Information Theory

Stefan Höst

March 9, 2012

Draft version

V001	First Draft Outline	March 2012

Dept. of Electrical and Information Technology
Lund University

Contents

Chapter 3

Probability

One of the basic insights when setting up a measure for information is that the observed quantity must have choices to contain any information. So, for us to set up a theory for information we must have a clear notation of probability theory. This chapter will give the parts of the probability theory that is needed throughout the rest of the document. It is not, in any way, a complete course in probability theory, for that we refer to e.g.[4] or some other standard text book in the area.

In short, a probability is a measure of how likely an event is to occur. It is represented by a number between zero and one, where zero means it will not happen and one that is certain to happen. The sum of the probabilities for all possible events is one, since it is certain that one of them will happen.

More formally, it was the Russian mathematician Kolmogorov who in 1933 set up the theory as we know it today. The **sample space** Ω is the set of all possible outcomes of the random experiment. In discrete probability theory the sample space is, in general, countably infinite,

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$$

An **event** is a subset of the sample space, $A \subseteq \Omega$. The event is said to occur if the outcome from the random experiment is found in the event. Examples of specific events are

- The impossible event $\emptyset$ (the empty subset of Ω)
- The certain event Ω (the full subset)

Each of the elements in Ω, $\omega_1, \omega_2, \dots, \omega_n$, i.e. the smallest nonempty subsets, are called elementary events or, sometimes, atomic events.

To each event there is assigned a **probability measure** P, where $0 \leq P \leq 1$, which is a measure of how probable the event is to happen. Some fundamental probabilities are

$$\begin{aligned} P(\Omega) &= 1 \\ P(\emptyset) &= 0 \\ P(A \cup B) &= P(A) + P(B), \text{ if } A \cap B = \emptyset \end{aligned}$$

where A and B are disjoint events. This imply that

$$P(A) = \sum_{\omega_i \in A} P(\omega_i)$$

An important concept in probability theory is how two events are related. This is described by the *conditional probability*. The probability of event A conditioned on the event B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

If the two events A and B are independent, the probability for A should not change if it is conditioned on B or not. Hence,

$$P(A|B) = P(A)$$

Applying this fact to the definition of conditional probabilities we can conclude that A and B are independent if and only if the probability of their intersection is equal to the product of the individual probabilities,

$$P(A \cap B) = P(A) \cdot P(B)$$

3.1 Random variable

A **random variable**, or stochastic variable, X is a function from the sample space into a specified set, most often the real numbers,

$$X : \Omega \to \mathcal{X}$$

where $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$, $k \leq n$, is the set of possible values. We denote the event $\{\omega | X(\omega) = x_i\}$ by $\{X = x_i\}$.

To describe how the probabilities for the random variable is distributed we use, for the discrete case, the *probability function*

$$p_X(x) = P(X = x)$$

and in the continuous case the *density function* $f_X(x)$. The *distribution function* for the variable can now be defined as,

$$F_X(x) = P(X \leq x) \tag{3.1}$$

which we can derive as[1]

$$F_X(x) = \begin{cases} \displaystyle\sum_{k \leq x} p_X(k) & X \text{ discrete} \\ \\ \displaystyle\int_{-\infty}^{x} f_X(\nu) d\nu & X \text{ continuous} \end{cases}$$

[1] Often in this chapter we will give the result both for discrete and continuous variables, as done here. If one of the equations are omitted this does not mean it does not exist. In most cases it is straight forward to get the discrete or continuous counterpart.

In many occasions later in the document we will omit the index $\cdot_X$, if it is clear what variable is used.

To consider how two or more random variables are jointly distributed we can view a vector of random variables. This vector will then represent a multi-dimensional random variable and can be treated the same way. Hence, we can consider (X, Y), where X and Y are random variables with possible outcomes $\mathcal{X} = \{x_1, x_2, \ldots, x_M\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$, respectively. Then the set of joint outcomes is

$$\mathcal{X} \times \mathcal{Y} = \Big\{(x_1, y_1), (x_1, y_2), \ldots, (x_M, y_N)\Big\}$$

Normally we write the *joint probability function* $p_{X,Y}(x, y)$ for the discrete case, meaning the event $P(X = x \text{ and } Y = y)$, and the corresponding *joint density function* is denoted $f_{X,Y}(x, y)$.

The *marginal distribution* can be derived from the joint distribution as

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

and

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)dy$$

for the discrete and continuous case, respectively. This leads us to consider how dependent the two variables are. We say that the *conditional probability distribution* is defined by

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

and

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

This gives a probability distribution for the variable X when the outcome for Y is known. Repeating this iteratively, we conclude the *chain rule for probabilities* for the probability of an n-dimensional random variable as

$$\begin{aligned} P(X_1 \ldots X_{n-1} X_n) &= P(X_n|X_1 \ldots X_{n-1})P(X_1 \ldots X_{n-1}) \\ &= P(X_1)P(X_2|X_1)P(X_3|X_1X_2) \ldots P(X_n|X_1 \ldots X_{n-1}) \\ &= \prod_{i=1}^{n} P(X_i|X_1 \ldots X_{i-1}) \end{aligned} \tag{3.2}$$

Combining the above results we conclude that two random variables X and Y are *statistically independent* if and only if

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

or

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

for all x and y.

3.2 Expectation and variance

Consider an example where the random variable X is the outcome of a throw with a fair dice. The probabilities for each outcome is equivalent, i.e.

$$p_X(x) = \frac{1}{6}, \quad x = 1, 2, \dots, 6$$

The arithmetic mean of the outcomes is

$$\overline{X} = \frac{1}{6}\sum_x x = 3.5$$

Now, instead have a look at a counterfeit dice where there is a small weight place close to the number one and let Y be the corresponding random variable. Simplifying the results we assume that the number one will never show and number six will show twice as often as before. That is, $p_Y(1) = 0$, $p_Y(i) = \frac{1}{6}, i = 2, 3, 4, 5$ and $p_Y(6) = \frac{1}{3}$. The probabilities for the other numbers are unchanged. The arithmetic mean above will still be the same 3.5, but it does not say very much about the actual result. Therefore, we introduce the *expected value*, which is a mean weighted with the probabilities for the outcomes,

$$E[X] = \sum_x x p_X(x) \tag{3.3}$$

In the case with the fair dice the expected value is the same as the arithmetic mean, but for the manipulated dice it becomes

$$E[Y] = 0 \cdot 1 + \tfrac{1}{6} \cdot 2 + \tfrac{1}{6} \cdot 3 + \tfrac{1}{6} \cdot 4 + \tfrac{1}{6} \cdot 5 + \tfrac{1}{3} \cdot 6 \approx 4.33$$

A slightly more general definition of the expected value is as follows.

Definition 1 *Let $g(X)$ be a real-valued function of a random variable X, the* ***expected value*** *(or mean) of $g(X)$ is for a discrete variable*

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

and for a continuous variable

$$E\big[g(X)\big] = \int_{\mathbb{R}} g(\nu) f_X(\nu) d\nu$$

□

Going back to our true and counterfeit dice, we can use the function $g(x) = x^2$. This will lead to the so called second order moment for the variable X. In the case for the true dice we get

$$E[X^2] = \sum_{i=1}^{6} \frac{1}{6} i^2 \approx 15.2$$

and for the counterfeit

$$E[Y^2] = \sum_{i=2}^{5} \frac{1}{6} i^2 + \frac{1}{3} 6^2 = 21$$

Later on, in Section 5.1, we will make use of the weak law of large numbers. It states that the arithmetic mean of a vector consisting of independent identically (i.i.d.) distributed random variables will approach the expected value as the length of the vector grows. In this sense the expected value is a most natural definition of the mean for the outcome of a random variable.

The expected value for the multi-dimensional variable is derived similarly, with the joint distribution. In the case of a 2-dimensional vector (X, Y) it becomes,

$$E\big[g(X,Y)\big] = \begin{cases} \displaystyle\sum_{x,y} g(x,y) p_{X,Y}(x,y) & \text{discrete case} \\ \\ \displaystyle\int_{\mathbb{R}^2} g(\nu,\mu) f_X(\nu,\mu) d\nu d\mu & \text{continuous case} \end{cases}$$

It can also be that one variable is discrete and one is continuous, so the formula has one sum and one integral.

From the definition of the expected value and basic calculus it is easy to verify that the expected value is a linear mapping, i.e. that

$$\begin{aligned} E[aX + bY] &= \sum_{x,y} (ax + by) p_{XY}(x,y) \\ &= \sum_{x,y} ax p_{XY}(x,y) + \sum_{x,y} by p_{XY}(x,y) \\ &= a \sum_{x} x \sum_{y} p_{XY}(x,y) + b \sum_{y} y \sum_{x} p_{XY}(x,y) \\ &= a \sum_{x} x p_X(x) + b \sum_{y} y p_Y(y) \\ &= aE[X] + bE[Y] \qquad (3.4) \end{aligned}$$

At the end of this section there will be stated a more general version of this theorem.

We can also verify that if X and Y are independent, then the expectation of their product equals the product of their expectations,

$$\begin{aligned} E[XY] &= \sum_{x,y} xy p_{X,Y}(x,y) = \sum_{x,y} xy p_X(x) p_Y(y) \\ &= \sum_{x} x p_X(x) \sum_{y} x p_Y(y) = E[X]E[Y] \qquad (3.5) \end{aligned}$$

While the expected value is a measure of the weighted mean for the outcome of a random variable, we also need a measure of its variation. This value is called the *variance* and is defined as the expected value of the squared distance to the mean.

Definition 2 *Let X be a random variable with expected value $E[X]$. The variance of the variable is*

$$V[X] = E\big[(X - E[X])^2\big]$$

□

The variance can often be derived from

$$\begin{aligned} V[X] &= E\big[(X - E[X])^2\big] = E\big[X^2 - 2XE[X] + E[X]^2\big] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - E[X]^2 \end{aligned}$$

where $E[X^2]$ is the second order moment of X. In many descriptions the expected value is denoted by m, but we have here chosen to preserve the notation $E[X]$. Still, it should be regarded as a constant in the derivations.

Often the *standard deviation*, σ_X, is used as a measure of the variation of the variable. This is the square root of the variance,

$$\sigma_X^2 = V[X]$$

For the true and counterfeit dice described earlier in this section we get the variance

$$\begin{aligned} V[X] &= E[X^2] - E[X]^2 \approx 2.9 \\ V[Y] &= E[Y^2] - E[Y]^2 \approx 2.2 \end{aligned}$$

The corresponding standard deviations are

$$\begin{aligned} \sigma_X &= \sqrt{V[X]} \approx 1.7 \\ \sigma_Y &= \sqrt{V[Y]} \approx 1.5 \end{aligned}$$

To get more understanding about the variance we first need to define one of its relatives, the covariance function. This can be seen as a measure of the dependencies between two random variables.

Definition 3 *Let X and Y be two random variable with expected value $E[X]$ and $E[Y]$. The covariance between the variables is*

$$\mathrm{Cov}(X, Y) = E\big[(X - E[X])(Y - E[Y])\big]$$

□

We can easily rewrite the covariance to get

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E\big[(X - E[X])(Y - E[Y])\big] \\ &= E\big[XY - XE[Y] - E[X]Y + E[X]E[Y]\big] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

From this result, and from (3.5), we see that when X and Y are independent we get zero covariance,

$$\mathrm{Cov}(X,Y) = 0 \tag{3.6}$$

Going back to the variance we can now derive the variance for the combination $aX + bY$ as

$$\begin{aligned} V[aX+bY] &= E\big[(aX+bY)^2\big] - E[aX+bY]^2 \\ &= E\big[a^2X^2 + b^2Y^2 + 2abXY\big] - \big(aE[X] + bE[Y]\big)^2 \\ &= a^2E[X^2] + b^2E[Y^2] + 2abE[XY] \\ &\quad - a^2E[X]^2 - b^2E[Y]^2 - 2abE[X]E[Y] \\ &= a^2V[X] + b^2V[Y] + 2ab\mathrm{Cov}(X,Y) \end{aligned} \tag{3.7}$$

With use of (3.6) we can also conclude that if X and Y are independent we get

$$V[aX+bY] = a^2V[X] + b^2V[Y] \tag{3.8}$$

Sometimes it is suitable to have a normalized random variable. We can get that by considering

$$\tilde{X} = \frac{X-m}{\sigma}$$

where $E[X] = m$ and $V[X] = \sigma^2$. Then the expectation and variance becomes

$$\begin{aligned} E[\tilde{X}] &= \frac{1}{\sigma}\big(E[X] - m\big) = 0 \\ V[\tilde{X}] &= \frac{1}{\sigma^2}E[(X-m)^2] = 1 \end{aligned}$$

To conclude the part about expected value and variance, we summarize and give more general versions of (3.4), (3.7) and (3.8). The results can be shown very similar to what has been done in the description above.

