machine-type communications (mMTCs).

- Theoretical insights: Polar codes provide a new perspective on channel polarization, which has led to further developments in coding theory, including polar-like code constructions.

1.9.1 How Do Polar Codes Compare with Other Codes?

Polar codes have unique features, but they also have limitations when compared to other coding schemes, such as LDPC codes, turbo codes, and convolutional codes. Below is a comparison based on various attributes:

Table 1.2 Comparison of polar codes with other coding schemes.

Feature	Polar codes	LDPC codes	Turbo codes	Convolutional codes
Capacity achievement	Achieves channel capacity asymptotically	Approaches capacity with iterative decoding	Approaches capacity for large block lengths	Not capacity achieving
Complexity	$O(L \cdot n \cdot \log n)$	$O(I \cdot n \cdot d_c)$	$O(I \cdot n \cdot 2^m)$	$O(n \cdot 2^m)$
Decoding algorithm	SC list (SCL) decoding	Message passing (belief propagation) algorithm	SISO iterative decoding (BCJR)	Viterbi algorithm
Performance at short lengths	Excellent	Inferior, widely used for long codes	Inferior, used for long codes	Good
Adoption in standards	5G control channels	5G traffic channels, Wi-Fi, DVB-S2, etc.	3G, 4G LTE	Legacy systems like 3G GSM
Design complexity	Explicit construction with fixed parameters	Requires optimization of parity-check matrix	Requires careful interleaver design	Simple design, no interleaver needed
Flexibility in code rate	Requires puncturing/shortening	Highly flexible	Flexible with interleavers	Limited
Error floor	No inherent error floor	May suffer from error floors at high SNR	May suffer from error floors at high SNR	No inherent error floor
Hardware efficiency	Efficient for SC decoding, low for SCL decoding	Moderate, depends on iterative decoder design	Moderate to high	Highly efficient

The parameters used in Table 1.2 are as follows. The code length is denoted by n, the number of iterations in interactive decoding is denoted by I (typically ranges from 10 to 50, depending on the target performance and latency constraints), L is the list size in list decoding algorithm (number of paths retained during decoding, typically small in practical systems; e.g. $L = 8, 16, 32$, to balance complexity and performance), d_c is the average column weight of the parity-check matrix (average number of connections per variable node, or the average degree of variable nodes), m is the constraint length (number of memory elements in the component convolutional codes). The complexity of decoding convolutional codes (standalone or as component codes in Turbo codes) grows exponentially with the constraint length m, making Viterbi decoding suitable for short constraint lengths (≤ 10). However, this makes turbo decoding more computationally expensive than LDPC decoding for large block lengths due to higher complexity in component decoders. Furthermore, hardware optimizations such as parallelism in belief propagation algorithm can significantly reduce the decoding latency of LDPC codes. It is worth noting that the decoding complexity of polar codes in Table 1.2 does not consider simplified and fast decoding techniques using special nodes, described in Chapter 9.

A comprehensive performance comparisons under various decoding algorithms and code parameters can be observed in Chapter 13.

Exercises

1 Using the relation $p(x, y) = p(x)p(y|x)$, show the following chain rule:

$$H(X, Y) = H(X) + H(Y|X).$$

2 Given a BSC with crossover probability $p = 0.01$. If an independent-symbol binary source with probabilities 0.25 and 0.75 is transmitted over this channel, calculate $H(X/Y)$ and $I(X, Y)$.

3 Given the joint distribution of (X, Y) as follows:

Y	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

Calculate $H(X)$ and $H(X|Y = 1)$.

4 Suppose that we have two random variables, X and Y, where X is the outcome of a fair coin toss (heads or tails), and Y is the outcome of a roll of a fair die (1–6).

a) Calculate the mutual information between X and Y.
b) Change the random variable X to the outcome of a biased coin toss (heads with probability 0.7 and tails with probability 0.3) and recalculate mutual information between X and Y.

5 Using the information inequality, show that $I(X;Y) = 0$ only if X and Y are independent.

6 A binary erasure channel has an erasure probability of 0.2. What is the capacity of the channel?

7 A channel with a bandwidth $B = 25\,\text{kHz}$ and an SNR of 18 dB is perturbed by AWGN. What is the capacity of this channel in bits per second?

8 What is the Shannon limit if we employ an ideal 2-PAM modulation scheme?

Bibliographic Notes

In his 1948 paper, *A Mathematical Theory of Communication* Shannon [1948], Claude E. Shannon laid the foundation for modern information theory. He introduced the concept of channel capacity and developed the channel coding theorem. In addition to his work on channel capacity and channel coding, Shannon also made significant contributions to the theory of entropy and mutual information. Shannon's work has had a profound impact on the design of communication systems. Gallager in his classic text, Information Theory and reliable communication Gallager [1968], delves into the fundamentals of information theory, discrete entropy, and various coding techniques. It provides a comprehensive treatment of the mathematical aspects of information theory. Furthermore, Cover and Thomas' book, *Elements of Information Theory* Cover and Thomas [2006], is widely used as a reference for understanding information theory. It covers topics such as entropy, mutual information, and channel capacity, providing both theoretical insights and practical applications.

Investigating the lower bound for the error performance in channel coding dates back to the work of Shannon et al. [1967] and earlier. Among recent results available in the literature, an approach called normal approximation (NA) is commonly used to identify the limit of reliable communication for a block code of length n. This limit for the code rate R_{NA} for a wide range of channels is stated as Polyanskiy et al. [2010]. Unlike Shannon [1959] based on the geometric construction associated with the sphere packing, the normal approximation (NA) limit of Polyanskiy et al. [2010] is based on statistical properties and can then be applied to

any channel model. More refined and simple evaluations of this approach based on integral expressions for various channels are provided by Erseghe [2016].

Jacobs and Berlekamp [1967], Jacobs and Wozencraft [1965], and Forney [1968] have shown the relationship between cutoff rate, code rate, error probability, and complexity of sequential and list decoding.

For a discussion of the Bhattacharyya parameter on the AWGN channel, the readers can refer to Li and Yuan [2013].

Bibliography

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

Tomaso Erseghe. Coding in the finite-blocklength regime: Bounds based on Laplace integrals and their asymptotic approximations. *IEEE Transactions on Information Theory*, 62(12):6854–6883, 2016.

G Forney. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Transactions on Information Theory*, 14(2):206–220, 1968.

Robert G Gallager. *Information theory and reliable communication*, volume 588. Springer, 1968.

I Jacobs and E Berlekamp. A lower bound to the distribution of computation for sequential decoding. *IEEE Transactions on Information Theory*, 13(2):167–174, 1967.

Irwin Mark Jacobs and John M Wozencraft. *Principles of communication engineering*, 1965.

Huijun Li and Jinhong Yuan. A practical construction method for polar codes in AWGN channels. In *IEEE 2013 Tencon-Spring*, pages 223–226. IEEE, 2013.

Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Claude E Shannon. Probability of error for optimal codes in a Gaussian channel. *Bell System Technical Journal*, 38(3):611–656, 1959.

Claude E Shannon, Robert G Gallager, and Elwyn R Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels. I. *Information and Control*, 10(1):65–103, 1967.