Federal University of Goiás

Institute of Physics

Quantum Pequi Group

The Postulates of Classical Information Theory

Monography

Ronaldo Ferreira Costa

Goiânia - 2021
Brazil

FEDERAL UNIVERSITY OF GOIÁS
INSTITUTE OF PHYSICS

The Postulates of Classical Information Theory

Ronaldo Ferreira Costa

Monograph presented to the Institute of Physics of the Federal University of Goiás as partial requirement to obtain the Bachelor's degree in Physics .

ADVISOR: *Prof. Dr. Lucas Chibebe Céleri*

Goiânia - 2021

Dedication

My Mom and Dad

Acknowledgments

Many people collaborated with the realization of this work. First, my parents, who, even without understanding anything I do, never denied support and always encouraged me to continue with my studies.
Secondly, my advisor, because without the discussions and several meetings we had, this work would not be possible.
My closest friends, who I couldn't fail to mention: Vinição, Miltin, Cattany, Gustavão, Luiz Henrique, Arthur, Dudu, Luciano.
I can't forget the Whatsapp group "Gangue Ué" and all the discord people who were present at opportune moments.
CNPq for the scientific initiation scholarship that was the basis of this work.

ABSTRACT

The present work aims to discuss the basic postulates of classical information theory, know as the Shannon coding theorems. The firs postulate states that information can be compressed up to a certain quantity, called Shannon entropy. The second postulates establishes the maximum rate of information transmission, the channel capacity. These postulates were proposed by Shannon in 1948 and gave birth to the field known today as Information Theory , with a huge influence not only in communication theory, but in several areas of science and engineering.

Contents

Chapter 1

Introduction

Information theory is the science of storing, processing, transmitting and using information [1] [2].

All beings communicate in some way. Here we will focus on the communication of human beings. First, we will establish a timeline up to the present day.

The emergence of language in the context of communication is dated between 200000 and 5000 B.C. Writing has its first records dated around 5000 B.C. In the meantime, there was also communication through objects or even smoke, which is used until today to warn of an emergency. Longer communication systems such as letters date their appearance from Persia in the mid 6th century B.C. However, this was limited to the same continents until the 15th century, when a great advance was made with navigation. Another giant step in communication came with the advent of the telegraph, dated 1838, which uses Morse code as its means of communication, which in turn uses short phrases and frequent symbols instead of long phrases with symbols not often used - keep this phrase in mind, it will be very useful later on. The telephone and the radio, launched in 1848 and 1896, respectively. Television in 1927. The Internet, which we can't do without nowadays, in 1983, and mobile Internet networks in the 1980s, and so on.

Communication theory was pioneered by Claude Elwood Shannon who was a mathematician and electronic engineer. Shannon was a researcher at Bell Laboratories from 1941 to 1972 and it was while he was there, in 1948, that he developed the precursor work of all information theory called *A Mathematical Theory of Communication*[3]. In this work the fundamental problem of communication is presented, which is how to reproduce a selected message exactly the same or approximately the same at another point. From this problem two fundamental questions arise: What is the maximum compression of information and what is the maximum communication rate for a channel? Shannon answered these questions as the entropy and the channel capacity, respectively. His work forms the basis of information theory, which stands on the pillar of these two theorems and has consequently applications in: Computer Science, Physics, Mathematics, Economics and so on.

The goal of the present monograph is to present the two main postulates of information theory as well as the definitions and theorems necessary for a minimal understanding, along with some consequences.

Chapter 2

Classical Information Theory

2.1 Introduction

Information theory rests on two fundamental pillars of communication theory: data compression and transmission of information [3]. The theory is fathered by Claude Elwood Shannon (1916-2001), as mentioned in the Introduction. Shannon had ideas that went well beyond his time. Its two main theorems, together with the proposed applications, only reinforce this fact. In this chapter, we will present a short view of some of the ideas that cannot be missed for the comprehension of the text that follows.

2.2 The problem of data compression

We live in the information age, where our biggest vehicle of communication is the internet. Here we are in contact with data compression formats, where the most common are **JPG** (used to compress photographic images), **PNG** (data format to generate images), **ZIP** (a file compatibility format) among others. All of these file formats have algorithms that govern data input and output [2].

Common sense can lead us to think that it is possible to compress the output of the information source to arbitrarily small size. For example, compressing 1 GB into 1 byte. However, when looking more closely at the problem of data compression we can see that it is not possible to compress the output of the information source to some arbitrarily small size. This fact was proved by Shannon and constitute his first coding theorem [2][4].

2.3 Information coding: from source to data reading

A question that we can raise is: for which scheme we would have the best data compression rate?

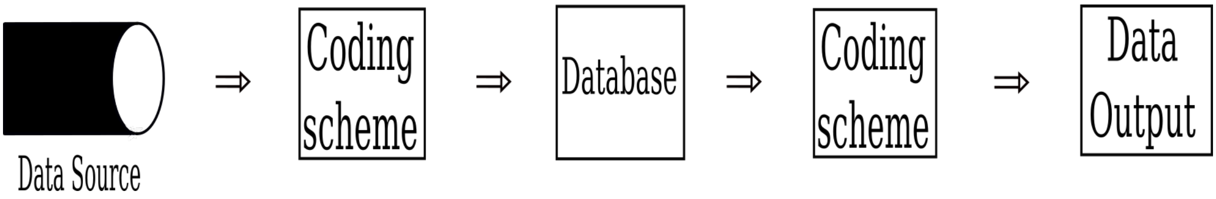

Figure 2.1: The figure represents a transmission of a classical information. It consists of the source of information, followed by the coding scheme that translates the information into codewords (usually compressed). This information is then transmitted to the receiver that can read it after the application of the decoding scheme.

Shannon answered this question as follows: If we consider that our source of information is very general and add the idea of the set of typical sequences [1], we have the most general representation possible. To illustrate this, we invited Alice and Bob to assist us. Alice wants to send a message to Bob, however, due to the degree of importance and secrecy that this message carries, they decide to use a channel that is noiseless, that is, if Alice sends "0", "0" will arrive, if she sends "1" the "1" will arrive. Remember that both Alice and Bob want to use the channel as little as possible, since it is expensive.

Let's assume that Alice has a random source of information that sends Bob four symbols, denoted by $\{a, b, c, d\}$ with the following probability distribution

$$\begin{aligned} \Pr\{a\} &= \frac{1}{2} \\ \Pr\{b\} &= \frac{1}{8} \\ \Pr\{c\} &= \frac{1}{4} \\ \Pr\{d\} &= \frac{1}{8}, \end{aligned} \tag{2.1}$$

so that the source will emit any of these symbols with the given probabilities.

The noiseless channel only accepts bits, that is, for each symbol a binary set will be associated, such that

$$a \to 00,\ b \to 01,\ c \to 10,\ d \to 11$$

where the binary representation of the symbols is called a *codeword*.

[1] This will be explained latter.