Theorem 1 *Given a set of N random variables X_n and scalar constants α_n, $n = 1, 2, \dots, N$, the sum*

$$Y = \sum_{n=1}^{N} \alpha_n X_n$$

has the expected value

$$E[Y] = \sum_{n=1}^{N} \alpha_n E[X_n]$$

and the variance

$$V[Y] = \sum_{n=1}^{N} \alpha_n^2 V[X_n] + 2\sum_{m<n} \alpha_n \alpha_m \mathrm{Cov}(X_n, X_m)$$

If the random variables X_n all are independent, then

$$V[Y] = \sum_{n=1}^{N} \alpha_n^2 V[X_n]$$

□

We have already seen a distribution for the outcome of a dice. We will finish this section by considering the geometric distribution and the Gaussian distributions, both often used in technical contents.

Example 1 [Geometric distribution]
The *geometric distribution* is a discrete distribution that can be explained from a set of coin flips. Assume that the probability for head and tail is given by $P(\text{head}) = \alpha$ and $P(\text{tail}) = 1 - \alpha$. Let K be the number of flips until a tail shows up. The probability for this number to be k is

$$p_K(k) = \alpha^{k-1}(1-\alpha), \quad k = 1, 2, \ldots$$

and the density function

$$F_K(k) = \sum_{i=1}^{k} \alpha^{k-1}(1-\alpha) = (1-\alpha)\sum_{n=0}^{k-1} \alpha^n = (1-\alpha)\frac{1-\alpha^k}{1-\alpha} = 1 - \alpha^k$$

We can directly see that this is indeed a probability distribution if we let $k \to \infty$ for the density function to get $F_K(k) \to 1$. Another way is to sum over all possible probabilities,

$$\sum_{k=1}^{\infty} \alpha^{k-1}(1-\alpha) = (1-\alpha)\sum_{n=0}^{\infty} \alpha^n = (1-\alpha)\frac{1}{1-\alpha} = 1$$

Here, and in the remaining of this example we will make use of the well known sums, for $|\alpha| < 1$,

$$\sum_{n=0}^{\infty} \alpha^n = \frac{1}{1-\alpha}, \quad \sum_{n=0}^{\infty} n\alpha^n = \frac{\alpha}{(1-\alpha)^2}, \quad \sum_{n=0}^{\infty} n^2\alpha^n = \frac{\alpha(1+\alpha)}{(1-\alpha)^3} \tag{3.9}$$

We can also find the expected value and the second order moment as

$$\begin{aligned} E[K] &= \sum_{k=1}^{\infty} k\alpha^{k-1}(1-\alpha) = (1-\alpha)\sum_{n=0}^{\infty}(n+1)\alpha^n \\ &= (1-\alpha)\Big(\frac{\alpha}{(1-\alpha)^2} + \frac{1}{1-\alpha}\Big) = \frac{1}{1-\alpha} \\ E[K^2] &= \sum_{k=1}^{\infty} k^2\alpha^{k-1}(1-\alpha) = (1-\alpha)\sum_{n=0}^{\infty}(n+1)^2\alpha^n \\ &= (1-\alpha)\Big(\frac{\alpha(1+\alpha)}{(1-\alpha)^3} + 2\frac{\alpha}{(1-\alpha)^2} + \frac{1}{1-\alpha}\Big) = \frac{1+\alpha}{(1-\alpha)^2} \end{aligned}$$

Hence, the variance is

$$V[K] = E[K^2] + -E[K]^2 = \frac{1+\alpha}{(1-\alpha)^2} - \Big(\frac{1}{1-\alpha}\Big)^2 = \frac{\alpha}{(1-\alpha)^2}$$

For $\alpha = \frac{1}{2}$, i.e. a fair coin, we obtain

$$\begin{aligned} p_k(k) &= \big(\tfrac{1}{2}\big)^k \\ E[K] &= 2 \\ E[K^2] &= 6 \\ V[K] &= 2 \end{aligned}$$

The next example treats the Gaussian distribution.

Example 2 [Gaussian distribution] The Gaussian distribution, or Normal distribution is a very central distribution in probability theory. It is also a very common distribution to use for modelling continuous channels. The distribution is often denoted $X \sim \mathcal{N}(m, \sigma)$ and the density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{3.10}$$

To show that this is actually a density function, and to derive the expectation and variance, we start with a version where we center the function at 0, $Y \sim \mathcal{N}(0, \sigma)$, with density function

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}$$

To show that this is a density function, i.e. that $\int_{-\infty}^{\infty} f_Y(y)dy = 1$ we consider the squared function,

$$\begin{aligned} \Big(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy\Big)^2 &= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}} dy \int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma^2}} dz \\ &= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{y^2+z^2}{2\sigma^2}} dydz \\ &= \frac{1}{2\pi\sigma^2} \int_{0}^{2\pi} \int_{0}^{\infty} re^{-\frac{r^2\cos^2\phi + r^2\sin^2\phi}{2\sigma^2}} drd\phi \\ &= \frac{1}{2\pi} \int_{0}^{2\pi} \int_{0}^{\infty} \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} drd\phi \\ &= \frac{1}{2\pi} \int_{0}^{2\pi} \Big[-e^{-\frac{r^2}{2\sigma^2}}\Big]_0^{\infty} d\phi = \frac{1}{2\pi} \int_{0}^{2\pi} d\phi = 1 \end{aligned}$$

where we used the standard variable change to polar coordinates. The expectation and the second order moment can be derived as

$$
\begin{aligned}
E[Y] &= \int_{-\infty}^{\infty} y\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}dy = \frac{\sigma^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\frac{y}{\sigma^2}e^{-\frac{y^2}{2\sigma^2}}dy\\
&= \frac{\sigma^2}{\sqrt{2\pi\sigma^2}}\Big[-e^{-\frac{y^2}{2\sigma^2}}\Big]_{-\infty}^{\infty} = 0\\
E[Y^2] &= \int_{-\infty}^{\infty} y^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}dy = \frac{\sigma^2}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} y\frac{y}{\sigma^2}e^{-\frac{y^2}{2\sigma^2}}dy\\
&= \frac{\sigma^2}{\sqrt{2\pi\sigma^2}}\Big(\Big[-ye^{-\frac{y^2}{2\sigma^2}}\Big]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-\frac{y^2}{2\sigma^2}}dy\Big)\\
&= \sigma^2\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}dy = \sigma^2
\end{aligned}
$$

To make the same derivations in the more general case $X \sim \mathcal{N}(m,\sigma)$ we will make use of the variable change $y = x - m$ (implying $dy = dx$ and $x = y + m$). Then $Y \sim \mathcal{N}(0,\sigma)$, and we get

$$
\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}dy = 1
$$

so, it is a density function. The expectation and second order moment is

$$
\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}dx = \int_{-\infty}^{\infty}(y+m)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}dy\\
&= E[Y] + m\int_{-\infty}^{\infty} f_Y(y)dy = m\\
E[X^2] &= \int_{-\infty}^{\infty} x^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}dx = \int_{-\infty}^{\infty}(y+m)^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}dy\\
&= E[Y^2] + 2mE[Y] + m^2 = \sigma^2 + m^2
\end{aligned}
$$

and the variance

$$
V[X] = E[X^2] - E[X]^2 = \sigma^2
$$

3.3 Weak law of large numbers

The weak law of large numbers will play a central role when proving the source coding theorem, which sets a bound on the compression ratio, as well as the channel coding theorem, which sets a limit on the capability of reliable communication. Those two theorems are very important results in information theory. In this section first the Markov's inequality and then Chebyshev's inequality will be stated. These are famous bounds in statistics. The latter will be used to show the weak law of large numbers.

The first inequality is stated in the following theorem. Here we will only consider the case for discrete random variables but the result holds also for continuous variables.

Theorem 2 (Markov's inequality) *Let X be a non-zero random variable with finite expected value. Then, for any non-zero integer a*

$$P(X > a) \leq \frac{E[X]}{a}$$

□

To show this we start with the expected value,

$$E[X] = \sum_{k=0}^{\infty} kp(k) \leq \sum_{k=a+1}^{\infty} kp(k) \leq \sum_{k=a+1}^{\infty} ap(k) = a \sum_{k=a+1}^{\infty} p(k) = aP(X > a)$$

which gives the result.

If we instead of X consider the squared distance to the expected value $E[X]$ we get the following result.

$$P\Big(\big(X - E[X]\big)^2 > \epsilon^2\Big) \leq \frac{E[(X - E[X])^2]}{\epsilon^2} = \frac{V[X]}{\epsilon^2}$$

Equivalently this can be stated as in the following theorem, which is Chebyshev's inequality.

Theorem 3 (Chebyshev's inequality) *Let X be a non-zero random variable with finite expected value $E[X]$. Then, for any positive ϵ*

$$P\Big(\big|X - E[X]\big| > \epsilon\Big) \leq \frac{V[X]}{\epsilon^2}$$

□

As stated previously this can be used to give a first proof of the weak law of large numbers. Consider a sequence of independent identically distributed (i.i.d.) random variables, X_i, $i = 1, 2, \ldots, n$. The arithmetic mean for the sequence can be viewed as a new random variable,

$$Y = \frac{1}{n} \sum_{i=1}^{n} X_i$$

From Theorem 1 we see that the expected value and variance of Y can be expressed as

$$E[Y] = E[X]$$
$$V[Y] = \frac{V[X]}{n}$$

Applying Chebyshev's inequality yields

$$P\Big(\Big|\frac{1}{n} \sum_{i=1}^{n} X_i - E[X]\Big| > \epsilon\Big) \leq \frac{V[X]}{n\epsilon^2}$$

As n grows the right hand side will tend to zero, bounding the arithmetic mean close to the expected value. Stated differently we get

$$\lim_{n\to\infty} P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\Big| < \epsilon\Big) = 1$$

which gives the *weak law of large numbers*. This probabilistic convergence is denoted $\overset{p}{\to}$ and is read out as *convergence in probability*. This gives the following way to express the relation.

Theorem 4 (Weak law of large numbers) *Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with finite expectation $E[X]$. Form the arithmetic mean as $Y = \frac{1}{n}\sum_i X_i$. Then Y converges in probability to $E[X]$,*

$$Y \overset{p}{\to} E[X], \quad n \to \infty$$

□

It should be noted that the proof given above requires that the variance is finite, but it is possible to show also the more general theorem given here.

To see how the convergence towards the expectation works in practice we give an example with a binary vector.

Example 3 Consider a length n vector of i.i.d.binary random variables with $p_X(0) = 1/3$ and $p_X(1) = 2/3$, $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$. The probability for a vector to have k ones is

$$P(k \text{ 1s in } \boldsymbol{X}) = \binom{n}{k}\Big(\frac{2}{3}\Big)^k\Big(\frac{1}{3}\Big)^{n-k} = \binom{n}{k}\frac{2^k}{3^n}$$

In Figure 3.1 the probability distribution of number of 1s in a vector of length $n = 5$ is shown both as a table and in a graphical version. We see here that it is most likely that we get a vector with 3 or 4 1s. Also it is interesting to observe that the it is less probable to get all five 1s.

In Figure 3.2 the distribution for the number of 1s is shown when the length of the vector is increased to 10, 50, 100 and 500. With increasing length it becomes more evident that the most likely vector has about $L \cdot E[X]$.

