Mathematical Foundations of Information Theory / Discrete Stochastics and Information Theory

Wolfgang Woess, Joshua Erde

Department of Mathematics,
TU Graz.

Contents

Preface

The course is based on the lecture notes of Wolfgang Woess.

1 Introduction to Probability Theory

Probability is a mathematical phenomenon that we see in every day life that we perhaps intuitively understand. As a motivating example, consider what is called the *law of large numbers* - if we toss a fair coin 1000 times every day, then each day we will get heads *about* 500 times. Of course, we won't get exactly 500 heads, but the *deviations* we observe, over the repeated trials, should be *small*. Similarly, if we roll a fair die many times, the relative frequency of the outcome "6" will be approximately 1/6. From a certain philosophical viewpoint, this is what we mean when we say "The probability of rolling a 6 is 1/6".

More generally, the law of large numbers says that if we have some random experiment, whose outcome is a real number, and we repeat the experiment many times, then the average of the outcomes should *converge* to some specific, deterministic number, which is the *expected* outcome of the experiment.

In some ways this is intuitive, in other ways almost tautological, but what we want then from a theory of probability is a set of axioms which behaves like how we experience probability in the real world, and so in particular statements like the law of large numbers should follow as a mathetmatical theorem from these axioms.

1.1 Probability Spaces

Definition 1.1. A *probability space* is a triple $(\Omega, \mathcal{A}, \mathbb{P})$, where

1. Ω is a non-empty set, the *sample space*,
2. $\mathcal{A}$ is a σ-algebra, that is, a collection of subsets of Ω such that
   (i) $\Omega \in \mathcal{A}$,
   (ii) $A \in \mathcal{A} \Rightarrow A^c = \Omega \setminus A \in \mathcal{A}$,
   (iii) if $A_n \in \mathcal{A}$ for $n = 1, 2, \ldots$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.
3. $\mathbb{P}$ is a *probability measure* on $\mathcal{A}$, that is, a function $\mathbb{P} : \mathcal{A} \to [0, 1]$ such that
   (i) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$,
   (ii) if $A_n \in \mathcal{A}$ are pairwise disjoint for $n = 1, 2, \ldots$, that is, $A_n \cap A_m = \emptyset$ for all $m \neq n$, then
   $$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

We should think of the sample space Ω as consisting of all possible outcomes of some random experiment.

Example 1.2. (a) Our experiment is a coin toss. Our two outcomes are "Heads" and "Tails" and so our sample space is $\Omega = \{\text{Heads}, \text{Tails}\}$. We could equally 'encode' the outcomes as Heads $= 1$ and Tails $= 0$, in which case our sample space is $\Omega = \{0, 1\}$.

(b) Our experiment is again a coin toss, but we don't just measure the side that the coin lies on, but also its position on the ground, which is some point (x, y) in the plane with the coin-tosser standing at the origin, as well as the number m of times that the coin rotates whilst in the air. Then, a possible sample space would be

$$\Omega = \{(\ell, x, y, m) \colon \ell \in \{0, 1\}, (x, y) \in \mathbb{R}^2, m \in \mathbb{N}_0\}.$$

(c) Our experiment is sequence of n coin tosses, and we measure the sequence of outcomes. We can take our sample space to be

$$\Omega = \{0, 1\}^n$$

sequences of 0s and 1s of length n, which we call *bitstrings*, where the kth element of the sequence is the outcome of the kth coin toss.

(d) Our (theoretical) experiment is an infinite sequence of coin tosses. Our sample space is then

$$\Omega = \{0, 1\}^{\mathbb{N}}$$

all infinite sequences of 0s and 1s (an uncountabke set!).

The function $\mathbb{P}$ then tells us, for an *event*, a particular subset of the possible outcomes, how likely it is that this event occurs, that is, how likely it is that the outcome lies in this subset.

It turns out, for complicated mathematical reasons, even if the sample space Ω is something familiar like the real numbers $\mathcal{R}$ or the unit interval $[0, 1]$, there is no way to define a consistent notion of measure that will assign a probability to *every* subset of Ω - very weird sets exist! For this reason we have to restrict ourselves to some 'well-behaved' collection of sets, this σ-algebra $\mathcal{A}$. However, this is no great restriction, as we can choose $\mathcal{A}$ such that any event that you can actually physically describe will lie inside $\mathcal{A}$.

When Ω is *countable*, one can usually take $\mathcal{A} = \mathcal{P}(\Omega)$, the *power set* of Ω, consisting of all subsets of Ω. However, when Ω is uncountable, such as $\Omega = \mathbb{R}$, then there is no way to define *any* function $\mathbb{P}$ satisfying the definition of a probability space with $\mathcal{A} = \mathcal{P}(\mathbb{R})$.

For the most part we will work with *discrete probability spaces*, those where Ω is countable, and so avoid these difficulties. In this case, if Ω is countable and $\mathcal{A} = \mathcal{P}(\Omega)$, then the probability measure is determined by the measure of the *elementary events* $\omega \in \Omega$ since

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathcal{P}(\{\omega\}) \qquad \text{for all } A \in \mathcal{P}(\Omega).$$

Definition 1.3. Given a logical expression Φ concerning the elements of Ω, we will write $[\Phi]$ for the event

$$A = \{\omega \in \Omega \colon \Phi \text{ is true for } \omega\},$$

and if $A \in \mathcal{A}$ (which will usually be the case) we define

$$\mathbb{P}[\Phi] := \mathbb{P}(A).$$

Example 1.4. (a) Our experiment is to roll two fair dice. Our sample space is

$$\Omega = \{(i, j) \colon 1 \leq i, j \leq 6\}$$

and we can take our σ-algebra to be $\mathcal{A} = \mathcal{P}(\Omega)$.

An event we could consider is the event that the total value of the two dice is 11, that is, we consider the event $A = [\text{the total value of the dice is 11}]$, or in other words

$$A = \{(i, j) \in \Omega \colon i + j = 11\}.$$

(b) Our experiment is as in Example 1.2 (b). We consider the event

$$\begin{aligned} A &= [\text{the coin lands at distance at most } r \text{ from the coin tosser}] \\ &= \{(\ell, x, y, m) \colon x^2 + y^2 \leq r^2\}. \end{aligned}$$

Similarly the event

$$\begin{aligned} A &= [\text{the coin spins at least 3 times and lands on Heads}] \\ &= \{(\ell, x, y, m) \colon \ell = 1, m \geq 3\}. \end{aligned}$$

One can easily deduce the following properties from Definition 1.1.

Proposition 1.5. *(i)* $\emptyset \in \mathcal{A}$,

(ii) If $A_n \in \mathcal{A}$ *for* $n = 1, 2, \ldots$, *then* $\bigcap_{n=1}^{\infty} A_n \in \mathcal{A}$,

(iii) If $A, B \in \mathcal{A}$ *then* $A \cup B, A \cap B \in \mathcal{A}$ *and*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

(iv) If $A \in \mathcal{A}$, *then* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,

(v) If $A, B \in \mathcal{A}$ *and* $A \subseteq B$, *then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

The next lemma is fundamental.

Lemma 1.6 (Continuity of the probability measure)**.** *If* $(A_n \colon n \in \mathbb{N})$ *is an increasing sequence, that is* $A_n \subseteq A_{n+1}$ *for all* $n \in \mathbb{N}$, *with* $A_n \in \mathcal{A}$ *for all* n, *then*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{N \to \infty} \mathbb{P}(A_N).$$

Similarly, if $(A_n \colon n \in \mathbb{N})$ *is an decreasing sequence, that is* $A_n \supseteq A_{n+1}$ *for all* $n \in \mathbb{N}$, *with* $A_n \in \mathcal{A}$ *for all* n, *then*

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_i\right) = \lim_{N \to \infty} \mathbb{P}(A_N).$$

Proof. □

Definition 1.7 (Conditional probability, Independence)**.** Given two events $A, B \in \mathcal{A}$ we define

$$\mathbb{P}(A \mid B) = \begin{cases} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, & \text{if } \mathbb{P}(B) > 0, \\ 0, & \text{if } \mathbb{P}(B) = 0. \end{cases}$$

Given two logical expression Φ_1 and Φ_2, where $A = \{\omega \in \omega \colon \Phi_1 \text{ is true for } \omega\}$ and $B = \{\omega \in \omega \colon \Phi_2 \text{ is true for } \omega\}$, we will write

$$\mathbb{P}[\Phi_1 \mid \Phi_2] = \mathbb{P}(A \mid B).$$

We say A and B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. A sequence (finite or infinite) of events $(A_n \colon n \in I)$ is called *(mutually) indepenent* if for all choices of indices $J \subseteq I$

$$\mathbb{P}\left(\bigcap_{i \in J} A_j\right) = \prod_{i \in J} \mathbb{P}(A_j).$$

We note that this condition is stronger than asking for *pairwise* independence of the events A_i, A_j for $i, j \in I$.

Example 1.8. Suppose we flip two fair coins, so that $\Omega = \{0,1\}^2$, $\mathcal{A} = \mathcal{P}(\Omega)$. We can check that each outcome is equally likely, and so for all events $A \subseteq \Omega$

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{4}.$$

Let us consider the events

$$\begin{aligned} A_1 &= [\text{the first coin lands Heads}], \\ A_2 &= [\text{the second coin lands Heads}], \\ A_3 &= [\text{both coins land on the same side}], \end{aligned}$$

so that $A_1 = \{(1,0),(1,1)\}, A_2 = \{(0,1),(1,1)\}$ and $A_3 = \{(0,0),(1,1)\}$ and $A_i \cap A_j = \{(1,1)\}$ for all i, j.

In particular $\mathbb{P}(A_i) = \frac{2}{4} = \frac{1}{2}$ for all $i \leq 3$ and $\mathbb{P}(A_i \cap A_j) = \frac{1}{4} = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j)$ for all $i, j \leq 3$, and hence all pairs of events are independent.

However, $A_1 \cap A_2 \cap A_3 = \{(1,1)\}$ and so $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{1}{4}$, but

$$\prod_{i=1}^{3} \mathbb{P}(A_i) = \left(\frac{1}{2}\right)^3 = \frac{1}{8} \neq \frac{1}{4}.$$

Hence, whilst these three events are pairwise independent, we shouldn't intuitively think of the sequence as being independent. Indeed, if we know that both A_1 and A_2 happens, then it is already determined that A_3 must happen!

1.2 Random Variables

Definition 1.9 (Discrete Random Variable, Distribution)**.** Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a *discrete random variable* is a function $X \colon \Omega \to \mathcal{X}$, where $|\mathcal{X}|$ is countable, such that for every $B \subseteq \mathcal{X}$ the set

$$X^{-1}(B) = \{\omega \in \Omega \colon X(\omega) \in B\} = [X \in B],$$

is a member of the σ-algebra $\mathcal{A}$.

That is, a discrete random variable is some function of our sample space which takes values in some discrete set $\mathcal{X}$, such that for any possible subset of $\mathcal{X}$, the probability that X lies in this subset is well-defined.

The *distribution* of X is the function $P_X \colon \mathcal{P}(\mathcal{X}) \to [0,1]$ given by

$$P_X(B) = \mathbb{P}[X \in B].$$

When One can check that $(\mathcal{X}, \mathcal{P}(\mathcal{X}), P_X)$ is a probability space. If $\mathcal{X} \subseteq \mathbb{R}$ we say that X is a discrete *real* random variable.

We can think of random variables as "functions of chance". When our probability space models the outcome of some random experiment, a random variable extracts some aspect of the experiment which can be measured.

A lot of very natural random variables are not discrete, for example when the observable is not a discrete, but a continuous quantity, and there is a corresponding theory of *continuous* random variables. However, this won't be relevant until the very last part of the course, and dealing with them formally is a bit more involved. In particular, unless otherwise stated, every random variable we consider in the course will be discrete, and we will only state the relevant results for discrete random variables. However, in almost all cases, analogous statements can be shown to hold for continuous random variables.

So, a random variable X assigns to each outcome ω in the sample space an element $X(\omega) = x \in \mathcal{X}$. It is important then to keep track of the difference between the random variable X and one of the possible values x that X can take.

Example 1.10. Suppose we toss a sequence of n fair coins, so that we have a sample space $\Omega = \{0,1\}^n$ (and since Ω is finite we can take $\mathcal{A} = \mathcal{P}(\Omega)$). Since the coin is fair, each possible outcome, each sequence $\omega \in \Omega$, is equally likely to occur, and so $\mathbb{P}(\{\omega\}) = 2^{-n}$ for all ω, and $\mathbb{P}(A) = |A|2^{-n}$ for all $A \in \mathcal{A}$.

Now we can look at some random variables, functions from Ω to $\mathbb{R}$, which are observable quantities from this experiment. For example I could define $X_k(\omega)$ to be the kth element in the sequence ω, the outcome of the kth coin toss.

Then X_k is a discrete random variable, it takes values in $\mathcal{X}_k = \{0,1\}$, and we can calculate the distribution

$$P_{X_k}(\{0\}) = \mathbb{P}[X_k = 0] = \mathbb{P}[\text{The } k\text{th coin toss is tails}] = \frac{1}{2},$$

and similarly $P_{X_k}(\{1\}) = \frac{1}{2}, P_{X_k}(\emptyset) = 0, P_{X_k}(\{0,1\}) = 1$.

Or we could define $S_n(\omega)$ to be the sum of the elements of ω, or in other words the number of heads thrown. Again, S_n is a discrete random variable, taking values in $\mathcal{S}_n = \{0, \ldots, n\}$. In

this case for each $k \in \{0, \dots, n\}$ we can calculate $P_{S_n}(\{k\}) = \mathbb{P}[S_n = k]$. Indeed, this is just a combinatorial exericise

$$\begin{aligned}\mathbb{P}[S_n = k] &= |[S_n = k]|2^{-n} \\ &= |\{\omega \colon \omega \text{ has precisely } k \text{ zeroes}\}|2^{-n} \\ &= \binom{n}{k} 2^{-n}.\end{aligned}$$

It is then easy to see that for every $A \subseteq \{0, \dots, n\}$

$$P_{S_n}(A) = \sum_{k \in A} P_{S_n}(\{k\})$$

Definition 1.11 (Discrete density function)**.** Given a discrete random variable X taking values in $\mathcal{X}$ the *discrete density function* $p_X \colon \mathcal{X} \to [0, 1]$ is defined by

$$p_X(x) = \mathbb{P}[X = x] = P_X(\{x\}).$$

This, $p_X(x) \neq 0$ if and only if $x = x_i$ for some $i \in I$. In particular,

$$1 = \mathbb{P}[X \in \mathcal{X}] = \sum_{x \in \mathcal{X}} p_X(x),$$

and for any $B \subseteq \mathcal{X}$

$$P_X(B) = \sum_{x \in B} p_X(x),$$

and so the discrete density function and the distribution of X determine one another. For this reason, we will often refer to the discrete density function *as* the distribution of the random variable.

We will often think of the discrete density function as a vector $\boldsymbol{p} \in \mathbb{R}^{\mathcal{X}}$ with $||\boldsymbol{p}||_1 = 1$. Conversely, for any such vector $\boldsymbol{p}$ there is a random variable X whose density function satisfies $p_X = \boldsymbol{p}$.

During the course we will normally just introduce random variables by specifying their distributions or discrete density function, rather than making reference to any specific probability space.

Example 1.12. (a) Given $q \in [0, 1]$ a *Bernoulli random variable* $\mathrm{Ber}(q)$ takes values in $\{0, 1\}$ and has distribution given by

$$p_{\mathrm{Ber}(q)}(1) = p \qquad \text{and} \qquad p_{\mathrm{Ber}(q)}(0) = 1 - p,$$

so that $p_{\mathrm{Ber}(q)} = (1 - q, q)$. We can think of this random variable as the outcome of a biased coin flip.

(b) Given an event A, the *indicator random variable* of the event A

$$\mathbb{1}_A(\omega) = \begin{cases} 0 & \text{if } \omega \notin A, \\ 1 & \text{if } \omega \in A. \end{cases}$$

In particular, if $\mathbb{P}(A) = q$, then $p_{\mathbb{1}_A} = (1 - q, q)$, and so $\mathbb{1}_A$ has the same distribution as $\mathrm{Ber}(q)$.

(c) Given $n \in \mathbb{N}$ and $q \in [0,1]$ a *binomial random variable* $\mathrm{Bin}(n,q)$ takes values in $\{0,\ldots,n\}$ and has distribution given by

$$p_{\mathrm{Bin}(n,q)}(k) = \binom{n}{k} q^k (1-q)^{n-k}.$$

We can think of this random variable as counting the number of heads in a sequence of n consecutive flips of a random coin.

However, in this way, if I have two random variables X and Y, given just in terms of their distributions, this doesn't necessarily tell us 'the whole story'. Indeed, if X and Y are both observables from the same random experiment, then their values may be related in some way - if X is the height of a random person on the street and Y is the weight, then for most outcomes, most people, the values of X and Y will be *positively correlated*, if X is large then Y is more likely to be large and vice versa.

Definition 1.13 (Joint distribution)**.** If we have two discrete random variables X and Y defined on the same probability space, the *joint discrete density function* (which again we will usually refer to as the joint distribution) is defined as

$$p_{X,Y}(x,y) = \mathbb{P}[X = x, Y = y],$$

and from the joint density function we can reconsruct the *marginal* density functions of X and Y, which are given by

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \qquad \text{and } p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x,y).$$

Note that there can be many different joint density functions $p_{X,Y}$ with the same marginal density functions p_X and p_Y.

More generally, if $X_1,\ldots,X_n$ are all discrete random variables, defined on the same probability space, taking values in sets $\mathcal{X}_1,\ldots,\mathcal{X}_n$, then the 'vector' of random variables $(X_1,\ldots,X_n)$ is also a discrete random variable, which takes values in some subset of the product set $\mathcal{X}_1 \times \ldots, \times \mathcal{X}_n$. In this case, the *joint discrete density function* is defined as

$$p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n],$$

and the marginal distributions are defined in the obvious way.

Definition 1.14 (Conditional distribution)**.** Given jointly distributed discrete random variables X and Y, and some value $x \in \mathcal{X}$ with $p_X(x) > 0$, the *conditional density function* (conditional distribution) of Y, given that $X = x$, is

$$p_{Y|X}(y|x) = \mathbb{P}[Y = y | X = x] = \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{p_{X,Y}(x,y)}{\sum_{y' \in \mathcal{Y}} p_{X,Y}(x,y')}.$$

Note that, p_X and $p_{Y|X}$ together determine the joint distribution $p_{X,Y}$ and hence also the marginal density function p_Y.

When the random variables that we are dealing with are clear from the context, we will often drop the subscripts in the notation above and simply write expressions likes $p(x), p(x,y)$ or $p(y|x)$.

Example 1.15. Suppose we toss three coins and we let X be the number of heads in the first two coin tosses and Y be the number of heads in the last two coin tosses.

Then we can calculate the joint distribution of X and Y :

(x, y)	0	1	2
0	1/8	1/8	0
1	1/8	1/4	1/8
2	0	1/8	1/8

The marginal distribution p_X is given by the sum of the rows, which is $p_X = (1/4, 1/2, 1/4)$ and the marginal distribution p_Y is given by the sum of the columns, which is $p_Y = (1/4, 1/2, 1/4)$. Note that, as expected, both are distributed as $\mathrm{Bin}(3, 1/2)$.

The conditional distribution $p_{X|Y}$ is then :

$(x\|y)$	0	1	2
0	1/2	1/4	0
1	1/2	1/2	1/2
2	0	1/4	1/2

Definition 1.16 (Independence)**.** A sequence of discrete random variables $X_1, \ldots, X_n$ are *independent*, if for any sequence of subsets $B_1 \subseteq \mathcal{X}_1, \ldots, B_n \subseteq \mathcal{X}_n$ the events

$$[X_1 \in B_1], \ldots, [X_n \in B_n]$$

are independent. In particular

$$\mathbb{P}[X_1 \in B_1, \ldots, X_n \in B_n] = \prod_{i=1}^{n} \mathbb{P}[X_i \in B_i].$$

One can check that it is equivalent to show that the joint density of the sequence is equal to the product of the marginal distributions, that is

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i) \qquad \text{whenever } x_i \in \mathcal{X}_i \text{ for all } i. \tag{1.1}$$

An infinite sequence $(X_n)_{n\in\mathbb{N}}$ of discerete random variables is *independent* if the sequence $X_1, \ldots, X_n$ is independent for all n, or equivalently if (1.1) holds for all $n \in \mathbb{N}$.

