Computation and Information

Before quantum computers

Contents

Foreword 5

I Computation

II Information

1 Measures of information 11

1.1 Surprising information 11

1.2 Refresher on probability theory 12

1.3 Axiomatic derivation of entropy 14

1.4 Entropy 15

1.5 Joint entropy, conditional entropy and mutual information 16

1.6 Further reading 19

2 Source coding 21

2.1 Transmitting the weather forecast 21

2.2 Formal problem 22

2.2.1 Data Compression 22

2.3 Huffman codes 24

2.4 Typical sequences 24

2.5 Shannon's theorem for source coding 24

2.6 Further reading 24

3 Channel coding 25
3.1 The communications problem 25
3.2 Linear codes 26
3.3 Hamming codes 26
3.4 Detection, correction and minimum distance 26
3.5 Random codes 26
3.6 Shannon's second theorem 26
3.7 The capacity of some basic channels 28
3.7.1 Binary Symmetric Channel 28
3.7.2 Binary Erasure Channel 28
3.8 Further reading 29
3.9 Exam problems 29
Bibliography 31
Index 33

Foreword

We are in the middle of the so-called informatoin society. Most of our interactions are mediated in one or another way by communication and computation devices which also majoritarily are connected to the internet. This information society facilitates many of our tasks, from looking for a neearby supermarket to learning about the best treatment for whatever illness we have or think we have (don't). It guides our choice of job, our application for job positions and allows us to communicate with distant relatives and also with friends that might be a few meters away from us. All this wealth of possibilities capable of satisfying our pst needs and drivign force of needs we did ot have a few years ago, relies on mathematical tools and models that did not exist until very recently. In fact some of these tools were thought to be impossible until they wre proved wrong. Even while part of the information society, we are also in the middle of a revolution, that is taking us from the world of classical information to a new world were information follows the strange laws of quantum mechanics. We expect much of this new world, amazing speed ups in computation, that will allow to discover new drugs, produce unhackable communication systems or improve our observations of the distant universe. However, what does it mean to speed up, what are the limits of the classical world? In which sense are some problems untractable by the computers we know today? Sure, it is only a matter of buying enough computation power and then we could also achieve the feats mentioned before, or perhaps is there some catch? It turns out, that there are problems that no computation device can solve, problems that can be computed but are beyond the reach of quantum computers. In fact, most of the problems that classical computers can deal with, can also be dealt with by quantum computers with no speed up. It is only for a limited number, of highly structured problems that quantum ocmputers promise to offer amazing improvement. However, still for those problems the speed up is in terms of how the computation resources necessary grow with the size of the problem. What is even information?

Computation

II Information

1 **Measures of information** 11

1.1 Surprising information
1.2 Refresher on probability theory
1.3 Axiomatic derivation of entropy
1.4 Entropy
1.5 Joint entropy, conditional entropy and mutual information
1.6 Further reading

2 **Source coding** 21

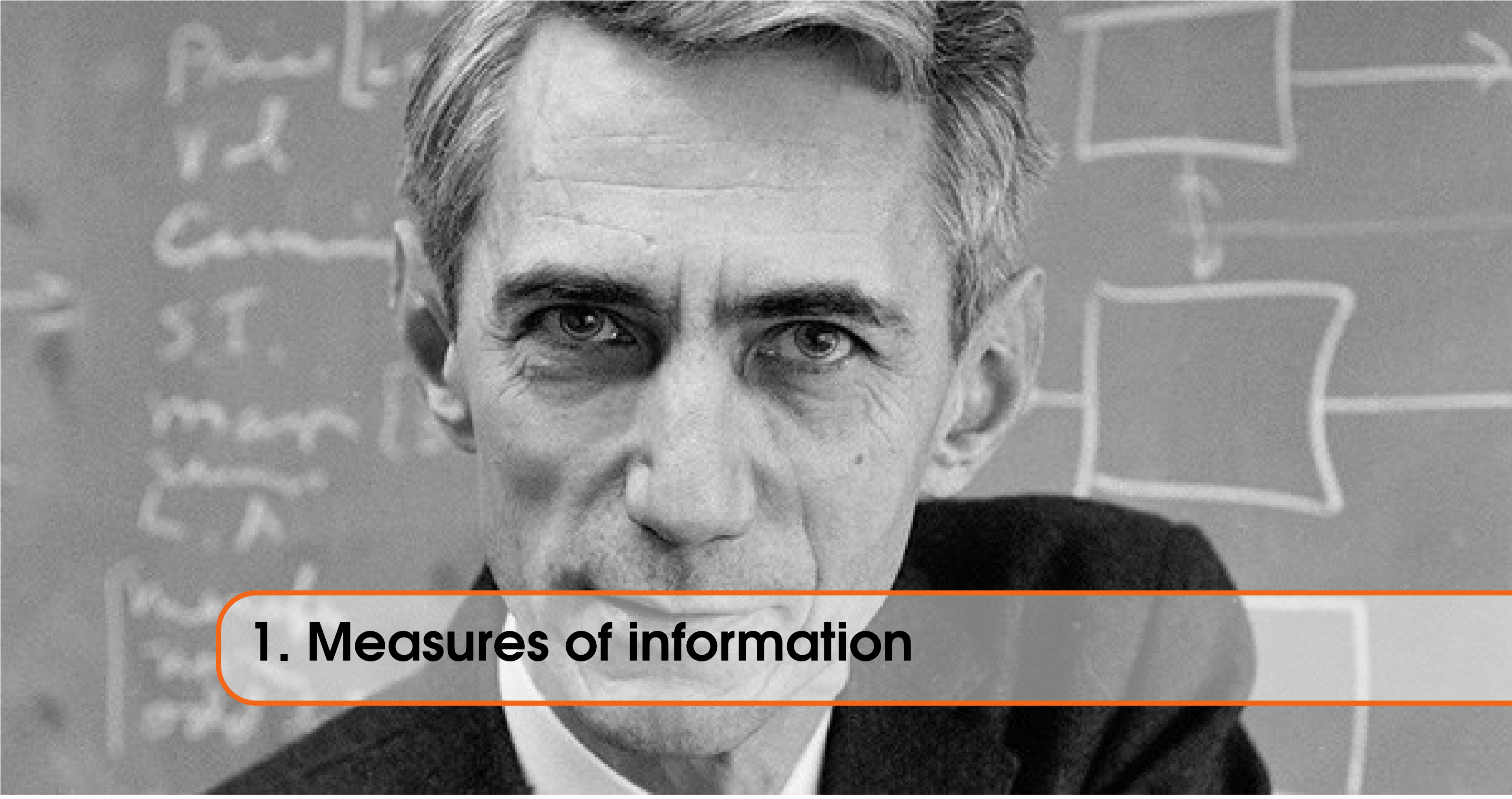
2.1 Transmitting the weather forecast
2.2 Formal problem
2.3 Huffman codes
2.4 Typical sequences
2.5 Shannon's theorem for source coding
2.6 Further reading

3 **Channel coding** 25

3.1 The communications problem
3.2 Linear codes
3.3 Hamming codes
3.4 Detection, correction and minimum distance
3.5 Random codes
3.6 Shannon's second theorem
3.7 The capacity of some basic channels
3.8 Further reading
3.9 Exam problems

Bibliography 31

Index 33



1. Measures of information

This is the age of information. We open documents, consume media, exchange text messages, perform videoconferences or watch the weather forecast to name a few examples. But what is exactly information? Can we quantify it? Is it possible to say that one weather forecast contains more information than another? In this chapter, we will learn that this is indeed possible. Our exposition, follows to a great extent the visionary paper of Claude Shannon in 194X.