The measure that defines the efficiency of the coding scheme is the size of the codeword. In the example of Alice with Bob, the codeword is two bits long. However, as we have an associated probability, we can devise a scheme in which it is possible to have a smaller codeword. Consider the following code scheme

$$a \to 0,\ b \to 110,\ c \to 10,\ d \to 111$$

It is important to note here that the code scheme can be uniquely decoded. Continuing with our example, Alice sends bob a string given by

$$0011010111010100010$$

and bob can see that the sequence is

$$0\ \ 0\ \ 110\ \ 10\ \ 111\ \ 0\ \ 10\ \ 10\ \ 0\ \ 0\ \ 10$$

and can determine that the message is

$$aabcdaccaac.$$

From that, we can calculate the size of the expected code scheme with the aid of the probability distribution. By looking at the message, we can create the following table:

Symbol	Codeword	Bit size
a	0	1
b	110	3
c	10	2
d	111	3

Table 2.1: Symbols, codewords and bit size of the considered sequence.

and then we calculate the expected size of this code scheme

$$\frac{1}{2}(1) + \frac{1}{8}(3) + \frac{1}{4}(2) + \frac{1}{8}(3) = \frac{7}{4} = 1,75.$$

Therefore, one can be concluded from the example that the intuitive scheme (the first part of the example), when given a probability distribution and a random data source, is not the smallest expected size of the code scheme. In other words, it is not the optimal code. A code based on the construction of the codewords accordingly with the probability distribution of the source presents a better performance. In fact, it is possible to prove that this code is the optimal one.

2.4 How to measure information

Let the random variable $X = \{\mathcal{X}, p_X(x)\}$ whose symbols $\mathcal{X} = \{a, b, c, d, \cdots\}$ occurs with probability $p_X(x)$, with $x \in \mathcal{X}$. The measure of the information content in the occurrence of symbol x is given by [3, 2]

$$i(x) \equiv \log_2 \left(\frac{1}{p_X(x)} \right) = -\log_2 p_X(x). \tag{2.2}$$

The base two of the logarithm is chosen as a convention, in such a way information is measured in bits. It has the required property of being high for low probabilities and of low for high probabilities. In other words, the lower the probability, the greater the surprise may be if the event occurs and vice versa. It is worth remembering that the information content enjoys the property of additivity for independent events. See appendix B for a proof based on the Shannon postulates.

Suppose the font produces two symbols, denoted by x_1 and x_2 that correspond to the random variables X_1 and X_2, respectively. The probability for this event, as long as the symbols are uncorrelated, is $p_{X_1X_2}(x_1, x_2) = p_{X_1}(x_1)\, p_{X_2}(x_2)$. The information that is contained x_1 and x_2 is additive because

$$\begin{aligned} i(x_1, x_2) &= -\log_2 \left[p_{X_1X_2}(x_1, x_2)\right] \\ &= -\log_2 \left[p_{X_1}(x_1) p_{X_2}(x_2)\right] \\ &= -\log_2 \left[p_{X_1}(x_1)\right] - \log_2 \left[p_{X_2}(x_2)\right], \end{aligned} \tag{2.3}$$

which shows the additivity of the information for independent events.

Thus, the average information associated with the source is

$$H(X) = \sum_{x \in \mathcal{X}} p_X(x) i(x) = -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x), \tag{2.4}$$

which is "the Shannon ***entropy*** of the information source", a key quantity in information theory.

In order to illustrate what we saw, let us compute the entropy associated with the code scheme given in Eq. (2.1)

$$\begin{aligned} H(X) &= -\sum_{x\in\mathcal{X}} p_X(x)\log_2\left[p_X(x)\right] \\ &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} \\ &= \frac{1}{2}(1) + \frac{1}{8}(3) + \frac{1}{4}(2) + \frac{1}{8}(3) \\ &= \frac{7}{4}. \end{aligned} \tag{2.5}$$

Therefore, the case used by Alice and Bob is a particular one in which all the probabilities are reciprocal powers of two. The entropy is equal to the expected length of the codeword.

2.5 Shannon's source coding theorem

In this section we present the first Shannon coding theorem, whose proof is presented latter.

Let X be the random variable describing a general information source. In order to encode the information source, we can associate a codeword to each symbol in the alphabet $\mathcal{X}$, as shown in table 2.1.

The innovative idea that Shannon had was to let the source emit a large number of achievements and encode the data emitted as a large block, instead of encoding each symbol, as in the example given in the preceding section. This technique is called "block coding". Shannon also allowed a small error in the coding scheme and, mainly, to show that this error disappears when the size of this block becomes very large'.

Let us consider that the source is allowed to emit the following sequence

$$x^n \equiv x_1\; x_2\; x_3\; x_4\; \cdots x_n, \tag{2.6}$$

in which n represents an arbitrarily large number and indicates the size of the emitted data block, x_i $\forall i = 1, ..., n$ denotes the i-th symbol emitted. Figure 2.2 shows this scheme.

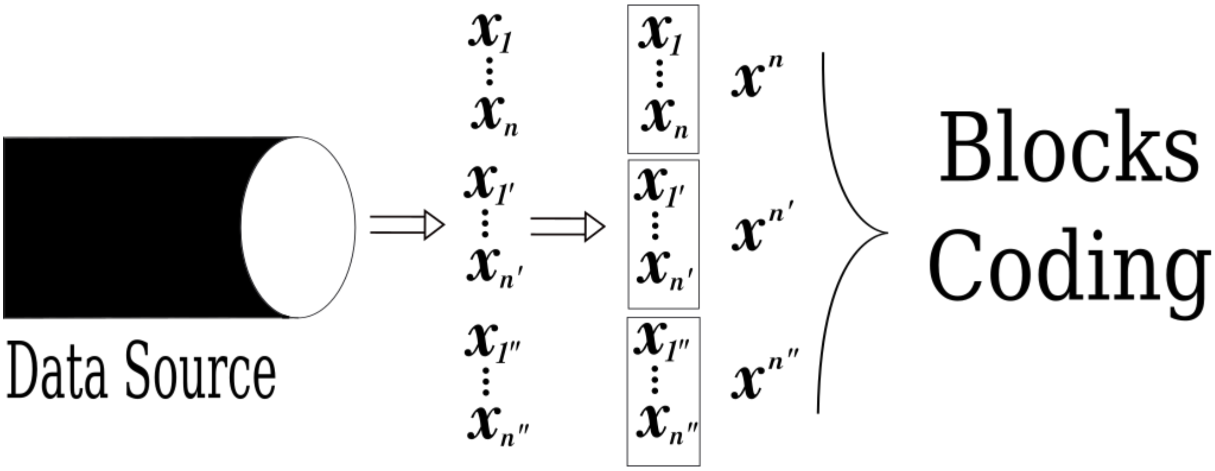

Figure 2.2: First Shannon coding theorem. The information source emits several blocks of symbols that are then encoded.