1.3 Markov's and Chebyshev's inequality

Definition 1.17 (Expectation)**.** The *expectation* or *expected value* or *mean* of a discrete real random variable X is

$$\mathbb{E}(X) := \sum_{x\in\mathcal{X}} x \cdot \mathbb{P}[X = x] = \sum_{x\in\mathcal{X}} x \cdot p_X(x),$$

if the sum converges. Otherwise we informally say that the expectation is infinite.

If $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ is the underlying probability space, it can sometimes be simpler to use the formula

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}).$$

When $\mathcal{X} \not\subseteq \mathbb{R}$, so that X is not a real random variable, it does not make sense to talk about the expectation of X. However, for any function $g\colon \mathcal{X} \to \mathbb{R}$, we have that $g(X)$ is a real random variable (defined as $g(X)(\omega) = g(X(\omega))$ for all ω) whose expectation we can compute as

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

Lemma 1.18 (Linearity of expectation)**.** *Let X and Y be jointly distributed discrete real random variables with finite expectations and let $c \in \mathbb{R}$. Then*

(i) $\mathbb{E}(c) = c$,

(ii) $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$,

(iii) $\mathbb{E}(c \cdot X) = c \cdot \mathbb{E}(X)$,

(iv) $x \geq 0$ *for all* $x \in \mathcal{X} \Rightarrow \mathbb{E}(X) \geq 0$.

Proof. □

However, it is not true in general that $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$! This does however hold in the special case where X and Y are independent.

Lemma 1.19. *Let X and Y be independent discrete real random variables with finite expectations. Then $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$.*

Proof. □

Another important quantity comes from considering how far a random variable deviates from it's expectation.

Definition 1.20 (Variance)**.** Let X be a discrete real random variable with $\mu = \mathbb{E}(X) < \infty$. The *variance* of X is defined as

$$\mathbb{V}(X) = \mathbb{E}((X - \mu)^2).$$

It can easily be shown, using the linearity of expectation, that

$$\mathbb{V}(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2.$$

In general, the variance is also not linear, but again for independent random variables, we do have a nice formula.

Lemma 1.21. *Let X and Y be independent discrete real random variables with finite expectations and variances and let $c \in \mathbb{R}$. Then $\mathbb{V}(cX) = c^2\mathbb{V}(X)$ and $\mathbb{V}(X \pm Y) = \mathbb{V}(X) + \mathbb{V}(Y)$.*

Proof. (Exercise) □

It turns out that we can get some control over the distribution of a random variable X just by controlling its expectation or variance. In practise, since many of the random variables that arise are 'simple' combinations of 'simple' random variables, it it often possible to calculate, estimate or bound the expectation or variance of these random variables, and in this way obtain information about their distributions.

In particular, given an event A, it is easy to calculate the expectation and variance of the indicator random variable $\mathbb{1}_A$. Indeed

$$\mathbb{E}\left(\mathbb{1}_A\right) = 1 \cdot \mathbb{P}[\mathbb{1}_A = 1] + 0 \cdot \mathbb{P}[\mathbb{1}_A = 0] = \mathbb{P}(A),$$

and since $\mathbb{1}_A^2 = \mathbb{1}_A$, we see that

$$\mathbb{V}(\mathbb{1}_A) = \mathbb{E}(\mathbb{1}_A^2) - \left(\mathbb{E}(\mathbb{1}_A)\right)^2 = \mathbb{E}(\mathbb{1}_A) - \left(\mathbb{E}(\mathbb{1}_A)\right)^2 = \mathbb{P}(A) - \mathbb{P}(A)^2.$$

Lemma 1.22 (Markov's inequality)**.** *Let X be a non-negative discrete real random variable such that $0 < \mathbb{E}(X) < \infty$ and let $a > 0$. Then*

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. □

A simple, but powerful consequence of Markov's inequality is Chebyshev's inequality.

Corollary 1.23 (Chebyshev's inequality)**.** *Let X be a real random variable such that $\mathbb{E}(X), \mathbb{V}(X) < \infty$ and let $a > 0$. Then*

$$\mathbb{P}\big[|X - \mathbb{E}(X)| \geq a\big] \leq \frac{\mathbb{V}(X)}{a^2}.$$

Proof. □

1.4 Convergence of Random Variables and the Law of Large Numbers

Definition 1.24 (Convergence of random variables)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real random variables and X a random variable, all defined on the same probability space.

(i) $X_n \longrightarrow X$ *in probability* if for every $a > 0$

$$\lim_{n\to\infty} \mathbb{P}\big[|X_n - X| > a\big] = 0.$$

(ii) $X_n \longrightarrow X$ *almost surely* if

$$\mathbb{P}[X_n \to X] = 1,$$

that is, if

$$\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1.$$

We will see that convergence almost surely implies convergence in probability, and so the second is a stronger notion of convergence. We note that converse is not true! There exist sequences of random variables which converge in probability but no almost surely.

Example 1.25. Let us consider a sequence of independent random variables $(X_n)_{n\in\mathbb{N}}$ each taking values in $\{0, 1\}$ such that

$$\mathbb{P}(X_n = i) = \begin{cases} \frac{1}{n} & \text{if } i = 1 \\ 1 - \frac{1}{n} & \text{if } i = 0. \end{cases}$$

For all $a > 0$, it is clear that

$$\mathbb{P}(|X_n - 0| \geq a) \leq \mathbb{P}(X_n = 1) = \frac{1}{n}$$

and hence, X_n tend to 0 in probability.

On the other hand, $X_n \to 0$ if and only there is some N such that $X_n = 0$ for all $n \geq N$. However for any fixed N, since the X_n are independent,

$$\begin{aligned} \mathbb{P}[\forall n \geq N : X_n = 0] &= \lim_{M\to\infty} \mathbb{P}[\forall N \leq n \leq M : X_n = 0] \\ &\leq \lim_{M\to\infty} \prod_{n=N}^{M} \left(1 - \frac{1}{n}\right) \\ &\leq \lim_{M\to\infty} \exp\left(-\sum_{n=N}^{M} \frac{1}{n}\right) \\ &= 0, \end{aligned}$$

where we used that $1 - x \leq e^{-x}$, which holds for all x, and also that $\sum_{n=N}^{M} \frac{1}{n} \approx \ln M - \ln N$.

Hence,

$$\mathbb{P}[X_n \to 0] \leq \sum_{N=1}^{\infty} \mathbb{P}[\forall n \geq N : X_n = 0] = 0.$$

Whilst convergence in probability is not enough to guarantee convergence almost surely, it does guarantee the existence of an almost surely convergent subsequence.

Theorem 1.26. *Let $(X_n)_{n\in\mathbb{N}}$ and X be as above. If $X_n \to X$ in probability, then there is a subsequence $(n_k)_{k\in\mathcal{N}}$ such that $X_n \to X$ almost surely.*

Theorem 1.27 (Weak law of large numbers)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with finite mean $\mathbb{E}(X_n) = \mu < \infty$ and finite variance $\mathbb{V}(X_n) = \sigma^2 < \infty$ (the same for all n). Then*

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \to \mu \text{ in probability.}$$

Proof. □

As we saw, convergence in probability is weaker than convergence almost surely, and at times we will need the stronger statement that the *sample mean* converges to the mean almost surely.

Theorem 1.28 (Strong law of large numbers)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed (i.i.d) random variables (that is, each X_n has the same distribution) with finite mean $\mathbb{E}(X_n) = \mu < \infty$. Then*

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \to \mu \text{ almost surely.}$$

The proof of this theorem is a bit beyond the focus of the course, however we will at time want to relate the two notions of convergence, and will prove a few short lemmas on this topic.

One useful thing for a notion of convergence is that limits should be unique, although here we only have uniqueness up to a set of measure 0.

Lemma 1.29. *Let $X, (X_n)_{n\in\mathbb{N}}$ be jointly distributed real random variables. If $X_n \to X$ and $X_n \to X'$ in probability, then $\mathbb{P}[X = X'] = 1$.*

Proof. □

Finally, let us prove that convergence almost surely implies convergence in probability.

Theorem 1.30. *Let $X, (X_n)_{n\in\mathbb{N}}$ be jointly distributed real random variables. If $X_n \to X$ almost surely, then $X_n \to X$ in probability.*

Moreover, if we write $U_k = \sup\{|X_n - X| : n \geq k\}$, then $X_n \to X$ almost surely if and only if $U_k \to 0$ in probability.

Proof. □

2 Discrete Entropy

2.1 Hartley's formula and Shannon's formula

Information theory deals with the mathematical problems that arise in the *storage*, *transformation* and *transmission* of information.

We would like to have some sort of theory that measures the *informational content* of some data, which in some way should not depend on the particular form the data takes

Example 2.1. Suppose I have written down secretly a number from 0 to 31 and you wish to identify the number asking only yes/no questions.

It is, intuitively, clear that the 'best' question to start with is "Is your number at most 15?", since either answer will reduce the number of possibilities by $\frac{1}{2}$. In a similar fashion each question can reduce the number of possibilites by $\frac{1}{2}$, and so after 5 questions you can also identify the number.

Considering a question as a unit of information, we might say that this hidden number then contains 5 of these units, which we will call *bits*, of information.

If we think of encoding the numbers from 0 to 31 in binary, each number corresponds to a sequence in $\{0,1\}^5$, and the questions that we ask correspond to asking about the value of the kth digit in the sequence.

Hartley made this idea formal in 1928 when he defined the notion of the *uncertainty* of a uniform random sample.

Definition 2.2 (Hartley's formula)**.** Suppose some element is chosen from a collection U_N of N different elements, with each being equally likely. The *uncertainty* of this random element (which one can think of the *informational cost* to identify the element) is given by

$$H(U_N) = \log_2 N.$$

This can be justified in terms of the follow (heuristic) *axiomatic requirements*

(A) $H(U_2) = 1$,

(B) $H(U_{N+1}) \geq H(U_N)$,

(C) $H(U_{N\cdot M}) = H(U_N) + H(U_M)$.

The first two are relatively intuitive - the amount of information needed to identify one of two elements is a single question or bit (alternatively, this is just some arbitrary choice to normalise this measure with respect to the units we've chosen). Furthermore, clearly there is more information needed to identify an element from a larger set.

For the third we can think of grouping our elements into N disjoint groups consisting of M elements

$$U_{N\cdot M} = U_M^{(1)} \cup \ldots \cup U_M^{(N)}.$$

In order to identify one element from $U_{N\cdot M}$ we could identify first the group $U_M^{(i)}$ that the element lies in, and so identify a uniformly chosen unknown group from a collection of N many groups, and then identify the unknown element of $U_M^{(i)}$, which is equally likely to be any of these elements. Hence, the cost to identify this element is at most $H(U_N) + H(U_M)$.

However, conversely, suppose we choose our random element by first choosing a random group, and then choosing a random element of our group. If we can identify the random element, we can identify both of these random choices, and so the cost to identify this element must be at least $H(U_N) + H(U_M)$.

In fact, Rényi showed that these three properties uniquely determine Hartley's formula.

Lemma 2.3. *The function $H(U_N) = \log_2 N$ is the unique function satisfying properties (A)–(C).*

Proof. □

Suppose now that our elements are not equally likely to be chosen, but that the kth element is instead chosen with some probability p_k. Can we justify, using the previous heuristic, what the informational cost of identifying the chosen element is?

Well, in some sense the 'cost' to identify the hidden element does not change, we still might need to identify any one of N elements. However, if one of the p_ks were much larger than all the others, so in almost every case the kth element is chosen, it would be much more sensible to start by asking "is the hidden element the kth element"? In the worst case we would have to ask more questions, but *on average* we'd identify the element with many fewer questions!

So, it makes sense instead to consider the *expected uncertainty*, or the expected informational cost to identify the unknown element. The following is a *heuristic* argument for how we should define this quantity.

Let us assume that the probabilities p_k are all rational, otherwise we can take some rational approximations and argue "in the limit". Instead of an element from U_N where the kth element is chosen with probability p_k, I could choose an element from a larger set

$$U_M = U_{M_1} \cup \ldots \cup U_{M_N}.$$

where U_{M_i} contains $p_i M$ elements, for some large M such that all these numbers are integers. There is then a clear equivalence between identifying the group U_{M_i} in which this element lies, and of identifying the element in the original problem.

By a similar argument as before, the expected informational cost of identify the random element of U_M, which should be $\log_2 M$ by Hartley's formula, should be given by the expected cost to identify the correct group U_{M_i}, which is the quantity we are interested in, which we denote by H_1, plus the expected cost to identify the correct element of this group, which we denote by H_2. Now the element lies in M_i with probability p_i, and if the element lies in U_{M_i} then the informational cost to identify it is $\log_2 M_i$ by Hartley's formula. Hence, the expected cost is

$$H_2 = \sum_i p_i \log_2 M_i = \sum_i p_i \log_2 p_i M = \sum_i p_i \log_2 p_i + \sum_i p_i \log_2 M = \log_2 M + \sum_i p_i \log_2 p_i.$$

Since $H_1 + H_2 = \log_2 M$, it follows that

$$H_1 = -\sum_i p_i \log_2 p_i$$

which is known as *Shannon's formula*. In the following section we will make this informal discussion mathematically rigorous.

2.2 Entropy

The idea of entropy originated in statistical mechanics. Roughly, given a thermodynamic system, such as a gas or a liquid, if we know some global properties of the system, e.g temperature, volume, energy, there are many different *microstates*, that is configurations of the individual particles within the system, which are consistent with these measurements.

As an example imagine flipping 1000 coins. We have a global measurement, the number of heads, but for each particular value for this, there are many different configurations of the specific states each of the 1000 coins landed in which achieve this number of heads.

Under a broad assumption that each of these microstates are equally likely, Boltzmann defined entropy of the system to be $k_B \log(\#$ of microstates) where k_B is some constant. Gibbs generalized this to microstates with unequal probabilities and gave the formula

$$S = -k_B \sum p_i \log(p_i),$$

where S is the *entropy*, p_i is the probability of the ith microstates, and the sum ranges over all the microstates. This reduces to Boltzmann's formula when the p_i are equal.

The second law of thermodynamics states that the entropy of an isolated system never decreases, and so such systems naturally 'tend' towards the state with maximum entropy, known as thermodynamic equilibrium. This was an attempt to formalise the idea that there is a natural 'direction' to natural processes, for example to explain why heat is transferred from hotter objects to cooler objects, rather than the other way round (which would not by itself contradict the conservation of energy in a process).

In the early 20th Century Hartley and Shannon found that similar equations arise naturally in the study of information theory, and at the suggestion of Von Neumann, Shannon also named it *entropy*.

"You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage." - John von Neumann

Definition 2.4 (Entropy)**.** Let X be a discrete random variable taking values in a *finite* set $\mathcal{X}$, and let

$$p(x) = p_X(x) = \mathbb{P}[X = x]$$

be the distribution of X. The *entropy* of X, which we will also call the entropy of the distribution p, is defined as

$$H(X) = H(p) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x). \tag{2.1}$$

If we enumerate $\mathcal{X} = \{x_1, \ldots, x_n\}$ and set $p_k = p(x_k)$ then we might also use the following notation, that $p = (p_1, \ldots, p_n)$ and

$$H(p) = H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Remark 2.5. *For ease of notation, it will often be convenient to define*

$$0 \log 0 := 0,$$

whenever it appears in such a sum.

Example 2.6. Suppose the X is uniformly distributed on a set $\mathcal{X} = \{x_1, \dots, x_n\}$ of size n, so that $p_k = p(x_k) = \frac{1}{n}$ for each k.

In this case

$$H(X) = H(p) = H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) = -\sum_{i=1}^{n} p_i \log_2 p_i = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -\log_2 \frac{1}{n} = \log_2 n.$$

Another way to think of entropy is as a measure of the expected amount of information we gain from learning the value of X. Indeed, suppose we have some function $g(x)$ which measure the information we gain from learning that X takes the value x. We clearly gain more information by knowing that a low probability event happens, so this function $g(x)$ should a decreasing function of $p(x)$. In fact, as we will see later, there are other natural assumptions to make about $g(x)$ which, similar to Hartley's formula, imply that the only 'reasonable' choice for this function g is to take $g(x) = -\log_2 p(x)$.

In this way, we can view (2.1) as an expectation - we have the weighted sum over some probability distribution of a quantity, where this quantity is the function $g : \mathcal{X} \to \mathbb{R}$ given by $g(x) = -\log_2 p(x)$ (note that this is a deterministic function, even though it encodes the probability distribution of the random variable X), then (2.1) can be rewritten as

$$H(X) = \sum_{x \in \mathcal{X}} p(x) g(x) = \mathbb{E}(g(X)) = \mathbb{E}(-\log_2 p(X)),$$

and it represents the expected amount of information we gain from learning the value of X.

Let us collect a few basic facts about the entropy function, some of which are obvious and some of which we will prove formally later.

Remark 2.7. *(1) $H(X) \geq 0$, with equality if and only if X is constant.*

(2) $H(X)$ doesn't depend on the values *of the random variable X, just the distribution of probabilities between these values. In other words, if we relabel the outcomes, that is, if we take some bijection $f : \mathcal{X} \to \mathcal{X}'$ and let $X' = f(X)$, then $H(X') = H(X)$.*

In other words if $(p_1, \dots, p_n)$ and $(p'_1, \dots, p'_n)$ are the same up to some permutation, then

$$H(p_1, \dots, p_n) = H(p'_1, \dots, p'_n).$$

(3) The function $p_1 \mapsto H(p_1, 1 - p_1)$ is continuous for $p_1 \in [0, 1]$. Furthermore this function is symmetric, takes values 0 at $p_1 = 0, 1$ and is maximised at $p_1 = \frac{1}{2}$ where it takes the value 1.

(4) More generally, for fixed n,

$$\max\{H(p) \colon p = (p_1, \dots, p_n)\} = H\left(\frac{1}{n}, \dots, \frac{1}{n}\right) = \log_2 n,$$

which the last equality is equivalent to Hartley's formula.

In other words, the uniform distribution has the maximum expected uncertainty, or the maximum expected information.

A discrete random variable X *determines* another jointly distributed random variable Y if there is an function $f : \mathcal{X} \to \mathcal{Y}$ such that $Y = f(X)$

Lemma 2.8. *Let X and Y be jointly distributed discrete random variables taking finitely many values such that X determines Y. Then $H(Y) \leq H(X)$.*

Proof. □

In particular, if X determines Y and Y determines X, then $H(X) = H(Y)$.

Given jointly distributed discrete random variable X and Y, taking values in finite sets $\mathcal{X}$ and $\mathcal{Y}$, as mentioned the random vector $Z = (X, Y)$ is again a discrete random variable and we can define the *joint entropy* of X and Y as the entropy of Z. That is, since for any $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$

$$\mathbb{P}[Z = z] = \mathbb{P}[X = x, Y = y] = p_{X,Y}(x, y),$$

we can calculate

$$H(X, Y) := H(Z) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x, y) \log_2 p(x, y)$$

Similarly, given $x \in \mathcal{X}$ we can consider the conditional distribution of Y, given that $X = x$, which we recall is

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

assuming that $p_X(x) > 0$. We can thus write the entropy of this conditional distribution as

$$H(Y \mid X = x) := -\sum_{y\in\mathcal{Y}} p(y|x) \log_2 p(y|x) = -\sum_{y\in\mathcal{Y}} \frac{p(x, y)}{p(x)} \log_2 \frac{p(x, y)}{p(x)}.$$

Example 2.9. Suppose X and Y are both distributed on $\{1, 2, 3\}$ and have joint distribution given by

	X=1	X=2	X=3
Y=1	1/8	1/4	0
Y=2	1/8	1/8	1/4
Y=3	0	1/8	0

so that $p_X = (1/4, 1/2, 1/4)$ and $p_Y = (3/8, 1/2, 1/8)$. In this case we can calculate

$$H(X) = \frac{1}{4} \log_2 4 + \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 = \frac{3}{2}.$$

$$H(Y) = \frac{3}{8} \log_2 \frac{8}{3} + \frac{1}{2} \log 2 + \frac{1}{8} \log_2 8 = 2 - \frac{3}{8} \log_2 3.$$

$$H(X, Y) = 4 \cdot \frac{1}{8} \log_2 8 + 2 \cdot \frac{1}{4} \log_2 4 = \frac{5}{2}.$$

$$H(Y|X = 1) = \frac{1}{2} \log_2 2 + \frac{1}{2} \log_2 2 + 0 \log_2 0 = 1.$$

Definition 2.10 (Conditional entropy)**.** The conditional entropy of a discrete random variable Y given a discrete random variable X, both taking finitely many values, is the average value of $H(Y|X = x)$ with respect to the possible values of X

$$\begin{aligned} H(Y \mid X) &= \sum_{x \in \mathcal{X}} p_X(x) H(Y \mid X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \\ &= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)} \end{aligned}$$

So in the previous example we can calculate

$$H(Y \mid X) = \frac{1}{8}\log_2 2 + \frac{1}{8}\log_2 2 + \frac{1}{4}\log_2 2 + \frac{1}{8}\log_2 4 + \frac{1}{8}\log_2 4 + \frac{1}{4}\log_2 1 = 1.$$

We can think of $H(Y|X)$ as the expected amount of information contained in the random variable Y if we already know the value of X. In particular, if $H(X, Y)$ represents the expected total information in both X and Y, then since discovering the value of X and Y is the same as first discovering the value of X, and then discovering the value of Y, heuristically it should be the case that $H(X, Y) = H(X) + H(Y \mid X)$, and indeed in the example above

$$\frac{5}{2} = H(X, Y) = H(X) + H(Y \mid X) = \frac{3}{2} + 1.$$

It should heuristically be true that conditioning can only decreases the entropy, and indeed this is the case. Later we will show a far more general statement.