1.1 Surprising information

Let us warm up with a series of questions. The premise of all of them is the following: For some reason you want to know the weather forecast for tomorrow, but don't want to watch it or read it yourself. However, in order to get some information that allows you to choose appropriately your clothing for the next day, you ask a friend to send you a message to your phone every night with a summary of the forecast. You agree on a very simple encoding for the message, your friend will send a 1 if the forecast predicts precipitations and 0 otherwise. Think about these questions and come back to them after reading the whole chapter.

Location	Days of rain	Days with no rain
Rotterdam[1]	153	212
Atacama desert	5	360

Table 1.1: Summary of precipitations in the year 2018.

Exercise 1.1 Let us assume that you are living in the Atacama desert where it rarely rains and you receive a 0. How much information does this message carry?

Exercise 1.2 Now let us assume that you live in the Netherlands where it does rain quite often, but certainly not every day. You also receive a 0, does the message contain information? What about if you receive a 1? Does the 1 message contain more or less information than the message 0? ■

Exercise 1.3 Finally, let us assume that you live in the Netherlands and you happen to know that it is mid August. Does the message 0 carry the same information as in winter? ■

Before you continue reading, pause for a moment and think what is common in your answers.

We have posed these questions to suggest a relation between the amount of information a message provides and how surprising it is. We will make this connection stronger in the rest of these chapter.

1.2 Refresher on probability theory

A basic understanding of probability theory is essential for the material that follows. Let us review the fundamental concepts and definitions together with the notation that we will use here. As we will only deal with discrete probability distributions, the definitions that follow are not fully general but sufficient for our purposes. If you have troubles following this section and doing the exercises please go back to your undergraduate text on the topic.

Given a finite set $\mathscr{X}$, we call a probability distribution a function $p : \mathscr{X} \to [0,1]$, that is a function from the elements of $\mathscr{X}$ to the closed interval in the real line between zero and one, with the condition that $\sum_{x\in\mathscr{X}} p(x) = 1$. Note that it follows automatically from our definition that for all $x \in \mathscr{X}$ $p(x) \geq 0$.

■ **Example 1.1** Let $\mathscr{X} = \{\text{tails},\text{heads}\}$ we could define a probability distribution function p such that $p(\text{heads}) = 0.3$ and $p(\text{tails}) = 0.7$. ■

■ **Example 1.2** An important example is the uniform distribution. Given a finite set $\mathscr{X}$, a uniform distribution on the set $\mathscr{X}$ is a function p that for all $x \in \mathscr{X}$ assigns the value

$$p(x) = \frac{1}{|\mathscr{X}|} \, .$$

where, we denote by $|\cdot|$ the number of elements in the set. ■

We define an ensemble X as the tuple of a probability distribution p_X together with its domain $\mathscr{A}_X$. Generally, we will refer to $\mathscr{A}_X$ as the sample space of X and to its elements as events. In information theory, X typically models an object in a communications setup (see **??**), in this context it is common to call $\mathscr{A}_X$ the alphabet of X and refer to its elements as letters.

Note that we can extend the definition of p_X to any subset $\mathscr{S} \subseteq \mathscr{A}_X$:

$$p_X(\mathscr{S}) = \sum_{x\in\mathscr{S}} p_X(x) \tag{1.1}$$

■ **Example 1.3** Let X be an ensemble with alphabet $\mathscr{A}_X = \{1,2,3\}$ and with p_X the uniform distribution. Then if $\mathscr{S} = \{1,2\}$, $p(\mathscr{S}) = 1/3 + 1/3 = 2/3$ ■

Abusing notation, we will also call event any subset of the alphabet of an ensemble. For this particular case of set we will drop the caligraphic notation for sets. Let a and b be two events in $\mathscr{X}$, we define $a \cup b$ and $a \cap b$ as the union and intersection of a and b. $a \cup b$ is the event that contains all outcomes belonging to a, to b and to both, we will also denote the event $a \cup b$ by a or b. $a \cap b$ is the event that contains all outcomes belonging to both a and b, we will also denote the event $a \cap b$ by a and b. Two events are disjoint if their intersection is null.

Given an ensemble X and two events a, b we say that they are independent if:

$$p_X(a \text{ and } b) = p_X(a)p_X(b) \tag{1.2}$$

Let a and b be two events with non zero probability. We call $p_X(a|b) = p_X(a \text{ and } b)/p_X(b)$ the conditional probability of a given that b occurs. It follows that if and only if a and b are independent $p_X(a|b) = p_X(a)$.

In the following, we will use the explicit notation $p_X, \mathscr{A}_X$ for the probability distribution of ensemble X and its alphabet whenever confusion can arise but we will drop the subscript whenever possible.

Exercise 1.4 Let X be an ensemble modelling two fair coins. Identify two events a, b that are independent and verify that $p_X(a|b) = p_X(a)$ ■

Solution. A ■

Given two alphabets $\mathscr{A}_X, \mathscr{A}_Y$ we can define a joint ensemble on them with sample space or alphabet the direct product: $\mathscr{A}_{XY} = \mathscr{A}_X \times \mathscr{A}_Y$. We can associate, as well, a probability distribution function to map all tuples (x, y) to $[0, 1]$.

The probability of an event in the joint ensemble is equally defined as the sum of the probability of the individual events. In particular, we can define for every $x \in \mathscr{X}$ the probability of $p_X(x)$ as the sum of $p_{XY}(x, y)$ for all $y \in \mathscr{Y}$:

$$p_X(x) = \sum_y p_{XY}(x, y) \tag{1.3}$$

and equivalently $p_Y(y)$:

$$p_Y(y) = \sum_x p_{XY}(x, y) \tag{1.4}$$

■ **Example 1.4** Consider n repetitions of an experiment, each repetition can be modelled by ensemble X and events in different experiments are independent. We can model the set of n repetitions via the joint ensemble $X_1 \dots X_N$, where X_i is the ensemble associated with the i-th experiment, and joint the probability distribution is given by:

$$p_{X_1 \dots X_N}(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} p_X(x_i) \tag{1.5}$$

■

A random variable V on the ensemble X is a numerical function from the elements of $\mathscr{A}_X$ to (typically) the real line. That is, a function $V : \mathscr{A}_X \to \mathscr{A}_V$, where $\mathscr{A}_V$ is a finite subset of the reals. The random variable V induces an ensemble with alphabet $\mathscr{A}_V$ and probability distribution p_V where p_V is given by:

$$p_V(v) = \sum_{x \in \mathscr{A}_X : V(x) = v} p_X(x) \tag{1.6}$$

for all $v \in \mathscr{A}_V$.

The mean or expectation of a random variable is given by:

$$\mathbb{E}[V] = \sum_{x \in \mathscr{A}_X} p_X(x) V(x) = \sum_{v \in \mathscr{A}_V} p_V v \tag{1.7}$$

1.3 Axiomatic derivation of entropy

Let us now try to understand what type of functions can quantify information in a satisfactory way. Let us make this investigation more precise. In particular, suppose that given some ensemble X we observe the occurrence of an event $x \in \mathscr{A}_X$. As we informally argued in the introduction, the information we gain seems to be related to the likelihood of the event we observed. But how can we make this intuition quantitative?

A function that quantifies information will be a function from a subset of $\mathscr{A}_X$ to the reals. Let us call this function h. Then given some event x, $h(x)$ will be some number that will quantify the information we learn. Let us discuss what properties an ideal information quantifier should have.