While we are dealing with the arithmetic mean of i.i.d. random variables it would not be fair not to mention the central limit theorem. This and the law of large numbers are the two main limits in probability theory. However, we will here give the theorem without a proof (as in many basic probability courses). The result is that the arithmetic mean of a sequence of i.i.d. random variables will be normal distributed, independent of the distribution of of the variables.

Theorem 5 (Central limit theorem) *Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with finite expectation $E[X] = m$ and finite variance $V[X] = \sigma^2$. Form the arithmetic mean as $Y =$*

k	$P(\#1 \text{ in } \boldsymbol{x}) = \binom{5}{k}\frac{2^k}{3^5}$
0	0.0041
1	0.0412
2	0.1646
3	0.3292
4	0.3292
5	0.1317

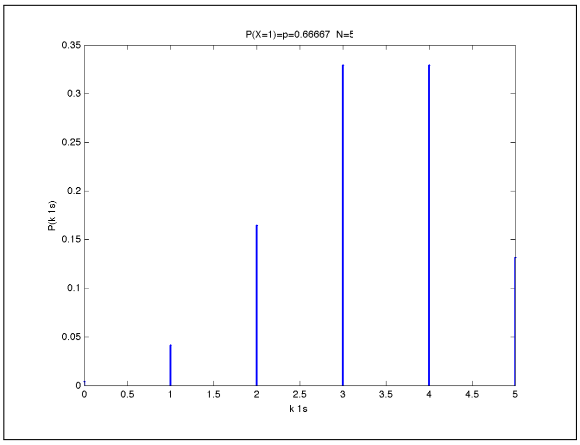


Figure 3.1: Probability distribution for k 1s in a vector of length 5, when $p(1) = 2/3$ and $p(0) = 1/3$.

$\frac{1}{n}\sum_i X_i$. Then, as n goes to infinity, Y becomes distributed according to a normal distribution, i.e.

$$\frac{Y - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \quad n \to \infty$$

□

3.4 Jensen's inequality

In many applications, like optimization, functions that are convex are desirable since they have a structure that is easy to work with. A convex function is defined in the following.

Definition 4 (Convex function) *A function $g(x)$ is **convex** in the interval $[a, b]$ if, for any $x_1, x_2 \in [a, b]$ and any λ, $0 \le \lambda \le 1$,*

$$g\big(\lambda x_1 + (1-\lambda)x_2\big) \le \lambda g(x_1) + (1-\lambda)g(x_2)$$

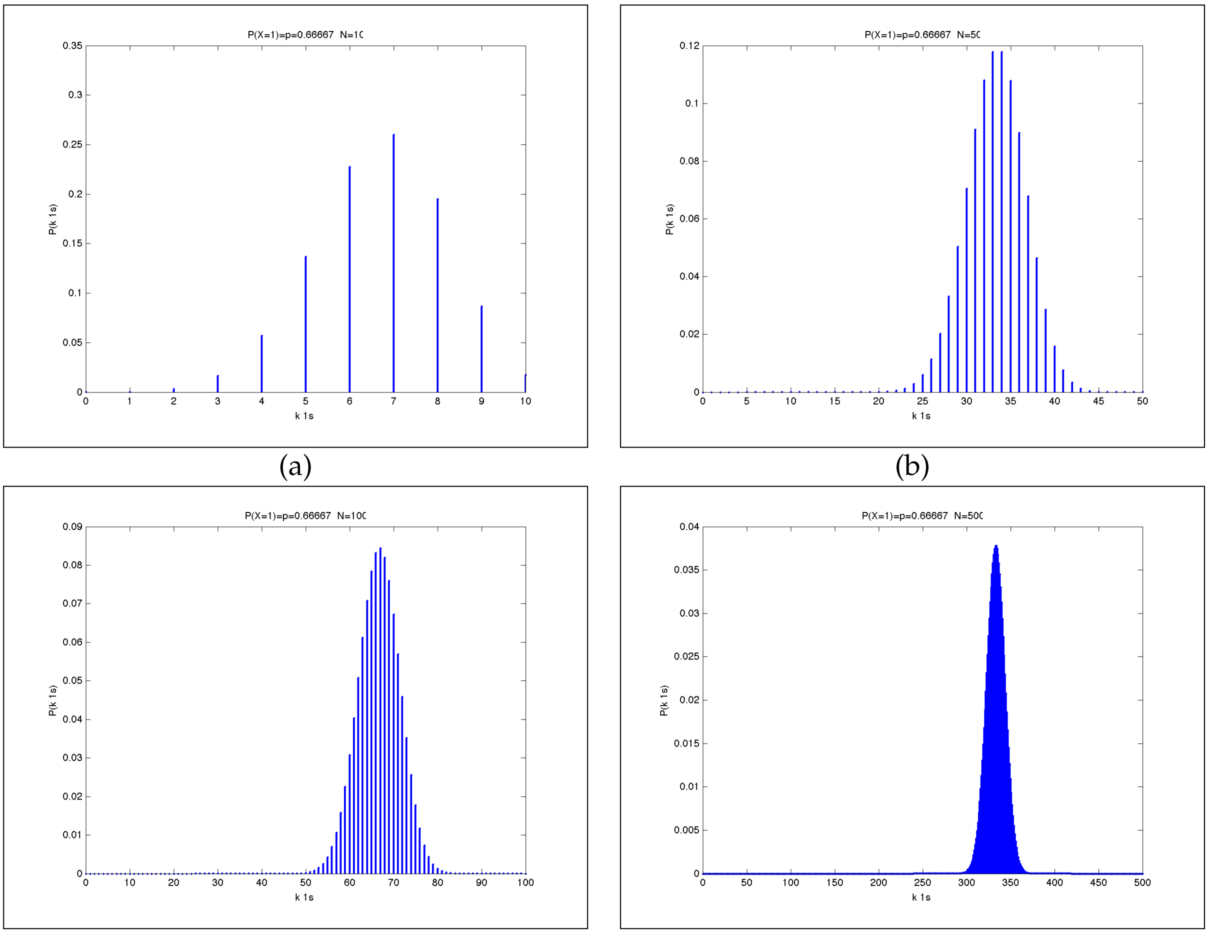
*Similarly, a function $g(x)$ is **concave** in the interval $[a, b]$ if $-g(x)$ is convex in the same interval.*

□

The inequality in the definition can be viewed graphically as in Figure 3.3. Let x_1 and x_2 be two numbers in the interval $[a, b]$, such that $x_1 < x_2$. Then, we can also mark the functions values, $g(x_1)$ and $g(x_2)$, on plot of $g(x)$. For λ in $[0, 1]$, we get that

$$x_1 \le \lambda x_1 + (1-\lambda)x_2 \le x_2$$

with equality to the left if $\lambda = 1$ and to the right if $\lambda = 0$. Then we can also mark the corresponding function value $g\big(\lambda x_1 + (1-\lambda)x_2\big)$. While $\lambda x_1 + (1-\lambda)x_2$ is a value between



(a) (b)

(c) (d)

Figure 3.2: Probability distributions for k 1s in a vector of length 10 (a), 50 (b), 100 (c) and 500 (d), when $p(1) = 2/3$ and $p(0) = 1/3$.

x_1 and x_2, the corresponding value between $g(x_1)$ and $g(x_2)$ is $\lambda g(x_1) + (1-\lambda)g(x_2)$, i.e. the coordinates

$$\big(\lambda x_1 + (1-\lambda)x_2, \lambda g(x_1) + (1-\lambda)g(x_2)\big)$$

describes a straight line between $\big(x_1, g(x_1)\big)$ and $\big(x_2, g(x_2)\big)$ for $0 \le \lambda \le 1$. From this reasoning we see that a convex function is typically shaped like a bowl. Similarly a concave function is the opposite, shaped like a hill.

Example 4 The functions x^2 and e^x are typically convex functions. On the other hand, the functions $-x^2$ and $\log x$ are concave functions.

In the literature the names convex $\cup$ and convex $\cap$ are also used instead of convex and concave.

Since λ and $1-\lambda$ can be interpreted as a binary probability function, both $\lambda x_1 + (1-\lambda)x_2$ and $\lambda g(x_1) + (1-\lambda)g(x_2)$ can be viewed as expected values. A more general way of putting this is Jensen's inequality.

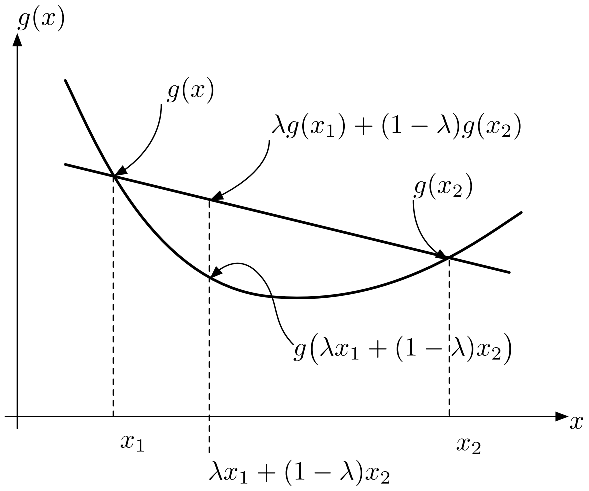

$g(x)$
$g(x)$
$\lambda g(x_1) + (1-\lambda)g(x_2)$
$g(x_2)$
$g\big(\lambda x_1 + (1-\lambda)x_2\big)$
x
x_1
x_2
$\lambda x_1 + (1-\lambda)x_2$

Figure 3.3: A graphical view of the definition of convex functions.

Theorem 6 (Jensen's inequality) *If $f(x)$ is a convex function and X a random variable we have*

$$E[f(X)] \geq f(E[X])$$

If $f(x)$ is a concave function and X a random variable we have

$$E[f(X)] \leq f(E[X])$$

□

Jensen's inequality is so important that a more thorough outline of the proof is required. Even though it will only be shown for the discrete case here, it is also valid in the continuous case. As stated prior to the theorem, the binary case follows directly from the definition of convexity.

To show that the theorem holds also in for distributions with more than two cases we will use induction. Assume that $a_1, a_2, \dots, a_n$ are positive numbers such that $\sum_i a_i = 1$ and that

$$f\big(\sum_i a_i x_i\big) \leq \sum_i a_i f(x_i)$$

where $f(x)$ is a convex function. Furthermore, let $p_1, p_2, \dots, p_{n+1}$ be a probability distribution for X. Then

$$\begin{aligned}
f(E[X]) &= f\Big(\sum_{i=1}^{n+1} p_i x_i\Big) = f\Big(p_1 x_1 + \sum_{i=2}^{n+1} p_i x_i\Big) = f\Big(p_1 x_1 + (1-p_1)\sum_{i=2}^{n+1} \frac{p_i}{1-p_1} x_i\Big) \\
&\overset{(a)}{\leq} p_1 f(x_1) + (1-p_1) f\Big(\sum_{i=2}^{n+1} \frac{p_i}{1-p_1} x_i\Big) \overset{(b)}{\leq} p_1 f(x_1) + (1-p_1)\sum_{i=2}^{n+1} \frac{p_i}{1-p_1} f(x_i) \\
&= \sum_{i=1}^{n+1} p_i f(x_i) = E\big[f(X)\big]
\end{aligned}$$

where inequality (a) comes from the convexity of f and (b) from the induction assumption. What is left to show is that $a_i = \frac{p_i}{1-p_1}$ satisfy the requirements above. Clearly, since p_i is a probability we have that $a_i \geq 0$. The second requirements follows from

$$\sum_{i=2}^{n+1} \frac{p_i}{1-p_1} = \frac{1}{1-p_1}\sum_{i=2}^{n+1} p_i = \frac{1}{1-p_1}(1-p_1) = 1$$

which completes the proof.

Example 5 Since $f(x) = x^2$ is a convex function we can see that $E[X^2] \geq E[X]^2$, which shows that the variance is a non-negative function.