Lemma 2.11. *For any two jointly distributed discrete random variables X and Y taking finitely many values $H(Y) \geq H(Y \mid X)$.*

Proof. □

Theorem 2.12 (Chain rule)**.** *For any two jointly distributed discrete random variables X and Y taking finitely many values*

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y).$$

Proof. □

More generally, using Theorem 2.12 it can be shown by induction that the following holds.

Theorem 2.13 (Chain rule)**.** *For any sequence of discrete random variables $X_1, X_2, \ldots, X_n$ taking finitely many values*

$$H(X_1, \ldots, X_n) = \sum_{k=1}^{n} H(X_k \mid X_{k_1}, \ldots, X_1).$$

In fact, the chain rule holds in a slightly more general form, for conditional entropies, which can be proved in much the same way.

Lemma 2.14 (Conditional Chain Rule)**.** *For any three jointly distributed random variables X, Y and Z taking finitely many values*

$$H(X, Y \mid X) = H(X \mid Z) + H(Y \mid X, Z)$$

Example 2.15. Suppose Z takes values in $\{1, \ldots, n\}$ and has probability distribution $(p_1, \ldots, p_n)$. Let us define two new random variables : $X = Z + \mathbb{1}_{[Z=1]}$ and $Y = \mathbb{1}_{[Z=1]}$. In particular, $p_X = (p_1 + p_2, p_3, \ldots, p_n)$ and $p_Y = (1 - p_1, p_1)$.

Now, since $Z = X - Y$, the pair (X, Y) determine Z, and clearly X and Y are determined by Z, and so

$$H(X,Y) = H(Z) = H(p_1, \ldots, p_n), \qquad H(X) = H(p_1{+}p_2, p_3, \ldots, p_n), \qquad H(Y) = H(1{-}p_1, p_1).$$

Now, if $X = x \geq 3$, then $Y = 0$ and so $H(Y \mid X = x) = 0$. If $X = 2$, which happens with probability $p_X(2) = p_1 + p_2$, then Z is either 1 or 2, with probabilities p_1 and p_2, and so Y is either 1 or 0 with the same probabilites, that is,

$$p_{Y|X}(1 \mid 2) = \frac{p_1}{p_1 + p_2} \qquad \text{and} \qquad p_{Y|X}(0 \mid 2) = \frac{p_2}{p_1 + p_2}.$$

Hence we can calculate,

$$H(Y|X) = \sum_{i=2}^{n} p_X(i) H(X|Y = i) = (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right),$$

and the chain rule $H(X, Y) = H(X) + H(Y|X)$ in this case implies

$$H(p_1, \ldots, p_n) = H(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right), \tag{2.2}$$

which holds for any probability distribution $(p_1, \ldots, p_n)$ (if we interpret the second term as 0 when $p_1 + p_2 = 0$).

For Hartley's formula, there was a heuristic collections of axioms that determine what we should expect from a measure of uncertainty that in fact determined Hartley's formula as the unique way to capture these axioms mathematically. It turns out that there is a similar axiomatic basis for Shannon's formula, in terms of some natural axioms that any measure of expected uncertainty should satisfy.

Theorem 2.16. *Let $\mathcal{B} = \{(p_1, \ldots, p_n) \colon n \in \mathbb{N}, p_k \geq 0$ for all $k \leq n, p_1 + \ldots + p_n = 1\}$ be the set of all finite probability distributions. Suppose that we have some function $H : \mathcal{B} \to \mathbb{R}$ which satisfies the following axioms:*

(I) H is transposition invariant : if $1 \leq i < j \leq n$ then

$$H(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_n) = H(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_n).$$

(II) Normalisation : $H(1/2, 1/2) = 1$.

(III) Continuity : The function $p_1 \to H(p_1, 1-p_1)$ is continuous

(IV) Equation (2.2) *holds for all $p \in \mathcal{B}$ with $n \geq 2$.*

Then

$$H(p_1, \ldots, p_n) = -\sum_{k=1}^{n} p_k \log_2 p_k.$$

Proof. For mathematical students only. □

In order to do this we will need the following variant of Lemma 2.3

Proposition 2.17. *The function $H(U_N) = \log_2 N$ is the unique function satisfying properties (A),(C) of Lemma 2.3 and the following variant of property (B):*

(B)* $\lim_{N\to\infty} H(U_{N+1}) - H(U_N) = 0$.

2.3 Kullback-Leibler Divergence and Mutual Information

Definition 2.18. Let p and q be probability distributions on the same finite set $\mathcal{X}$. The *relative entropy* or *Kullback-Leibler Divergence* of p with respect to q is

$$D(p \parallel q) = \sum_{x\in\mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} = \mathbb{E}\left(\log_2 \frac{p(X)}{q(X)}\right),$$

where X is some random variable with distribution p.

Remark 2.19. *Again here we need some convenition to deal with the cases where the quantity $p(x)\log_2 \frac{p(x)}{q(x)}$ is not defined. If $p(x) = 0$ then we define*

$$0 \log_2 \frac{0}{q(x)} = 0 \text{ for any } q(x) \geq 0,$$

and if $p(x) \neq 0, q(x) = 0$ we define

$$p(x) \log_2 \frac{p(x)}{q(x)} = \infty \text{ for any } p(x) > 0.$$

In particular, if there is any $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p \parallel q) = \infty$.

This quantity is also sometimes called the *Kullbakc-Liebler distance*, however one should be careful that this function does not behave as we would expect a distance function to behave - in particular, it is not always *symmetric* and it does not satisfy the *triangle inequality*. In fact, it is not even obvious that this quantity is *non-negative*, although we will later show that this is the case.

Example 2.20. Let $\mathcal{X} = \{0, 1\}$, $p = (p_1, p_2)$ and $q = (q_1, q_2)$, with $p_1 + p_2 = q_1 + q_2 = 1$. Then

$$D(p \parallel q) = p_1 \log_2 \frac{p_1}{q_1} + p_2 \log_2 \frac{p_2}{q_2}.$$

For example, for $p = (1/2, 1/2)$ and $q = (1/4, 3/4)$ we can compute

$$D(p \parallel q) = \frac{1}{2} \log_2 2 + \frac{1}{2} \log_2 \frac{2}{3} = 1 - \frac{1}{2}, \log_2 3$$

and

$$D(q \parallel p) = \frac{1}{4} \log_2 \frac{1}{2} + \frac{3}{4} \log_2 \frac{3}{2} = \frac{3}{4} \log_2 3 - 1.$$

Definition 2.21. Let X and Y be two jointly distributed discrete random variables taking values in finite sets $\mathcal{X}$ and $\mathcal{Y}$. The *mutual information* of X and Y is defined as

$$I(X \ ; \ Y) = D(p_{X,Y} \parallel p_X \otimes p_Y)$$

where $p_X \otimes p_Y(x, y) = p_X(x)p_Y(y) = \mathbb{P}[X = x]\mathbb{P}[Y = y]$.

In other words, if we thinkg of the Kullback-Liebler divergence as a distance between distributions, the mutual information of X and Y measures how far their joint distribution is from the joint distribution of independent copies of X and Y, and so we can think of the mutual information as a measure of dependence between random variables.

Plugging Definition 2.18 into Definition 2.21 we get the following explicit formula for the mutual information

$$I(X \ ; \ Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{X,Y}(x, y) \log_2 \frac{p_{X,Y}(x, y)}{p_X(x)p_y(y)}.$$

In particular, if X and Y are independent, then the term inside the log is always one, and so $I(X \ ; \ Y) = 0$. The larger the mutual information, the further in some sense X and Y are from being independent.

Lemma 2.22. *Let X and Y be two jointly distributed discrete random variables taking finitely many values. Then*

$$I(X \ ; \ Y) = H(X) + H(Y) - H(X, Y) = H(Y) - H(Y \mid X) = H(X) - H(X|Y) = I(Y \ ; \ X).$$

In particular, $I(X \ ; \ X) = H(X)$.

Proof. □

Example 2.23 (Secure encryption)**.** Suppose we have a set of *messages* $\mathcal{M}$ that we wish to encrypt and a set of *keys* $\mathcal{K}$ that we can use to encrypt these messages. That is, every pair $m \in \mathcal{M}$ and $k \in \mathcal{K}$ of a message and a key can be used to generate some encrypted text $c \in \mathcal{C}$, or *ciphertext*.

Normally we have some (pseudo)-random method of generating keys $k \in \mathcal{K}$, which determines some random variable K on $\mathcal{K}$, and there is some underlying distribution M on the messages $\mathcal{M}$. An *encryption scheme* for M is a pair of random variables K and C, representing the key

and the encrypted text such that K and C together determine M. This last condition is just saying that we can decrypt the message given the key and the ciphertext.

A *classical encryption scheme* would consist of some deterministic function $e : \mathcal{M} \times \mathcal{K} \to \mathcal{C}$ such that for each $k \in \mathcal{K}$ the function $e(\cdot, k) \to \mathcal{C}$ is injective, and then taking $C = e(M, K)$.

What does it mean for an encryption scheme to be *secure*? We want that someone who doesn't know the key cannot infer any information about the message from the ciphertext. To put this in terms of entropy, we want that there is no mutual information between C and M.

Definition 2.24 (Perfectly secure encryption scheme)**.** An encryption scheme K, C for M is *perfectly secure* if $I(M\ ;\ C) = 0$.

There is an obvious example of a perfectly secure encryption scheme which is known as a *one-time pad*. We assume (essentially wlog) that $\mathcal{M} = \{0,1\}^n$ and that we have a uniformly distributed set of keys on the same set $\mathcal{K} = \{0,1\}^n$. We take then a classical encryption scheme where $e(m, k) = m + k$ where addition is taken in $\mathbb{Z}_2^n$.

Theorem 2.25. *The one time pad is perfectly secure.*

Proof. (Exercise) □

However this clearly isn't a very efficient method of encryption, since it requires the two parties to share a key which is as large as the message itself. However Shannon showed that this is essentially necessary for a secure encryption scheme, in the sense that, in an perfectly secure encryption scheme the set of keys must contain as least as much information as the messages.

Theorem 2.26. *If K, C is a perfectly secure encryption scheme for M then $H(K) \geq H(M)$.*

Proof. □

As a more concrete example, if both M and K are uniformly distributed then Theorem 2.26 says that

$$\log_2 |\mathcal{K}| = H(K) \geq H(M) = \log_2 |\mathcal{M}|.$$

That is, $|\mathcal{K}| \geq |\mathcal{M}|$ and so we need at least as many different keys as we have messages.

As with entropy, we can extend the concept of mutual information to conditional spaces.

Definition 2.27 (Conditional mutual information)**.** Let X, Y, Z be three jointly distributed discrete random variables taking finitely many values. The *conditional mutual information* of X and Y given Z is defined as

$$I(X\ ;\ Y \mid Z) = \sum_{z \in \mathcal{Z}} p_Z(z) I(X\ ;\ Y \mid Z = z)$$

Let us briefly clarify the meaning of the above definition. Suppose $p_{X,Y,Z}(x,y,z)$ is the joint distribution of the three random variables. We can define the joint distribution of X and Y conditioned on Z as

$$p_{X,Y|Z}(x,y|z) = \frac{p_{X,Y,Z}(x,y,z)}{p_Z(z)}.$$

Then

$$I(X\,;\,Y \mid Z=z) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{X,Y|Z}(x,y|z)\log_2 \frac{p_{X,Y|Z}(x,y,z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)},$$

so that

$$I(X\,;\,Y \mid Z) = \sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} p_{X,Y,Z}(x,y,z)\log_2 \frac{p_{X,Y|Z}(x,y,z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}.$$

Lemma 2.28. *Let X, Y, Z be three jointly distributed discrete random variables taking finitely many values. Then*

$$\begin{aligned} I(X\,;\,Y \mid Z) &= H(X \mid Z) + H(Y \mid Z) - H(X,Y \mid Z) \\ &= H(X \mid Z) - H(X \mid Y,Z) = H(Y \mid Z) - H(Y \mid X,Z). \end{aligned}$$

Proof. □

As a consequence it is easy to deduce the following variant of the chain rule for mutual information.

Theorem 2.29 (Chain rule for mutual information)**.** *Let $X_1, \dots X_n$ and Y be jointly distributed discrete random variables taking finitely many values. Then*

$$I(X_1,\dots,X_n\,;\,Y) = \sum_{k=1}^{n} I(X_k\,;\,Y \mid X_{k-1},\dots,X_1).$$

Proof. Exercise - Use Theorem 2.13 and induct on n. □

So far we have only proved various equalities about entropy, just by rearranging the formulas. At various points it will be useful to be able to *estimate*, that is, bound from above or below, entropies and related quantities, and a particularly useful tool for this come from *Jensen's inequality*. To state this inequality we will require a little background from analysis.

Definition 2.30 (Convex and concave)**.** Let $I \subseteq \mathbb{R}$ be an open interval. A function $f\colon I \to \mathbb{R}$ is *convex* if for every $x, y \in I$ and $\lambda \in (0,1)$

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \tag{2.3}$$

We can think of this as taking a weighted average $z = \lambda x + (1-\lambda)y$ of the points x and y, which lies somewhere between x and y. f is convex if the value of the function at this point is smaller than the same weighted average of $f(x)$ and $g(y)$.

Geometrically, this asserts that the line between $f(x)$ and $f(y)$ lies above the graph of the function f between x and y.

We say f is *strictly convex* if (2.3) is strict for any $x \neq y$. Similarly we say f is *concave* or *strictly concave* if the inequality in (2.3) is reversed.

Theorem 2.31 (Jensen's inequality)**.** *Let $f\colon I \to \mathbb{R}$ be a convex function on an open interval I and let X be a real random variable taking values in I. If $\mathbb{E}(X)$ and $\mathbb{E}(f(X))$ exist, then*

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}(X)).$$

Furthermore, if f is strictly convex then the inequality is strict unless X is almost surely constant.

Proof. □

Probably the most important application of Jensen's inequality in information theory is the following:

Theorem 2.32 (Information Inequality)**.** *Let p and q be distributions on a finite set $\mathcal{X}$. Then $D(p \parallel q) \geq 0$ with equality if and only if $p = q$.*

Proof. □

Let us note then some immediate, and incredibly useful, corollaries of this Theorem.

Corollary 2.33. *Let X, Y, Z and $X_1, \ldots, X_n$ be jointly distributed discrete random variables taking values in a finite set.*

1. *$I(X\,;\,Y) \geq 0$, with equality if and only if X and Y are independent,*
2. *$H(X \mid Y) \leq H(X)$, with equality if and only if X and Y are independent,*
3. *$I(X\,;\,Y \mid Z) \geq 0$, with equality if and only if X and Y are independent conditional upon Z,*
4. *$H(X_1, \ldots, X_n) \leq H(X_1) + \ldots + H(X_n)$ with equality if and only if the X_k are mutually independent.*

Proof. □

Another important consequence of Theorem 2.32 is the following.

Lemma 2.34 (Log sum inequality)**.** *Let $a_1, b_1, \ldots, a_n, b_n \geq 0$ and let $a = \sum_{k=1}^n a_k$ and $b = \sum_{k=1}^n b_k$. Then*

$$\sum_{k=1}^{n} a_k \log_2 \frac{a_k}{b_k} \geq a \log_2 \frac{a}{b}.$$

Proof. □

As a corollary we get a weird looking statement asserting a sort of multivariable concavity of the Kullback-Liebler divergence.

Corollary 2.35. *Let $p^{(1)}, p^{(2)}, q^{(1)}$ and $q^{(2)}$ be probability distributions on the same finite set $\mathcal{X}$ and let $\lambda \in (0,1)$.*

$$D(\lambda \cdot p^{(1)} + (1-\lambda) \cdot p^{(2)} \parallel \lambda \cdot q^{(1)} + (1-\lambda) \cdot q^{(2)}) \leq \lambda D(p^{(1)} \parallel q^{(1)}) + (1-\lambda) D(p^{(2)} \parallel q^{(2)})$$

Proof. □

As a simple corollary we find that the entropy function is also concave.

Corollary 2.36. *Let p, q be probability distributions on a finite set $\mathcal{X}$ and let $\lambda \in (0,1)$. Then*

$$H(\lambda p + (1-\lambda) q) \geq \lambda H(p) + (1-\lambda) H(q).$$

Proof. □

Remark 2.37. *From the inequality*

$$D(p \parallel u) = \log_2 |\mathcal{X}| - H(p) \geq 0,$$

we can conclude from Theorem 2.32 that $H(p) \leq \log_2 |\mathcal{X}|$ with equality if and only if p is the uniform distribution.

Suppose we are given the conditional distribution of a random variable Y with respect to a random variable X, but no the distribution of X or Y. That is, we have the function

$$p_{Y|X}(y, x) = p(y|x),$$

where for each x we can think of the function $p(\cdot|x)$ as a distribution on $\mathcal{Y}$.

Then, any probability distribution p_X on $\mathcal{X}$ gives rise to a joint distribution $p_{X,Y}$ via

$$p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y, x),$$

from which we can also derive the distribution $p(y) = \mathbb{E}_{p_X} p(y|X) = \sum_{x \in \mathcal{X}} p(x) p(y|x)$.

In other words, if we fix ahead of time the conditional distribution, then the joint distribution (X, Y), and so in particular the distribution of Y, is determined by the distribution of X. We can think of this as the evolution of a random process in time, where X and Y represent the outcome at two specific times. In order to calculate the probability that we saw the outcome (x, y) we need to know the probability that at the first time we had the outcome x, and then the outcome that the process developed from x to become y at the second time.

We can think of this in terms of a *transition matrix* P, whose columns are indexed by $\mathcal{X}$ and rows by $\mathcal{Y}$ and whose entry $P_{xy} = p(y|x)$ is the probability that we observe y after observing x, the probability that our process evolves from x to become y. The rows of this matrix $p(\cdot|x)$ correspond to the conditional distribution of Y given that $X = x$, and so given a distribution p_X on $\mathcal{X}$ we can compute the marginal distribution on $\mathcal{Y}$ as $p_Y = p_X P$ and the joint distribution (as an $\mathcal{X} \times \mathcal{Y}$ matrix) can be seen to be the product $\mathrm{diag}(p_X) \cdot P$ of a diagonal matrix with entries $p_X(x)$ together with P, so that that xth row is $p_X(x) p(\cdot|x) = p_{X,Y}(x, \cdot)$.

Theorem 2.38. *Suppose $p_{Y|X}$ is the conditional distribution of some discrete random variable Y taking values in a finite set $\mathcal{Y}$ with respect to some unknown discrete random variable X taking values in a finite set $\mathcal{X}$. Then the function $f(p_X) = I(X\ ;\ Y) = D(p_{X,Y} \parallel p_X \otimes p_Y)$ is concave, that is, for any two distributions p and q on $\mathcal{X}$ and $\lambda \in (0,1)$*

$$f(\lambda \cdot p + (1-\lambda)\cdot q) \geq \lambda f(p) + (1-p)f(q).$$

Conversely, if p_X is known, then the function $F(p_{Y|X}) = I(X\ ;\ Y)$ is convex.

Proof. □

Definition 2.39. Let X, Y and Z be jointly distributed discrete random variables taking values in a finite set. The triple (X, Y, Z) is called a *Markov(ian) triple*, which we write as $X \to Y \to Z$, if for all x, y with $\mathbb{P}[X = x \mid Y = y] > 0$

$$\mathbb{P}[Z = z \mid X = x, Y = y] = \mathbb{P}[Z = z \mid Y = y]$$

If we think of $X \to Y \to Z$ as the evolution of some random process over time, then this says that if we know the "present", the value of Y, then the future evolution does not depend on the past.