- The measure should be non-negative, that is, an event gives either none or some information, but it can not give negative information. That is, for all events $x \in \mathscr{A}_X$ we require:

$$h(x) \geq 0 \tag{1.8}$$

- Suppose that we buy two lottery tickets in two different lottery games, event x is: "our first ticket wins a prize", event y is: "our second ticket does not win a prize". We expect these two events to be independent and the information content of knowing both events should be the sum of the information of the individual events. The occurrence of two independent events should yield the same information that the occurrence of the single events would provide an observer. If we let h be an information measuring function

$$p_X(x \text{ and } y) = p_X(x)p_X(y) \Rightarrow h(x \text{ and } y) = h(x) + h(y) \tag{1.9}$$

- Following our discussion about information and surprise, we want h to quantify less probable events with a larger value than more probable events. For any two ensembles X, Y and events $x \in \mathscr{A}_X$ and $y \in \mathscr{A}_Y$, we require:

$$p_X(x) < p_Y(y) \Rightarrow h(x) > h(y) \tag{1.10}$$

- The final condition is that we don't want that arbitrarily small changes in probability lead to a change in the information quantity, i.e. h should be a continuous function.

It turns out that there is a very limited set of functions that verify these properties. Given some ensemble X, the unique family of functions is of the form:

$$h(x) = -\log_\lambda p_X(x) \tag{1.11}$$

where $x \in \mathscr{A}_X$ and with $\lambda > 1$ for the measure to be positive. Choosing different values of λ allows us to measure information with different units.

There are some common choices of λ that give rise to well known units of information: if we let $\lambda = 2$, the unit of information is called bit. When $\lambda = 3$ information is measured in trits, for $\lambda = 10$ the unit is called a digit and when $\lambda = e$ nat. Unless stated otherwise, in the following we will assume that $\lambda = 2$ and will let $\log = \log_2$.

Definition 1.3.1 Given an ensemble X the information measured in bits of an event $S \subset \mathscr{A}_X$ is given by:

$$h(\mathscr{S}) = -\log p_X(\mathscr{S}) \tag{1.12}$$

Exercise 1.5 Let X be an ensemble modelling a fair coin, that is with alphabet $\mathscr{A}_X = \{\text{heads}, \text{tails}\}$ and with p_X the uniform distribution. What is the information of the event heads and of the event tails? ■

Solution. As p_X is uniform, we have that $p(\text{heads}) = p(\text{tails}) = 1/2$. Hence:

$$h(\text{heads}) = -\log(1/2) = 1 \text{ bit}$$

and

$$h(\text{heads}) = -\log(1/2) = 1 \text{ bit} .$$

■

Let us end this section by checking that all our desired conditions hold. First since the log function is continuous and monotonically increasing in the range $(0,1]$ it holds that h is also continuous and monotonically decreasing in the range. Finally, if two events a,b are independent, $p(a \text{ and } b) = p(a)p(b)$ and in consequence

$$\begin{aligned} h(a \text{ and } b) &= -\log\left(p(a \text{ and } b)\right) && (1.13)\\ &= -\log\left(p(a)p(b)\right) && (1.14)\\ &= -\log(p(a)) - \log(p(b)) && (1.15)\\ &= h(a) + h(b) && (1.16) \end{aligned}$$

1.4 Entropy

We define the entropy of an ensemble as the average information content it provides:

Definition 1.4.1 Let X be an ensemble, the entropy of the ensemble is defined as:

$$H(\mathbf{X}) = -\sum_x p(x) \log p(x) \tag{1.17}$$

where we take the convention that $0 \log 0 = 0$, i.e. adding a zero-probability event to a source does not affect its entropy.

We can rewrite the definition of entropy as the expectation of the random variable $h(X)$. That is a random variable that associated each event with the negative logarithm of its probability:

$$H(\mathbf{X}) = -\sum_x p(x) \log p(x) = E(-\log p(\mathbf{X})) \tag{1.18}$$

Note that entropy only depends on the values of the probabilities. In the following we will sometimes be interested in the entropy a probability distribution independently of an ensemble. We will use the notation $H(p_1, \ldots, p_n)$ to indicate the probability distribution. Let us now investigate some basic properties of entropy that we will use through this course.

Exercise 1.6 Show that entropy can not be negative.

$$H(\mathbf{X}) \geq 0$$

■

Solution.

$$0 \leq p(x) \leq 1 \Rightarrow -\log p(x) \geq 0 \Rightarrow H(X) \geq 0 \tag{1.19}$$

■

The following is known as Jensen's inequality and will be of use in the following. See [3] for a proof.

Theorem 1.4.1 — Jensen's inequality. Let f be a concave function and X a random variable. Then:

$$f(E(\mathbf{X})) \geq E(f(\mathbf{X}))$$

Lemma 1.4.2 The distribution that maximizes entropy for any alphabet is the uniform distribution.

$$H(p_1,...,p_n) \leq \log n$$

Proof.

$$\begin{aligned}
H(p_1,...,p_n) - \log n &= \sum_{i=1}^{n} p_i \log\frac{1}{p_i} - \sum_{i=1}^{n}\frac{1}{n}\log n \\
&= \sum_{i=1}^{n} p_i \log\frac{1}{p_i} - \log n\sum_{i=1}^{n}\frac{1}{n} \\
&= \sum_{i=1}^{n} p_i \log\frac{1}{p_i} - \log n\sum_{i=1}^{n} p_i \\
&= \sum_{i=1}^{n} p_i \log\frac{1}{p_i} - \sum_{i=1}^{n} p_i \log n \\
&= \sum_{i=1}^{n} p_i \log\frac{1}{np_i} \\
&\leq \log\sum_{i=1}^{n}\frac{1}{n} = 0
\end{aligned} \tag{1.20}$$

where the second equality follows from the fact that a probability distribution adds up to one and the last inequality holds from log being a concave function and applying Jensen's inequality. ■

1.5 Joint entropy, conditional entropy and mutual information

We will now explore three information measures that derive from entropy as we defined it in the previous section. The first measure is joint entropy, which is a direct application of the definition of entropy to a joint source.

Definition 1.5.1 Given two ensembles **X** and **Y** the entropy of the joint ensemble XY is given by:

$$H(\mathbf{XY}) = -\sum_{x,y} p(x,y)\log p(x,y) \tag{1.21}$$

Exercise **??** suggests that the information content depends on the context. The second information measure that we introduce is conditional entropy. First, we can extend in a straightforward way the reasoning in Sec. **??** to define an information measure conditional on the knowledge of some event y. It can analogously be proved that a conditional information measure is of the form:

$$h(a|b) = -\log p(a|b) \tag{1.22}$$

Let XY be a joint ensemble, we can define the conditional entropy of X given the event y as the average conditional information:

$$H(\mathbf{X}|y) = \sum_x p(x|y)h(x|y) \tag{1.23}$$

and the conditional entropy of X given ensemble Y:

$$H(\mathbf{X}|\mathbf{Y}) = \sum_y H(X|y) \tag{1.24}$$

Exercise 1.7 Show that $H(X|Y) = H(XY) - H(Y)$. ■

Let us investigate some basic properties of the conditional entropy.