Let X^n be the random variable associated with the sequence x^n. It is crucial to assume that the symbols in each sequence are *independently and identically distributed* (i.i.d), that is, that each random variable X_i has the same probability distribution as the random variable X and we use the index i only to track each symbol x_i. So, by assuming it is i.i.d., the probability of any sequence emitted is given as

$$\begin{aligned} p_{X^n}(x^n) &= p_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) \\ &= p_{X_1}(x_1)p_{X_2}(x_2)\cdots p_{X_n}(x_n) \\ &= p_X(x_1)p_X(x_2)\cdots p_X(x_n) \\ &= \prod_{i=1}^{n} p_X(x_i). \end{aligned} \tag{2.7}$$

Now, let us write the letters of the alphabet as $\mathcal{X}$ as $a_1, \ldots, a_{|\mathcal{X}|}$[2] in order to distinguish the letters from the realizations. We will then denote the number of occurrence of the letter a_i in the sequence x^n by $N(a_i|x^n)$ such that $i = 1, \ldots, |\mathcal{X}|$. For instance, if $x^n = acaabcadcca$ we have $N(a|x^n) = 5$, $N(b|x^n) = 1$, $N(c|x^n) = 4$ and $N(d|x^n) = 1$.

Given this definition, we can rewrite Eq. 2.7

$$p_{X^n}(x^n) = \prod_{i=1}^{n} p_X(x_i) = \prod_{i=1}^{n} p_X(a_i)^{N(a_i|x^n)} \tag{2.8}$$

It is worth mentioning that the length n of the emitted sequence is allowed to be extremely large,

[2] $|\mathcal{X}|$ is the cardinality of the set.

so that it is much larger than the alphabet $|\mathcal{X}|$, $n \gg |\mathcal{X}|$. Therefore, we get a simplification with respect to Eq. 2.7 since we have a smaller number of factors.

Since we are interested only in the number of occurrences, the order of the letters does not matter. Then

$$x^n \to \underbrace{a_1 \dots a_1}_{N(a_1|x^n)} \;\; \underbrace{a_2 \dots a_2}_{N(a_2|x^n)} \;\; \cdots \;\; \underbrace{a_{|\mathcal{X}|} \dots a_{|\mathcal{X}|}}_{N(a_{|\mathcal{X}|}|x^n)},$$

given that the probability calculation is invariant by permutations.

Therefore, Eq. 2.8 characterizes the probability of occurrence of any sequence x^n. Everything we talked about so far is applied to a specific x^n sequence emitted by the information source.

Now, we will present how a random sequence behaves. It is worth noting that in the case in question, the sample average of the information content of the random sequence x^n will be considered[3].

$$-\frac{1}{n}\log_2[p_{X^n}(x^n)] = -\frac{1}{n}\log_2\left[\prod_{i=1}^{|\mathcal{X}|} p_{X^n}(a_i)^{N(a_i|x^n)}\right] \tag{2.9}$$

so that we can rewrite

$$-\frac{1}{n}\log_2\left[\prod_{i=1}^{|\mathcal{X}|} p_{X^n}(a_i)^{N(a_i|x^n)}\right] = -\sum_{i=1}^{|\mathcal{X}|}\frac{N(a_i|x^n)}{n}\log_2[p_X(a_i)]. \tag{2.10}$$

In other words, the sample mean can be rewritten by

$$-\frac{1}{n}\log_2[p_{X^n}(x^n)] = -\sum_{i=1}^{|\mathcal{X}|}\frac{N(a_i|x^n)}{n}\log_2[p_X(a_i)]. \tag{2.11}$$

Now, the function $N(a_i|\cdot)$, being the number of appearances of the letter a_i in the random sequence x^n, is a random variable which represents the distribution for the letters a in a given alphabet $\mathcal{X}$. As n becomes large, the laws of large numbers states that it is extremely likely that given a random sequence, its empirical distribution $N(a_i|x^n)/n$ approaches the true distribution $p_X(a_i)$. Note that the reciprocal is not true, it is very unlikely that a random sequence will satisfy this property.

In the next section we will define the typical sequence that will be latter employed in the proof of the first Shannon coding theorem.

[3] Dividing the information content by n, we have the sample mean.

2.6 Typical sequences and Asymptotic Equipartition Property (AEP)

A random sequence of length n is given by

$$X^n = (X_1, X_2, \ldots, X_n),$$

while one of its possible realization is

$$x^n = (x_1, x_2, \ldots, x_n).$$

Remembering that n represents a large number. The joint probability density function is given by the product of the individual probabilities, which, recalling the expression already mentioned, is given by

$$p_{X^n}(x^n) = \prod_{i=1}^{n} p(x_i) = p(x^n)$$

in which the total number of possible sequences will be $|\mathcal{X}|^n$.

Considering the set of all possible sequences x^n, our goal is to identify a subset that contains most of the probability. Such subset, called typical, indeed exists and it is exponentially smaller than the total set. Therefore, in an encoding protocol, we just need to codify the typical set. This is the central idea behind data compression.

Let us start with the following theorem.

Theorem 2.6.1. *If* $X_1, X_2, \ldots, X_n$ *is i.i.d* $\sim p(x)$ *so*

$$-\frac{1}{n}\log_2 p(x_1, x_2, \ldots, x_n) \quad \longrightarrow \quad H(X) \;\; in \;\; probability.$$

Proof. Since the functions of independent random variables are also independent random variables we can write

$$\begin{aligned} -\frac{1}{n}\log_2 p(x_1, x_2, \ldots, x_n) &= -\frac{1}{n}\sum_i \log_2(X_i) \\ &= E[-\log_2 p(X)] \quad (in \;\; probability) \\ &= H(X) \end{aligned}$$

□

A brief exposition of what is meant by the asymptotic equipartition property is necessary. AEP derives directly from the weak form of the law of large numbers —which fundamentally states that the average of the results obtained by a large number of trials is close to the population mean— that is $-(1/n)\log p(X_1, X_2, \ldots, X_n)$, when n is large, tends to entropy of X. So, the probability $p(X_1, X_2, \ldots, X_n)$ assigned to a sequence has the limit 2^{-nH} [2].

Based on the AEP, we aim to identify the subset $A_\epsilon^{(n)} \subset |\mathcal{X}|^n$ that contains most of the probability. In other words, for arbitrarily small ϵ we want to find the set such that

$$p[X^n \in A_\epsilon^{(n)}] = \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \geq 1 - \epsilon \tag{2.12}$$

Definition 2.6.1. *The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is a set of strings $(x_1, x_2, \ldots, x_n) \in |\mathcal{X}|^n$ defined by*

$$A_\epsilon^{(n)} = \left\{ x^n \in |\mathcal{X}|^n : 2^{-n(H(x)+\epsilon)} \leq p_{X^n}(x^n) \leq 2^{-n(H(x)-\epsilon)} \right\}.$$

Equivalent, the set of all sequences $x^n \in |\mathcal{X}|^n$ must obey

$$\underbrace{H(X)}_{} - \epsilon \leq \underbrace{-\frac{1}{n} \log_2 pX^n(x^n)}_{} \leq \underbrace{H(X)}_{} + \epsilon. \tag{2.13}$$

$$entropy \quad empirical\ entropy \quad entropy \tag{2.14}$$

This is the set of all strings whose probability is approximately equal to the expected probability.

The typical set is one that contains almost all probability. In addition, all sequences that are typical have approximately the same probability, which in summary is the asymptotic equipartition property.