For a Markov triple we can write the joint distribution as the product

$$p_{X,Y,Z}(x,y,z) = p_X(x)\cdot p_{Y|X}(y \mid x)\cdot p_{Z|X,Y}(z \mid x,y) = p_X(x)\cdot p_{Y|X}(y \mid x)\cdot p_{Z|Y}(z \mid y). \quad (2.4)$$

There is another nice equivalent description in terms of the conditional distributions.

Lemma 2.40. *(X, Y, Z) is a Markovian triple if and only if X and Z are conditionally independent, given Y. That is, if whenever $p_Y(y) > 0$,*

$$p_{X,Z|Y}(x, z \mid y) = p_{X|Y}(x \mid y)p_{Z|Y}(z \mid y).$$

Proof. □

Note that the condition in Lemma 2.40 is symmetric in X and Z. In other words, we see that $X \to Y \to Z$ is a Markovian triple if and only if $Z \to Y \to X$ is as well.

If we think about our random process as some method of processing some data, our input data X is encoded or transmitted say, and we have as output data Y. Then, if forget about our original data X, there should be no way to increase the amount of information about X that the output Y contains by further processing (via some deterministic, or random, process).

As an example, we can think about the tranmission of some message X from Alice to Bob, who receives the transmitted message Y (perhaps some random noise has been added in the channel). Without access to the message X, there should be no way that Bob can deduce more information about X than what is contained in Y.

Theorem 2.41 (Data processing inequality)**.** *If $X \to Y \to Z$, then $I(X\ ;\ Z) \leq I(X\ ;\ Y)$.*

Proof. □

Note that, since $Z \to Y \to X$ is also a Markovian triple, we can also deduce from Theorem 2.41 that $I(Z\ ;\ Y) \geq I(Z\ ;\ X) = I(X\ ;\ Z)$, and so

$$I(X\ ;\ Z) \leq \max\{I(X\ ;\ Y), I(Y\ ;\ Z)\}.$$

Suppose, in the previous example, Alice transmit a message X to Bob, who receives Y, and Bob wishes to reconstruct the message X from Y, via some process. Perhaps Alice is sending pictures of some letters, and Bob receives slightly perturbed pictures, and so has to guess which letter fits the picture best. This might be some deterministic process, or maybe when Bob is unsure he makes some random choice, weighted by how likely he thinks each letter is. In this way Bob makes a (potentially random) guess $Z = \hat{X}$ based on Y, and we have a Markov triple $X \to Y \to Z$.

How accurate can Bob be? There are many ways we could measure this, but one way would be to look at the probability that Bob's guess is correct, the probability that $\hat{X} = X$.

Now, from the Data processing inequality, we know that if lots of information is lost in transmission, and so $I(X\ ;\ Y)$ is small, it shouldn't be the case that $\hat{X}$ is well-correlated with $\hat{X}$, since $I(X\ ;\ \hat{X})$, which is a measure of dependence, is also small. So, we should expect to be able to bound the probability of success, as some function of the mutual information $I(X\ ;\ Y)$, and the following inequality makes this precise (in fact, we bound instead the probability of failure, as a function of the conditional entropy $H(X \mid Y)$).

Theorem 2.42 (Fano's inequality)**.** *Let $X \to Y \to \hat{X}$ be a Markov triple, where we think of $\hat{X}$ as an estimation of X on the basis of Y. Define $p_{err} = \mathbb{P}[\hat{X} \neq X]$. Then*

$$H(p_{err}, 1 - p_{err}) + p_{err} \log_2 |\mathcal{X}| \geq H(X \mid \hat{X}) \geq H(X \mid Y).$$

In particular

$$p_{err} \geq \frac{H(X \mid \hat{X}) - 1}{\log_2 |\mathcal{X}|} \geq \frac{H(X \mid Y) - 1}{\log_2 |\mathcal{X}|}.$$

Proof. □

It is reasonable to ask why we mention also the weaker bounds in terms of $H(X \mid Y)$ rather than $H(X \mid \hat{X})$. This is useful if we're interested in an *a priori estimate*, one that only depends on the message X and the transmitted message Y, and not the method of reconstruction $\hat{X}$. This bound then holds for *every* possible $\hat{X}$ - no matter how Bob attempts to guess the message X, he must always have a failure probability of at least this quantity.

Remark 2.43. *All the material in this section extends straightforwardly to arbitrary discrete random variables, respectively probability distributions, taking values in countable sets.*

That is, given a random variable X taking values in $\mathcal{X} = \{x_k \colon x \in \mathbb{N}\}$ with distribution $p = (p_1, p_2, \ldots)$ we can define the entropy

$$H(X) = H(p) = -\sum_{k=1}^{\infty} p_k \log_2 p_k,$$

which may also take the value $+\infty$.

3 Entropy Rate and Asymptotic Equipartition

3.1 Entropy Rate

Definition 3.1 (Stochastic process, state space)**.** A *stochastic process* in *discrete time* is a sequence $(X_n)_{n\in\mathbb{N}}$ of jointly distributed random variables. The *state space* of the process is the set $\mathcal{X}$ of possible values which the X_n can take.

Example 3.2. Pick a random page of a book and let X_n be the nth letter on the page.

Let $X_1 = 100$ be constant, and let X_n be the bankroll after n spins of a roulette wheel of a gambler who bets his entire stake on red each time.

Let $X_1 = (0,0) \in \mathbb{Z}^2$ and let X_n be the position after n steps of a 'random walk', where in each time step we choose uniformly at random one of the four neighbours in the grid of our current position and move there.

Given a stochastic process $(X_n)_{n\in\mathbb{N}}$ we can think of $H(X_1,\dots,X_n)$ as the total amount of information in the system at time n. This is clearly an increasing function of n. What we're interested in is the rate at which this function is increasing.

Definition 3.3 (Entropy rate)**.** If $(X_n)_{n\in\mathbb{N}}$ is a stochastic process whose state space is finite, then the *entropy rate* or *asymptotic entropy* of the stochastic process is defined as

$$h := \lim_{n\to\infty} \frac{1}{n} H(X_1,\dots,X_n),$$

if the limit exists. The unit of h is *bits per time unit.*

We will see that the entropy rate represents a theoretical limit on how efficiently we can encode the data stream $(X_n)_{n\in\mathbb{N}}$. Conversely, we will find for a broad class of processes, we can achieve this theoretical limit asymptotically, using the idea of asymptotic equipartition.

We can think of h as the average amount of new information introduced in each step of the stochastic process. Indeed, by the chain rule (Theorem 2.13)

$$\frac{1}{n} H(X_1,\dots,X_n) = \frac{1}{n} \sum_{k=1}^{n} H(X_k \mid X_1,\dots,X_{k-1}), \tag{3.1}$$

where $H(X_k \mid X_1,\dots,X_{k-1})$ is the amount of information introduced at the kth step.

What we will find is that the entropy rate represents a theoretical limit on how efficiently we can encode the data stream $(X_n)_{n\in\mathbb{N}}$. Conversely, the idea of asymptotic equipartition is that for a broad class of processes, we can achieve this theoretical limit asymptotically.

Lemma 3.4. *If $(X_n)_{n\in\mathbb{N}}$ is a stochastic process whose state space is finite and the limit $h' = \lim_{k\to\infty} H(X_k \mid X_1,\dots,X_{k-1})$ exists, then h exists and $h = h'$.*

Proof. □

For i.i.d sequences, it is trivial to compute h using Lemma 3.4

Lemma 3.5. *If $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d discrete random variables taking values in a finite set, then the entropy rate h exists and is equal to $H(X_1)$.*

Proof. □

Definition 3.6. A stochastic process $(X_n)_{n\in\mathbb{N}}$ is *stationary* if for every $\ell, k \in \mathbb{N}$ the two random vectors

$$(X_1, \ldots, X_\ell) \qquad \text{and} \qquad (X_{k+1}, \ldots, X_{k+\ell})$$

have the same distribution. In other words, for every choice of elements $x_1, \ldots, x_\ell \in \mathcal{X}$ in the state space,

$$\mathbb{P}[X_1 = x_1, \ldots, X_\ell = x_\ell] = \mathbb{P}[X_{k+1} = x_1, \ldots, X_{k+\ell} = x_\ell].$$

Example 3.7. The simplest example of stationary processes are sequences of independent and identically distributed random variables. For example, if we repeatedly roll a dice and let X_n be the value of the nth roll.

As another example, suppose we have two biased coins with different probabilities p and q of heads, and I choose randomly, say with probability $\frac{1}{2}$ one of the coins to flip, and then let X_n be the (bit) value of the nth coin toss. Then the X_n are not independent, if p is very close to one and q is very close to 0, then if $X_1 = 1$, it's very likely that I picked the first coin and so very likely that $X_2 = 1$ as well. However, it is easy to show that this process is stationary.

The random walk on $\mathbb{Z}^2$ from the previous example is not stationary - X_0 is deterministic, whereas X_1 is uniformly distributed on $\{(\pm 1, 0), (0, \pm 1)\}$.

Lemma 3.8. *If $(X_n)_{n\in\mathbb{N}}$ is a stationary process with a finite state space, then the entropy rate h exists.*

Proof. □

Whilst Lemma 3.8 asserts the existence of the entropy rate for stationary processes, it is non-constructive - it does not provide us a formula to calculate h. It is reasonable to ask if there is a broader class of stochastic processes (than i.i.d) for which we can compute h explicitly.

3.2 Time-homogeneous Markov Chains

Definition 3.9. A stochastic process $(X_n)_{n\geq 0}$ with finite state space $\mathcal{X}$ is a *Markov chain* (MC) if for all $n \in \mathbb{N}$ and for all $x_0, \ldots, x_n \in \mathcal{X}$,

$$\mathbb{P}[X_n = x_n \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}] = \mathbb{P}[X_n = x_n \mid X_{n-1} = x_{n-1}] := p_n(x_n|x_{n-1}),$$

whenever $\mathbb{P}[X_1 = x_0, \ldots, X_{n-1} = x_{n-1}] > 0$.

A Markov chain is *time-homogeneous* if $p_n(y|x) := p(y|x)$ does not depend on n. In this case the matrix $P = \Big(p(y|x)\Big)_{x,y\in\mathcal{X}}$ is the *transition matrix* of the time-homogeneous Markov chain and the distribution of X_0

$$\nu(x) = \mathbb{P}[X_0 = x]$$

is the *initial distribution* or *starting distribution.*

We note that the transition matrix P is always a *stochastic* matrix - the entries are all non-negative and each row sums to one, that is for all $x \in \mathcal{X}$

$$\sum_{y \in \mathcal{X}} p(y|x) = 1.$$

We can think of a Markov chain as a *memoryless* stochastic process - given the state of the process at some time n, the future distribution does not depend on the past. In particular, it is easy to check that each consecutive triple is Markovian and so

$$X_0 \to X_1 \to \ldots \to X_n.$$

Example 3.10. Suppose we're playing some board game with a number of possible states $\mathcal{X}$. Each turn we roll a dice and play according to some fixed strategy, so that the probability that we move from a state x to a state y in any particular turn is fixed. The state X_n of some player is then a time-homogeneous Markov chain.

A random walk is also an example of a time-homogeneous Markov chain - if we are currently at a vertex x the probability that we move to a vertex y only depends on the current state, and not the history of the walk.

Lemma 3.11. *If $(X_n)_{n\geq 0}$ is a Markov chain, then for any $k \in \mathbb{N}$ $((X_n, X_{n+1}, \ldots, X_{n+k}))_{n\geq 0}$ is a Markov chain.*

Proof. Exercise. □

Definition 3.12 (The (di)graph of a Markov chain)**.** Given a time-homogeneous Markov chain $(X_n)_{n\geq 0}$ with state space $\mathcal{X}$, we can draw an associated (weighted) (di)graph whose vertex set is $\mathcal{X}$ and for any two states x and y we draw an arc from x to y with weight $p(y|x)$ if $p(y|x) > 0$.

We can think of the Markov chain as a simple random walk on this graph, where the probability of moving from state x to y is given by the weight of the arc from x to y and the distribution of the starting vertex is given by X_0.

Example 3.13. We can think of the following simplified model of the evolution of the weather. Our stochastic process has three state $\mathcal{X} = \{\text{sun, rain, snow}\} = \{N, R, S\}$

Our transition matrix is given as follows

$$P = \begin{array}{c|ccc} & \text{N} & \text{R} & \text{S} \\ \hline \text{N} & 0 & 1/2 & 1/2 \\ \text{R} & 1/4 & 1/2 & 1/4 \\ \text{S} & 1/4 & 1/4 & 1/2 \end{array}.$$

So, we never have two sunny days in a row - if a day is sunny then the next day is equally likely to be rainy or snowy. On rainy or snow days the next day has probability $1/2$ to have the same weather, and probability $1/2$ to change to one of the other options uniformly.

In this case the digraph of this Markov chain has vertex set $V = \{N, R, S\}$ and arcs

$$e_1 = (N, R),\ e_2 = (N, S),\ e_3 = (R, N),\ e_4 = (R, R),$$
$$e_5 = (R, S),\ e_6 = (S, N),\ e_7 = (S, R),\ e_8 = (S, S).$$

with weights

$$w(e_1) = 1/2,\ w(e_2) = 1/2,\ w(e_3) = 1/4,\ w(e_4) = 1/2,$$
$$w(e_5) = 1/4,\ w(e_6) = 1/4,\ w(e_7) = 1/4,\ w(e_8) = 1/2.$$

.

By the Markovian property it is relatively easy to write down the joint distribution of $(X_0, \ldots, X_n)$ as

$$\mathbb{P}[X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n] = \nu(x_0)p(x_1|x_0)p(x_2|x_1)\ldots p(x_n|x_{n-1}), \tag{3.2}$$

and from this it is also clear what the conditional distribution of X_n, given X_0 is.

Lemma 3.14. *Let $(X_n)_{n\geq 0}$ be a Markov chain with initial distribution ν and transition matrix P. Then*

$$p^{(n)}(y|x) := \mathbb{P}[X_n = y \mid X_0 = x] = (P^n)_{xy},$$

that is, the matrix given by $\left(p^{(n)}(y|x)\right)_{x,y\in\mathcal{X}}$ is the nth power of P.

If we consider ν as a row vector, then $p_{X_n} = \nu P^n$, that is

$$p_{X_n}(y) = \sum_{x\in\mathcal{X}} \nu(x)p^{(n)}(y|x).$$

Proof. □

We will show that for a natural class of time-homoegenous Markov chains the entropy rate exists, and can be easily calculated, and furthermore is independent of the choice of the initial distribution ν.

The existence of the entropy rate would be clear if the Markov chain were stationary, by Lemma 3.8. However, whilst is easy to verify that every stationary Markov chain is time-homogeneous (exercise). The converse is not true in general, but will hold for sensible choices of initial distribution.

Lemma 3.15. *Let $(X_n)_{n\geq 0}$ be a time-homogeneous Markov chain with initial distribution ν and transition matrix P. Then the Markov chain is stationary if and only if $\nu P = \nu$, that is, only if ν is an eigenvector of P with eigenvalue one.*

Proof. □

In this case we call ν a *stationary distribution* for P, or for the Markov chain.

Lemma 3.16. *Let $(X_n)_{n\geq 0}$ be a time-homogeneous Markov chain with a stationary initial distribution ν. Then the entropy rate exists and is given by*

$$h = \sum_{x\in\mathcal{X}} \nu(x) H(p(\cdot|x)),$$

where

$$H(p(\cdot|x)) = -\sum_{y\in\mathcal{X}} p(y|x)\log_2 p(y|x)$$

is the entropy of the probability vector which is the row of the transition matrix P indexed by x.

Proof. □

We note that there is a trivial right eigenvector of P with eigenvalue one given by the all ones vector $\mathbf{1} = (1,\ldots,1)$. Indeed, since P is stochastic, for all $x \in \mathcal{X}$

$$(P\mathbf{1})_x = \sum_{y\in\mathcal{X}} p(y|x) = 1.$$

More generally, a function $f : \mathcal{X} \to \mathbb{R}$ is called *harmonic* (with respect to P) if, when viewed as a column vector, it satisfies $Pf = f$. Since

$$(Pf)_x = \sum_{y\in\mathcal{X}} p(y|x)f(y) = \sum_{y:\ p(y|x)>0} p(y|x)f(y),$$

we can think of this as saying that the weighted average of the function f over the neighbourhood of x in the digraph of the Markov chain is equal to $f(x)$.

Hence, since the left and right eigenvalues of a matrix agree, there must be *some* vector ν which is a left eigenvector of P with eigenvalue one. It remains to show that ν is a probability vector. Arranging that ν sums to one is trivial, any linear scaling of an eigenvector lies in the same eigenspace, however it is not obvious that ν is non-negative.

Lemma 3.17. *Let $(X_n)_{n\geq 0}$ be a time-homogeneous Markov chain with a finite state space $\mathcal{X}$ and transition matrix P. Then there is at least one stationary probability distribution ν for P.*

Proof. □

Note that the same argument would apply to any accumulation point μ of the sequence μ_n. Can we say when this stationary distribution is unique?

Definition 3.18. Let $(X_n)_{n\geq 0}$ be a time-homogeneous Markov chain with a finite state space $\mathcal{X}$ and transition matrix P. The Markov chain, and transition matrix, are called *irreducible* if for every pair $x, y \in \mathcal{X}$ there is some $n \in \mathbb{N}$ such that $p^{(n)}(y|x) > 0$. That is, for any pair of states, there is some n such that we can transition from one state to the other in n steps with positive probability.

If we think about the associated digraph of the Markov chain, then irreducibility is equivalent to the property that this digraph is strongly connected - for any pair of vertices x and y there is a directed path from x to y.

Proposition 3.19. *Let $(X_n)_{n\geq 0}$ be a irreducible time-homogeneous Markov chain with a finite state space $\mathcal{X}$ and transition matrix P. Then there is a unique stationary distribution ν and furthermore $\nu(x) > 0$ for all $x \in \mathcal{X}$.*

Proof. □

Corollary 3.20. *Let $(X_n)_{n\geq 0}$ be a irreducible time-homogeneous Markov chain with a finite state space $\mathcal{X}$ and transition matrix P. Then*

$$\lim_{n\to\infty} \frac{1}{n} \left(\mu + \mu P + \mu P^2 + \ldots + \mu P^{n-1}\right)$$

exists and is equal to the unique stationary distribution ν.

Proof. □

Corollary 3.21. *Let $(X_n)_{n\geq 0}$ be a irreducible time-homogeneous Markov chain with a finite state space $\mathcal{X}$ and transition matrix P. Then for any initial distribution μ, the entropy rate of the Markov chain exists and is equal to the entropy rate of the Markov chain under its unique stationary distribution (see Lemma 3.16).*

Proof. □

Definition 3.22 (Return time)**.** If $(X_n)_{n\geq 0}$ is a Markov chain with a finite state space $\mathcal{X}$, then for any $x \in \mathcal{X}$ we define

$$\tau^x = \inf\{n \geq 1 \colon X_n = x\},$$

which is the first time the chain is in state x after the start (where we define $\inf \emptyset = \infty$). Note that τ_x is a random variable! If $X_0 = x$ then we call τ^x the *return time* to x. A state x is called *recurrent* if the Makrov chain returns to x almost surely, that is, if

$$\mathbb{P}[\tau^x < \infty \mid X_0 = x] = 1,$$

and it is *positive recurrent* if in addition the return time has finite expectation, that is,

$$\mathbb{E}\left(\tau^x \mid X_0 = x\right) < \infty.$$

Theorem 3.23 (Ergodic Theorem for Markov chains)**.** *Let $(X_n)_{n\geq 0}$ be a irreducible, time-homogeneous Markov chain with a finite state space $\mathcal{X}$. Then every state $x \in \mathcal{X}$ is positive recurrent, and the (unique) stationary distribution ν is given by*

$$\nu(x) = \frac{1}{\mathbb{E}\left(\tau^x \mid X_0 = x\right)}.$$

Furthermore, for any initial distribution μ and any function $f \colon \mathcal{X} \to \mathbb{R}$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) = \sum_{x\in\mathcal{X}} f(x)\nu(x) \qquad \textit{almost surely.}$$

The expression on the right hand side is a determinstic quantity, the *space average* of the value $f(x)$ - the average of f over the state space $\mathcal{X}$ under the stationary distribution ν. On

the left hand side we have a *random* quantity, the *time average* of f over the trajectory of the random process X_n and the theorem asserts that the two are almost surely equal.