Clearly, the above example comes as no surprise since it is clear from the definition of variance. A somewhat more interesting result from Jensen's inequality is the so called *log-sum inequality*. The function $f(t) = t\log t$ is in fact a convex function. Then, if α_i forms a probability distribution, it follows from Jensen's inequality that

$$\sum_i \alpha_i t_i \log t_i \geq \big(\sum_i \alpha_i t_i\big)\log\big(\sum_i \alpha_i t_i\big) \tag{3.11}$$

This can be used to get

$$\begin{aligned}\sum_i a_i \log\frac{b_i}{a_i} &= \textstyle\sum_j b_j \displaystyle\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\log\frac{a_i}{b_i}\\ &\geq \textstyle\sum_j b_j \displaystyle\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\log\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i} = \sum_i a_i \log\frac{\sum_i a_i}{\sum_j b_j}\end{aligned}$$

where we identified $\alpha_i = \frac{a_i}{\sum_j b_j}$ and $t_i = \frac{a_i}{b_i}$ in (3.11). Summarizing, we have the following theorem.

Theorem 7 (log-sum inequality) *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be non-negative numbers. Then*

$$\sum_i a_i \log\frac{a_i}{b_i} \geq \left(\sum_i a_i\right)\log\frac{\sum_i a_i}{\sum_i b_i}$$

□

3.5 Stochastic Processes

So far, we have assumed that the source symbols are independent in time. But it is often useful to also consider how sequences depend in time, i.e. a dynamic system. We will here look at an example to see that in texts we have dependencies in time and that one letter is clearly dependent of the surrounding letters. This example comes directly from the work of Claude Shannon [15].

Shannon assumed an alphabet with 27 symbols, i.e. 26 letters and 1 space. To get the 0th order approximation a sample text was generated with equal probability for the letters.

Example 6 [0th order approximation] Choose letters from the English alphabet with equal probability.

```
XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD
QPAAMKBZAACIBZLHJQD
```

Clearly, this text does not have much in common with normal written English. So, instead, he counted the number of occurrences per letter in normal English texts, and estimated the probabilities. The probabilities are given by Table 3.1.

$\mathcal{X}$	P	$\mathcal{X}$	P	$\mathcal{X}$	P
A	8.167	J	0.153	S	6.327
B	1.492	K	0.772	T	9.056
C	2.782	L	4.025	U	2.758
D	4.253	M	2.406	V	0.978
E	12.702	N	6.749	W	2.360
F	2.228	O	7.507	X	0.150
G	2.015	P	1.929	Y	1.974
H	6.094	Q	0.095	Z	0.074
I	6.966	R	5.987		

Table 3.1: Probabilities in percent for the letters in English text.

Then, according to those probabilities, a sample text for the 1st order approximation can be generated. In the next example such text is shown. Here, the text have more of a structure of English text, but still far from readable.

Example 7 [1st order approximation] Choose the symbols according to their normal probability (12 % E, 2 % W, etc.):

```
OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI
ALHENHTTPA OOBTTVA NAH BRL
```

The next step is to extend the distribution so the probability depends on the previous letter, i.e. the probability for the letter at time t becomes $P(S_t|S_{t-1})$. In the next example such text is given.

Example 8 [2nd order approximation] Choose the letters conditioned on the previous letter:

```
ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN
D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY
TOBE SEACE CTISBE
```

Similarly, in the next example the 3rd order approximation conditions on the two previous letters. We see here that the text becomes more like English, even if it still dos not make any sense.

Example 9 [3rd order approximation] Choose the symbols conditioned on the two previous symbols:

```
IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID
PONDENOME OF DEMONSTRURES OF THE REPTAGIN IS
REGOACTIONA OF CRE
```

If instead of letters, the source of the text is generated from probabilities for words. The first order approximation uses the unconditioned probabilities.

Example 10 [1st order word approximation] Choose words independently (but with probability):

```
REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME
CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE
TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE
MESSAGE HAD BE THESE
```

If the probabilities for words are conditioned on the previous word we can get a much more readable text. Still without any direct meaning, of course.

Example 11 [2nd order word approximation] Choose words conditioned on the two previous word:

```
THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER
THAT THE CHARACTER OF THIS POINT IS THEREFORE
ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO
EVER TOLD THE PROBLEM FOR AN UNEXPECTED
```

The above examples show that it in many situations it is important to view sequences instead of individual symbols. In probability theory this is denoted a *random process*, or a stochastic process. In a general form it can be defined as follows.

Definition 5 (Random process) *A* ***discrete random process*** *is a sequence of random variables, $\{X_i\}_{i=1}^n$, defined on the same sample space.* □

There can be an arbitrary dependency among the variables, and the process is characterized by the joint probability function

$$P\big(X_1, X_2, \dots, X_n = x_1, x_2, \dots, x_n\big) = p(x_1, x_2, \dots, x_n),\ n = 1, 2, \dots$$

Since there is a dependency in time we need also more general parameters corresponding to the second order moment and the variance of the random variable, depending also on the time shift. The *auto correlation* function reflects the correlation in time and is defined as

$$r_{XX}(n, n+k) = E\big[X_n X_{n+k}\big]$$

If the mean and the autocorrelation function are time independent, i.e. for all n

$$\begin{aligned} E[X_n] &= E[X] \\ r_{XX}(n, n+k) &= r_{XX}(n) \end{aligned}$$

the process is said to be *wide sense stationary* (WSS). The relation with the second order moment function is that $r_{XX}(k) = E[X^2]$. The same relation for the variance comes with the *autocovariance function*, defined for a WSS process as

$$c_{XX}(k) = E\big[(X_n - E[X])(X_{n+k} - E[X])\big]$$

for all n, We can easily see that $c_{XX}(k) = r_{XX}(k) - E[X]^2$ and that $c_{XX}(0) = V[X]$.

The class of WSS processes is a very powerful when modelling. However, sometimes we want an even stronger condition on the time invariance. We say that a process is *stationary*, or time-invariant, if the probability distribution does not depend on the time shift. That is, if

$$P\big(X_1, \dots, X_n = x_1, \dots, x_n\big) = P\big(X_{l+1}, \dots, X_{l+n} = x_1, \dots, x_n\big)$$

for all n and time shifts ℓ. Clearly this is a subclass of WSS processes.

3.6 Markov process

A widely used class of the discrete stationary random processes is the class of Markov processes. Here, the probability for a symbol depends only on the previous symbol. With this simplification we get a system that is relatively easy to handle from a mathematical point of view, while still having time dependency in the sequence to be a powerful modelling tool.

Definition 6 (Markov chain) *A* ***Markov chain****, or* ***Markov process****, is a stationary random process with unit memory, i.e.*

$$P\big(x_n|x_{n-1},\dots,x_1\big) = P\big(x_n|x_{n-1}\big)$$

for all x_i. □

Using the chain rule for probabilities (3.2) we conclude that for a Markov chain the joint probability function is

$$\begin{aligned} p(x_1,x_2,\dots,x_n) &= \prod_{i=1}^{n} p(x_i|x_{i-1}) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}) \end{aligned}$$

The unit memory of a Markov chain gives that we can characterize it by

- A finite set of **states**

  $$\mathcal{X} = \{x_1,x_2,\dots,x_r\}$$

  where the state determines everything about the past. This represents the unit memory of the chain.

- A **state transition matrix**

  $$P = [p_{ij}]_{i,j\in\{1,2,\dots,r\}} \text{ where } p_{ij} = p(x_j|x_i)$$

  and $\sum_j p_{ij} = 1$.

The behaviour of a Markov chain can be visualized in a **state transition graph** consisting of states and edges, labeled with probabilities. The following example explains how this is related.

Example 12 A three state Markov chain is described by the three states

$$\mathcal{X} = \{s_1,s_2,s_3\}$$

and a state transition matrix

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

From P we see that conditioned that we have state s_1 as the previous state, the probability for s_1 is $1/3$, s_2 is $2/3$ and s_3 is 0. This can be viewed as transitions in a graph from state s_1 to the other states, see Figure 3.4. Similarly, the other rows in the state transition matrix describe the probabilities for transitions from the other states.

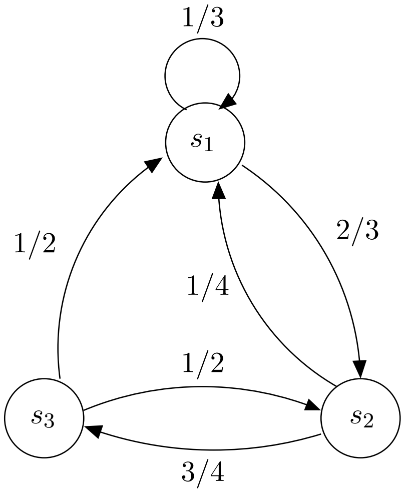


Figure 3.4: A state transition graph of a three state Markov chain.

For a Markov chain with r states the transition matrix will be a $r \times r$ matrix with the transition probabilities. Let the initial probabilities for the states at time 0 be

$$\boldsymbol{w}^{(0)} = \begin{pmatrix} w_1^{(0)} & \cdots & w_r^{(0)} \end{pmatrix}$$

where $w_i^{(0)} = P(X_0 = i)$. Then the probability for being in state j at time 1 becomes

$$w_j^{(1)} = P(X_1 = j) = \sum_i P(X_1 = j | X_0 = i) P(X_0 = i) = \sum_i p_{ij} w_i^{(0)}$$

Hence, the vector describing the state probabilities at time 1 becomes

$$\boldsymbol{w}^{(1)} = \boldsymbol{w}^{(0)} P$$

where P is the state transition matrix. Similarly, letting $\boldsymbol{w}^{(n)}$ be the state probabilities at time n, we get

$$\boldsymbol{w}^{(n)} = \boldsymbol{w}^{(n-1)} P = \cdots = \boldsymbol{w}^{(0)} P^n$$

In the following example we view state transition matrices from time 1 to time 2, 4, 8 and 16 respectively. Here we see that the columns of the matrix becomes more and more independent of the starting distribution.

Example 13 Continuing with the Markov chain from Example 12. The state transition matrix there is

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

It shows the probabilities for the transitions from time 0 to time 1. If instead we are considering P^n we have the transition probabilities from time 0 to time n. Below we derive the transition probabilities from time 1 to time 2, 4, 8 and 16.

$$P^2 = PP = \begin{pmatrix} \frac{20}{72} & \frac{16}{72} & \frac{36}{72} \\ \frac{33}{72} & \frac{39}{72} & 0 \\ \frac{21}{72} & \frac{24}{72} & \frac{27}{72} \end{pmatrix} \approx \begin{pmatrix} 0.2778 & 0.2222 & 0.5000 \\ 0.4583 & 0.5417 & 0 \\ 0.2917 & 0.3333 & 0.3750 \end{pmatrix}$$

$$P^4 = P^2P^2 = \begin{pmatrix} \frac{1684}{5184} & \frac{1808}{5184} & \frac{1692}{5184} \\ \frac{1947}{5184} & \frac{2049}{5184} & \frac{1188}{5184} \\ \frac{1779}{5184} & \frac{1920}{5184} & \frac{1485}{5184} \end{pmatrix} \approx \begin{pmatrix} 0.3248 & 0.3488 & 0.3264 \\ 0.3756 & 0.3953 & 0.2292 \\ 0.3432 & 0.3704 & 0.2865 \end{pmatrix}$$

$$P^8 = P^4P^4 \approx \begin{pmatrix} 0.3485 & 0.3720 & 0.2794 \\ 0.3491 & 0.3721 & 0.2788 \\ 0.3489 & 0.3722 & 0.2789 \end{pmatrix}$$

$$P^{16} = P^8P^8 \approx \begin{pmatrix} 0.3488 & 0.3721 & 0.2791 \\ 0.3488 & 0.3721 & 0.2791 \\ 0.3488 & 0.3721 & 0.2791 \end{pmatrix}$$

Already for 16 steps in the graph the numbers in each column are equal, for this accuracy with four decimals (actually, we have this already for P^{12}). This means that for 16 steps or more in the graph the probability for a state is independent of the starting state. Hence, with any starting distribution $\boldsymbol{w}$ we get

$$\boldsymbol{w}^{(16)} = \boldsymbol{w}P^{16} = \begin{pmatrix} 0.3488 & 0.372 & 0.2791 \end{pmatrix}$$

With higher accuracy in the derivations we need to keep multiplying longer but eventually we will reach a stage where it does not change.