Exercise 1.8 Show that the conditional entropy is non-negative.

$$H(\mathbf{X}|\mathbf{Y}) \geq 0$$

■

Solution. $H(\mathbf{X}|\mathbf{Y})$ is a sum of entropies, which are positive by Lem. 1.6, weighed by the probabilities of each event which are also positive. ■

Exercise 1.9 Show that the entropy of the random variable $\mathbf{X}$ given any random variable $\mathbf{Y}$ is not greater than the entropy of $\mathbf{X}$.

$$H(\mathbf{X}|\mathbf{Y}) \leq H(\mathbf{X})$$

■

Solution.

$$\begin{aligned}
H(\mathbf{X}|\mathbf{Y}) - H(\mathbf{X}) &= \sum_y p(y) \sum_x p(x|y) \log \frac{1}{p(x|y)} - \sum_x p(x) \log \frac{1}{p(x)} \\
&= \sum_y \sum_x p(x,y) \log \frac{1}{p(x|y)} + \sum_{x,y} p(x,y) \log p(x) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)}{p(x|y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \\
&\leq \log \sum_{x,y} p(x)p(y) = 0
\end{aligned} \tag{1.25}$$

■

Exercise 1.10 Given random variables $\mathbf{X}$ and $\mathbf{Y}$ if $\mathbf{X} = f(\mathbf{Y})$:

$$H(\mathbf{X}|\mathbf{Y}) = 0$$

■

Solution. If $\mathbf{X} = f(\mathbf{Y})$, then given $\mathbf{Y}$ we know $\mathbf{X}$ with absolute certainty, in other words, given $\mathbf{Y}$ there is just one possible outcome.

$$\begin{aligned} H(\mathbf{X}|\mathbf{Y}) &= \sum_y p(y) H(\mathbf{X}|y) \\ &= 0 \end{aligned} \tag{1.26}$$

■

Exercise 1.11 Show that the following relation holds for any two ensembles XY:

$$H(\mathbf{XY}) = H(\mathbf{X}) + H(\mathbf{Y}|\mathbf{X})$$

■

Solution.

$$\begin{aligned} H(\mathbf{XY}) &= -\sum_{x,y} p(x,y) \log p(x,y) \\ &= -\sum_x p(x) \sum_y p(y|x) \log p(x) p(y|x) \\ &= -\sum_x p(x) \log p(x) \sum_y p(y|x) \\ &\quad -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\ &= H(\mathbf{X}) + H(\mathbf{Y}|\mathbf{X}) \end{aligned} \tag{1.27}$$

■

The third information measure that we introduce is the mutual information:

Definition 1.5.2 Given a joint ensemble XY, we define the mutual information between X and Y by:

$$I(X;Y) = H(X) + H(Y) - H(XY)$$

The mutual information $I(\mathbf{X};\mathbf{Y})$ is a measure of the information shared between the two variables $\mathbf{X}$ and $\mathbf{Y}$. Let us make this intuition more precise:

Exercise 1.12 Show that for any ensemble X: $I(X;X) = H(X)$. ■

Exercise 1.13 Show that $I(X;Y) = 0$ if and only if X and Y are independent. ■

Exercise 1.14 Show that $I(X;Y) \geq 0$ ■

Fig. 1.1 shows the relationship between the four measures that we have defined: entropy, joint entropy, conditional entropy and mutual information.

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ &= H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) \\ &= I(\mathbf{Y};\mathbf{X}) \end{aligned} \tag{1.28}$$

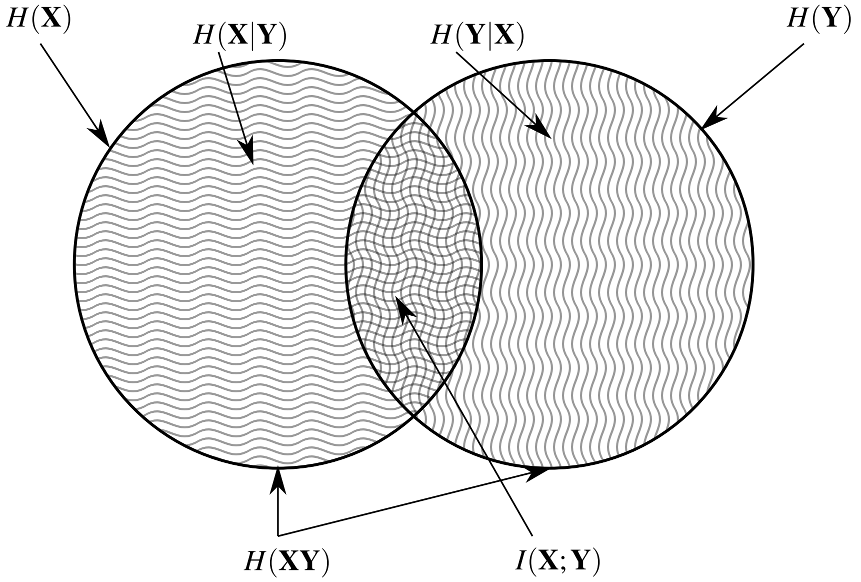
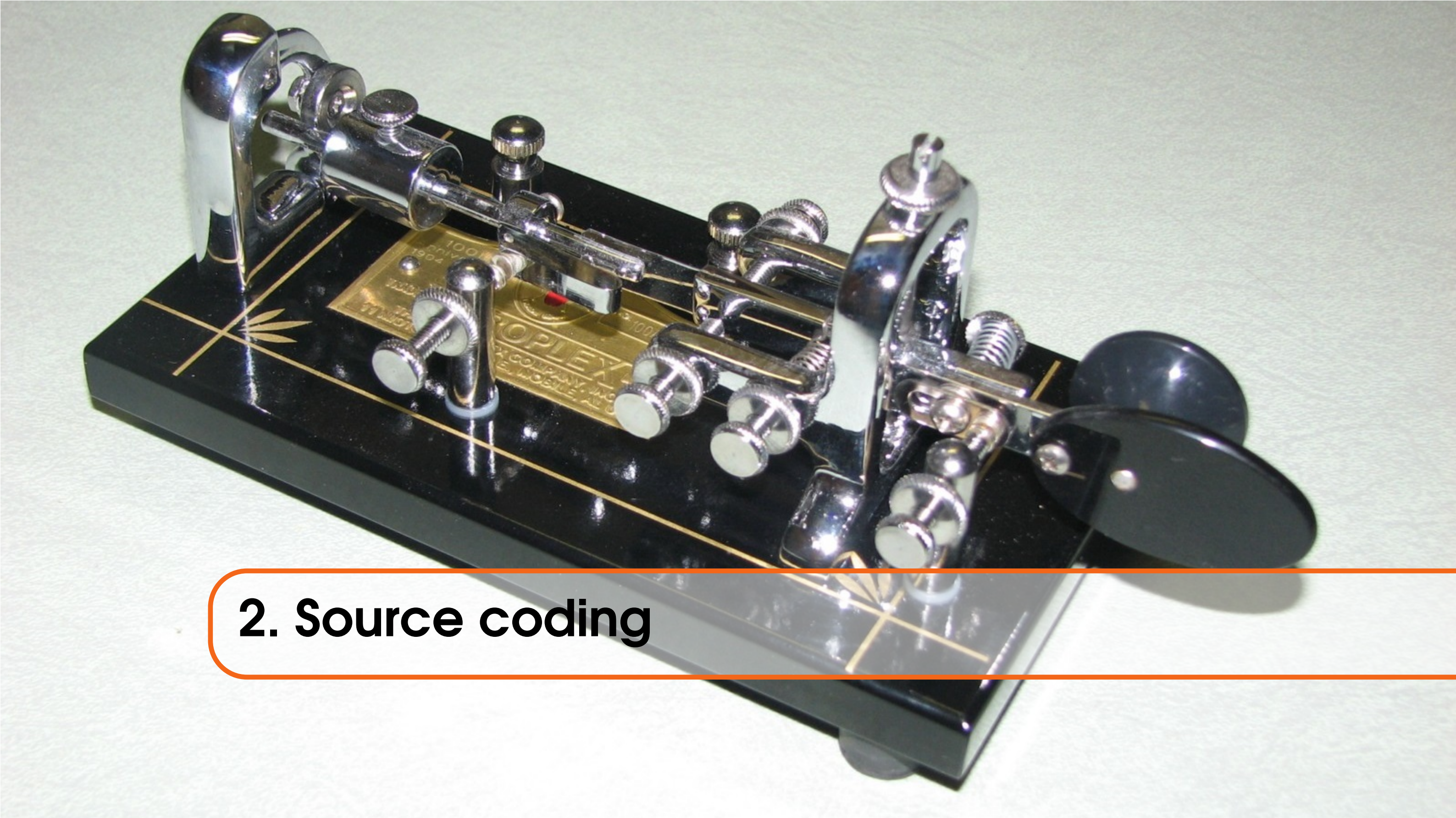

$H(\mathbf{X})$
$H(\mathbf{X}|\mathbf{Y})$
$H(\mathbf{Y}|\mathbf{X})$
$H(\mathbf{Y})$
$H(\mathbf{XY})$
$I(\mathbf{X};\mathbf{Y})$

Figure 1.1: Graphical representation of the information measures.