In applications, we are often interested in the value of the right hand side, however if the state space is large then it can be hard, or inefficient to compute the space average directly. On the other hand, the time average can be approximated by simulating the Markov chain for a large number of steps and calculating the time average. In this way we get a tool for approximating the sum on the right hand side, which is know as the *Markov chain Monte Carlo* method.

3.3 The Asymptotic Equipartition Property

Let $(X_n)_{n\geq 1}$ be a stochastic process with a finite state space $\mathcal{X}$. For each $n \in \mathbb{N}$ we can consider the joint distribution $p_n = p_{X_1,\ldots,X_n}$ on $\mathcal{X}^n$, that is

$$p_n(x_1, \ldots, x_n) = \mathbb{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n].$$

Note that the sequence of distributions $(p_n)_{n\geq 1}$ (which are deterministic functions on $\mathcal{X}^n$), determines all the probabilistic characteristics of the stochastic process.

Suppose that the entropy rate of the stochastic process

$$h = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n\to\infty} \frac{1}{n} H(p_n)$$

exists. Since for any random variable X we can write

$$H(X) = \mathbb{E}(-\log_2 p_X(X)),$$

as the expected value of the deterministic function $-\log_2 \circ p_X$ applied to the random variable X, we can apply this to the random vector $(X_1, \ldots, X_n)$ to conclude that

$$\frac{1}{n} H(X_1, \ldots, X_n) = \mathbb{E}\left(-\frac{1}{n} \log_2 p_n(X_1, \ldots, X_n)\right).$$

Then the entropy rate h is the limit of the expected value of the random variables $Y_n = -\frac{1}{n} \log_2 p_n(X_1, \ldots, X_n)$. A much stronger property than the limit of the expectation existing would be that the random variables themselves converge (in probability or almost surely) to some limiting random variable with finite expectation h.

Definition 3.24 (Asymptotic equipartition property)**.** Let $(X_n)_{n\geq 1}$ be a stochastic process with a finite state space $\mathcal{X}$ whose entropy rate h exists. We say X has the *asymptotic equipartition property* (AEP), if

$$-\frac{1}{n} \log_2 p_n(X_1, \ldots, X_n) \longrightarrow h \text{ almost surely, as } n \to \infty.$$

The asymptotic equipartition property makes a very strong *prediction* about the observed outcome $(x_n)_{n\geq 1}$ of the stochastic process - with very high probability the quantity $-\frac{1}{n} \log_2 p_n(x_1, \ldots, x_n)$ will be close to the entropy rate h, where the error in probability and in approximation to h is tending to 0 with n.

Example 3.25. Let $\mathcal{X} = \{0, 1\}$ and let $(X_n)_{n\geq 1}$ be i.i.d with distribution $\mathrm{Ber}(\theta)$, that is,

$$\mathbb{P}[X_n = 1] = \theta \qquad \text{and} \qquad \mathbb{P}[X_n = 0] = 1 - \theta,$$

where $0 < \theta < 1$. For any bitstring $(x_1, \ldots, x_n) \in \{0, 1\}^n$, let

$$s_n = x_1 + \ldots + x_n,$$

be the total number of ones, so that $n - s_n$ is the total number of zeroes. Let us write S_n for the random variable $X_1 + \ldots + X_n$.

Since the sequence is i.i.d we can compute

$$p_n(x_1, \ldots, x_n) = \theta^{s_n}(1-\theta)^{n-s_n} \qquad \text{and so} \qquad p_n(X_1, \ldots, X_n) = \theta^{S_n}(1-\theta)^{n-S_n}.$$

Hence,

$$-\frac{1}{n}\log_2 p_n(X_1, \ldots, X_n) = -\frac{1}{n}\log_2\left(\theta^{S_n}(1-\theta)^{n-S_n}\right) = -\frac{S_n}{n}\log_2\theta - \left(1 - \frac{S_n}{n}\right)\log_2(1-\theta).$$

On the other hand, since the X_n are i.i.d, by Lemma 3.5 the entropy rate h exists and is equal to

$$h = H(X_1) = H(\theta, 1-\theta) = -\theta\log_2\theta - (1-\theta)\log_2(1-\theta).$$

So, in this case, the statement that $(X_n)_{n\geq 1}$ satisfies the AEP would be that almost surely

$$-\frac{S_n}{n}\log_2\theta - \left(1 - \frac{S_n}{n}\right)\log_2(1-\theta) \longrightarrow -\theta\log_2\theta - (1-\theta)\log_2(1-\theta).$$

Or, in other words, the AEP is equivalent to the statement that $\frac{S_n}{n} \to \theta$ almost surely, which is the strong law of large numbers.

In fact the argument above works for general for i.i.d sequences.

Lemma 3.26. *Let $(X_n)_{n\geq 1}$ be an i.i.d stochastic process with a finite state space $\mathcal{X}$. Then $(X_n)_{n\geq 1}$ satisfies the AEP, where the entropy rate $h = H(X_1)$.*

Proof. □

There are perhaps two natural questions to ask at this point :

(I) Which classes of stochastic process have the AEP (does this include a nice large natural class)?

(II) What is the practical application of knowing that we have the AEP?

Theorem 3.27. *Let $(X_n)_{n\geq 0}$ be a irreducible, time-homogeneous Markov chain with a finite state space $\mathcal{X}$. Then for any initial distribution, $(X_n)_{n\geq 0}$ satisfies the AEP, where the entropy rate h is given by the formula in Corollary 3.21 / Lemma 3.16.*

Proof. □

What can we conclude from the fact that the AEP holds? Since convergence almost surely implies convergence in probability, if the AEP holds then for any $\varepsilon > 0$

$$\mathbb{P}\left[\left|-\frac{1}{n}\log_2 p_n(X_1,\dots,X_n) - h\right| < \varepsilon\right] \to 1. \tag{3.3}$$

In other words, there is some *deterministic set* inside the set of trajectories $\mathcal{X}^n$, which we can specify ahead of time in terms of the deterministic function p_n and the quantity h, such that with very high probability the trajectory of the process lies inside this set.

Definition 3.28 (Typical set)**.** Given a stochastic process $(X_n)_{n\geq 1}$ which satisfies the AEP with entropy rate h. For every n and (small) $\varepsilon > 0$ the *typical set* is given by

$$A_\varepsilon^{(n)} = \left\{\boldsymbol{x} = (x_1,\dots,x_n) \in \mathcal{X}^n \colon \left|-\frac{1}{n}\log_2 p_n(x_1,\dots,x_n) - h\right| < \varepsilon\right\}.$$

The following properties of typical sets follow immediately from the definitions.

Proposition 3.29. *Given a stochastic process $(X_n)_{n\geq 1}$ which satisfies the AEP with entropy rate h. Then for all (small) $\varepsilon > 0$ the typical set has the following properties:*

(a) There exists $N(\varepsilon)$ such that for all $n \geq N(\varepsilon)$

$$\mathbb{P}[(X_1,\dots,X_n) \in A_\varepsilon^{(n)}] > 1 - \varepsilon.$$

(b) For all $\boldsymbol{x} = (x_1,\dots,x_n) \in A_\varepsilon^{(n)}$,

$$2^{-n(h+\varepsilon)} < p_n(\boldsymbol{x}) < 2^{-n(h-\varepsilon)}.$$

(c) The size of the typical set satisfies

$$(1-\varepsilon)2^{n(h-\varepsilon)} < \left|A_\varepsilon^{(n)}\right| < 2^{n(h+\varepsilon)},$$

where the second inequality holds for all n, and the first for all $n \geq N(\varepsilon)$.

Proof. □

So, from (a) we see that p_n is almost concentrated on the set $A_\varepsilon^{(n)}$, and from (b) and (c) we see furthermore that it is almost *equidistributed* on this set (and this is where the name asymptotic equipartition property comes from)!

In general, the fact that p_n is concentrated on this 'smaller' set $A_\varepsilon^{(n)}$ will be most useful when $\left|A_\varepsilon^{(n)}\right| \ll |\mathcal{X}^n|$, which in light of (c) will be the case when

$$2^{n(h+\varepsilon)} \ll |\mathcal{X}^n| \iff h + \varepsilon < \log_2 |\mathcal{X}|,$$

in which case $A_\varepsilon^{(n)}$ will be exponentially smaller than $\mathcal{X}^n$. When the X_n are i.i.d and uniformly distributed, then $h = \log_2 |\mathcal{X}|$, and the typical set consists of almost all of the possible trajectories.

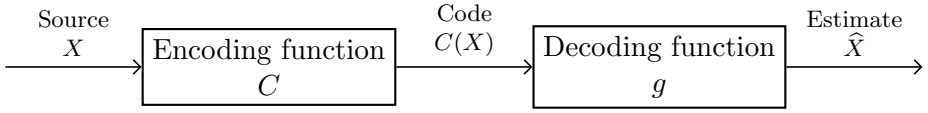
Whilst there are possibly many more possible trajectories in $\mathcal{X}^n$ than typical sequences in $A_\varepsilon^{(n)}$, it is vanishingly unlikely that the observed trajectory lies outside of $A_\varepsilon^{(n)}$, and so these non-typical trajectories play no significant role in the analysis of the process.

4 Data compression and Codes

4.1 Block codes

Suppose we have a set of elements $\mathcal{X}$, for example the alphabet of some language, and we wish to encode an element $x \in \mathcal{X}$, using a binary string or in general the elements of some finite set Σ. It is clear that we can represent each element of $x \in \mathcal{X}$ by a unique binary string of length $n = \lceil \log |\mathcal{X}| \rceil$ and so we can encode an arbitrary element using at most n bits of information, but equally we need at least 2^n elements to uniquely encode each element of $\mathcal{X}$.

However if there is some distribution, given by a random variable X on $\mathcal{X}$ which we call a *source*, in which some elements are more likely to appear than others, then it might be that we can exploit this to find an encoding whose length is shorter on average, or one which is deterministically shorter, but has some (small) probability of error.



Definition 4.1 (Encoding scheme)**.** An encoding scheme is a triple (X, C, g) where X is some discrete random variable taking values in a finite set $\mathcal{X}$, C is a *code*, that is a mapping

$$C \colon \mathcal{X} \to \bigcup_{n=1}^{\infty} \{0,1\}^n := \{0,1\}^+,$$

and $g : \{0,1\}^+ \to \widehat{\mathcal{X}} := \mathcal{X} \cup \{\perp\}$ is a *decoding function.*

That is, the source X is encoded as $C(X)$, and is decoded by the decoding function to $\widehat{X} = g(C(X))$. If C is injective, then we can take the decoding function g to be any function such that $C \circ g$ is the identity on $\mathcal{X}$ and then $X = \widehat{X}$ and this is a *lossless* encoding, otherewise we will also be interested in the *error probability* $p_{\text{err}} = \mathbb{P}[\widehat{X} \neq X]$.

In general, we might wish to encode not just one element $x \in \mathcal{X}$, but a string of elements $(x_1, \ldots, x_n)$, which are then generated according to some stochastic process $(X_n)_{n \geq 1}$. In this case we can ask whether it is more efficient to encode each element individually, or instead to encode longer strings.

To begin with, we will consider the case where we encode the whole string, and so in general our source will be a random vector $(X_1, \ldots, X_n)$, and we are interested in finding a coding $C^{(n)} \colon \mathcal{X}^n \to \{0,1\}^+$, either lossless or with error probability tending to zero, which is particularly *efficient*, either in terms of the average number of symbols we use in a lossless encoding, or perhaps in terms of the maximum number of symbols we use in a lossy encoding.

More precisely, for a lossless encoding we want to minimise

$$L^{(n)} = L^{(n)}\left(C^{(n)}\right) = \mathbb{E}\left(\frac{1}{n}\ell(C^{(n)}(X_1, \ldots, X_n))\right),$$

where $\ell(\cdot)$ measures the length of a binary string. The AEP gives us a powerful method of *data compression*, which allows us to encode messages with a close to optimal rate.

Theorem 4.2. *Let $(X_n)_{n\geq 1}$ be a stochastic process which satisfies the AEP with rate h. Then there is a lossless encoding $C^{(n)} \colon \mathcal{X}^n \to \{0,1\}^+$ such that*

$$\lim_{n\to\infty} L^{(n)} \leq h.$$

Proof. □

Of course, if we just insist the encoding is lossless, then one can construct an encoding function minimising $L^{(n)}$ by greedily assigning the most likely strings $\boldsymbol{x} \in \mathcal{X}^n$ to the shortest strings in $\{0,1\}^+$. However, the code we constructed in Theorem 4.2 has some extra properties which we will see later are useful.

On the other hand, if we consider lossy encoding, we can hope to minimise even the maximum number of symbols we use. The simplest case would be to try to minimise the number of symbols used in some *block code*, which encodes sequences of length n as bitstrings of some fixed length m. In this case, given a block code $C^{(n)} \colon \mathcal{X}^n \to \Sigma^m$ we say it has *rate* $r^{(n)} = \frac{m}{n}$.

Theorem 4.3 (Shannon's source coding theorem for block codes)**.** *Let $(X_n)_{n\geq 1}$ be a stochastic process which satisfies the AEP with rate h. Suppose $r^{(n)}$ is a sequence of rates with $\lim_{n\to\infty} r^{(n)} = r$.*

- *If $r < h$ then for any block code $C^{(n)} : \mathcal{X}^n \to \{0,1\}^m$ with rate $r^{(n)}$, $\lim_{n\to\infty} p_{err}^{(n)} > 0$;*
- *If $r > h$ then there exists some block code $C^{(n)} : \mathcal{X}^n \to \{0,1\}^m$ with rate $r^{(n)}$ such that $\lim_{n\to\infty} p_{err}^{(n)} = 0$*

4.2 Variable length codes

However, Theorem 4.2 does not lead to efficient construction of codes, and the codes are not particular useful in application. More useful are codes which assign codewords to each element of $\mathcal{X}$ and encode a string $(X_1, \ldots, X_n)$ by concatenating the code words. It is much easier to encode messages using such a code, and under certain conditions, also much easier to decode. However, we will see it is still possible to construct codes of this sort which are essentially as efficient as that of Theorem 4.2.

In what follows we will deal with more general alphabets than binary. Given a set Σ let us write Σ^* for the set of finite strings (words) of elements of Σ

$$\Sigma^* = \{w = a_1 a_2 \dots a_n \colon n \geq 0, a_i \in \Sigma\},$$

where $\ell(w) := n$ is the *length* of the *word* w. When $n = 0$ we have the empty word $\varepsilon \in \Sigma^*$, and we will write $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$ for the set of non-empty words. Given a word $w = a_1 \dots a_n$ for any $k \leq n$ we say $v = a_1 \dots a_k$ is a *prefix* of w.

Definition 4.4 (Source code)**.** A *source code* is a mapping

$$C \colon \mathcal{X} \to \Sigma^+,$$

where $C(x)$ is the *codeword* of $x \in \mathcal{X}$. We define an extension of C to $\mathcal{X}^+$, which we still denote by $C \colon \mathcal{X}^+ \to \Sigma^+$, by concatenation via

$$C(x_1 \dots x_k) = C(x_1) \dots C(x_k).$$

Given a source code and a discrete random variable X taking values in $\mathcal{X}$, the *expected code length* is defined as

$$L_C = \mathbb{E}\big(\ell(C(X))\big) = \sum_{x \in \mathcal{X}} \ell(C(x)) p_X(x).$$

Example 4.5. Let $\mathcal{X} = \{a, b, c, d\}$ and let $\Sigma = \{0, 1\}$ and suppose X has distribution

$$p(a) = \frac{1}{2}, \qquad p(b) = \frac{1}{4}, \qquad p(c) = p(d) = \frac{1}{8}.$$

On possible source code would be

$$C(a) = 00, \qquad C(b) = 10, \qquad C(c) = 10, \qquad C(d) = 11.$$

In this case all codewords have length two, and so it is clear that $L_C = 2$.

However, a better code would be as follows:

$$C^*(a) = 0, \qquad C^*(b) = 10, \qquad C^*(c) = 110, \qquad C^*(d) = 111,$$

where we can compute that

$$L_C = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = \frac{7}{4}.$$

Of course, minimising L_C is a trivial optimisation problem - we simply assign the shortest codewords to the most likely elements of $\mathcal{X}$. However, the extension of C to Σ^+ might cause ambiguities when there are distinct sequences of elements $(x_1, x_2, \dots, x_k)$ and $(x'_1, x'_2, \dots, x'_k) \in \mathcal{X}^+$ such that

$$C(x_1) \dots C(x_k) = C(x'_1) \dots C(x'_k).$$

Definition 4.6 (Unique decodability)**.** (a) A source code $C \colon \mathcal{X} \to \Sigma^+$ is called *non-singular* if the mapping is injective, that is, if different elements have different codewords.

(b) C is called *uniquely decodable* if the extension $C \colon \mathcal{X}^+ \to \Sigma^+$ is non-singular.

(c) C is called *prefix-free*, or *instantaneous*, if no codeword is a prefix of another codeword, that is, whenever $x \neq y \in \mathcal{X}$ then $C(x)$ is not a prefix of $C(y)$.

It is clear that a uniquely decodable code must be non-singular. It is also easy to see that a prefix-free code is uniquely decodable - if $C(x_1)\dots C(x_k) = C(x_1\dots x_k) = C(x'_1)\dots, x'_m) = C(x'_1)\dots C(x'_m)$ then since $C(x_1)$ is not a prefix of $C(x'_1)$ and vice versa, it follows that $C(x'_1) = C(x_1)$ and hence $x_1 = x'_1$, since C is non-singular. The result then follows by induction.

Prefix-free codes are called instantaneous as they can be decoded in 'real-time'. If Alice encodes a word $x_1\dots x_n$ by $C(x_1\dots x_n)$ and transmits the codeword to Bob, then for any k Bob can decode the prefix $x_1\dots x_k$ as soon as he has recevied the prefix $C(x_1)\dots C(x_k)$ of the codeword.

Example 4.7. Let $\mathcal{X} = \{a, b, c, d\}$ and let $\Sigma = \{0, 1\}$ as before.

(a) The code
$$C(a) = 0, \qquad C(b) = 010, \qquad C(c) = 01, \qquad C(d) = 10,$$
is non-singular, but it is not uniquely decodable, since $C(ad) = 010 = C(b)$.

(b) If C is a prefix code and $\overleftarrow{C}$ is the code obtained by reversing each codeword, then $\overleftarrow{C}$ is still uniquely decodable (since reversing a string is an involution on Σ^+). However, $\overleftarrow{C}$ is not in general prefix-free.

For example, the code C^* from Example 4.5 is prefix-free, and so
$$C(a) = 0, \qquad C(b) = 01, \qquad C(c) = 011, \qquad C(d) = 111,$$
is uniquely decodable, however it is not prefix-free, as $C(a) = 0$ is a prefix of $C(b) = 01$.

(c) Let
$$C(a) = 10, \qquad C(b) = 00, \qquad C(c) = 11, \qquad C(d) = 110.$$
It can be shown (exercise) that C is uniquely decodable, but it is not *instantaneous*, in that there are arbitrarily long words $w \in \mathcal{X}^+$ such that Bob cannot decode any prefix of w before having received the entire codeword $C(w)$. For example $cb\dots b$ and $db\dots b$ both encode to $110\dots0$, with the only difference being the parity of the string of 0s, which Bob cannot discover until the last element is transmitted.

So, rather than asking to minimise the expected length of *any* code, it is reasonable to restrict our attention to codes that are uniquely decodable, so that sequences of codewords can be unambiguously decoded. Stronger still would be to insist that our code is prefix-free, in which case we should expect our code to have to be longer. Rather surprisingly, it turns out that the expected length of the shortest uniquely decodable code in fact coincides with the expected length of the shortest prefix-free code, and that both are controlled by the entropy of X.

A useful tool will be be Kraft-McMillan inequality, which bounds from below the length of the codewords in a prefix-free code.

Lemma 4.8. *[Kraft-McMillan inequality]*

(1) Let $C\colon \mathcal{X} \to \Sigma^+$ be a prefix-free code with $|\Sigma| = D \geq 2$. Then

$$\sum_{x\in\mathcal{X}} D^{-\ell(C(x))} \leq 1.$$

(2) Let $\{\ell_x\colon x \in \mathcal{X}\}$ be a (multi)-set of numbers such that

$$\sum_{x\in\mathcal{X}} D^{-\ell_x} \leq 1,$$

then there exists a prefix-free code C such that $\ell(C(x)) = \ell_x$ for each $x \in \mathcal{X}$.