As seen[2] in Example 13 we will reach an asymptotic distribution $\boldsymbol{w} = (w_1, \dots, w_r)$ such that

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{w} \\ \vdots \\ \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} w_1 & \cdots & w_r \\ w_1 & \cdots & w_r \\ \vdots & \ddots & \vdots \\ w_1 & \cdots & w_r \end{pmatrix}$$

In this limit we also can see that

$$\begin{pmatrix} \boldsymbol{w} \\ \vdots \\ \boldsymbol{w} \end{pmatrix} = P^n = P^{n+1} = P^nP = \begin{pmatrix} \boldsymbol{w} \\ \vdots \\ \boldsymbol{w} \end{pmatrix} P$$

Considering one row in the left matrix we conclude that

$$\boldsymbol{w} = \boldsymbol{w}P$$

which is the *stationary distribution* of the system. We are now ready to state a theorem for the relationship of the stationary and the asymptotic distribution.

[2] We have omitted the formal proof in this text.

Theorem 8 *Let $w = \begin{pmatrix} w_1 & w_2 & \cdots & w_r \end{pmatrix}$ be an asymptotic distribution of the state probabilities. Then*

- $\sum_j w_j = 1$
- $\boldsymbol{w}$ *is a* ***stationary distribution****, i.e.* $\boldsymbol{w}P = \boldsymbol{w}$
- $\boldsymbol{w}$ *is a unique stationary distribution for the source.*

□

Clearly the first statement is fulfilled since $\boldsymbol{w}$ is a distribution. The second has already been shown above. The third, on uniqueness, we still need to prove.

Assume that $\boldsymbol{v} = \begin{pmatrix} v_1 & \cdots & v_r \end{pmatrix}$ is a stationary distribution, i.e. it fulfills $\sum_i v_i = 1$ and $\boldsymbol{v}P = \boldsymbol{v}$. Then, as $n \to \infty$, the equation $\boldsymbol{v} = \boldsymbol{v}P^n$ can be written

$$\begin{pmatrix} v_1 & \dots & v_r \end{pmatrix} = \begin{pmatrix} v_1 & \dots & v_r \end{pmatrix} \begin{pmatrix} w_1 & \cdots & w_j & \cdots & w_r \\ \vdots & & \vdots & & \vdots \\ w_1 & \cdots & w_j & \cdots & w_r \end{pmatrix}$$

This imply that

$$v_j = \begin{pmatrix} v_1 & \cdots & v_r \end{pmatrix} \begin{pmatrix} w_j \\ \vdots \\ w_j \end{pmatrix} = v_1 w_j + \cdots + v_r w_j = w_j \underbrace{(v_1 + \cdots + v_r)}_{=1} = w_j$$

That is, $\boldsymbol{v} = \boldsymbol{w}$, which proves uniqueness.

To derive the stationary distribution we start with the equation, $\boldsymbol{w}P = \boldsymbol{w}$. Equivalently we can write $\boldsymbol{w}P - \boldsymbol{w} = \mathbf{0}$, and

$$\boldsymbol{w}(P - I) = \mathbf{0}$$

However, since $\boldsymbol{w} \neq \mathbf{0}$ we see that $P - I$ does not have full rank and we need at least one more equation. For this we can use $\sum_j w_j = 1$. Hence the equation system we should solve is

$$\begin{cases} \boldsymbol{w}(P - I) = \mathbf{0} \\ \sum_j w_j = 1 \end{cases}$$

In the next example we show the procedure.

Example 14 Again use the state transition matrix from Example 12,

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Starting with $\boldsymbol{w}(P-I)=\mathbf{0}$ we have

$$\boldsymbol{w}\left(\begin{pmatrix}\frac{1}{3} & \frac{2}{3} & 0\\ \frac{1}{4} & 0 & \frac{3}{4}\\ \frac{1}{2} & \frac{1}{2} & 0\end{pmatrix}-I\right)=\boldsymbol{w}\begin{pmatrix}-\frac{2}{3} & \frac{2}{3} & 0\\ \frac{1}{4} & -1 & \frac{3}{4}\\ \frac{1}{2} & \frac{1}{2} & -1\end{pmatrix}=\mathbf{0}$$

In $P-I$ column 2 plus column 3 equals column 1. Therefore, exchange column 1 with the equation $\sum_j w_j = 1$,

$$\boldsymbol{w}\begin{pmatrix}1 & \frac{2}{3} & 0\\ 1 & -1 & \frac{3}{4}\\ 1 & \frac{1}{2} & -1\end{pmatrix}=\begin{pmatrix}1 & 0 & 0\end{pmatrix}$$

This is solved by

$$\boldsymbol{w}=\begin{pmatrix}1 & 0 & 0\end{pmatrix}\begin{pmatrix}1 & \frac{2}{3} & 0\\ 1 & -1 & \frac{3}{4}\\ 1 & \frac{1}{2} & -1\end{pmatrix}^{-1}=\begin{pmatrix}1 & 0 & 0\end{pmatrix}\begin{pmatrix}\frac{15}{43} & \frac{16}{43} & \frac{12}{43}\\ \frac{42}{43} & -\frac{24}{43} & -\frac{18}{43}\\ \frac{36}{43} & \frac{4}{43} & -\frac{40}{43}\end{pmatrix}=\begin{pmatrix}\frac{15}{43} & \frac{16}{43} & \frac{12}{43}\end{pmatrix}$$

Derived with four decimals we get the same as earlier when the asymptotic distribution was discussed,

$$\boldsymbol{w}\approx\begin{pmatrix}0.3488 & 0.372 & 0.2791\end{pmatrix}$$

Chapter 4

Information Measure

Information theory is a mathematical theory of communication, that has its base in probability theory. It involves specifying what information is and setting up measurements for information, and had its birth with Shanon's landmark paper from 1948 [15]. The theory give answer to two fundamental questions:

- What is information?
- What is communication?

Even if we think we know what information and communication is, it is not the same as defining it in a mathematical point. As an example one can consider an electronic copy of Claude Shannon's paper in pdf format. The one considered here has the file size 2.2MB. It contains a lot of information about the subject, but how can it be measured? One way to look at it is to compress it as much as we can. In zip-format the same file has the size 713kB. So if we should quantify the amount of information in the paper, the pdf version contains at least 1.5MB that is not necessary for the pure information in the text. Is then the number 713kB a measure of the contained information? From a mathematical point of view, we will see that it is definitely closer to the truth. However, from a philosophical point of view, it is not certain. We can compare this text with a text of the same size containing only randomly chosen letters. If they have the same file size, do they contain the same amount of information? These semantic doubts are not considered in the mathematical model. Instead the question is how much information is needed to describe the text.

4.1 Information

In his paper Shannon did set up a mathematical theory for information and communication, based on probability theory. He gave a quantitative measure of the amount of information stored in one variable and gave limits of how much information that can be transmitted from place to another over a given communication channel.

However, already twenty years earlier, in 1928, Hartley stated that a symbol can contain information only if it has multiple choices [6]. That is, the symbol must be a random variable. Hartley argued that if one symbol, a, has k alternatives, a vector of n independent such symbols $(a_1, \dots, a_n)$ has k^n alternatives. To form a measure of information, one must notice that if the symbol a has the information I, then the vector should have the information kI. The conclusion of this was that an appropriate information measure should be the logarithm of the number of alternatives,

$$I_H(a) = \log k$$

In that way

$$I_H(a_1, \dots, a_n) = n \log k$$

Example 15 Consider a the outcome of a throw with a fair dice. It has 6 alternatives, and hence, the information according to Hartley is[1]

$$I_H(\text{Dice}) = \log_2 6 \approx 2.585 \text{bit}$$

In this example Hartley's information measure makes sense, since it is the number of bits needed to point out one of the six alternatives. But there can be other situations where it runs into problem, like in the next example.

Example 16 Let the variable X be the outcome of a counterfeit coin, with the probabilities $P(X = \text{Head}) = p$ and $P(X = \text{Tail}) = 1 - p$. According to Hartley the information is

$$I_H(X) = \log_2(2) = 1 \text{bit}$$

In the previous example the measurement makes sense if the two outcomes are equally likely. If we instead consider the case when p is very small, and flip the coin several times after each other. Since p is small we would expect to have most of the outcomes to be Tail and only a small fraction to be Head. Hence, the normal case in such vector of outcomes would be to have Tail, meaning there is not much information in this outcome. In the rare cases we get Head instead, there would be much more information. In other words, even if Hartley's information measure was ground breaking at its time, it lacked consideration of the outcome distribution.

In Shannon's information measure, introduced 20 years later, the base of the information quantity is the probability distribution of a random variable. The information achieved about the outcome of one variable by observing another, can be viewed as the difference between the unconditioned and the conditioned probability for that outcome.

[1] Hartley did not specify the basis of the logarithm, but using the binary base, the information has the unit bits. In this way it specifies the number of bits required to distinguish the alternatives.

Definition 7 *The information about the event $\{X = x\}$ from the event $\{Y = y\}$, denoted $I(X = x; Y = y)$, is*

$$I(X = x; Y = y) = \log_b \frac{P(X = x|Y = y)}{P(X = x)},$$

where we assumed that $P(X = x) \neq 0$ och $P(Y = y) \neq 0$. □

If nothing else is stated, we will assume the the logarithmic base $b = 2$ to achieve the unit bit. This unit was first used in Shannon's paper, but it was John W. Tukey who coined the expression.

Example 17 The outcome of a dice is reflected by the two random variables

$$\begin{aligned} X &= \text{ Number of pips} \\ Y &= \text{ Odd or even number} \end{aligned}$$

The information achieved about the event $X = 3$ from the event $Y = \text{Odd}$ is

$$I(X = 3; Y = \text{Odd}) = \log \frac{P(X = 3|Y = \text{Odd})}{P(X = 3)} = \log \frac{1/3}{1/6} = \log 2 = 1\text{bit}$$

In other words, by knowing that the number of pips is odd, which split the set of outcomes in half, we gain one bit information about the event that the number is 3.

We can see from the following derivation that the information achieved from one event about another is a symmetric measurement.

$$I(A; B) = \log \frac{P(A|B)}{P(A)} = \log \frac{P(A, B)}{P(A)P(B)} = \log \frac{P(B|A)}{P(B)} = I(B; A)$$

Example 18 [cont'd] The information from the event $X = 3$ about the event $Y = \text{Odd})$ is

$$I(Y = \text{Odd}; X = 3) = \log \frac{P(Y = \text{Odd}|X = 3)}{P(\text{Odd})} = \log \frac{1}{1/2} = \log 2 = 1\text{bit}$$

The knowledge about $X = 3$ gives us full knowledge about the outcome of Y, which is a binary choice with two equally sized parts. To specify one of the two outcomes of Y it is required one bit.

The mutual information can be bounded by

$$-\infty \leq I(X = x; Y = y) \leq -\log P(Y = y) \tag{4.1}$$

To see this we consider the variations of $P(X = x|Y = y)$, which is a value between 0 and 1. The two end cases give

$$\begin{aligned} P(X = x|Y = y) = 0 &\Rightarrow I(X = x; Y = y) = \log \frac{0}{P(Y = y)} = \log 0 = -\infty \\ P(X = x|Y = y) = 1 &\Rightarrow I(X = x; Y = y) = \log \frac{1}{P(Y = y)} = -\log P(Y = y) \end{aligned}$$

Notice that since $0 \leq P(Y=y) \leq 1$ the value $-\log P(Y=y)$ is positive. If $I(X=x;Y=y)=0$ the events $X=x$ and $Y=y$ are statistically independent since

$$I(X=x;Y=y)=0 \Rightarrow \frac{P(X=x|Y=y)}{P(Y=y)}=1$$

We are now ready to consider the self information, i.e. the information achieved about an event by observing the same event.