1.6 Further reading

The mathematical foundations of information theory were to a certain extent developped single handedly by Claude Shannon. His original paper [] developped the framework and also solved some of the most important problems. The text has not aged with time and remains a greatly written and accessible introduction to the field. A second excellent source for digging deeper into the material is the book of Cover and Thomas [], it is the reference of the field and widely used in most introductory courses on information theory. Chapters X and Y expand on the material covered in this chapter.

We will not prove it here, for a complete discussion on axiomatic derivations of entropy and information please refer to [1, 2, 4, 5].



2. Source coding

In the previous chapter we posed in an abstract way a series of conditions that information measures should possess. We built on top of those conditions and found a series of information measures satisfying them. Moreover, in a certain sense those measures are unique.

In this chapter we will begin a journey to show that not only entropy is a good measure for information, but also that it carries a strong operational meaning. In fact, we will show that matching our intuition, if a source has a certain entropy, i.e. it contains a certain amount of information. Then, then the length of the message that we need to send to some other party in order to transmit the source needs to be at least the entropy of the source.

We will also begin to investigate the difference between proofs of existance and constructive methods. This will be a common pattern in information theory, though of limited importance in our introductory course. In fact, it is in many circumstances, possible to prove that codes with desired properties must exist by relying on random ensembles of codes, while actually pointing an example is more complicated.

2.1 Transmitting the weather forecast

Let us first go back to our example from the previous chapter. It will give us enough material to formalize the discussion. Let us suppose again that we have a distant but commited friend that has agreed to send us a daily message summarizing the weather forecast. However, we want a little bit more information that in the previous chapter and now we want our friend to tell us whether tomorrow will be sunny, will snow, will rain or will GRANIZAR. Let us investigate different ways in which our friend can communicate this information to us:

A first simple way would be to send a message containing the appropriate word: "sun", "rain", "snow", "GRAN". However, since there are only four possible messages, we could instead just send a number: "0" for rain, "1" for sun, 2 for snow and 3 for GRAN.

Definition 2.1.1 Code

We could be more sophisticated. Let us assume that the four events occur with the following probability: it rains with probability $1/2$, it snows with probability $1/4$, and both sunny days and

GRAn days happen with probability $1/8$. We could assign the following words to each event: rain 0, snow 10, sun 110 and GRAN 111. That is, we have assigned more likely events shorter words? Why is this better?

Definition 2.1.2 Given a discrete random variable X taking events in the finite alphabet $\mathscr{X}$ and a code C from $\mathscr{X}$ to the code alphabet $\mathscr{X}$, we define the average length of the code as follows:

$$l(C) = \sum_{x\in\mathscr{X}} p(x)|C(x)| \tag{2.1}$$

where $|C(x)|$ is the length of the codeword of event x.

Exercise 2.1 With the definition of average legnth, we find that the first code taking each event to a two bit word, has average length 2 while the second code with variable length has average length 1.375 which is significantly shorter. ■

Can we do better than 1.375? We can indeed, consider the code that assigns the events rain and sun the word 0 and the events snow and GRAN the word 1. This code has average length 1. However, what is the problem? The problem is that if we receive the message 0, we can not know whether tomorrow is going to rain or to be sunny. The code is not-singular.

Definition 2.1.3 A code $C : \mathscr{X} \mapsto \mathscr{Y}$ is singular if $\forall x, y \in \mathscr{X}$ with $x \neq y$ $C(x) \neq C(y)$.

However, in order to recover the information content, it is not enough for a code to be singular. Consider the following code: rain 0, sun 1, gran 01 and snow 10. And suppose that our friend wants to send the forecast of two consecutive days. If we receive the message 010, we could decode it as 01 and 0 with the meaning first day gran second day rain or we could decode it as 0 and 10 with the meaning rain and snow. The code can not be uniquely decoded.

Definition 2.1.4 Uniquely decodable

In the example, it was fairly easy to spot that the code was not uniquely decodable. However, for some codes it is not so simple.

Exercise 2.2 Let $C = \{asdfsadf\}$ is C uniquely decodable? (Hint: if you can not find the solution, continue reading and return to this exercise once you have understood the Sardinas-Patterson method) ■

Let us now investigate a method to find if a code is uniquely decodable or not. This method was proposed by Sardinas and Patterson [] and is also known as the method of the dangling suffixes.

2.2 Formal problem

2.2.1 Data Compression

Once presented an information measure, we review some of its operational interpretations, in particular its relation with the theoretical limits for data compression.

A source code C in alphabet $\mathscr{Y}$ for a random variable taking values in $\mathscr{X}$ is a function $c : \mathscr{X} \longrightarrow \mathscr{Y}^*$, where $\mathscr{Y}^*$ is any finite direct product of $\mathscr{Y}$. We call $c(x)$ the codeword for x and $l(x) = |c(x)|$ the length of the codeword for x.

The length of a source code C is the sum of the lengths of every codeword in C weighted by its relative frequency.

$$L(C(\mathbf{X})) = \sum_x p(x)l(x) \tag{2.2}$$

A source code is uniquely decodable if any concatenation of codewords can only be generated by a unique concatenation of source symbols. In other words, the source symbols generating the codewords can be recovered with no possible equivocation.

The next inequality on the length of uniquely decodable source codes was first proved by McMillan [7] however the following proof is a simpler version by Karush [3, 6].

Theorem 2.2.1 The length of a uniquely decodable source code C for a random variable $\mathbf{X}$ taking values in alphabet $\mathscr{Y}$ verifies:

$$\sum_x \frac{1}{|\mathscr{Y}|^{l(x)}} \leq 1$$

Proof. Let $c(x_1, x_2, \ldots, x_k)$ be a concatenation of codewords of aggregated length $l(x_1, x_2, \ldots, x_k) = \sum_{i=1}^k l(x_i)$. Since C is uniquely decodable for any aggregated length k, no more than $|\mathscr{Y}|^k$ different concatenation of codewords can be generated.