Remark 4.9. *In fact, the bound in (1) can be shown to hold for uniquely decodable codes.*

Proof. □

One useful thing to notice is that Kraft's inequality allows us to rephrase the problem of finding codes of minimal expected length as an integer optimisation problem - Given the distribution $p\colon \mathcal{X} \to [0,1]$ we wish to find $\ell = (\ell_x)_{x\in\mathcal{X}} \in \mathbb{N}_0^X$ which

$$\text{minimises } \sum_{x\in\mathcal{X}} \ell_x p(x) \text{ subject to the constraint } \sum_{x\in\mathcal{X}} D^{-\ell_x} \leq 1.$$

This problem can be solved algorithmically in a number of ways.

For example, we can solve the corresponding continuous optimisation problem, noting that we may strengthen the constraint to $\sum_{x\in\mathcal{X}} D^{-\ell_x} = 1$ in this case, using Lagrange multipliers. This gives an approximate solution ℓ_x and if you 'round up', the values $\lceil \ell_x \rceil$ will satisfy the constraint from Kraft's inequality and you can use the implicit algorithm therein to build a code which is close to optimal.

An alternative method uses entropy, and can also give a nearly matching upper bound. Since we are working over a general alphabet Σ of size D, which might not always be equal to two, it makes sense to consider a slightly different notion of entropy, as follows

$$H_D(X) = -\sum_{x\in\mathcal{X}} p(x) \log_D p(x) = -\frac{1}{\log_2 D} H(X).$$

Theorem 4.10. *[Source coding theorem for symbol codes] Let $C\colon \mathcal{X} \to \Sigma^+$ be a prefix-free source code to an alphabet Σ of size D and let X be a discrete random variable taking values in $\mathcal{X}$. Then*

$$L_C \geq H_D(X),$$

with equality if and only if $p(x) = D^{-\ell(C(x))}$ for all $x \in \mathcal{X}$.

Conversely there exists a prefix-free code C such that

$$L_C \leq H_D(X) + 1.$$

Proof. □

Remark 4.11. *Since Kraft's inequality holds for uniquely decodable codes, the first inequality in Theorem 4.10 holds for any uniquely decodable codes. In particular, Theorem 4.2 is optimal if we insist the code is uniquely decodable (and in fact, it is easy to check that the code we constructed there is even prefix-free).*

If we are transmitting then a sequence $(X_n)_{n\geq 1}$ of i.i.d elements distributed according to X_1, then we can see that the average number of symbols used to encode each element of $\mathcal{X}$ is given by L_C, and so Theorem 4.10 leads to a code with the same asymptotic rate as that given by Theorem 4.2 (in the case where $D = |\Sigma| = 2$), since the entropy rate of an i.i.d sequence is given by $H(X_1)$.

However, in general, if we are transmitting a sequence of elements from $\mathcal{X}$, which come now from some stochastic process $(X_n)_{n\geq 1}$, even if the process if time-homogeneous, so that each X_i has the same distribution, it might not be the case that a code which minimises the expected length of each individual codeword, is the one which minimises the length of an encoded message $(X_1, \dots, X_n)$ of longer length. Moreover, we may be able achieve a smaller expected length of codeword *per symbol transmitted*, so an encoding scheme with a smaller rate, if we group our symbols together and encode the elements of $\mathcal{X}^n$ rather than of $\mathcal{X}$.

In this case, given a joint distribution $(X_1, \dots, X_n)$ on $\mathcal{X}$ and a code $C^{(n)} \colon \mathcal{X}^n \to \Sigma^+$ we could ask about the expected codeword length per symbol, or the rate

$$L^{(n)} = \mathbb{E}\left(\frac{1}{n}\ell(C^{(n)}(X_1, \dots, X_n)\right).$$

If we insist that the code $C^{(n)}$ is prefix-free, then Theorem 4.10 implies that the optimal expected length L^* satisfies

$$H_D(X_1, \dots, X_n) \leq n \cdot L^{(n)} \leq H_D(X_1, \dots, X_n) + 1,$$

and hence

$$\frac{1}{n}H_D(X_1, \dots, X_n) \leq L^{(n)} \leq \frac{1}{n}H_D(X_1, \dots, X_n) + \frac{1}{n}.$$

Therefore, if the limit

$$h_D := \lim_{n\to\infty} \frac{1}{n}H_D(X_1, \dots, X_n)$$

exists, then we have a natural bound for this quantity.

However, this limit is just precisely, up to a multiplicative factor of $\log_2 D$, the entropy rate of the process $(X_n)_{n\geq 1}$. Hence we see that the code from Theorem 4.2 does indeed have an optimal rate, and we get another process by which we can construct such codes.

Theorem 4.12. *Let $(X_n)_{n\geq 1}$ be a stochastic process whose entropy rate h exists. Then the minimal expected rate of a prefix-free code satisfies*

$$\lim_{n\to\infty} L^{(n)} = h_D := \frac{h}{\log_2 D}.$$

4.3 Huffman Codes

Given a source X, Theorem 4.10 tell us that for any prefix-free source code C on an alphabet Σ of size D, $L_C \leq H_D(X)$, and conversely, gives us via Kraft's inequality (Lemma 4.8) a way

to construct a prefix-free C such that $L_C \leq H_D(X) + 1$, however in general these codes will not be optimal.

However, Huffman gave a simple algorithm which, given a distribution p on $\mathcal{X}$ with $|\mathcal{X}| \geq 2$, produces an optimal prefix-free binary code.

Whilst this works for all alphabet sizes, let us focus now on the case $\Sigma = \{0, 1\}$. We can think of Σ^* as the *infinite binary tree*, whose root corresponds to the empty string ε, and where each vertex labelled w has two *children* labelled $w0$ and $w1$, which we call *siblings*. Given a string $w \in \Sigma^+$, let us write w' for its sibling.

When is a string w a prefix of another string v? Precisely when the unique path from the root ε to v passes through w. In particular, if we have a prefix-free code, we can build a finite subtree T whose leaves are precisely the codewords by taking the union of these paths from the codewords to the root. For any other vertex in T, at least one child also lies in T.

Huffman's algorithm works by constructing, for each distribution p on $\mathcal{X}$, an appropriate subtree $T(p)$, whose leaves are labelled by the elements of $\mathcal{X}$.

The algorithm is *recursive* - if $N := |\mathcal{X}| = 2$, then we take $T(p)$ to be the tree consisting of the root and its two children.

Otherwise, we start by sorting the elements $\mathcal{X} = \{x_1, \ldots, x_N\}$ such that $p(x_1) \geq p(x_2) \geq \ldots \geq p(x_N)$. Note that this ordering is not necessarily unique, and this may change the output of the algorithm.

We define a new set $\mathcal{X}' = \{x'_1, \ldots, x'_{N-1}\}$, and a probability distribution p' on $\mathcal{X}'$ by

$$p'(x'_i) = \begin{cases} p(x_i) & \text{if } i \leq N-2 \\ p(x_{N-1}) + p(x_N) & \text{if } i = N-1. \end{cases}$$

We build $T(p'(x'_1), \ldots, p'(x'_{N-1}))$ and we form $T(p(x_1), \ldots, p(x_N))$ by taking the leaf labelled x'_{N-1} and adding its two children as leaves, labelled x_{N-1} and x_N. All other leaves labelled x'_i with $i \leq N-2$ we label with x_i.

We call a code constructed in this way a *Huffman code*.

Alternatively we can think of building the tree starting from the leaves up - we start with an independent set of vertices labelled $p(x_1)$ to $p(x_N)$ and recursively we choose the two vertices in the forest which have no parent and have the smallest labels p_1 and p_2 and we add a new vertex, which is joined as a parent to these two vertices, and has label $p_1 + p_2$. We continue until there is a unique vertex in the forest with no parent. By construction this graph is a binary tree, and by choosing an arbitrary $\{0, 1\}$ labelling of the edges from a vertex to its children we can assign to each leaf a string in $\{0, 1\}^+$, giving us a prefix-free code.

Example 4.13. Suppose $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5\}$ and

$$p(x_1) = 0.4, \qquad p(x_2) = 0.2, \qquad p(x_3) = 0.15, \qquad p(x_4) = 0.15, \qquad p(x_5) = 0.1.$$

So, in the first step the vertices with the smallest labels are x_4 and x_5 and so we would merge them to a new vertex, which we might call $x_{4,5}$ with label $p(x_{4,5}) = p(x_4) + p(x_5) = 0.25$. Now

the vertices without parents are x_1, x_2, x_3 and $x_{4,5}$, with labels

$$p(x_1) = 0.4, \qquad p(x_{4,5}) = 0.25, \qquad p(x_2) = 0.2, \qquad p(x_3) = 0.15,$$

and so the vertices with the smallest labels are x_2 and x_3, and in the next step we merge them to a new vertex $x_{2,3}$ with label $p(x_{2,3}) = p(x_2) + p(x_3) = 0.35$. Now the vertices without parents are $x_1, x_{2,3}$ and $x_{4,5}$ with labels

$$p(x_1) = 0.4, \qquad p(x_{2,3}) = 0.35, \qquad p(x_{4,5}) = 0.25,$$

and so the vertices with the smallest labels are $x_{2,3}$ and $x_{4,5}$, and in the next step we merge them to a new vertex $x_{2,3,4,5}$ with label $p(x_{2,3,4,5}) = p(x_{2,3}) + p(x_{4,5}) = 0.6$. Now the vertices with parents are x_1 and $x_{2,3,4,5}$ with labels

$$p(x_{2,3,4,5}) = 0.6, \qquad p_1 = 0.4,$$

and so the vertices with the smallest labels are $x_{2,3,4,5}$ and x_1, and in the last step we merge them to a new vertex $x_{1,2,3,4,5}$ with label $p(x_{1,2,3,4,5}) = 1$.

If we label the edges so that the edge labelled 1 goes the the child with the smaller label, then we end up with the following code $C\colon \mathcal{X} \to \Sigma^*$

$$C(x_1) = 1, \qquad C(x_2) = 000, \qquad C(x_3) = 001, \qquad C(x_4) = 010, \qquad C(x_5) = 011.$$

In the example above we can calculate that the expected code length is then

$$L_C = 0.4 + 3 \cdot (0.2 + 0.15 + 0.15 + 0.1) = 2.2$$

and the entropy of p is ≈ 2.15. So, this code is definitely close to the theoretical limit of $H(p)$, and in fact, it can be shown that no other binary code will do better than the Huffman code. Given a source X, let us say a prefix-free binary code C if *optimal* if $L_C \leq L_{C'}$ for any other prefix-free binary code C'.

Theorem 4.14. *Huffman codes are optimal binary codes.*

Let us start by showing the following

Lemma 4.15. *Let X be a source on $\mathcal{X}$ and let $C\colon \mathcal{X} \to \{0,1\}^+$ be an optimal prefix-free binary code. Then*

(1) If $p(x) > p(y)$, then $\ell(C(x)) \leq \ell(C(y))$;

(2) Let $\ell_{max} = \max\{\ell(w)\colon w \in C(\mathcal{X})\}$ and

$$W = \{w \in C(\mathcal{X})\colon \ell(w) = \ell_{max}\},$$

then for all $w \in W$, its sibling $w' \in W$.

Proof. □

Proposition 4.16. *Let X be a source on $\mathcal{X}$ and let $C\colon \mathcal{X} \to \{0,1\}^*$ be an optimal prefix-free binary code. If $\mathcal{X} = \{x_1, \dots, x_N\}$ is some ordering of $\mathcal{X}$ such that $p(x_1) \geq p(x_2) \dots \geq p(x_N)$ then there is some permutation π on $\mathcal{X}$ such that $C' = C \circ \tau$ is an optimal prefix-free binary code such that*

$$C'(x_{N-1}), C'(x_N) \in W \qquad \text{and they are siblings,}$$

where W is as in Lemma 4.15 (2).

Proof. □

We call codes satisfying the conclusion of Proposition 4.16 *canonical* for the ordering $\mathcal{X} = \{x_1, \dots, x_N\}$.

At this point we are ready to prove that Huffman codes are optimal.

Proof of Theorem 4.14. □

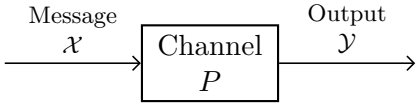
5 Information Channels

A *channel* is a way to model the transmission of some message. We have some set $\mathcal{X}$ of messages which are to be transmitted, potentially in some encrypted form, and a set $\mathcal{Y}$ of possible outputs, messages received by the other party, where the output might not depend deterministically on the message due to some inherent *noise* in the channel, which might randomly change the output, or even some randomness in the encryption process.

Definition 5.1 (Discrete channel)**.** A *discrete (memoryless) channel*

$$\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$$

consists of two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a stochastic transition matrix $P = (p(y|x))_{x\in\mathcal{X}, y\in\mathcal{Y}}$.



That is, the rows of P are index by $\mathcal{X}$, the columns by $\mathcal{Y}$ and each row is a conditional probability distribution $p(\cdot|x)$ on $\mathcal{Y}$, that is $p(y|x) \geq 0$ for all $y \in \mathcal{Y}$ and

$$\sum_{y\in\mathcal{Y}} p(y|x) = 1.$$

We think of the distribution $p(\cdot|x)$ as being the distribution of the output of the channel, the received message $y \in \mathcal{Y}$ when x is the input.

Example 5.2. (a) The *binary symmetric channel* has $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ where each message has a ε chance of resulting in the wrong output and so $P = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$.

(b) The *binary erasure channel* has $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, \perp\}$ and each message has $(1-\varepsilon)$ chance of being transmitted correctly and a ε chance of being 'lost', and outputting $\perp$, and so $P = \begin{pmatrix} 1-\varepsilon & 0 & \varepsilon \\ 0 & 1-\varepsilon & \varepsilon \end{pmatrix}$

Definition 5.3 (Channel extension)**.** Given a channel $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$ the *nth channel extension* is the channel

$$\mathcal{C}^n = (\mathcal{X}^n, P_n, \mathcal{Y}^n)$$

where

$$(P_n)_{\boldsymbol{x},\boldsymbol{y}} = p_n(\boldsymbol{y}|\boldsymbol{x}) = p_n(y_1, \ldots, y_n|x_1, \ldots, x_n) = \prod_{k=1}^{n} p(y_k|x_k).$$

In other words, the nth channel extension is the channel we get by sending n consecutive, independent messages over the channel $\mathcal{C}$.

Typically, we are interested in the behaviour of the channel when the input arrives as some $\mathcal{X}$-valued random variable X, with some distribution p_X. In this case the output Y is also a

random variable, which inherits the randomness from X, as well as some of the randomness inherent in the channel described by P.

For a fixed channel, and so a fixed P, we might hope to choose the input distribution p_X in some optimal way. Let $\mathcal{M}(\mathcal{X})$ be the collection of all probability distributions on $\mathcal{X}$. In other words, if $\mathcal{X} = \{x_1, \ldots, x_n\}$, then we can think of $\mathcal{M}$ as consisting of all vectors $(p_1, \ldots, p_n) \in \mathbb{R}^n$ with non-negative entries and sum one, with $p(x_i) = p_i$.

One measure of the 'quality' of the channel is how well the message is preserved, and one way to measure this would be to measure how much information about the message X is contained in the random variable Y, motivating the following definition.

Definition 5.4 (Channel capacity)**.** Given a channel $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$ the *channel capacity* is defined as

$$\text{cap}(\mathcal{C}) = \max\{I(X\,;\,Y)\colon p_X \in \mathcal{M}(\mathcal{X})\}.$$

Remark 5.5. *Note that, p_X and P together determine $p_{X,Y}$ and hence p_Y and $I(X\,;\,Y)$. Indeed,*

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x) = p_X(x)\,(P)_{x,y} \qquad \textit{and} \qquad p_Y(y) = \sum_{x\in\mathcal{X}} p_{X,Y}(x,y).$$

Hence, we can consider the function from $\mathcal{M}(\mathcal{X}) \to \mathbb{R}$ given by $p_X \mapsto I(X\,;\,Y)$. This is then a continuous function, on a compact set $\mathcal{M}(\mathcal{X}) \subseteq \mathbb{R}^{\mathcal{X}}$ and so it indeed achieves some maximum, which is then the channel capacity. Note that this maximum is not necessarily achieved by a unique distribution p_X!

Note that, by Corollary 2.33 and Lemma 2.22,

$$0 \le I(X\,;\,Y) = H(X) - H(X \mid Y) \le H(X) \le \log_2 |\mathcal{X}|,$$

and so

$$0 \le \textit{cap}(\mathcal{C}) \le \log_2 |\mathcal{X}|.$$

Example 5.6. (a) *Noiseless binary channel :* Suppose we take the binary symmetric channel withb $\varepsilon = 0$, that is $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

It is easy to verify that in this case $X = Y$ and so for any p_X, $I(X\,;\,Y) = I(X\,;\,X) = H(p_X)$, and in particular

$$\text{cap}(\mathcal{C}) = \max\{H(p_X)\colon p_X \in \mathcal{M}(\mathcal{X})\} = \log_2 |\mathcal{X}| = 1,$$

which is achieved only for the uniform distribution $(1/2, 1/2)$. Hence $\text{cap}(\mathcal{C}) = 1$.

(b) *Channel with non-overlapping outputs :* More generally, for any transition matrix P where X is determined by Y, we have by Lemma 2.8

$$I(X\,;\,Y) = H(X) - H(X|Y) = H(X) \le \log_2 \mathcal{X},$$

where equality is again achieved uniquely by the uniform distribution on X. Hence $\text{cap}(\mathcal{C}) = \log_2 \mathcal{X}$.

(c) *Noisy typewriter :* Suppose $\mathcal{X} = \{x_1, \ldots, x_{2N}\}$ is a set of letters on a rather improbable circular typewriter, where when I go to type a letter x_i, I miss and hit the letter x_{i+1} with probability $\frac{1}{2}$ (with addition mod $2N$). Hence $\mathcal{Y} = \mathcal{X}$ and the transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \ldots & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & \ldots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ \frac{1}{2} & 0 & 0 & \ldots & \frac{1}{2} \end{pmatrix}.$$

Given any distribution p_X we can calculate

$$H(Y|X) = \sum_{j=1}^{2N} p_X(x_j) H(Y \mid X = x_j) = \sum_{j=1}^{2N} p_X(x_j) H(1/2, 1/2) = 1,$$

and so

$$I(X\,;\,Y) = H(Y) - H(Y \mid X) = H(Y) - 1 \leq \log_2 2N - 1 = \log_2 N,$$

and equality is achieved whenever p_Y is uniform. Hence $\mathrm{cap}(\mathcal{C})$ will be $\log_2 N$ if there is some p_X such that p_Y is uniform, and it is easy to verify that when p_X is uniform, so is p_Y. Hence $\mathrm{cap}(\mathcal{C}) = \log_2 N$.

However, in this case there are multiple optimal distributions. For example if X is uniformly distributed on the odd elements, or uniformly distributed on the even elements, then Y is again uniformly distributed on $\mathcal{Y}$. More generally, any convex combination of these two distributions acheives the optimal capacity.

A small remark here is that, when X is distributed on the even or odd elements, then this example reduces to a channel with non-overlapping outputs.

(d) *Binary symmetric channel :* Recall that $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $P = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$.

In this case we can again compute

$$\begin{aligned} I(X\,;\,Y) &= H(Y) - H(X \mid Y) = \\ &= H(Y) - H(\varepsilon, 1-\varepsilon) \\ &\leq 1 - H(\varepsilon, 1-\varepsilon), \end{aligned}$$

and it is easy to verify that if X is uniformly distributed, then so is Y, and so equality is achieved. Hence $\mathrm{cap}(\mathcal{C}) = 1 - H(\varepsilon, 1-\varepsilon)$.

(e) *Binary erasure channel :* Recall that $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, \perp\}$ and $P = \begin{pmatrix} 1-\varepsilon & 0 & \varepsilon \\ 0 & 1-\varepsilon & \varepsilon \end{pmatrix}$.