The advantage that is derived from this definition is that we can characterize the size of the typical set in terms of the entropy $H(X)$ starting from the law of large numbers, resulting in the following theorem

Theorem 2.6.2 (Direct coding theorem). *For all $n \geq N_\epsilon$ we have*

$$p[X^n \in A_\epsilon^{(n)}] \geq 1 - \epsilon \tag{2.15}$$

so the upper limit is given by

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)} \tag{2.16}$$

and the lower limit by

$$|A_\epsilon^{(n)}| \geq (1 - \epsilon) 2^{n(H(X)-\epsilon)}. \tag{2.17}$$

Proof. Since the probability is normalized, $\sum_{x\in|\mathcal{X}|} p(x^n) = 1$, we can write

$$\begin{aligned} 1 &\geq \sum_{x^n\in A_\epsilon^{(n)}} p(x^n) \\ &\geq \sum_{x^n\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\ &= |A_\epsilon^{(n)}|2^{-n(H(X)+\epsilon)}, \end{aligned} \tag{2.18}$$

remembering that $|A_\epsilon^{(n)}|$ denotes the number of elements in the set $A_\epsilon^{(n)}$. Then, the upper limit on the size of the typical set is

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}. \tag{2.19}$$

Now, let us move to the lower limit. For a sufficiently large n we have

$$p[A_\epsilon^{(n)}] > 1-\epsilon, \tag{2.20}$$

from what we get

$$\begin{aligned} 1-\epsilon < p[A_\epsilon^{(n)}] &= \sum_{x^n\in A_\epsilon^{(n)}} p(x^n) \\ &\leq \sum_{x^n\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \\ &= |A_\epsilon^{(n)}|2^{-n(H(X)-\epsilon)}. \end{aligned} \tag{2.21}$$

Therefore, we conclude that $1-\epsilon \leq |A_\epsilon^{(n)}|2^{-n(H(X)-\epsilon)}$, thus implying the following lower bound on the size of the typical set

$$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}. \tag{2.22}$$

□

What can we learn from this? Since the typical set has asymptotically all the probability, we can consider only such set in the encoding scheme and simply disregard the sequences that are not typical. The error will disappear asymptotically. That was Shannon's key idea. In the following section we will show that the size of the typical set is, in general, much smaller than the set of all possible sequences.

2.7 Representation of some consequences of AEP and Data Compression

Given $X_1, X_2, \dots, X_n$ random variables *i.i.d.*, obtained from the probability density function $p(x)$, we will look for short descriptions for sequences of random variables. Every sequence in $|\mathcal{X}|^n$ will belong to one of the two sets: the typical set $A_\epsilon^{(n)}$ and the complementary set $A_\epsilon^{(n)\ c}$.

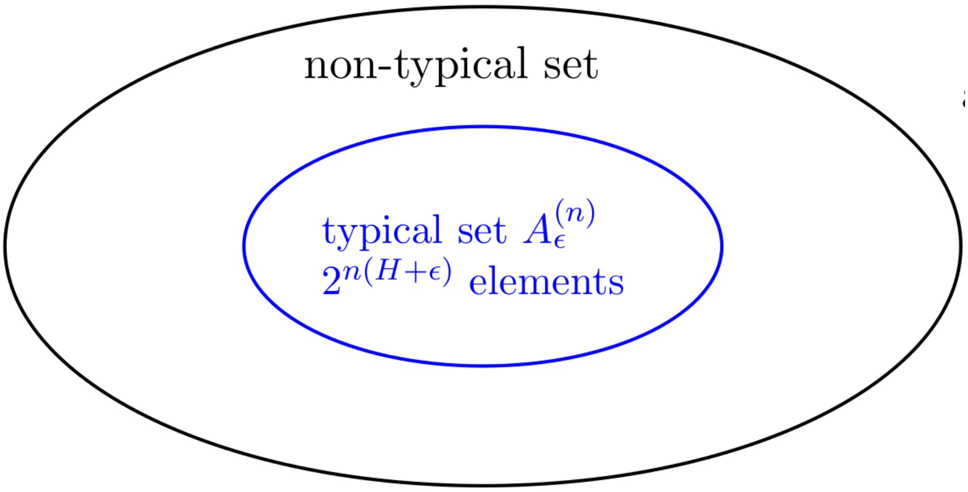

Figure 2.3: Illustration of typical set with respect to $\mathcal{X}^n$.

When ordering all the elements of the set according to some sort of ordering (cardinality, alphabetical or even lexicographical order), we can then represent the typical sequence $A_\epsilon^{(n)}$ of the set by the index that is used in the sequence of the set. Noticing that there are $2^{n(H+\epsilon)}$ or less typical sequences in $A_\epsilon^{(n)}$, no more is needed than $n(H+\epsilon)+1$ bits. The extra bit is necessary because occasionally $n(H+\epsilon)$ may not be an integer. The sequence is usually prefixed by an initial value in which cardinality is respected, for example, in $\mathbb{R}_+$ this value would be 0, which would still favor a total length limited by $A_\epsilon^{(n)}$. It is worth mentioning that the important quantity is $n(H)$ because this is the quantity that describes, in the asymptotic limit, the size of the typical sequences.

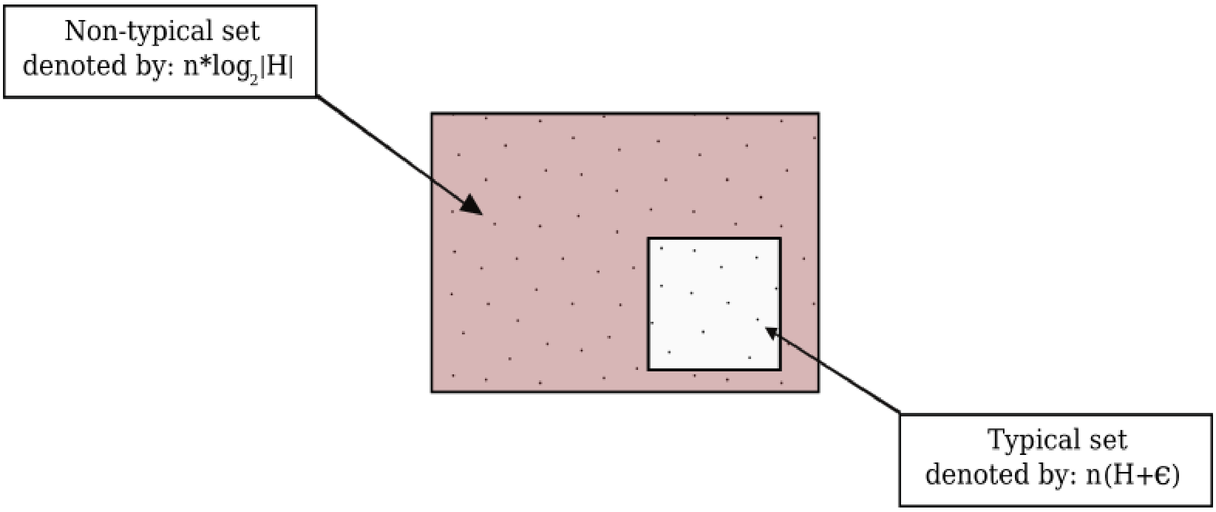

Figure 2.4: Representation of the source code encoding using typical sequences.