We can consider the related expression on the aggregated length:

$$\begin{aligned} \left(\sum_x \frac{1}{|\mathscr{Y}|^{l(x)}}\right)^n &= \sum_{x_1} \frac{1}{|\mathscr{Y}|^{l(x_1)}} \sum_{x_2} \frac{1}{|\mathscr{Y}|^{l(x_2)}} \cdots \sum_{x_n} \frac{1}{|\mathscr{Y}|^{l(x_n)}} \\ &= \sum_{x_1, x_2, \ldots, x_n} \frac{1}{|\mathscr{Y}|^{l(x_1)+l(x_2)+\cdots+l(x_n)}} \\ &= \sum_{x_1, x_2, \ldots, x_n} \frac{1}{|\mathscr{Y}|^{l(x_1+x_2+\cdots+x_n)}} \qquad (2.3) \end{aligned}$$

which can also be written as the sum for all possible lengths i of the number T_i of concatenation of n codewords:

$$\begin{aligned} \left(\sum_x \frac{1}{|\mathscr{Y}|^{l(x)}}\right)^n &= \sum_{i=1}^{nl_{max}} \frac{T_i}{|\mathscr{Y}|^i} \\ &\leq \sum_{i=1}^{nl_{max}} \frac{|\mathscr{Y}|^i}{|\mathscr{Y}|^i} \\ &\leq nl_{max} \qquad (2.4) \end{aligned}$$

where $l_{max} = \max_x l(x)$. And taking the n-th root in both sides:

$$\sum_x \frac{1}{|\mathscr{Y}|^{l(x)}} \leq (nl_{max})^{1/n} \qquad (2.5)$$

now since the limit $\lim_{n\to\infty} (nl_{max})^{1/n} = 1$ and the result holds for all n:

$$\sum_x \frac{1}{|\mathscr{Y}|^{l(x)}} \leq 1 \qquad (2.6)$$

■

We finish this brief overview of source coding with Th. 2.2.2 [3], it shows that the length of a uniquely decodable code is lower bounded by the entropy of the random variable.

Theorem 2.2.2 The length of a uniquely decodable code taking values from finite alphabet $\mathscr{Y}$ for random variable $\mathbf{X}$ is lower bounded by the entropy of $\mathbf{X}$.

$$L \geq H_{|\mathscr{Y}|}(\mathbf{X})$$

Proof.

$$\begin{aligned} H_{|\mathscr{Y}|}(\mathbf{X}) - L &= \sum_x p(x)\log_{|\mathscr{Y}|}\frac{1}{p(x)} - \sum_x p(x)l(x) \\ &= \sum_x p(x)\log_{|\mathscr{Y}|}\frac{1}{p(x)} - \sum_x p(x)\log_{|\mathscr{Y}|}|\mathscr{Y}|^{-l(x)} \\ &= \sum_x p(x)\log_{|\mathscr{Y}|}\frac{|\mathscr{Y}|^{-l(x)}}{p(x)} \\ &\leq \log_{|\mathscr{Y}|}\sum_x |\mathscr{Y}|^{-l(x)} \\ &\leq \log_{|\mathscr{Y}|} 1 = 0 \end{aligned} \tag{2.7}$$

where the first inequality is again an application of Jensen's result Th. 1.4.1 and the second one results from applying McMillan's Th. 2.2.1. ■

This result can be extended to consider a source code C for a sequence of n random variables iid from the ensemble $\mathbf{X}$. That is, a source $\mathbf{X}$ as we defined it in Sec. 1.4. Let $R = L(C(\mathbf{X^n}))/n$, the length per symbol of C, be the encoding rate of the source $\mathbf{X}$. It is trivial to see that the rate is also lower bounded by the entropy of the source:

$$\begin{aligned} R &= L(C(\mathbf{X^n}))/n \\ &\geq H_{|\mathscr{Y}|}(\mathbf{X^n})/n \\ &= H_{|\mathscr{Y}|}(\mathbf{X}) \end{aligned} \tag{2.8}$$

2.3 Huffman codes

2.4 Typical sequences

2.5 Shannon's theorem for source coding

2.6 Further reading

The references of the previous chapter are also relevant for this one. In particular sections X and Y of Shanon's original paper [] and chapters A and B in Cover and Thomas []. A popular introduction to source coding is provided by xkcd (the comic stripe) [], in the post you will recognize the image that opens the chapter, perhaps you can explain it's relation with the topic.

3. Channel coding

In the previous chapter we investigated the source coding problem. That is, given a source and some noiseless means of communication, the problem is to encode the source in such a way that we minimize the usage of the noiseless communications channel while we allow the receiver to recover the message. Here we will investigate a dual problem, the problem of transmitting a uniform source over a noisy channel. This problem, is the problem that your mobile phone faces each time that it wants to exchange information with the nearest base station, or the problem that your ADSL router faces to transmit information over fiber. It is also the same problem that your computer faces when it wants to store information on a disk in such a way that it can be recovered at a later time. This is as we see a very relevant problem. We won't have the time to dig into great depth, but the menu should allow you to get a solid understanding of the mathematical foundations and a glimpse about how we tackle it currently. In the figure above, you can see the XXXX space XXX. As you can imagine, the bandwidth of the communications channel linking it with earth is very limited, the information we want to exchange precious and the channel rather noisy. Space exploration has been one unexpected driver of progress for pushing the limits of error correcting codes. (why error correting?)

3.1 The communications problem

Let us first of all, depict the building blocks of an idealized communications problem. Our description parallels the one of Shannon [8], see in Fig. 3.1 a graphical representation. The figure shows five entities: an information source, a transmitter, a noise source, a receiver, and a destination. The communications scheme works as follows:

First the information source generates a message $\mathbf{m}$ from a set of possible messages M. Then, the transmitter takes $\mathbf{m}$ and encodes it into n channel symbols. We define the coding rate R as:

$$R = \frac{\log M}{n} \tag{3.1}$$

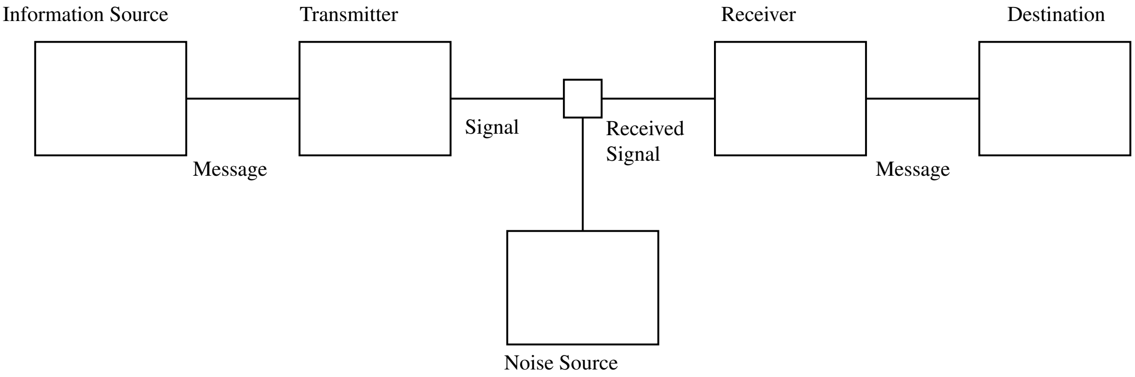


Figure 3.1: This figure reproduces the communications system diagram introduced by Shannon [8].

Exercise 3.1 Suppose that we want to transmit a message from a set of 16 messages that a source will choose uniformly at random. We have two channels available, both are noiseless, one allows to transmit one bit per channel use and we can use it once per second. The second channel allows to transmit two bits per channel use and we can use it once per minute. Over which channel can we communicate at a higher rate? How does time influence this number? ■

The channel is a physical medium of transmission. Mathematically, we can model it as a system taking symbols from input alphabet $\mathscr{X}$ to symbols of output alphabet $\mathscr{Y}$ and characterized by a transition probability matrix that maps the probability of every symbol y if symbol x is sent. The receiver tries to undo the encoding given the noisy received signal and at the end of the scheme the destination receives the $\hat{\mathbf{m}}$ possibly identical to $\mathbf{m}$.