It is easy to see that, for any distribution p_X,

$$\begin{aligned} \mathbb{P}[Y = \perp] &= \mathbb{P}[X = 0] \cdot \mathbb{P}[Y = \perp \mid X = 0] + \mathbb{P}[X = 1] \cdot \mathbb{P}[Y = \perp \mid X = 1] \\ &= \varepsilon p_X(1) + \varepsilon p_X(0) \\ &= \varepsilon \end{aligned}$$

and for any $i \in \{0, 1\}$

$$\mathbb{P}[X = i \mid Y = \perp] = \frac{\mathbb{P}[X = i, Y = \perp]}{\mathbb{P}[Y = \perp]} = p_X(i).$$

Furthermore, $(X|Y=i)$ is constant for $i \in \{0,1\}$. Hence

$$\begin{aligned} I(X\,;\,Y) &= H(X) - H(X|Y) \\ &= H(X) - \mathbb{P}[Y=1]H(X|Y=1) - \mathbb{P}[Y=0]H(X|Y=0) - \mathbb{P}[Y=\perp]H(X|Y=\perp) \\ &= H(X) - 0 - 0 - \varepsilon H(X) \\ &= (1-\varepsilon)H(X) \leq 1-\varepsilon. \end{aligned}$$

Again, it is easy to verify that equality is achieved only when $H(X)=1$, and so when p_X is uniform. Hence $\mathrm{cap}(\mathcal{C}) = 1-\varepsilon$.

(f) *Non-symmetric binary channel :* Suppose $\mathcal{X} = \mathcal{Y} = \{0,1\}$ but now the probability of error is different for 0 and 1, so that $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$.

The distribution of X can be described by a single parameter $p_X = (1-t, t)$ and so for fixed α and β the mutual information $I(X\,;\,Y)$ is a (continuous) function of t, where $t \in [0,1]$, which can be maximised using, for example, Lagrange multipliers (exercise).

Definition 5.7 (Weakly symmetric channel)**.** A channel $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$ is called *weakly symmetric* if

(i) All rows $p(\cdot|x)$ for $x \in \mathcal{X}$ are permutations of each other;

(ii) All columns $p(y|\cdot)$ for $y \in \mathcal{Y}$ have the same (constant) sum.

Proposition 5.8. *If $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$ is a weakly symmetric channel, then*

$$\mathit{cap}(\mathcal{C}) = \log_2 |\mathcal{Y}| - H(p(\cdot|x_0)),$$

where $x_0 \in \mathcal{X}$ is arbitrary.

Proof. □

Another nice property of the channel capacity is that the capacity of the nth channel extension is determined by the channel itself.

Lemma 5.9. *Let $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$ be a channel and let $\mathcal{C}^n$ be the nth channel extension. Then*

$$\mathit{cap}(\mathcal{C}^n) = n \cdot \mathit{cap}(\mathcal{C}).$$

Proof. □

5.1 Shannon's channel coding theorem

Let us try to give a practical meaning to the channel capacity.

Suppose Alice wishes to transmit a message from some input set $\mathcal{W}$ to Bob through a noisy channel $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$. For example, $\mathcal{W} \subseteq \mathcal{X}^+$ might be some set of phrases over the alphabet $\mathcal{X}$ and Alice will independently transmit the symbols one by one through the channel, where the

output $\mathcal{Y} = \mathcal{X}$ that Bob receives is a symbol that may or may not agree with the transmitted symbol.

Now, Bob would like to recreate the message, so he should have some function $g\colon \mathcal{Y}^+ \to \mathcal{W}$ which represents his *guess* of what the input was, given the observed output.

Since the channel is noisy, there is some chance Bob's guess will be incorrect. We could reduce this chance by sending a longer sequence of symbols, perhaps by sending the entire message twice, or via some more complicated *encoding* scheme. However, this comes then at the cost of a longer transmission. Ideally we would like to keep this error of failure small, using as few transmissions as possible.

Definition 5.10 ((M,n)-codes)**.** An (M,n)*-code* for the channel $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$ consists of the following:

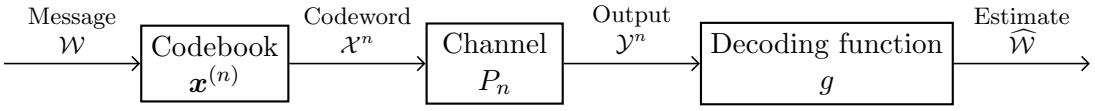
- A set $\mathcal{W}$ of *messages*, with $|\mathcal{W}| = M$,
- A mapping $\boldsymbol{x}^{(n)}\colon \mathcal{W} \to \mathcal{X}^n$, which we call the *codebook*,
- A function $g : \mathcal{Y} \to \widehat{\mathcal{W}}$, where $\widehat{\mathcal{W}} = \mathcal{W}$ or $\widehat{\mathcal{W}} = \mathcal{W} \cup \{\perp\}$, with the *erasure* (or dummy) *symbol* $\perp$.

Remark 5.11. *Note the word code here does not have the same meaning as in Section 4.*

The *rate* of an (M,n)-code is given by

$$R = \frac{\log_2 M}{n}.$$

The rate is then measured in bits per transmission



For example, if we have a binary channel, with $\mathcal{X} = \{0,1\}$, then in order to encode the elements of $\mathcal{W}$ by distict *codewords*, binary sequences of length n, we would need $n = \lceil \log_2 M \rceil$. In this case the rate would be equal to one.

However, since these binary sequences may be corrupted by the channel, the likelihood that Bob's guess is correct will be closely related to the probability that any bit is incorrectly transmitted, which may be very large. Hence, in order to make more accurate guesses, it might be necessary to take a larger n, and so longer transmissions, and enocde the messages in some *robust* manner, at the expense of decreasing the rate.

Definition 5.12 (Probability of error)**.** Given a channel $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$ and an (M,n)-code $(\mathcal{W}, \boldsymbol{x}^n, g)$ we define the *conditional probability of error*, given $w \in \mathcal{W}$, as the probability λ_w that if we encode the message w, transmit the encoded message across the channel and then decode, that we do not recover w. That is

$$\lambda_w^{(n)} = \sum_{\boldsymbol{y} \in \mathcal{Y}^n\colon g(\boldsymbol{y}) \neq w} p\left(\boldsymbol{y} | \boldsymbol{x}^{(n)}(w)\right).$$

The *maximal probability of error* is defined as

$$\lambda_{\max}^{(n)} = \max\Big\{\lambda_w^{(n)} \colon w \in \mathcal{W}\Big\}.$$

The *average probability of error* if defined as

$$p_{\text{err}}^{(n)} = \frac{1}{M}\sum_{w\in\mathcal{W}} \lambda_w^{(n)}.$$

In the above, it can be useful to think of a random variable W distributed uniformly on $\mathcal{W}$ (which is independent of the channel). The codebook transforms W into a random vector $\boldsymbol{x}^{(n)}(W) = (X_1, \dots, X_n)$, which is then transmitted through the channel, with output the random vector $(Y_1, \dots, Y_n)$. This output is then decoded as $\widehat{W} = g(Y_1, \dots, Y_n)$, which is a random variable distributed on $\widehat{\mathcal{W}}$. In this this case we can express

$$\lambda_w^{(n)} = \mathbb{P}\left[\widehat{W} \neq w \mid W = w\right] \qquad \text{and} \qquad p_{\text{err}} = \mathbb{P}\left[\widehat{W} \neq W\right].$$

Note that in the above

$$W \to (X_1, \dots, X_n) \to (Y_1, \dots, Y_n) \to \widehat{W}$$

is a Markovian quadruple.

The specific question we will be interested in is how *efficiently*, in terms of the rate of the code, can we achieve (arbitrarily) small maximum error probability. That is, given $\varepsilon > 0$, for what rate R can we achieve $\lambda_{\max}^{(n)} \leq \varepsilon$.

It is not apriori obvious that we would not need increasing rate, as a function of ε, to achieve smaller and smaller error probabilities. Indeed, if we fix M and n, then there are only finitely many choices of codebook $\boldsymbol{x}^{(n)}$ and decoding function g, and (for non-trivial channels) each lead to a *strictly positive* maximum probability of error $\lambda_{\max}^{(n)}$. In particular, the minimum over all (M, n)-codes of $\lambda_{\max}^{(n)}$ will also be strictly positive, and so if we fix M and n we cannot reduce the error probability arbitrarily.

However, it will turn out that for certain rates, if we allow the length of our codewords to grow, we can achieve *any* maximum error probability, however small.

Definition 5.13 (Achievable rates)**.** A real number $R > 0$ is an *achievable rate* for the channel $\mathcal{C}$ if there is a sequence of (M_n, n)-codes, with maximal probability of error $\lambda_{\max}^{(n)}$ such that the rate $R_n = \frac{\log_2 M_n}{n}$ satisfy

$$\lim_{n\to\infty} R_n = R \qquad \text{and} \qquad \lim_{n\to\infty} \lambda_{\max}^{(n)} = 0.$$

In other words, R is achievable if for any $\varepsilon > 0$ there is some (M_n, n)-code with rate $R_n > R - \varepsilon$ and $\lambda_{\max}^{(n)} < \varepsilon$.

It turns that the capacity of a channel controls the achievable rates.

Definition 5.14. The *achievable capacity* of a channel $\mathcal{C}$ is defined as

$$R^* = \sup\{R \colon R \text{ is an achievable rate for } \mathcal{C}\}.$$

Remark 5.15. *We note that this supremum is in fact an attained maximum.*

Theorem 5.16 (Shannon's channel coding theorem)**.** *For any channel $\mathcal{C}$ it achievable capacity R^* is equal to the channel capacity cap$(\mathcal{C})$.*

We will find that it is relatively easy to show that $R^* \leq \text{cap}(\mathcal{C})$, that is, every achievable rate R satisfies $R \leq \text{cap}(\mathcal{C})$. It will be rather more difficult to show that if $R < \text{cap}(\mathcal{C})$ then R is achievable, since to do so we will have to construct sequences of (M_n, n)-codes with rate tending to R.

Shannon's original proof, while mathematically ingenious, is merely an existence proof - it does not provide an explicit algorithm to construct such (M_n, n)-codes. It is one of the earliest examples of a proof via the *probabilistic method*.

Let us start by showing the 'easy' half of Shannon's channel coding theorem, that the achievable capacity of a channel is at most the channel capacity.

Proof that $R^* \leq cap(\mathcal{C})$. □

To prove the other direction, $R^* \geq \text{cap}(\mathcal{C})$, we have to produce a good sequence of (M_n, n)-codes, so we have to build clever codebooks and decoding functions that work well with the channel $\mathcal{C}$.

We start with the following lemma, which tells us that it is sufficient to bound the *average* probability of error, which is easier to work with, due to the nice equality $p_{\text{err}} = \mathbb{P}[\widehat{W} \neq W]$.

Lemma 5.17. *A real number $R > 0$ is an achievable rate for a channel $\mathcal{C}$ if and only if there is a sequence of (M_n, n)-codes, with average probability of error $p_{err}^{(n)}$ such that the rates $R_n = \frac{\log_2 M_n}{n}$ satisfy*

$$\lim_{n\to\infty} R_n = R \qquad \textit{and} \qquad \lim_{n\to\infty} p_{err}^{(n)} = 0.$$

Proof. □

Let us assume without loss of generality that $\mathcal{W} = \{1, \ldots, M_n\}$. We can think of the codebook $\boldsymbol{x}^{(n)} \colon \mathcal{W} \to \mathcal{X}^n$ as a large $(M_n \times n)$ matrix:

$$\boldsymbol{x}^{(n)} = \begin{pmatrix} x_1(1) & x_2(1) & \ldots & x_n(1) \\ x_1(2) & x_2(2) & \ldots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(M) & x_2(M) & \ldots & x_n(M) \end{pmatrix}$$

where $\boldsymbol{x}^{(n)}(i) = (x_1(i), x_2(i), \ldots, x_n(i))$ is the codeword corresponding to the ith message in $\mathcal{W}$. One could also think of this as a phyiscal book, where on each page there is some vector $(x_1(i), x_2(i), \ldots, x_n(i))$ which is the codeword corresponding to some message.

Alice will transmit one of these codewords across the channel $\mathcal{C}^n$ and Bob will receive a transmission $(y_1, y_2, \ldots, y_n)$. He then has to 'choose' which of the possible codewords $(x_1(i), x_2(i), \ldots, x_n(i))$ he thinks was transmitted, which corresponds to the decoding function g.

So, our aim is to choose a sensible codebook so that for a 'typical' transmission $(y_1, y_2, \ldots, y_n)$ there is a uniquely identifiable codeword $(x_1(i), x_2(i), \ldots, x_n(i))$ which is the likely input resulting in the output $(y_1, y_2, \ldots, y_n)$.

Shannon's ingenious idea was to choose a *random* codebook. That is, if we let $\mathcal{B}_n$ be the set of all possible codebooks, where it is easy to see that

$$|\mathcal{B}_n| = |\mathcal{X}|^{n \cdot M_n},$$

since we get to choose the $n \cdot M_n$ elements of the matrix, each of which is an element of $\mathcal{X}$.

Our codebook will then be a random variable $B^{(n)}$, which is distributed on $\mathcal{B}_n$. In other words, $B^{(n)}$ is a random $(M_n \times n)$ matrix:

$$B^{(n)} = \begin{pmatrix} X_1(1) & X_2(1) & \ldots & X_n(1) \\ X_1(2) & X_2(2) & \ldots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(M) & X_2(M) & \ldots & X_n(M) \end{pmatrix}$$

where each $X_i(w)$ is a random variable taking values in $\mathcal{X}$. In this way, given a fixed message $w \in \mathcal{W}$ the input and output to the channel $\mathcal{C}^n$ are random variables

$$\boldsymbol{X}^{(n)}(w) = (X_1(w), \ldots, X_n(W)) \qquad \text{and} \qquad \boldsymbol{Y} = (Y_1, \ldots, Y_n)^{(n)}.$$

In order to decode the transmission, we now need some way to identify which pairs $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ are likely to occur, or in other words, what are the 'typical' values taken by the random vector $(\boldsymbol{X}(w), \boldsymbol{Y})$.

Suppose we have a pair of jointly distributed random variables X, Y taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively with joint distribution $p_{X,Y}$ and marginal distributions p_X and p_Y. Given $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$ we write

$$p^{(n)}_{X,Y}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{k=1}^{n} p_{X,Y}(x_k, y_k)$$

for the joint distribution of n i.i.d copies $(X_1, Y_1) \ldots, (X_n, Y_n)$ of (X, Y). Analogously we have the (marginal) distributions of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$

$$p^{(n)}_X(\boldsymbol{x}) = \prod_{k=1}^{n} p_X(x_k) \qquad \text{and} \qquad p^{(n)}_Y(\boldsymbol{y}) = \prod_{k=1}^{n} p_Y(y_k).$$

Since i.i.d sequences of random variables have the AEP, we know from Lemma 3.26 that

$$\begin{aligned} -\frac{1}{n} \log_2 p^{(n)}_{X,Y}(X_1, \ldots, X_n, Y_1, \ldots, Y_n) &\to H(X, Y), \\ -\frac{1}{n} \log_2 p^{(n)}_X(X_1, \ldots, X_n) &\to H(X), \text{ and} \\ -\frac{1}{n} \log_2 p^{(n)}_Y(Y_1, \ldots, Y_n) &\to H(Y), \text{ almost surely.} \end{aligned}$$

Let us define then, the set of 'typical' sequences in $\mathcal{X}^n \times \mathcal{Y}^n$ which match these predictions up to some small deviation.

Definition 5.18 (Jointly typical sequences)**.** Given X,Y and $(X_1, Y_1), \ldots, (X_n, Y_n)$ as above and $\varepsilon > 0$, the set of *jointly typical sequences* is given by

$$\tilde{A}_\varepsilon^{(n)} = \left\{ (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n \colon \left\{ \begin{array}{rl} \left| -\frac{1}{n} \log_2 p_{X,Y}^{(n)}(X_1, \ldots, X_n, Y_1, \ldots, Y_n) - H(X,Y) \right| & < \varepsilon \\ \left| -\frac{1}{n} \log_2 p_X^{(n)}(X_1, \ldots, X_n) - H(X) \right| & < \varepsilon \\ \left| -\frac{1}{n} \log_2 p_Y^{(n)}(Y_1, \ldots, Y_n) - H(Y) \right| & < \varepsilon \end{array} \right\} \right.$$

We will use the jointly typical sequences to decode. Indeed, suppose Bob has access to the codebook $\boldsymbol{x}^{(n)}$ and he also knows which sequences are jointly typical.

Bob receives some transmission $\boldsymbol{y}$, and he knows that that is very likely that the input $\boldsymbol{x}$ is such that $(\boldsymbol{x}, \boldsymbol{y})$ is jointly typical. So, Bob can look through the codebook and see whether the pair $(\boldsymbol{x}^{(n)}(k), \boldsymbol{y})$ is jointly typical.

If there is a unique such pair, then it is reasonable to guess that the input was $\boldsymbol{x}^{(n)}(k)$, since given any other input it's very unlikely that the output was $\boldsymbol{y}$.

Of course, even if the codebook is chosen very carefully, it may still be the case that there is no codework such that $(\boldsymbol{x}^{(n)}(k), \boldsymbol{y})$ is jointly typical, or there are multiple. In these cases, Bob cannot make an unambiguous guess, and so he can just decode to the erasure symbol $\perp$.

What we will find is that the properties of jointly typical sequences mean that, when Bob makes a guess it is very accurate, and that if we choose our codebook at random, then as long as the rate is not too large, it is in fact unlikely that Bob cannot make an unambiguous guess.

Proposition 5.19. *Let X,Y,$(X_1, Y_1), \ldots, (X_n, Y_n)$, $\varepsilon > 0$ and $\tilde{A}_\varepsilon^{(n)}$ be as above. Then there exists $N(\varepsilon)$ such that:*

(a) For all $n \geq N(\varepsilon)$

$$\mathbb{P}\left[(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \in \tilde{A}_\varepsilon^{(n)} \right] > 1 - \varepsilon,$$

(b)

$$(1-\varepsilon) 2^{n\left(H(X,Y) - \varepsilon\right)} \leq \left| \tilde{A}_\varepsilon^{(n)} \right| \leq 2^{n\left(H(X,Y) + \varepsilon\right)},$$

where the first inequality holds for all $n \geq N(\varepsilon)$ and the second for all n.

(c) If $(X_1', Y_1'), \ldots (X_n', Y_n')$ are i.i.d random variables where X_i' and Y_i' are independently *distributed as X and Y, then*

$$(1-\varepsilon) 2^{-n\left(I(X\,;\,Y) + 3\varepsilon\right)} \leq \mathbb{P}\left[(X_1', \ldots, X_n', Y_1', \ldots, Y_n') \in \tilde{A}_\varepsilon^{(n)} \right] \leq 2^{-n\left(I(X\,;\,Y) - 3\varepsilon\right)},$$

where the first inequality holds for all $n \geq N(\varepsilon)$ and the second for all n.

Proof. □

It is now apparent why this encoding scheme should be effective. If we choose a message $w \in \mathcal{W}$, and transmit the (randomly chosen) codeword $\boldsymbol{X}^{(n)}(w)$ across the channel, then it is

very likely, by (a), that $\left(\boldsymbol{X}^{(n)}(w), \boldsymbol{Y}^{(n)}\right)$ forms a typical pair, and so Bob will have at least one candidate codeword.

However, whilst the output $\boldsymbol{Y}^{(n)}$ of the channel may depend on the input $\boldsymbol{X}^{(n)}(w)$, it is independent of the other (randomly chosen) codewords. In particular, for any $w' \neq w$ the pair $\boldsymbol{X}^{(n)}(w')$ and $\boldsymbol{Y}^{(n)}$ are independent, and so by (c) it is very unlikely that the pair $\left(\boldsymbol{X}^{(n)}(w'), \boldsymbol{Y}^{(n)}\right)$ forms a typical pair, and so the candidate codeword will likely be unique!

Proof that $R^* \geq cap(\mathcal{C})$. □

This is perhaps in some ways unsatisfactory, since whilst Theorem 5.16 asserts the existence of a good sequence of (M_n, n)-codes, whose rate is tending to R and whose maximum error probability can be made arbitrarily small, the proof is *non-constructive*.

Firstly, we used a probabilistic argument to deduce the existence of a sequence of codes whose average error probability is small, and secondly one can check that in Lemma 5.17 we also used a probabilistic argument to show that we can use a code whose average error probability is small to cosntruct a code whose maximum error probability is small.

Indeed, in Theorem 5.16 we essentially used the following 'trivial' statement:

Claim. If $a \in \mathbb{R}$, X is a real discrete random variable taking values in $\mathcal{X}$, $f\colon \mathcal{X} \to \mathbb{R}$ and $\mathbb{E}(f(X)) \leq a$, then there is some $x \in \mathcal{X}$ such that $f(x) \leq a$.