We can then summarize the encoding process: Each code is unique and easily decodable; we use a numbering in which we only take into account the elements of the set $A_\epsilon^{(n)}$, since they represent a smaller quantity in $|\mathcal{X}|^n$ and surprisingly, this is sufficient for an efficient description.

Chapter 3

Data Compression

3.1 Introduction to Lossless Data Compression

In the previous chapter, it was briefly presented how the coding, measurement, data reading and data compression works. The first change to be made in this chapter is to add content to the definition of entropy, which will involve establishing a fundamental limit for information compression, so that data compression is achieved by assigning short descriptions to the most frequent results and longer descriptions to the less frequent results. An example is the Morse code in which the symbol is more frequent in the alphabet is represented by a point.

3.1.1 Fundamental limits of compression: Kraft-McMillan Inequality

Our efforts will focus on building codes that are instantaneous and have a minimum length expected to describe a specific source. For this, it is necessary not to assign small codewords to all the symbols that come from the source and also, to leave them without any necessary prefix to decode it. Therefore, the set of possible codeword sizes for instantaneous codes should be limited by the Kraft inequality.

Lemma 3.1.1. *Consider an alphabet $\mathcal{X}$, with cardinality $|\mathcal{X}|$. Let $l(x)$ be a function denoting the length of a codeword $C(x)$ $\forall\, x \in \mathcal{X}$. There is an decodable encoding, with size and with codeword D if, and only if,*

$$\sum_{x\in\mathcal{X}} D^{l(x)} \leq 1 \tag{3.1}$$

that is, each code that is uniquely decoded satisfies the Kraft-McMillan inequality.[1]

Proof. Let $C(x)$ be a source of codes that is exclusively decodable. For a given sequence of

[1] In short, it is a generalization of Kraft inequality, as it generalizes to non-binary coding systems.

origin x^n, we will have the length of the extended codeword is given by

$$l(x^n) = \sum_{i=1}^{n} l(x_i) \tag{3.2}$$

and that, therefore, the extended codeword length function obeys the following

$$nl_{min} \leq l(x^n \leq nl_{max} \tag{3.3}$$

so that $l_{min} = min_{x \in \mathcal{X}}\ l(x)$ and $l_{max} = max_{x \in \mathcal{X}}\ l(x)$.

Let A_k also be the number of strings with the source in the font of size n, for which $l(x^n) = k$, i.e

$$A_k = \{x^n \in \mathcal{X}^n : l(x^n) = k\}. \tag{3.4}$$

As long as the code is exclusively decodable, the number of the source string with codewords of size k cannot exceed the number of D-arias of length k

$$A_k \leq D^k \tag{3.5}$$

that applying to extended codewords sizes, they must obey

$$\begin{aligned}
\sum_{x^n \in \mathcal{X}^n} D^{-l(x^n)} &= \sum_{x^n \in \mathcal{X}^n} \left\{ \sum_{k=1}^{\infty} [l(x^n) = k]\, D^{-k} \right\} && (3.6)\\
&= \sum_{k=1}^{\infty} \left\{ \sum_{x^n \in \mathcal{X}^n} [l(x^n) = k] \right\} D^{-k} && (3.7)\\
\sum_{x^n \in \mathcal{X}^n} D^{-l(x^n)} &= \sum_{k=1}^{\infty} A_k D^{-k} && (3.8)
\end{aligned}$$

and how they are only decodable

$$\begin{aligned}
\sum_{x^n \in \mathcal{X}^n} D^{-l(x^n)} &\leq \sum_{k=nl_{min}}^{nl_{max}} D^k D^{-k} && (3.9)\\
& leq \quad nl_{max} && (3.10)
\end{aligned}$$

and with that they must obey the following

$$\sum_{x^n \in \mathcal{X}^n} D^{-l(x^n)} = \sum_{x_1 \in \mathcal{X}} D^{-l(x_1)} \times \sum_{x_2 \in \mathcal{X}} D^{-l(x_2)} \times \sum_{x_3 \in \mathcal{X}} D^{-l(x_3)} \cdots \tag{3.11}$$

$$\sum_{x^n \in \mathcal{X}^n} D^{-l(x^n)} = \left[\sum_{x \in \mathcal{X}} D^{-l(x)} \right]^n \tag{3.12}$$

and that combining the results above

$$\left[\sum_{x \in \mathcal{X}} D^{-l(x)} \right]^n \leq n l_{max}. \tag{3.13}$$

If the code does not satisfy the Kraft-McMillan inequality, the left side will explode exponentially so that n becomes large and soon, that inequality will be validated. Therefore, the code must satisfy the Kraft-McMillan inequality. □

Corollary 3.1.1. *Given a decodable code exclusively for a font alphabet $\mathcal{X}$ it also satisfies the Kraft inequality.*

3.2 Shannon's First Theorem: Data Compression

To compress the data, Shannon created a protocol that is summarized in a scheme in which you can compress the output of an information source, as long as it is $i.i.d.$

Theorem 3.2.1. *For all uniquely decodable code that comes from a source that has iid properties, we have that the average size $\bar{l}(x)$ of each encoded symbol satisfies:*

$$H(x) \leq \bar{l}(x) \tag{3.14}$$

so that the entropy of an information source is specified by a discrete random variable x is the lowest achievable rate for compression.

Proof. Here we will use the log in base 2 because our focus is on computation and bits, however we will omit it for a better writing of the text. However, logarithms in the α base could be used, as $H_\alpha(x)$ indicates entropy relative to base α.

Be

$$-\sum_{\text{x}} p_X(x) \log \left[p_X(x) \right] \leq \bar{l} \tag{3.15}$$

$$\tag{3.16}$$

which is equivalent to

$$\begin{aligned} -\sum_{\mathrm{x}} p_X(x) \log\left[p_X(x)\right] \;&\leq\; \bar{l}(x) \cdot log(K)\ ;\ (K \geq 2). && (3.17) \\ & && (3.18) \end{aligned}$$

First, let's show that $H(x) - \bar{l}(x) \leq 0$, the way that

$$\begin{aligned} H(x) - \bar{l}(x)\ log(K) \;&=\; -\sum_{\mathrm{x}} p_X(x) \log\left[p_X(x)\right] && (3.19) \\ -\sum_{\mathrm{x}} p_X(x) \log\left[p_{\mathrm{X}}(x)\right] \;&=\; \sum_{\mathrm{x}} p_{\mathrm{X}}(x)\ log\left(\frac{K^{-l_{\mathrm{x}}}}{p_{\mathrm{X}}(x)}\right). && (3.20) \end{aligned}$$