We define C, the capacity of a channel, as the maximum mutual information for all possible input distributions:

$$C = \max_{p(x)} I(\mathbf{X};\mathbf{Y}) \tag{3.2}$$

3.2 Linear codes

3.3 Hamming codes

3.4 Detection, correction and minimum distance

3.5 Random codes

3.6 Shannon's second theorem

The capacity of a channel specifies the maximum rate at which a source can be reliably sent through a channel. On the other hand, no source with a rate over the capacity of the channel can be sent with a vanishing error probability.

A sketch of the proof would be as follows. Encoder and decoder share a code-book of 2^{nR} codewords chosen within the $2^{nH(\mathbf{X})}$ typical sequences [**Massey_77**]. The encoder sends a codeword $\mathbf{x}$ drawn with uniform probability. The decoder outputs a word $\hat{\mathbf{x}}$ jointly typical with the received word $\mathbf{y}$. It declares an error if $\mathbf{x}$, $\mathbf{y}$ are not jointly typical and a decoding error can occur if there exists $\mathbf{x}' \neq \mathbf{x}$ jointly typical with $\mathbf{y}$. We know by Eq. **??** that the probability of non-joint typicality for long enough n can be made as small as desired.

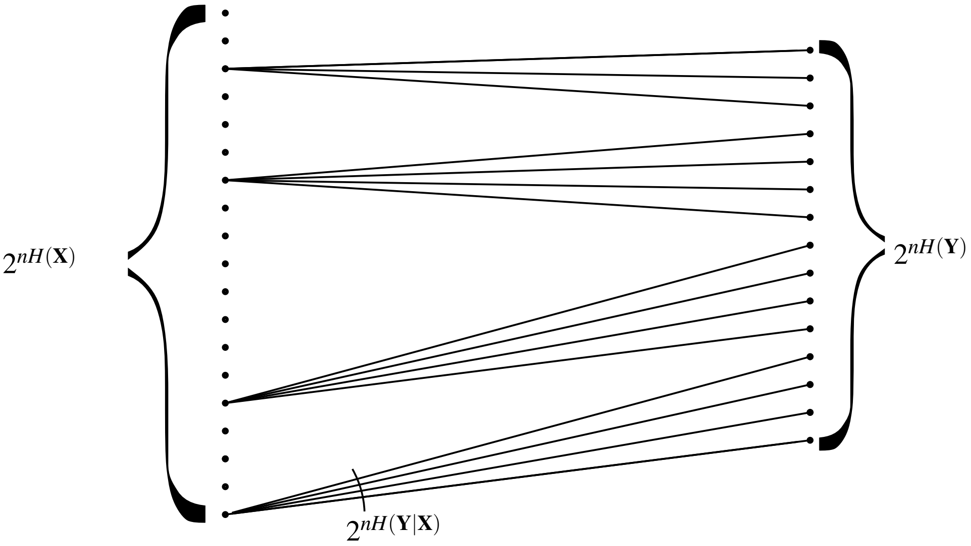


Figure 3.2: Graphical representation of the input and output typical sequences. A good encoding chooses as codewords a subset of the input typical sequences that produces disjoint sets of output typical sequences.

The intuition behind the achievability proof is simple. The decoder has access to two sets: the set of sequences jointly typical with $\mathbf{y}$, and the set of codewords. If the intersection is to be a single word, every codeword has to be jointly typical with a disjoint set of typical output words.

Approximately, every codeword is jointly typical with $2^{nH(\mathbf{Y}|\mathbf{X})}$ words. Then the number of jointly typical output words with input codewords is upper bounded by $2^{nR+nH(\mathbf{Y}|\mathbf{X})}$, where R is the coding rate. This number should be much smaller than the total number of typical sequences $2^{nH(\mathbf{Y})}$:

$$2^{nR+nH(\mathbf{Y}|\mathbf{X})} < 2^{nH(\mathbf{Y})}$$

which operating returns the expected result:

$$R < I(\mathbf{X};\mathbf{Y})$$

In conclusion, as long as the coding rate is below the mutual information between input and output for n long enough we can construct a code that allows the decoder to distinguish between codewords with a vanishing probability of error.

The converse statement follows from Fano's inequality [**Fano_61**]. The intuition behind this part is that if we think of an encoding that achieves a vanishing error probability, then necessarily $R < I(\mathbf{X};\mathbf{Y})$ [3].

Theorem 3.6.1 Fano's inequality

Proof. a ■

3.7 The capacity of some basic channels

3.7.1 Binary Symmetric Channel

In the BSC the binary elements or bits are either perfectly transmitted with probability $1-p$ or flipped with probability p.

Let us first find the mutual information between the input **X** and the output **Y** [3]::

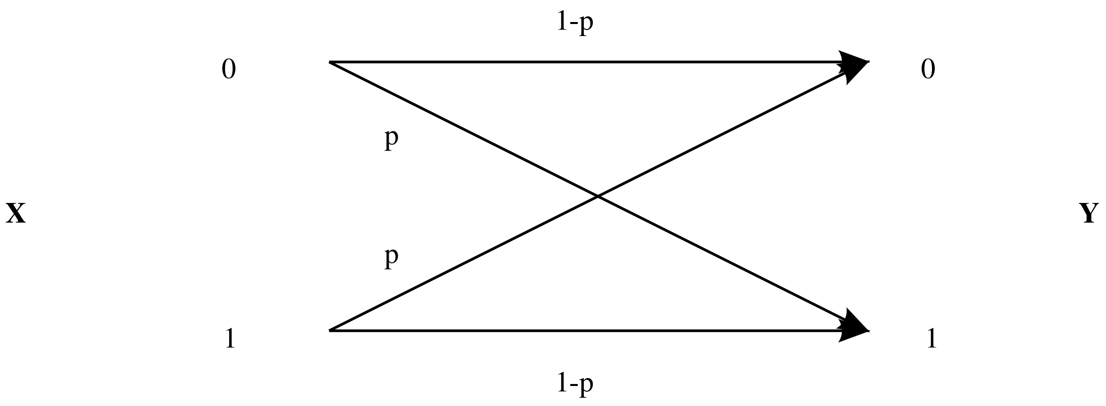


Figure 3.3: Binary Symmetric Channel.

$$\begin{aligned}
I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) && (3.3)\\
&= H(Y) - \sum_x p(x) H(\mathbf{Y}|x) && (3.4)\\
&= H(Y) - \sum_x p(x) H(p, 1-p) && (3.5)\\
&= H(Y) - H(p,1-p) \sum_x p(x) && (3.6)\\
&\leq 1 - H(p, 1-p) && (3.7)
\end{aligned}$$

We obtain the capacity by finding the maximum of the mutual information for all possible input distributions. It can be easily verified that the the uniform distribution reaches the upper bound in Eq. 3.7 and the capacity of the BSC is one minus the binary entropy of p.

3.7.2 Binary Erasure Channel

The BEC was introduced by Elias in his famous paper "Coding for Two Noisy Channels" [**Elias_55**]. The BEC has two input elements while the output alphabet is composed of three elements: 0, 1, and e, which stands for an erasure in the channel. In this channel the bits are either correctly transmitted with probability $1-p$, or are erased with probability p.