Definition 8 *Let $X=Y$. Then, the **self information** in the event $X=x$ is defined as*

$$I(X=x)=I(X=x;X=x)=\log\frac{P(X=x|X=x)}{P(X=x)}=-\log P(X=x)$$

□

Hence, $-\log \Pr(X=x)$ is the amount of information needed to determine that the event $X=x$ has occurred. The self information is always a non-zero quantity.

4.2 Entropy

The above quantities deal with the information in specific events. If we instead are consider the amount of information required in average to determine the outcome of a random variable, we need to consider the expected value of the self information. We get the following important definition.

Definition 9 *The **entropy**, which is a measure of the **uncertainty** of a random variable, is derived as*

$$H(X)=E_X\big[-\log p(x)\big]=-\sum_x p(x)\log p(x) \tag{4.2}$$

□

In the derivations, we use the convention that $0\log 0=0$, which stems from the corresponding limit value. Here, and hereafter, if nothing else is stated the logarithm uses the binary base, i.e. $\log_2$. To derive this we use that

$$x=a^{\log_a x}=b^{\log_b x}=a^{\log_a b\log_b x}$$

where we used that $b=a^{\log_a b}$. This leads to

$$\log_a x=\log_a b\log_b x \quad \Rightarrow \quad \log_b x=\frac{\log_a x}{\log_a b}$$

Especially, it is convenient to use

$$\log_2 x=\frac{\ln x}{\ln 2}=\frac{\log_{10} x}{\log_{10} 2}$$

In e.g. Matlab there is a command log2(n) to derive the binary logarithm.

Since p is a probability, and therefore between 0 and 1, we can directly say that $-\log p$ is a non-negative quantity. That also means the sum in (4.2) is non-zero,

$$H(X) \geq 0 \tag{4.3}$$

i.e. the uncertainty cannot be negative.

Example 19 Given a coin that could be counterfeit. The outcome from one flip has the sample space $\Omega = \{\text{Head}, \text{Tail}\}$. Denote the probabilities for the outcome by

$$\begin{aligned} P(\text{Head}) &= p \\ P(\text{Tail}) &= 1 - p \end{aligned}$$

The uncertainty of the random variable X is

$$H(X) = -p \log p - (1-p)\log(1-p)$$

and is shown in Figure 4.1. As an example, if $p = 0.1$ we get the entropy $H(X) = 0.469$. That is, in average it is needed 0.469 bit information to determine the outcome of a flip.

The entropy function of a binary choice, as in the previous example, is very important and has its own notation.

Definition 10 *The **binary entropy function** for the probability p is defined as*

$$h(p) = -p \log p - (1-p)\log(1-p)$$

□

It follows directly that the binary entropy function is symmetric in p. i.e. that

$$h(p) = h(1-p)$$

This symmetry can also be seen in Figure 4.1. From the figure we also see that it has a maximum value at $h(\frac{1}{2}) = 1$. In the case of a flip with a coin, the maximum corresponds to a fair coin and the uncertainty of the outcome is one bit. If the probability for Head is increasing a natural initial guess is that the outcome would be this, and the uncertainty decreases. Similarly, if the probability for Tail increases the uncertainty should decrease. At the and points, $p = 0$ or $p = 1$, the outcome is known and the uncertainty is zero, corresponding to $h(0) = h(1) = 0$.

Example 20 Let X be the outcome of a true dice. Then $P(X = x) = 1/6$, $x = 1, 2, \ldots, 6$. The entropy is

$$H(X) = -\sum_x \frac{1}{6}\log\frac{1}{6} = \log 6 = 1 + \log 3 \approx 2.5850 \text{ bit}$$

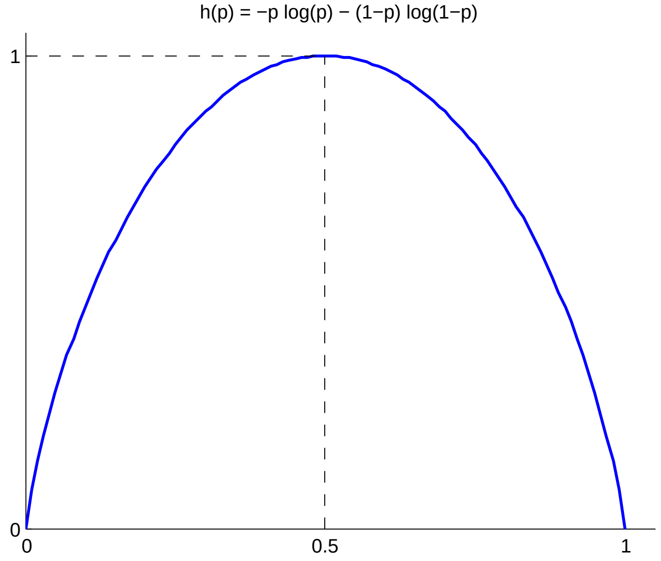


Figure 4.1: The Binary entropy function.

The definition of the entropy also yield vectorized random variables, such as (X, Y) with the joint probability function $p(x, y)$.

Definition 11 *The **joint entropy** is the entropy for a pair of random variables with the joint distribution* $p(x, y)$*,*

$$H(X,Y) = E_{XY}\big[-\log p(x,y)\big] = -\sum_x \sum_y p(x,y)\log p(x,y)$$

□

Clearly, in the general case with an N-dimensional vector the corresponding entropy function is

$$H(X_1,\dots,X_N) = -\sum_{x_1,\dots,x_N} p(x_1,\dots,x_N)\log p(x_1,\dots,x_N)$$

Example 21 Let X and Y be the outcomes from two independent true dice. Then the joint probability is $P(X,Y = x,y) = 1/36$ and the joint entropy

$$H(X,Y) = -\sum_{x,y}\frac{1}{36}\log\frac{1}{36} = \log 36 = 2\log 6 \approx 5.1699$$

We conclude that the uncertainty of the outcome of two dice is twice the uncertainty of one dice.

Instead consider the sum of the dice, $Z = X + Y$. The probabilities are shown in the following table

Z	2	3	4	5	6	7	8	9	10	11	12
$P(Z)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The entropy of Z is

$$\begin{aligned}H(Z) &= H\big(\tfrac{1}{36},\tfrac{2}{36},\tfrac{3}{36},\tfrac{4}{36},\tfrac{5}{36},\tfrac{6}{36},\tfrac{5}{36},\tfrac{4}{36},\tfrac{3}{36},\tfrac{2}{36},\tfrac{1}{36}\big)\\ &= -2\frac{1}{36}\log\frac{1}{36}-2\frac{2}{36}\log\frac{2}{36}-2\frac{3}{36}\log\frac{3}{36}\\ &\quad -2\frac{4}{36}\log\frac{4}{36}-2\frac{5}{36}\log\frac{5}{36}-\frac{6}{36}\log\frac{6}{36}\\ &= \dots = \frac{23}{18}+\frac{5}{3}\log 3-\frac{5}{18}\log 5 \approx 3.2744\end{aligned}$$

The uncertainty of the sum of the dice is less than the outcomes of the individual dice. This makes sense, since several outcomes of the pair X, Y give the same sum Z.

In (4.3) we saw that the entropy function is a non-negative function. To achieve an upper bound of it we first need to consider an important result, sometimes named the IT-inequality.

Lemma 9 *For every positive real number r*

$$\log(r) \leq (r-1)\log(e)$$

with equality if and only if $r = 1$. □

Proof: Graphically we consider the two functions $y_1 = r - 1$ and $y_2 = \ln r$ as shown in Figure 4.2. From this we conclude that $\ln r = r - 1$ at the point when $r = 1$. To formally show that in all other cases $\ln r < r - 1$ we notice that the derivative of $r - 1$ is always 1. Then, the derivative of $\ln r$ is

$$\frac{d}{dr}\ln r = \frac{1}{r} = \begin{cases} > 1, & r < 1 \Rightarrow \ln r < r-1\\ < 1, & r > 1 \Rightarrow \ln r < r-1\end{cases}$$

and we can conclude that $\ln r \leq r - 1$, with equality if and only if $r = 1$. Rewriting into binary logarithm completes the proof. ■

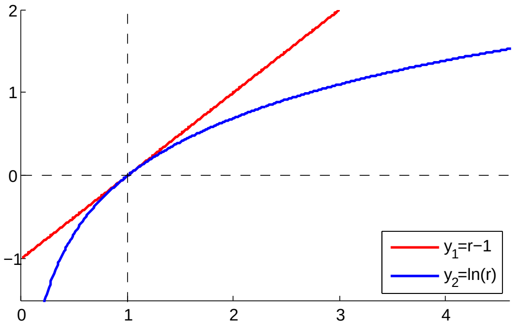


Figure 4.2: Graphical interpretation of the IT-inequality.

From what we have seen in the previous examples, and some intuition, we would guess that the maximum value of the entropy would occur when the outcomes have equal probabilities. Assume a random variable X has L outcomes, $\{x_1, \dots, x_L\}$, and that $P(X = x_i) = \frac{1}{L}$ for all of them. Then the entropy is

$$H(X) = -\sum_i \frac{1}{L} \log \frac{1}{L} = \log L$$

To show that this actually is a maximum value over all distributions we consider

$$\begin{aligned}
H(X) - \log L &= -\sum_x p(x) \log p(x) - \sum_x p(x) \log L \\
&= \sum_x p(x) \log \frac{1}{p(x)L} \\
&\leq \sum_x p(x) \Big(\frac{1}{p(x)L} - 1 \Big) \log e \\
&= \Big(\sum_x \frac{1}{L} - \sum_x p(x) \Big) \log e \\
&= (1 - 1) \log e = 0
\end{aligned}$$

where the inequality follows from the IT-inequality with $r = \frac{1}{p(x)L}$, implying we have equality if and only if $\frac{1}{p(x)L} = 1$. In other words, we have shown that

$$H(X) \leq \log L \tag{4.4}$$

with equality if and only if $p(x) = \frac{1}{L}$. Combining (4.3) and (4.4) we can state the following theorem.

Theorem 10 *If X is a random variable with L outcomes, $|\mathcal{X}| = L$, then*

$$0 \leq H(X) \leq \log L$$

with equality to the left if and only if there exists some i where $p(x_i) = 1$, and with equality to the right if and only if $p(x_i) = 1/L$ for all $i = 1, 2, \dots, L$. □

We are now ready to define a conditional entropy. The base is the conditional probability, say $p(x|y)$. Here we still have two random variables that need to be averaged and we get the following definition.