Now, using inequality:

$$\log_{\alpha} b \leq b - 1\ ; \alpha \in \mathbb{N} \tag{3.21}$$

We have to

$$H(x) - \bar{l}(x)\ log(K) \;\leq\; \sum_{\mathrm{x}} p_{\mathrm{X}}(x) log\left(\frac{K^{-l_{\mathrm{x}}}}{p_{\mathrm{X}}(x)} - 1\right) \tag{3.22}$$

i.e

$$\sum_{\mathrm{x}} p_{\mathrm{X}}(x) log\left(\frac{K^{-l_{\mathrm{x}}}}{p_{\mathrm{X}}(x)} - 1\right) \;=\; \sum_{\mathrm{x}} K^{-l_x} - \sum_{\mathrm{x}} p_{\mathrm{X}}(x). \tag{3.23}$$

As the code is being built and demonstrated on guarantee bases that it is only decodable and yet, we show that it satisfies the Kraft inequality and, therefore, we have to:

$$\begin{aligned} H(x) - \bar{l}(x)\ log(K) \;&\leq\; 1 - 1 \\ H(x) - \bar{l}(x)\ log(K) \;&\leq\; 0 && (3.24) \end{aligned}$$

and therefore

$$H(x) \leq \bar{l}(x). \tag{3.25}$$

□

Chapter 4

Channel Capacity

Let's start with a question that will guide us throughout the chapter: Is there a way to encode information for a channel that has noise, or in other words, for a noisy channel while still maintaining a good communication rate? The answer to this question is given by Shannon's second coding theorem[1]. However, we need to better understand what a communication is, that is, what is Alice sending a message to Bob or a more general case, what is communicating?

4.1 Brief Introduction to Communication

Communication derives from the Latin term *″comunicare‶*, which means sharing or making something common. Through the act of communicating, there is an instantaneous sharing of information making this fact the main part in communication.

Which means sharing or making something common. Through the act of communicating, there is an instantaneous sharing of information making this fact the main part in communication.

4.1.1 How does communication work?

Alice wants to communicate with Bob and for that she will use a channel. To put it better, we can say that ″A's physical acts induced a physical state in B‶. This communication takes place through a physical means and therefore, it is not totally free from interference, or as we will call it here, from noise. When Bob signals to Alice that he received the message and was able to understand it, it is a sign that communication was successful.

Our efforts turned to the most fundamental question, which is: Since I can send information through this channel, can I send as much as I think is necessary at once? The answer is no and as

[1] Here it is worth remembering that Shannon left researchers in the field of communication completely scared, because in 1948 it was not usual for engineers to use such a mathematical tool.

we will see later, the characterization of the capacity of a channel is directly with the logarithm of the number of distinguishable signals, as being the maximum of mutual information or the maximum of correlation.

4.1.2 Communication over a channel

Definition 4.1.1. *We define a discrete channel to be a system consisting of an input alphabet X and output alphabet Y and a probability transition matrix $p(y|x)$ that expresses the probability of observing the output symbol y given that we send the symbol x The channel is said to be memoryless if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs.*

Definition 4.1.2. *We define the ″ information ‶channel capacity of a discrete memoryless channel as :*

$$C = \max_{p(x)} I(X;Y). \tag{4.1}$$

As Shannon defines that the capacity of a channel to transmit information is equal to the operational capacity C_{op} of that channel, however the term channel capacity is more commonly used.

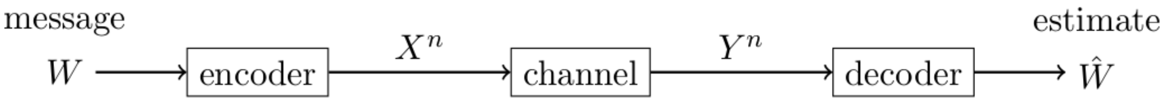

Figure 4.1: Communication scheme on a channel.

The message here denoted by $W \in \{1, 2, 3, \ldots, M\}$ it is one of a total of possible numbers or symbols that will be used for communication. As in the case of Data compression, an $i.i.d.$ source is used for all possibilities.

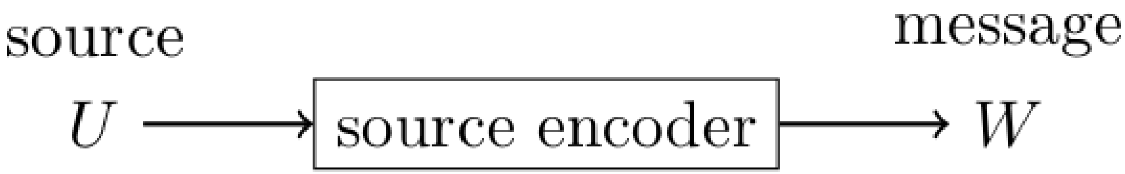

Figure 4.2: Graphical representation of the path that the information takes from the source to the entrance of the channel.

The figure above represents the *encoding scheme* (M, n) which is basically an encoder $\mathcal{E}$ - that receives the message from the source and encodes it - in which it maps the message W to a sequence of size n of channel entries X^n, as follows:

$$\mathcal{E} \ : \ \{1, 2, 3, \ldots, M\} \implies \mathcal{X}^n \tag{4.2}$$

and then, because it is encrypted, it is necessary to decode the message. For this, a decoder maps the sequences that were sent by the channel, of size n that have Y^noutputs in which is an estimate for the $\hat{W}$ message.

$$\mathcal{D} \ : Y^n \implies \{1, 2, 3, \ldots, M\}\,. \tag{4.3}$$

In other words, the channel is able to specify the transformations that were necessary for inputs to outputs, as follows:

$$\mathcal{P}\left[Y^n = y^n | X^n = x^n\right] = p_{Y^n|X^n}\left(y^n|x^n\right). \tag{4.4}$$

Another important factor is that the *channel has no memory*, that is, it does not retain any remnants of previous communications, so that if the outputs between the uses of the channel are completely conditionally independent, given an $i.i.d.$ input, we have:

$$p_{Y^n|X^n}\left(y^n|x^n\right) = \prod_{i=1}^{n} p_{Y|X}\left(y_i|x_i\right) \tag{4.5}$$

We will define at the rate R of a scheme (M, n) as follows:

$$R = \frac{\log_2 M}{n} \tag{4.6}$$

which denotes the bit rate per transmission.

Starting from 4.6, we can also define the number of messages for a given rate R with block size n denoted by:

$$M = 2^{nR} \tag{4.7}$$

It is worth mentioning that, on some occasions it is necessary to specify a rate code R for $\left(2^{nR}, n\right)$ instead of (m, n).