We can first find $H(\mathbf{X}|\mathbf{Y})$:

$$\begin{aligned}
H(\mathbf{X}|\mathbf{Y}) &= \pi(1-p) H(\mathbf{X}|\mathbf{Y}=0)\\
&\quad + (\pi p + (1-\pi)p) H(\mathbf{X}|\mathbf{Y}=e)\\
&\quad + (1-\pi)(1-p) H(\mathbf{X}|\mathbf{Y}=1) && (3.8)\\
&= p && (3.9)
\end{aligned}$$

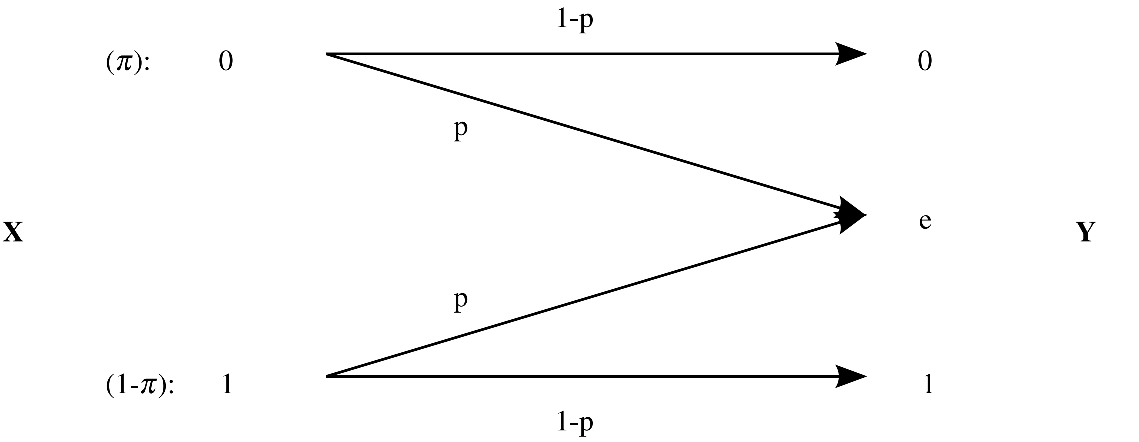


Figure 3.4: Binary Erasure Channel.

where $p(\mathbf{X}=0)=\pi$.The second equality holds from $H(\mathbf{X}|\mathbf{Y}=1)=H(\mathbf{X}|\mathbf{Y}=0)=0$ and $H(\mathbf{X}|\mathbf{Y}=e)=1$. We can now plug Eq. 3.8 in Eq. 1.28 and bound from above the mutual information:

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= H(\mathbf{X})-H(\mathbf{X}|\mathbf{Y}) && (3.10)\\ &= H(\pi,1-\pi)-p && (3.11)\\ &\le 1-p && (3.12)\end{aligned}$$

equality in Eq. 3.12 is achieved again by the uniform distribution. That is, for $\pi=\frac{1}{2}$.

It might seem that the capacity of a BSC that flips bits with probability p is greater than the capacity of a BEC that erases bits with probability p. Fig. 3.5 shows that it is the opposite situation. On the range $p\in(0,0.5)$, the capacity of the BEC is greater than the capacity of the BSC. Bits on the BEC are either perfectly known or perfectly unknown, however, it is not possible to distinguished flipped bits from correct bits in the BSC.

3.8 Further reading

3.9 Exam problems

One time pad?

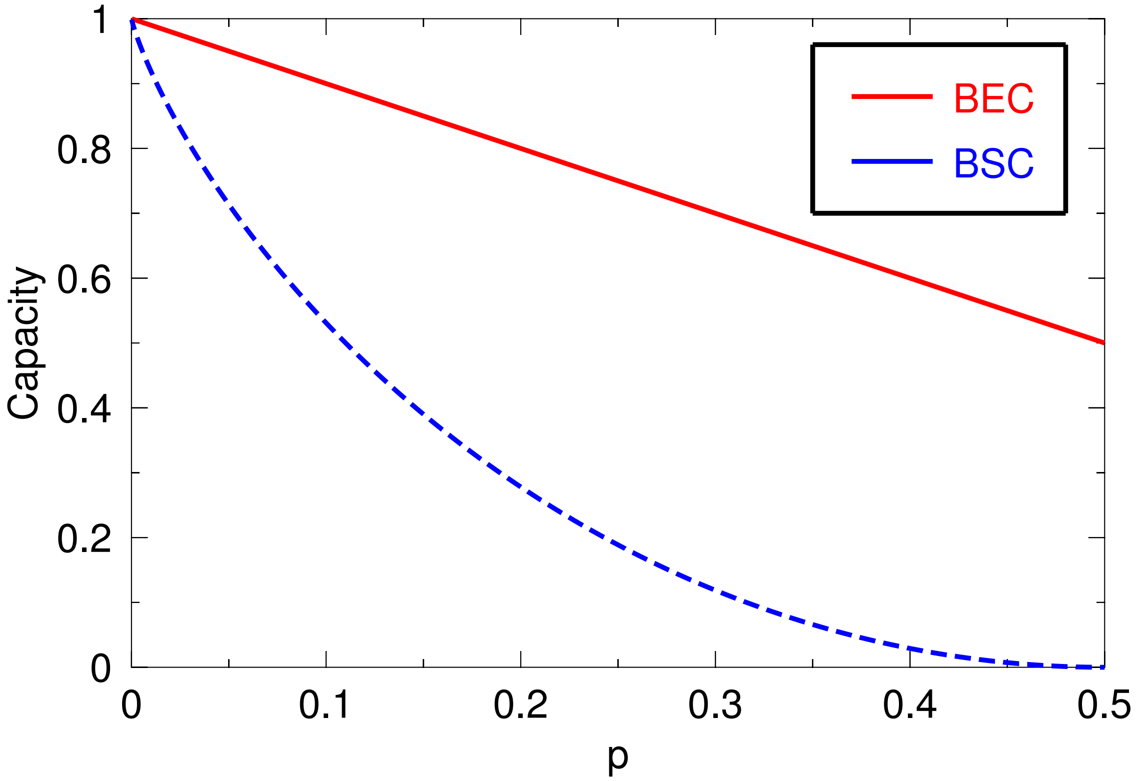


Figure 3.5: The capacity of the BEC and BSC.

Bibliography

[1] J. Aczel and Z. Daroczy. *On measures of information and their characterizations*. Academic Press, 1975 (cited on page 19).

[2] J. Aczél, B. Forte, and C. T. Ng. “Why the Shannon and Hartley Entropies Are Natural”. In: *Advances in Applied Probability* 6.1 (1974), pp. 131-146. ISSN: 00018678 (cited on page 19).

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Aug. 1991 (cited on pages 15, 23, 27, 28).

[4] I. Csiszár. “Axiomatic Characterizations of Information Measures”. In: *Entropy* 10 (2008), pages 261–273 (cited on page 19).

[5] A. Feinstein. *Foundations of Information Theory*. 1958 (cited on page 19).

[6] J. Karush. “A simple proof of an inequality of McMillan (Corresp.)” In: *IRE Transactions on Information Theory* 7.2 (Apr. 1961), page 118 (cited on page 23).

[7] B. McMillan. “Two inequalities implied by unique decipherability”. In: *IRE Transactions on Information Theory* 2 (Dec. 1956), pages 115–116 (cited on page 23).

[8] C. E. Shannon. “A mathematical theory of Communication”. In: *The Bell system technical journal* 27 (July 1948), pages 379–423 (cited on pages 25, 26).

Index

Axiomatic derivation of entropy, 14

Detection, correction and minimum distance, 26

Entropy, 15
Exam problems, 29

Formal problem, 22
Further reading, 19, 24, 29

Hamming codes, 26
Huffman codes, 24

Joint entropy, conditional entropy and mutual information, 16

Linear codes, 26

Random codes, 26
Refresher on probability theory, 12

Shannon’s second theorem, 26
Shannon’s theorem for source coding, 24
Surprising information, 11

The capacity of some basic channels, 28
The communications problem, 25
Transmitting the weather forecast, 21
Typical sequences, 24