Definition 12 *The* ***conditional entropy*** *for X conditioned on Y, with the joint probability $p(x, y)$, is*

$$H(X|Y) = E_{XY}\big[-\log p(x|y)\big] = -\sum_{x,y} p(x, y) \log p(x|y)$$

□

Using the chain rule for probabilities, $p(x,y) = p(x|y)p(y)$, the sum can be rewritten as

$$\begin{aligned} H(X|Y) &= -\sum_x \sum_y p(x,y) \log p(x|y) = -\sum_x \sum_y p(x|y)p(y) \log p(x|y) \\ &= \sum_y p(y) \Big(-\sum_x p(x|y) \log p(x|y) \Big) \end{aligned}$$

By introducing the entropy of X conditioned on the event $Y = y$,

$$H(X|Y=y) = -\sum_x p(x|y) \log p(x|y)$$

the conditional entropy can be derived as

$$H(X|Y) = \sum_y H(X|Y=y)p(y)$$

Example 22 The joint distribution of the random variables X and Y is given by

		Y	
	$p(x,y)$	0	1
X	0	0	$\frac{3}{4}$
	1	$\frac{1}{8}$	$\frac{1}{8}$

The marginal distributions of X and Y can be derived as $p(x) = \sum_y p(x,y)$ and $p(y) = \sum_x p(x,y)$, respectively. These yields

x	0	1
$p(x)$	$\frac{1}{8}$	$\frac{7}{8}$

y	0	1
$p(y)$	$\frac{3}{4}$	$\frac{1}{4}$

The individual entropies are

$$\begin{aligned} H(X) &= h(\tfrac{1}{8}) \approx 0.5436 \\ H(Y) &= h(\tfrac{1}{4}) \approx 0.8113 \end{aligned}$$

and the joint entropy

$$H(X,Y) = H(0, \tfrac{3}{4}, \tfrac{1}{8}, \tfrac{1}{8}) \approx 1.0613$$

To calculate the conditional entropy $H(X|Y)$ we first consider the conditional probabilities, derived by $p(x|y) = \frac{p(x,y)}{p(y)}$

		$P(X|Y=0)$	$P(X|Y=1)$
X	0	0	$\frac{1}{2}$
	1	1	$\frac{1}{2}$

Therefore,

$$\begin{aligned} H(X|Y=0) &= h(0) = 0 \\ H(X|Y=1) &= h(\tfrac{1}{2}) = 1 \end{aligned}$$

Putting things together we get the conditional entropy as

$$\begin{aligned}H(X|Y) &= H(X|Y=0)P(Y=0) + H(X|Y=0)P(Y=0)\\ &= h(0)\frac{3}{4} + h(1)\frac{1}{4} = \frac{1}{4}\end{aligned}$$

We can use the chain rule for probabilities again to achieve a corresponding chain rule for entropies. The joint entropy can be written as

$$\begin{aligned}H(X,Y) &= -\sum_{x,y} p(x,y)\log p(x,y) = -\sum_{x,y} p(x,y)\log p(x|y)p(y)\\ &= -\sum_{x,y} p(x,y)\log p(x|y) - \sum_{y} p(y)\log p(y)\\ &= H(X|Y) + H(Y) \qquad (4.5)\end{aligned}$$

Rewriting the result we get $H(X|Y) = H(X,Y) - H(Y)$. That is, the conditional entropy is the uncertainty of the pair (X,Y) when X is known. A slightly more general version of (4.5) can be stated as the chain rule for entropies in the following theorem.

Theorem 11 *Let $X_1,\dots,X_N$ be an N-dimensional random variable drawn according to $p(x_1,\dots,x_N)$. Then the chain rule for entropies state that*

$$H(X_1,\dots,X_N) = \sum_{i=1}^{N} H(X_i|X_1,\dots,X_{i-1})$$

□

Example 23 [Cont'd from Example 22] The joint entropy can alternatively be derived as

$$H(X,Y) = H(X|Y) + H(Y) = \frac{1}{4} + h(\tfrac{1}{4}) = \frac{9}{4} - \frac{3}{4}\log 3 \approx 1.0613$$

4.3 Mutual Information

The entropy was obtained by averaging the self information for a random variable. Similarly, the average mutual information achieved about X when observing Y can be defined as follows.

Definition 13 *The average **mutual information** between the random variables X and Y is defined as*

$$I(X;Y) = E_{X,Y}\Big[\log\frac{p(x|y)}{p(x)}\Big] = \sum_{x,y} p(x,y)\log\frac{p(x|y)}{p(x)} \qquad (4.6)$$

□

Utilizing that

$$\frac{p(x|y)}{p(x)} = \frac{p(x,y)}{p(x)p(y)}$$

an alternative definition can be made by considering the ratio between the joint and the marginal probabilities,

$$I(X;Y) = E_{X,Y}\Big[\log\frac{p(x,y)}{p(x)p(y)}\Big] = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{4.7}$$

This is a measure of how strongly the two variables are connected. From (4.7), it is clear that the mutual information is a symmetric measure,

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

Breaking up the logarithm in (4.6) or (4.7), it is possible to derive the mutual information from the entropies as

$$\begin{aligned} I(X;Y) &= E_{X,Y}\Big[\log\frac{p(x,y)}{p(x)p(y)}\Big] \\ &= E_{X,Y}\big[\log p(x,y) - \log p(x) - \log p(y)\big] \\ &= E_{X,Y}\big[\log p(x,y)\big] - E_X\big[\log p(x)\big] - E_Y\big[\log p(y)\big] \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

where in the last two equalities the chain rule was used.

Example 24 [Cont'd from Example 22] The mutual information can be derives as

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &\approx 0.5436 + 0.8113 - 1.0613 = 0.2936 \end{aligned}$$

Alternatively, we can use

$$I(X;Y) = H(X) - H(X|Y) \approx 0.5436 - 1/4 = 0.2936$$

Both the entropy and the mutual information are two very important measures of information. The entropy states how much information is needed to determine the outcome of a random variable. It will be shown later that this is a limit of how many bits is needed in average to describe the variable. In other words, this is a limit of how much a source can be compressed without any data being lost. The mutual information, on the other hand, describes the amount of information achieved about the variable X by observing the variable Y. In a communication system a symbol, X, is transmitted over a channel. The received symbol Y, which is a distorted version of X, is used by a receiver to estimate X. The mutual information is a measure of how much information that can be transmitted over this channel. It will lead to the concept of the channel capacity.

To get more knowledge about these quantities we introduce the relative entropy. It was first introduced by Kullback and Leibler [12].

Definition 14 *Given two probability distributions $p(x)$ and $q(x)$ for the same sample set $\mathcal{X}$. The* ***relative entropy****, or* ***Kullback-Leibler distance****, is defined as*

$$D(p||q) = E_p\Big[\log\frac{p(x)}{q(x)}\Big] = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

□

Example 25 Consider a binary random variable, $X \in \{0, 1\}$, where we set up two distributions. First we assume that the values are equally probable,

$$p(0) = p(1) = 1/2$$

and, secondly, we assume a skew distribution,

$$q(0) = 1/4 \text{ and } q(1) = 3/4$$

The relative entropy from p to q is then

$$\begin{aligned} D(p||q) &= \frac{1}{2}\log\frac{1/2}{1/4} + \frac{1}{2}\log\frac{1/2}{3/4} \\ &= \cdots = 1 - \frac{1}{2}\log 3 \approx 0.2075 \end{aligned}$$

On the other hand, if we consider the relative entropy from q to p we get

$$\begin{aligned} D(q||p) &= \frac{1}{4}\log\frac{1/4}{1/2} + \frac{3}{4}\log\frac{3/4}{1/2} \\ &= \cdots = \frac{3}{4}\log 3 - 1 \approx 0.1887 \end{aligned}$$

That is, the relative entropy is not a symmetric measure, and can not be treated as a distance in normal sense. However, when talking about optimal source coding, we will see that it is natural to talk about the relative entropy as a distance from one distribution to another. This is the reason for the alternative name Kullback-Leibler distance.

The relative entropy was introduced as a generalized information measure. We can see that the mutual information as defined earlier can be expressed as a special case of the relative entropy as

$$I(X;Y) = E\Big[\frac{p(x,y)}{p(x)p(y)}\Big] = D\big(p(x,y)\big|\big|p(x)p(y)\big)$$

The mutual information is the information distance from the joint distribution to the independent case, i.e.the information distance corresponding to the relation between X and Y.

Another aspect of the relative entropy to consider is the relationship with the entropy function. Consider a random variable with the possible outcomes in the set $\mathcal{X}$ with cardinality $|\mathcal{X}| = L$ and probability distribution $p(x)$, $x \in \mathcal{X}$. Let $u(x) = 1/L$ be the uniform distribution for the same set of outcomes. Then,

$$\begin{aligned} H(X) &= -\sum_x p(x)\log p(x) = \log L - \sum_x p(x)\log p(x)L \\ &= \log L - \sum_x p(x)\log\frac{p(x)}{u(x)} = \log L - D(p||u) \end{aligned} \tag{4.8}$$

where we see that the relative entropy from $p(x)$ to $u(x)$ is the difference between the entropy based on the true distribution and the maximum value of the entropy. Since the maximum value is achieved by the uniform distribution, we see that the relative entropy is some sort of measure of how much $p(x)$ diverge from the uniform distribution.

The relative entropy can be shown to only take positive values. Here we will use the IT-inequality to show this fact.

$$\begin{aligned} -D(p||q) &= -\sum_x p(x)\log\frac{p(x)}{q(x)} = \sum_x p(x)\log\frac{q(x)}{p(x)} \\ &\leq \sum_x p(x)\Big(\frac{q(x)}{p(x)} - 1\Big)\log e \\ &= \Big(\sum_x p(x) - \sum_x q(x)\Big)\log e = (1-1)\log e = 0 \end{aligned}$$

with equality when $\frac{q(x)}{p(x)} = 1$, i.e.when $p(x) = q(x)$, for all x. Alternatively, Jensen's inequality can be used to show the same result. We can express the result in a theorem.

Theorem 12 *Given two probability distributions $p(x)$ and $q(x)$ for the same sample set $\mathcal{X}$. Then the relative entropy is positive*

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x. □

Since the mutual information can be expressed as the relative entropy, we directly get the corollary below.

Corollary 13 *For any two random variables X and Y*

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

with equality if and only if they are independent. □

Since we can write the mutual information as

$$I(X;Y) = H(X) - H(X|Y)$$

and that this is non-zero, we conclude the following theorem.

Theorem 14 *For any two random variables X and Y*

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

with equality if and only if they are independent. □

Intuitively, this means that the uncertainty, in average, will not increase by observing some side information. If the two variables are independent, we will have the same uncertainty as before. Using this result together with the chain rule for entropy, Theorem 11, we can get the next result.

Theorem 15 *Let $X_1, X_2, \dots, X_n$ be an n-dimensional random variable drawn according to $p(x_1, x_2, \dots, x_n)$. Then*

$$H(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if all X_i are independent. □

That is, the uncertainty is minimized when considering a random vector as a whole, instead of individual variables. In other words, we should take relationship between the variables into account when minimizing the uncertainty.

4.3.1 Convexity of information measures

In Definition 4 the terminology of convex functions was defined. In areas like optimization this is an important property since it is much easier to find a global optima. Here we will show that the information measurements we have defined are convex (or concave) functions. We will begin with the relative entropy which will give the base for the other functions.

Our previous definition for a convex function is only for one dimensional functions. Therefore, we need to start with generalizing the definition. A straight forward way is to say that a multidimensional functions is convex if it is convex in all dimensions. This means that for a two dimensional function we get a surface that resembles a bowl. In equation forms we write the condition for convexity for a two dimensional function $g(x, y)$ as

$$\begin{aligned} g\big(\lambda(x_1, y_1) + (1-\lambda)(x_2, y_2)\big) &= g\big(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2\big) \\ &\leq \lambda g(x_1, y_1) + (1-\lambda)g(x_2, y_2) \end{aligned}$$

The relative entropy can now be written as

$$\begin{aligned} &D\big(\lambda p_1 + (1-\lambda)p_2 \big|\big| \lambda q_1 + (1-\lambda)q_2\big) \\ &\quad = \sum_x \Big(\lambda p_1(x) + (1-\lambda)p_2(x)\Big) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \\ &\quad \leq \sum_x \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + \sum_x (1-\lambda)p_1(x) \log \frac{(1-\lambda)p_1(x)}{(1-\lambda)q_1(x)} \\ &\quad = \lambda \sum_x p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda) \sum_x p_1(x) \log \frac{p_1(x)}{q_1(x)} \\ &\quad = \lambda D\big(p_1 \big|\big| q_1\big) + (1-\lambda) D\big(p_2 \big|\big| q_2\big) \end{aligned}$$

where the inequality is a direct application of the log-sum inequality in Theorem 7. Hence, we see that the relative entropy is a convex function in (p, q).

From (4.8) we know that the entropy can be expressed as $H_p(X) = \log L - D(p||u)$ where u is the uniform probability function and p the distribution used for calculating the entropy. Then

$$\begin{aligned}
H_{\lambda p_1+(1-\lambda)p_2}(X) &= \log L - D\big(\lambda p_1 + (1-\lambda)p_2\big|\big|u\big) \\
&\geq \log L - \lambda D\big(p_1\big|\big|u\big) - (1-\lambda)D\big(p_2\big|\big|u\big) \\
&= \lambda\Big(\log L - D\big(p_1\big|\big|u\big)\Big) + (1-\lambda)\Big(\log L - D\big(p_2\big|\big|u\big)\Big) \\
&= \lambda H_{p_1}(X) + (1-\lambda)H_{p_2}(X)
\end{aligned}$$

where the inequality follows from the convexity of the relative entropy. We see here that the entropy is a concave function in p.