Definition 4.1.3. *The capacity of a channel is the supreme of all achievable rates, that is, it is the smallest of the majorities. Thus, rates lower than capacity generate an arbitrarily small probability of error for sufficiently large block lengths.*

4.2 The second Shannon Theorem: Channel Coding.

There are countless channels through which a communication can be made, however, the case we will be dealing with, a discrete channel with no memory will be used (*DMC*), which is specified by an input alphabet $\mathcal{X}$, as it denotes the sets of symbols that the channel accepts as input and by an output $\mathcal{Y}$ alphabet that denotes the set of symbols that the channel can produce as a result and lastly, it has a distribution conditional probability so that:

$$P_{Y|X}(\cdot|x) \, \forall \, x \in \mathcal{X} \tag{4.8}$$

Such that all outputs are independent of the input, which rewritten

$$P_{Y^n|X^n}\left(y^n|x^n\right) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i). \tag{4.9}$$

The information capacity or in other words, the transmission of the information in a discrete channel without memory is defined as 4.1 that can be rewritten as

$$C = \max_{p(x)} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x)p(y|x)log\left(\frac{p(y|x)}{\sum_{x'} p(y|x')p(x')}\right). \tag{4.10}$$

It is worth mentioning that $I(X;Y)$ is concava in $p_{\mathrm{x}}(x)$, that is, it is possible to find a distribution that maximizes it.

Below, we present the channel coding theorem, or also known as Shannon's second theorem.

Theorem 4.2.1. *The operational capacity C_{op} of a discrete channel without memory is equal to the information capacity:*

$$C_{op} = \max_{p(x)} I(X;Y) \tag{4.11}$$

so that from this equality two statements follow:

- ***Achievability:*** *Every rate $R < C$ is achievable i.e., there exists a sequence of $(2^{nr}, n)$ coding schemes such that the maximum error probability $P^n_{e,max}$ converges to zero as the block-length n increases:*

$$R < C \implies R \textit{ is achievable};$$

- ***Converse:*** *Any sequence of $(2^{nr}, n)$ coding schemes with maximum error probability $P^n_{e,max}$ converging to zero as the block-length n increase must have rate*

$$R \textit{ is achievable} \implies R \leq C.$$

To demonstrate it will be necessary to help the following lemma:

Lemma 4.2.1. *For any input distribution $p_{X^n}(x^n)$, the mutual information between the input X^n and output Y^n f a discrete memoryless channel with capacity C obeys*

$$I(X^n; Y^n) \leq nC. \tag{4.12}$$

Proof. For any input distribution on X^n and observe that

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n) \tag{4.13}$$

that by the chain rule, we have to

$$I(X^n; Y^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_1^{i-1}, X^n). \tag{4.14}$$

Because the channel has no memory, you can write being

$$I(X^n; Y^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{4.15}$$

and that by the chain rule again

$$I(X^n; Y^n) = H(Y_i|Y_1^{i-1}) - \sum_{i=1}^{n} H(Y_i|X_i). \tag{4.16}$$

Starting from the definition that conditions entropy, there cannot be an increase in entropy, we have

$$\begin{aligned} I(X^n;Y^n) &\leq H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i). &(4.17)\\ I(X^n;Y^n) &\leq I(X_i;Y_i) &(4.18)\end{aligned}$$

and then we come to the definition

$$I(X^n;Y^n) \leq nC. \tag{4.19}$$

□

Now we would like to show that $R \leq C$ for any sequence of channel codes with vanishing probability of error, i.e, $P^{(e)} \to 0$ as $n \to \infty$. For this, we need Fano's inequality to prove the following lemma

Lemma 4.2.2. *Let $\hat{W}$ be an estimate of the message $W \in \{1, 2, 3, \ldots, 2^{nR}\}$. The (n) conditional entropy of W given $\hat{W}$ is related to the average error probability $P^{(e)} = \boldsymbol{P}(W \neq \hat{W})$ via the inequality*

$$H(W|\hat{W}) \leq 1 + P^{(e)}nR. \tag{4.20}$$

Proof.

$$\begin{aligned} P^{(e)} &= \mathbf{P}(W \neq \hat{W}) &(4.21)\\ P^{(e)} &\geq \frac{H(W|\hat{W}) - 1}{log\,(2^{nR})} &(4.22)\\ P^{(e)} &= \frac{H(W|\hat{W}) - 1}{(nR)}. &(4.23)\end{aligned}$$

□

Now, with these tools, we can prove theorem (4.2.1).

Proof. Starting with the fact that W is uniformly distributed and $i.i.d.$ on $\{1, \ldots, 2^{nR}\}$, we have

$$H(W) = nR \Leftrightarrow nR = H(W) \tag{4.24}$$

$$nR = H(W|\hat{W}) + I(W;\hat{W}) \tag{4.25}$$

that by Fano's Inequality we have:

$$nR \leq 1 + P^{(e)}nR + I(W;\hat{W}) \tag{4.26}$$

and using what we have already defined above,

$$nR \leq 1 + P^{(e)}nR + I(X^n;Y^n) \tag{4.27}$$

and at (4.2.1) we find

$$nR \leq 1 + P^{(e)}nR + nC \tag{4.28}$$

and dividing both sides by n and rearranging leads to

$$P^{(e)} \geq 1 - \frac{C}{R} - \frac{1}{nR}. \tag{4.29}$$

If $R > C$ then the right-hand side is strictly positive for large values of n and thus R is not achievable. We have shown that

$$R > C \implies R \text{ is not achievable} \tag{4.30}$$

which concludes the proof of the converse to the coding theorem. □

It is worth mentioning that the proof we use is given in the context of the so-called weak inverse, since the probability of the error is limited to zero, however, it does not rule out the probability that there is a small error. However, using more advanced techniques as the proof via random coding it is possible to show that if $R > C$, then the probability that there is an error converges to 1.

In summary, we saw that it is possible to reconstruct the input sequences in the output with an insignificant probability of err, that is, by mapping the source in input sequences with the appropriate spacing for the appropriate channel, we can transmit a message, or more generally, carry out a communication with a very low probability for the occurrence of an error.

Chapter 5

Conclusions and perspectives

The text presented the postulates of the classical communication theory that rests on the pillars of data compression and channel capacity, as well as the tools and definitions of the theory. It also shows that information has a limit to compression, which is entropy, and that communication over a channel has its limit defined by channel capacity.

This work provides the minimum basis to deepen the studies of information theory as well as other areas that are in its core, because there are applications in probability and statistics: as theorem limits and hypothesis testing; in economics: applying its concepts to the portfolio theory; in computer science: as the Kolmogorov complexity, in physics with the quantum generalization of the theorems shown and also in thermodynamics, in medicine with neuroscience among others.

Bibliography

[1] S. Herner, "Brief history of information science," *Journal of the American Society for Information Science*, vol. 35, no. 3, pp. 157–163, 1984.

[2] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[3] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[4] M. M. Wilde, *Quantum information theory*. Cambridge University Press, 2013.

Appendices

Appendix A

A Brief Introduction to Probability

Working with information theory requires knowledge of some areas of mathematics, especially probability theory along with statistics. This includes aspects of conditioning and expectation, discrete and continuous variables as well as some familiarity with multivariate Gaussian distributions, Markov chains - this in particular is sometimes covered in the course of statistical mechanics in some undergraduate programs - and convergence of random variables.

Under this assumption, some fundamental concepts will be presented for the understanding of the text.