Note that this follows from Lemma 1.18 (iv), by considering the random variable $a - f(X)$. This tells us that such an x exists, but gives us no *algorithmic* way to construct such an x.

Similarly in Lemma 5.17 we used the slightly more complicated claim:

Claim. If $a, X, \mathcal{X}, f$ are as above, then

$$\sum_{\substack{x \in \mathcal{X} \\ f(x) \leq 2a}} p_X(x) = \mathbb{P}[f(X) \leq 2a] \geq \frac{1}{2}.$$

This is essentially immediate from Markov's inequality (Lemma 1.22), since

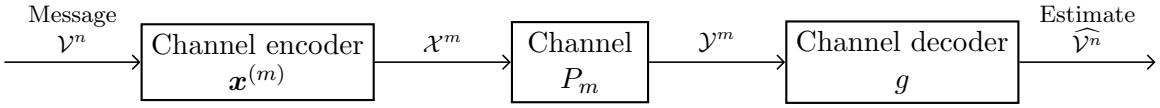
$$\mathbb{P}[f(X) \geq 2a] \leq \frac{\mathbb{E}(f(X))}{2a} \leq \frac{1}{2}.$$

However, again this does not lead to any particular algorithmic to identify $\{x \in \mathcal{X}\colon f(x) \leq 2a\}$.

One could of course run a brute-force search for appropriate (M_n, n)-codes, and there are also ways to *derandomize* the argument to give a constructive algorithm to find these codes, but neither are computationally efficient. Explicit constructions of such codes were a major open problem for a long time, finally being settled in the 90s with the invention of *turbo codes*, however we will not discuss these codes in any detail.

5.2 Source-channel separation theorem

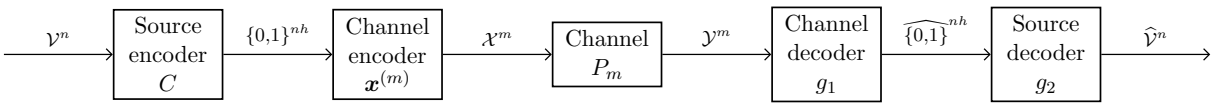
Suppose don't want to transmit a single message, but the (partial) output of some stochastic process $(V_n)_{n\in\mathbb{N}}$ (which satisfies the AEP with rate h) across the mth extension of some channel $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$, where we wish to minimise the *rate* $\frac{m}{n}$. Let us write $V^{(n)} = (V_1, \ldots, V_n)$.



As before, we transmit the message $V^{(n)}$ by first encoding it as some sequence X^m in $\mathcal{X}^m$ of length m, transmitting it across the channel with ouput Y^m, and then decoding the message to some estimate $\widehat{V^{(n)}}$.

If the capacity of the channel $\text{cap}(\mathcal{C}^m) = m \cdot \text{cap}(\mathcal{C})$ is smaller than the entropy rate nh of the process, then a similar argument as in Theorem 5.16 will tell us that we cannot transmit $V^{(n)}$ with vanishing error probability $p_{\text{err}}^{(n)} = \mathbb{P}\left[V^{(n)} \neq \widehat{V^{(n)}}\right]$, however we choose our channel encoding and decoding.

Conversely, if $m \cdot \text{cap}(\mathcal{C}) > nh$, then one can combine Theorem 4.3 and Theorem 5.16 to transmit the message with vanishing error probability by encoding the source and the channel separately, which we refer to as *source-channel separation*



That is, we choose first an encoding of the source $C : \mathcal{V}^n \to \{0,1\}^{nh}$ which we can decode with a vanishing probability of error (for ease of presentation here we have written nh, although in practise we need to take $n(h+\varepsilon)$ for some sufficiently small epsilon) which exists by Theorem 4.3. Since $R = \frac{nh}{m} < \text{cap}(\mathcal{C})$ is an achievable rate, by Theorem 5.16 there is a channel encoding $x^{(m)} : \{0,1\}^{nh} \to \mathcal{X}^m$ which can be transmitted across the channel and correctly decoded with a vanishing probability of error. The total probability of error is then (at most) the sum of the errors in the source and channel encoding, and so is also vanishing.

The following is an semi-formal statement of the above discussion.

Theorem 5.20 (Shannon source-channel separation theorem)**.** *Let $(V_n)_{n\in\mathbb{N}}$ be a stochastic process and $V^{(n)} = (V_1, \ldots, V_n)$ and let $\mathcal{C} = (\mathcal{X}, P, \mathcal{Y})$ be a channel.*

- *If $m \cdot cap(\mathcal{C}) < H\left(V^{(n)}\right)$, then $V^{(n)}$ cannot be transmitted across the mth extension channel $\mathcal{C}^m$ with a vanishing probability error.*

- *If $(V_n)_{n\in\mathbb{N}}$ satisfies the AEP with rate h and $m \cdot cap(\mathcal{C}) > nh$, then $V^{(n)}$ can be transmitted across the mth extension channel $\mathcal{C}^m$ with a vanishing probability error, and the source and channel coding can be done separately.*

Remark 5.21. *Note that, if $(V_n)_{n\in\mathbb{N}}$ satisfies the AEP with rate h, then $\frac{1}{n}H\left(V^{(n)}\right) \to h$ and so the first and second part of the theorem are complementary.*

6 Differential Entropy

6.1 Differential Entropy

Definition 6.1. Let X be a real, continuous random variable with density $f(x) = f_X(x)$, that is, for any $B \subseteq \mathbb{R}$

$$P_X(B) := \mathbb{P}[X \in B] = \int_B f(x)\,dx.$$

The *differential entropy* of X, respectively of the density function f, is defined as

$$\mathfrak{h}(X) = \mathfrak{h}(f) = -\int_{\mathbb{R}} f(x)\log_2 f(x)\,dx = \mathbb{E}(-\log_2 f(X)),$$

whenever the integral exists (in the sense of Lebesgue integration).

As with discrete entropy, we use the convention that $0\log_2 0 := 0$, and so we can think of the integral as being taken over the set $\{x\colon f(x) > 0\}$. An immediate observation is, by the translation invariance of the Lebesgue measure, for any $a \in \mathbb{R}$

$$\mathfrak{h}(X - a) = \mathfrak{h}(X).$$

Example 6.2. (a) *Continuous equidistribution on an interval $[a, b]$*

In this case

$$f_X = \frac{1}{b-a}\mathbb{1}_{[a,b]}$$

and so we can calculate

$$\mathfrak{h}(X) = -\int_a^b \frac{1}{b-a}\log_2\frac{1}{b-a}\,dx = \log_2(b-a).$$

Already here we see some fundamental differences to the discrete entropy function. When $b - a = 1$, the entropy is $\log_2 1 = 0$, and if $b - a < 1$ then the entropy is negative!

(b) *Normal distribution $N(\mu, \sigma^2)$*

By the comment above about translation invariance, we may assume that $\mu = 0$, in which case

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right) \qquad \text{and} \qquad \log_2 f(x) = -\frac{1}{\ln 2}\left(\frac{x^2}{2\sigma^2} + \ln\left(\sqrt{2\pi}\sigma\right)\right),$$

and we can calculate

$$\begin{aligned}
\mathfrak{h}(X) &= \frac{1}{\ln 2}\int_{\mathbb{R}} f(x)\left(\frac{x^2}{2\sigma^2} + \ln\left(\sqrt{2\pi}\sigma\right)\right)dx \\
&= \frac{1}{\ln 2}\left(\int_{\mathbb{R}} f(x)\frac{x^2}{2\sigma^2}\,dx + \int_{\mathbb{R}} \ln\left(\sqrt{2\pi}\sigma\right) f(x)\,dx\right) \\
&= \frac{1}{\ln 2}\left(\frac{\sigma^2}{2\sigma^2} + \ln\left(\sqrt{2\pi}\sigma\right)\right) \\
&= \frac{1}{\ln 2}\left(\frac{1}{2} + \ln\left(\sqrt{2\pi}\sigma\right)\right) \\
&= \frac{1}{2}\log_2(2\pi e\sigma^2).
\end{aligned}$$

Lemma 6.3. *Let X be a real, continuous random variable and let $a \in \mathbb{R}$ be non-zero. Then*

$$\mathfrak{h}(aX) = \mathfrak{h}(X) + \log|a|.$$

Proof. □

6.2 Discretization

Let us 'compare' in a way the differential entropy to the discrete entropy by way of *discretization*.

Suppose we have a random variable X on $\mathbb{R}$ which has a 'well-behaved' density function, say which is continuious on some open (bounded or unbounded) interval and is zero outside of the closure of that interval. In this case we can subdivide this interval into finitely or countably many disjoint intervals I_k, each of length $\delta > 0$.

Now, by the Mean Value Theorem for integrals, there is some $x_k = x_k(\delta) \in I_k^\circ$ (the interior of the interval) such that

$$\int_{I_k} f(x)\,dx = \delta f(x_k).$$

We can then define a discrete approximation to X, which we denote by X_δ defined as

$$X_\delta(\omega) = x_k, \text{ if } X(\omega) \in I_k.$$

X_δ is then a discrete random variable, where for each k $\mathbb{P}[X_\delta = x_k] = \mathbb{P}[X \in I_k] = \delta f(x_k)$.

We can calculate then the discrete entropy of the random variable X_δ.

$$\begin{aligned}
H(X_\delta) &= \sum_k -\delta f(x_k)\log_2(\delta f(x_k)) \\
&= \sum_k -\delta f(x_k)\log_2\delta - \delta f(x_k)\log_2 f(x_k)) \\
&= -\log_2\delta - \sum_k \delta f(x_k)\log_2 f(x_k)).
\end{aligned}$$

However, the latter sum is a Riemann sum for the integral $\int_{\mathbb{R}} f(x) \log_2 f(x)$, and so as $\delta \to 0$

$$H(X_\delta) \approx -\log_2 \delta + \mathfrak{h}(X),$$

where we have glossed over some technical details of convergence if the interval is unbounded.

In particular, taking $\delta = \frac{1}{n}$ we see that

$$H(X_{\frac{1}{n}}) \approx \log_2 n + \mathfrak{h}(X).$$

So, it is not the case that the differential entropy is merely the limit of the discrete entropy of a sequence of sufficiently fine discrete approximatiomns to our random variable, the entropy of these approximations will grow unboudnedly. However, we can recover an approximation to the differential entropy

$$\mathfrak{h}(X) \approx H(X_\delta) + \log_2 \delta,$$

by taking into account the approximation parameter δ.

6.3 Joint and conditional differential entropy

More generally, if we have a collection of jointly distributed random variables $X_1, \ldots, X_n$ we can think of them as a random vector $X = (X_1, \ldots, X_n)^T$ distributed on $\mathbb{R}^n$.

We have then a corresponding density function[1] $f_X : \mathbb{R}^n \to \mathbb{R}$ with respect to the Lebesgue measure.

Definition 6.4. Given a random vector $X = (X_1, \ldots, X_n)^T$ with joint density function f_X, the *joint differential entropy* is defined as

$$\mathfrak{h}(X) = \mathfrak{h}(f_X) := -\int_{\mathbb{R}^n} f(\boldsymbol{x}) \log_2 f(\boldsymbol{x})\, d\boldsymbol{x} \qquad \text{where } \boldsymbol{x} = (x_1, \ldots, x_n)^T,$$

when the integral exists in the sense of Lebesgue integration. As before, it is easy to see that the differential entropy is translation invariant.

Given two random vectors X and Y, with a joint density function $f_{X,Y}$ and a conditional density function $f_{X|Y}$ we have in general[2]

$$f_{X|Y}(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{f_{X,Y}(\boldsymbol{x}, \boldsymbol{y})}{f_Y(\boldsymbol{y})}$$

and we can define

$$\mathfrak{h}(X \mid Y = \boldsymbol{y}) := \mathfrak{h}(f_{X|Y}(\cdot \mid \boldsymbol{y}))$$

and then the *conditional differential entropy* as

$$\mathfrak{h}(X \mid Y) := -\int_{\mathbb{R}^n} \mathfrak{h}(X \mid Y = \boldsymbol{y}) f_Y(\boldsymbol{y})\, d\boldsymbol{y},$$

where $f_Y(\boldsymbol{y}) = \int_{\mathbb{R}_m} f_{X,Y}(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x}$, as always if the integral exists.

[1] Note that in general, even if $X_1, \ldots, X_n$ have density functions, it may be that X does not.

[2] Glossing over some technicalities.

It is relatively easy to verify that, if all the involved integrals are finite, then

$$\mathfrak{h}(X \mid Y) = -\int_{\mathbb{R}^m \times \mathbb{R}^n} f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}) \log_2 f_{X|Y}(\boldsymbol{x} \mid \boldsymbol{y})\, d\boldsymbol{x}\, d\boldsymbol{y} = \mathfrak{h}(X, Y) - \mathfrak{h}(Y).$$

The following lemma then follows inductively from the definitions as in the proof of Theorem 2.13.

Lemma 6.5 (Chain rule for differential entropy)**.** *Given a continuous random vector* $X = (X_1, \ldots, X_n)^t$

$$\mathfrak{h}(X) = \sum_{i=1}^{n} \mathfrak{h}(X_i \mid X_1, X_2, \ldots, X_{i-1}).$$

We also have the following multidimensional version of Lemma 6.3 for how the entropy of a random vector scales under invertible linear transformations.

Lemma 6.6. *Let* $X = (X_1, \ldots, X_n)$ *be a real, continuous random vector and let* A *be a non-singular* $(n \times n)$*-matrix and* $\boldsymbol{b} \in \mathbb{R}^n$*. Then*

$$\mathfrak{h}(AX + \boldsymbol{b}) = \mathfrak{h}(X) + \log_2 |\det A|.$$

Proof. □

Example 6.7. The general n-dimensional normal distribution is determined by two parameters $\boldsymbol{u}^T \in \mathbb{R}^n$ and a positive definite $(n \times n)$-matrix Σ and has density function

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \sigma^{-1} (\boldsymbol{x} - \boldsymbol{u})\right).$$

It can be calculated that $N(\boldsymbol{u}, \Sigma) = (X_1, \ldots, X_n)^T$ is such that

$$\mathbb{E}(X) = \boldsymbol{u}^T \text{ and } \Sigma = (\mathrm{Cov}(X_i, X_j))_{i \in [n], j \in [n]}.$$

This distribution can also be obtained as follows: We start with a vector $Y = (Y_1, \ldots, Y_n)^T$ of i.i.d standard normal random variables, a non-singular $(n \times n)$-matrix A and some vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. If we let

$$X = (X_1, \ldots, X_n)^T = AT + \boldsymbol{\mu},$$

then we can compute

$$\mathbb{E}(X) = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n))^T = \boldsymbol{\mu}$$

and the covariance matrix is given by

$$\Sigma = (\mathrm{Cov}(X_i, X_j))_{i \in [n], j \in [n]} = AA^T.$$

Finally, a computation will show that the density of X is given by

$$f_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \sigma^{-1} (\boldsymbol{x} - \boldsymbol{u})\right),$$

and so $X \sim N(\boldsymbol{u}, \Sigma)$.

Hence, it follows from Lemma 6.6 that

$$\mathfrak{h}(N(\boldsymbol{u}, \Sigma)) = \mathfrak{h}(Y_1, \ldots, Y_n) + \log_2 |\det A|.$$

Now, since $\Sigma = AA^T$, it follows that $|\det A| = \sqrt{|\det \Sigma|}$, and we will show shortly that for independent random variables $\mathfrak{h}(Y_1, \ldots, Y_n) = \sum_{i=1}^n \mathfrak{h}(Y_i)$ and hence

$$\mathfrak{h}(N(\boldsymbol{u}, \Sigma)) = \frac{n}{2} \log_2(2\pi e) + \frac{1}{2} \log_2 |\det \Sigma| = \frac{1}{2} \log_2 \left((2\pi e)^n |\det \Sigma|\right).$$

Definition 6.8. Given two density functions f and g on $\mathbb{R}^n$ we can define the *Kullback-Leibler Divergence* as

$$D(f \parallel g) := \int_{\mathbb{R}^n} f(\boldsymbol{x}) \log_2 \left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \, d\boldsymbol{x} \right)$$

under usual conventions about the value of $a \log_2 b$ when a or b equal 0.

It can be checked that if P_f and P_g are the associated probability measures, then $D((\parallel f), g)$ will be finite only if $g > 0$ holds P_f almost everywhere.

Theorem 6.9. *[Information Inequality] Let f and g be density functions on $\mathbb{R}^n$, then*

$$D(f \parallel g) \geq 0$$

with equality if and only if $f = g$ almost everywhere.

Proof. □

Definition 6.10. Given two random vectors X and Y, with a joint density function $f_{X,Y}$ we define

$$I(X \,;\, Y) = D(f_{X,Y} \parallel f_X \otimes f_Y).$$

One can argue in precisely the same way as in the discrete case that the following holds (whenevr all the relevant integrals are finite).

Lemma 6.11. *1. $I(X \,;\, Y) = \mathfrak{h}(X) - \mathfrak{h}(X \mid Y) = \mathfrak{h}(Y) - \mathfrak{h}(Y \mid X) == \mathfrak{h}(X) + \mathfrak{h}(Y) - \mathfrak{h}(X, Y)$;*

2. $I(X \,;\, Y) \geq 0$ with equality if and only if X and Y are independent;

3. $\mathfrak{h}(X, Y) \leq \mathfrak{h}(X) + \mathfrak{h}(Y)$ with equality if and only if X and Y are independent.

Remark 6.12. *Whilst we can only define the differential entropy for random variables with 'well-behaved' density functions, using our discretisation process we can define a sensible notion of mutual information for arbitrary continuous, real random vectors.*

Indeed, considering just the one dimensional case, given jointly distributed random variables X and Y we could consider the discrete approximations X_δ and Y_δ. Then, one can similarly show that

$$I(X_\delta \,;\, Y_\delta) = H(X_\delta) - H(X_\delta \mid Y_\delta) \approx \mathfrak{h}(X) - \log_2 \delta - (\mathfrak{h}(X \mid Y) - \log_2 \delta) = I(X \,;\, Y).$$

In the general case, one can show that the mutual information of X and Y can be obtained as the limit of the mutual information of a sequence of finer and finer discrete approximations to the pair (X, Y).

Theorem 6.13. *Let $X = (X_1, \ldots, X_n)^T$ be a continuous real random vector with $\mathbb{E}(X) = \mathbf{0}$ and with a positive definite covariance matrix $\Sigma = (Cov(X_i, X_j))_{i\in[n], j\in[n]}$. Then*

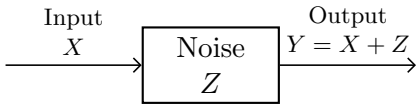
$$\mathfrak{h}(X) \leq \mathfrak{h}\left(N(\mathbf{0}, \Sigma)\right),$$

with equality if and only if $X \sim N(\mathbf{0}, \Sigma)$.

Proof. □

6.4 The Gaussian Channel

This is a simple channel where the input is some continuous, real random variable X. During transmission some Gaussian random noise Z with distribution $N(0, \sigma^2)$ is added to the signal, and this noise is *additive*, so that the ouput Y is given by $X + Z$.



As with discrete channels, we can consider the nth *channel extension*, where the input consists of a sequence $X_1, \ldots, X_n$ of independent random variables (which we will consider to be i.i.d copies of some fixed X) and there is an independent sequence of i.i.d copies $Z_1, \ldots, Z_n$ of the Gaussian noise Z, and the output is then a sequence $Y_1, \ldots, Y_k$ of i.i.d random variables where $Y_i = X_i + Z_i$ for each $i \in [k]$.

A typical assumption on the input distribution X is that $\mathbb{E}(X^2) \leq \tau^2$ for some pre-determined constant $\tau > 0$.

Definition 6.14. The *capacity* of the Gaussian channel with parameters σ^2 and τ^2 is defined as

$$\text{cap}\left(\sigma^2, \tau^2\right) = \max\left\{ I(X;Y) : \begin{array}{l} X \text{ a continuous, real random variable with } \mathbb{E}(X^2) \leq \tau^2 \text{ and} \\ Y = X + Z \text{ where } Z \sim N(0, \sigma^2) \text{ independent of } X \end{array} \right\}.$$

Theorem 6.15. *For any $\sigma^2, \tau^2 > 0$*

$$cap\left(\sigma^2, \tau^2\right) = \frac{1}{2}\log_2\left(1 + \frac{\tau^2}{\sigma^2}\right)$$

and the maximum is achieved when $X \sim N(0, \tau^2)$.