1.1 Probability Space

Let Ω be the sample space of all possible outcomes. Let $\mathcal{F}$ be the space of events defined on the sample space. We then have that $\mathbb{P}$ is a probability measure that satisfies the following

$$\mathbb{P}[\Omega] = 1 \quad \text{and} \quad \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] \quad \forall A, B \in \mathcal{F} \; \text{ with A} \cap B = \emptyset.$$

1.2 Notation

Random variables are denoted by capital letters, like: X, Y, Z. When these are deterministic values, that is, not random values, they will be denoted by lower case letters. The alphabet or symbol support is of a random variable is denoted by the calligraphic symbols given by $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$

1.3 Discrete Random Variables

The **probability mass function** or **(pmf)** of a discrete random variable is X, Y with alphabet or symbol support $\mathcal{X}, \mathcal{Y}$ is given by

$$p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y] \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

To simplify notation, it is common to write $p(x, y)$ where the association with the pair (X, Y) is implied by the argument of the function.

1.4 Independence

Events A and B are independent if and only if their joint probability equals the product of their probabilities

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

Random variables X and Y are independent if and only if the joint probability is equal to the product of the marginals

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

1.5 Expectation

The expected value of a random variable X is given by:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x).$$

The expected value of a function $f : \mathcal{X} \to \mathbb{R}$ is given by:

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x) p_X(x).$$

Here it is worth noting that the expectation $\mathbb{E}[f(X)]$ not a random variable. Instead, it is a functional of the distribution of X. Sometimes it is written as $\mathbb{E}_{p_X}[f]$ to make this relationship clear. The conditional expectation of X given a particular realization $\{Y = y\}$ is:

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x \mid y).$$

This is a deterministic, that is, non-random quantity that is a function of y.

1.6 Variance

The variance of a random variable X is given by

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2,$$

and the conditional variance is the variance of $X \sim p_{X|Y}(x \mid y)$,

$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y = y])^2 \mid Y = y\right].$$

1.7 Law of large numbers (LLN)

If a sequence of random variables $X_1, X_2, \ldots$ is i.i.d. with finite absolute first moment $\mathbb{E}\left[|X_1|\right] < \infty$, then the long-term average converges to the mean:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X]$$

with probability one as $n \to \infty$.

Appendix B

Why use the logarithm?

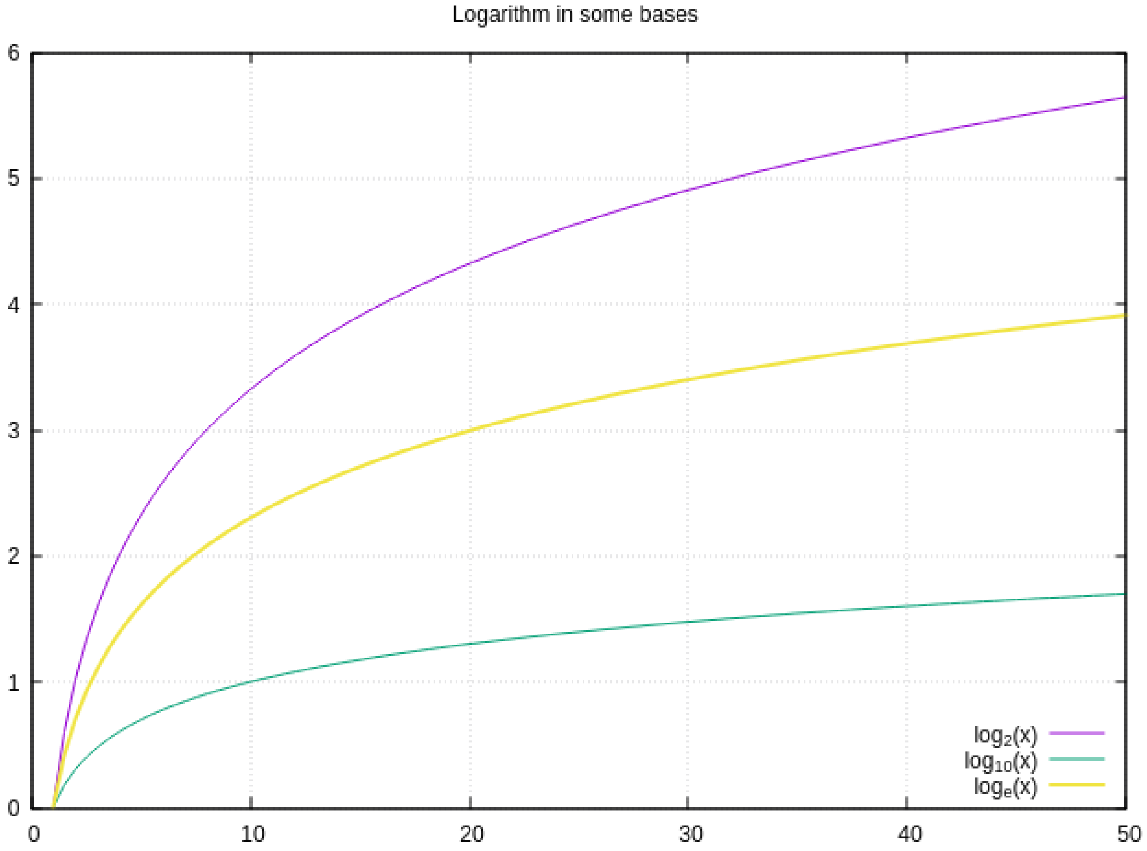

Figure 2.1: Representation of some functions.

Well, the first explanation we can give is that the logarithm function is increasing of its argument, so $\log(x)$ grows with (x).

It is worth noting that the base of the logarithm changes the value of the function, but the behavior is basically the same. The fundamental issue is that $\log(1) = 0$ is compatible with the notion of information that is associated with an event. An event i whose associated probability of occurrence is $p_i = 1$ does not carry any information with it. By the definition that Shannon gives us, the amount of information associated with an event is $I_i = \log\left(\frac{1}{1}\right) = \log(1) = 0$.

However, an event whose probability of occurrence is very small has a large amount of information associated with it, since $\frac{1}{p_i}$ grows as pi decreases and therefore $\log 1/p_i$ grows as p_i decreases.

Another explanation for the use of the logarithm function is

$$\log(ab) = \log a + \log b$$

that is, the logarithm of the product is equal to the sum of the logarithms. This property is important because it is based on it that we can calculate the amount of information associated with the occurrence of two events a and b . When we have two sources that emit signals with a given sequence and at a given time, we have that the signals we observe from the two sources are a and b who have an associated amount of information denoted by I_a and I_b. Therefore, the amount of information associated with the observation of the two signals is that it is equal to the sums of the amounts of information associated with events a and b individually, provided that a and b are independent and that is exactly what the logarithmic function presents

$$I_{ab} = \log\left(\frac{1}{p_{ab}}\right) = \log\left(\frac{1}{p_a}\frac{1}{p_b}\right) = \log\left(\frac{1}{p_a}\right) + \log\left(\frac{1}{p_b}\right) = I_a + I_b$$

We can then state that the amount of information associated with an event is an *additive measure*.

Therefore, we can use the logarithmic function in any base to calculate the amount of information, as long as we use the same base for all calculations. The base used determines the unit in which the amount of information is measured. If you use base 2, the unit is the bit. If you use base 10, the unit is not given a special name, so the generalization applies to everyone.