4.4 Entropy of sequences

When considering random processes and information theoretic measures it is natural to start with a well known lemma based on a three state Markov chain. It states that the information cannot increase by data processing, neither pre nor post. It can only be transformed to another representation (or be destroyed). In information theoretic terms it can be stated as the next lemma.

Lemma 16 (Data Processing Lemma) *If the random variables X, Y and Z form a Markov chain, $X \rightarrow Y \rightarrow Z$, we have*

$$\begin{aligned}
&I(X;Z) \leq I(X;Y) \\
&I(X;Z) \leq I(Y;Z)
\end{aligned}$$

□

For the considered Markov chain we have that, conditioned on Y, X and Z are independent, i.e.

$$P(XZ|Y) = P(X|Y)P(Z|XY) = P(X|Y)P(Z|Y)$$

where the second equality comes from the Markov condition. Starting with the first inequality of the theorem,

$$I(X;Z) = H(X) - H(X|Z) \leq H(X) - H(X|YZ) = H(X) - H(X|Y) = I(X;Y)$$

Similarly, the second inequality comes from

$$I(X;Z) = H(Z) - H(Z|X) \leq H(Z) - H(Z|XY) = H(Z) - H(Z|Y) = I(Z;Y)$$

which concludes the lemma.

The entropy function is an important measurement of the amount of information stored in a random variable. In the case when we are considering a random process and a sequence of random variables with some sort of correlation in between, we need a more general definition. In this section we will introduce two natural generalizations of the entropy function, and show that they are in fact equivalent. This function can in many cases be used in the same way for a random process as the entropy is used for a random variable.

A natural way to define the entropy per symbol for a sequence is by treating the sequence as a multi- dimensional random variable and averaging over the number of symbols. I the length of the sequence tend to infinity we get the following definition.

Definition 15 *The* ***entropy rate*** *of a stochastic process is*

$$H_\infty(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1 X_2 \dots X_n)$$

□

To see that this is in deed a generalization of the entropy measure we have from before we consider a sequence of i.i.d. variables in the next example.

Example 26 Consider a sequence of i.i.d. random variables with entropy $H(X)$. Then the entropy rate becomes the entropy function as defined earlier,

$$\begin{aligned} H_\infty(X) &= \lim_{n\to\infty} \frac{1}{n} H(X_1 \dots X_n) = \lim_{n\to\infty} \frac{1}{n} \sum_i H(X_i|X_1 \dots X_{i-1}) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_i H(X_i) = \lim_{n\to\infty} H(X) \frac{1}{n} \sum_i 1 = H(X) \end{aligned}$$

We can also define an alternative entropy rate for a stochastic process, where we consider the entropy of the nth variable in the sequence, conditioned on all the previous. As $n \to \infty$ we get

$$\overline{H}_\infty(X) = \lim_{n\to\infty} H(X_n|X_1 X_2 \dots X_{n-1})$$

To see how this variant relates to (**??**) we use the chain rule

$$\frac{1}{n} H(X_1 \dots X_n) = \frac{1}{n} \sum_i H(X_i|X_1 \dots X_{i-1})$$

The right hand side is the arithmetic mean of $H(X_i|X_1 \dots X_{i-1})$. By the law of large numbers, as $n \to \infty$ this will approach $\overline{H}_\infty(X)$. Hence, asymptotically as the length of the sequence grows to infinity the two definitions for the entropy rate are equal,

$$H_\infty(X) = \overline{H}_\infty(X)$$

Considering a stationary (time-invariant) random process, it can be seen that

$$H(X_n|X_1 \dots X_{n-1}) \le H(X_n|X_2 \dots X_{n-1}) = H(X_{n-1}|X_1 \dots X_{n-2})$$

where the last equality follows from the stationarity. Here we see that $H(X_n|X_1 \dots X_{n-1})$ is a decreasing function in n. As n decreases we get to a point

$$H(X_n|X_1 \dots X_{n-1}) \le \cdots \le H(X_2|X_1) \le H(X_1) = H(X) \le \log|\mathcal{X}|$$

Finally, since the entropy is a non-negative function we can state the following theorem.

Theorem 17 *For a stationary stochastic process the entropy rate is bounded by*

$$0 \le H_\infty(X) \le H(X) \le \log|\mathcal{X}|$$

□

In Figure 4.3 the relation between $\log|\mathcal{X}|$, $H(X)$, $H(H_n|X_1 \dots X_{n-1})$ and $H_\infty(X)$ is shown.

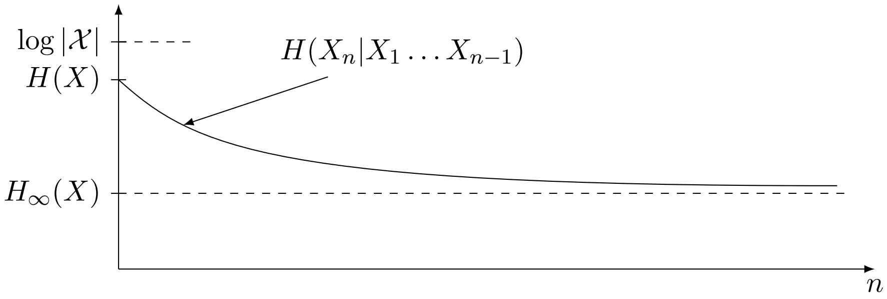


Figure 4.3: The relation between $H_\infty(X)$ and $H(X)$.

So far the entropy rate has been treated for the class of stationary random processes. If the subclass Markov chains are considered it is possible to say more on how to derive it. For the conditional entropy, the Markov condition can be written as $H(X_n|X_1 \dots X_{n-1}) = H(X_n|X_{n-1})$. Then, the entropy rate can be derived as

$$\begin{aligned} H_\infty(X) &= \lim_{n\to\infty} H(X_n|X_1 \dots X_{n-1}) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1) \\ &= \sum_{i,j} P(X_1 = x_i, X_2 = x_j) \log P(X_2 = x_j|X_1 = x_i) \\ &= \sum_i P(X_1 = x_i) \sum_j P(X_2 = x_j|X_1 = x_i) \log P(X_2 = x_j|X_1 = x_i) \\ &= \sum_i H(X_2|X_1 = x_i) P(X_1 = x_i) \end{aligned} \tag{4.9}$$

where

$$H(X_2|X_1 = x_i) = \sum_j P(X_2 = x_j|X_1 = x_i) \log P(X_2 = x_j|X_1 = x_i)$$

In (4.9) the transition probability is given by the state transition matrix

$$P = \Big[p_{ij} = P(X_2 = x_j|X_1 = x_i)\Big]_{x_i, x_j \in \mathcal{X}}$$

and the stationary distribution by $w_i = P(X_1 = x_j)$. In terminology of the Markov process we get the following theorem.

Theorem 18 *For a stationary Markov chain with stationary distribution $\boldsymbol{w}$ and transition matrix $P = [p_{ij}]$, the entropy rate can be derived as*

$$H_\infty(X) = \sum_i w_i H(X_2|X_1 = x_i)$$

where

$$H(X_2|X_1 = x_i) = -\sum_j p_{ij} \log p_{ij}$$

□

Example 27 The Markov chain shown in Figure 3.4 has the state transition matrix

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

In Example 14 the steady state distribution was calculated as

$$\boldsymbol{w} = \begin{pmatrix} \frac{15}{43} & \frac{16}{43} & \frac{12}{43} \end{pmatrix}$$

The conditional entropies are the entropies for each row in P,

$$\begin{aligned} H(X_2|X_1 = s_1) &= h(\tfrac{1}{3}) = \log 3 - \tfrac{2}{3} \\ H(X_2|X_1 = s_2) &= h(\tfrac{1}{4}) = 2 - \tfrac{3}{4}\log 3 \\ H(X_2|X_1 = s_3) &= h(\tfrac{1}{2}) = 1 \end{aligned}$$

and the entropy rate becomes

$$\begin{aligned} H_\infty(X) &= w_1 H(X_2|X_1 = s_1) + w_2 H(X_2|X_1 = s_2) + w_3 H(X_2|X_1 = s_3) \\ &= \tfrac{15}{43} h(\tfrac{1}{3}) + \tfrac{16}{43} h(\tfrac{1}{4}) + \tfrac{12}{43} h(\tfrac{1}{2}) \\ &= \tfrac{3}{43}\log 3 + \tfrac{34}{43} \approx 0.9013 \text{ bit/symbol} \end{aligned}$$

4.5 Random Walk

To be done.

Consider an undirected weighted graph where W_{ij} is the weight for the edge joining node x_i and x_j. Let

$$W_i = \sum_j W_{ij} \qquad W = \sum_{i,j:i<j} W_{ij}$$

denote the sum of the weights leaving node x_i, and the total weight, respectively. The probability for using the edge from node x_i is $p_{ij} = W_{ij}/W_i$.

- <1-> The stationary distribution is

$$\mu_i = \frac{W_i}{2W}$$

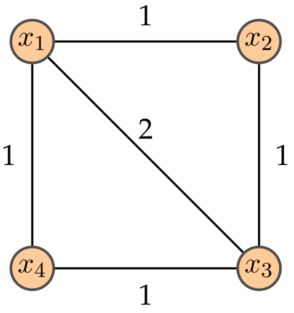
- <2-> The entropy rate

$$H_\infty(X) = H\Big(\dots, \frac{W_{ij}}{2W}, \dots\Big) - H\Big(\dots, \frac{W_i}{2W}, \dots\Big)$$

- <3-> If all non-zero edges have $W_{ij} = 1$,

$$H_\infty(X) = \log(2W) - H\Big(\dots, \frac{W_i}{2W}, \dots\Big)$$

Example 28 Consider a weighted graph



The conditional probabilities are

$$p_{ij} = \frac{W_{ij}}{\sum_j W_{ij}}$$

The state transition matrix

$$P = \begin{pmatrix} 0 & 1/4 & 2/4 & 1/4 \\ 1/2 & 0 & 1/2 & 0 \\ 2/4 & 1/4 & 0 & 1/4 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix}$$

With $W = 6$ and $W_1 = 4$, $W_2 = 2$, $W_3 = 4$, $W_4 = 2$ we get the stationary distribution

$$\boldsymbol{\mu} = \Big(\frac{4}{12}\ \frac{2}{12}\ \frac{4}{12}\ \frac{2}{12}\Big) = \Big(\frac{1}{3}\ \frac{1}{6}\ \frac{1}{3}\ \frac{1}{6}\Big)$$

and the entropy rate

$$\begin{aligned} H_\infty(X) &= H\Big(\frac{1}{12}, \frac{2}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{2}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}\Big) \\ &\quad - H\Big(\frac{4}{12}, \frac{2}{12}, \frac{4}{12}, \frac{2}{12}\Big) \\ &\approx 3.25 - 1.92 = 1.33 \end{aligned}$$

Chapter 5

SourceCodingTheorem

Text...

5.1 Asymptotic Equipartition Property

Asymptotic Equipartition Property (AEP) is a very useful tool in information theory. The basic idea is that a series of i.i.d. events is viewed as a vector of events. The probability for each vector becomes very small as the length grows, giving that it is pointless of considering the probabilities for each vector. Instead it is the probability for the distribution of events within the vector that is important. From the law of large numbers we can see that it is only a fraction of the possible outcomes that is the probable ones. To see this, consider 100 consecutive tosses with a fair dice. The probability for each of the length 10 vectors is

$$P(x_1, \dots, x_{100}) = \left(\frac{1}{2}\right)^{100} \approx 8 \cdot 10^{-31}$$

So, the probability for each of the vectors in the outcome is very small and it is not very likely that the vector with 100 Heads will occur. However, since there are $\binom{50}{100} \approx 10^{29}$, the probability for getting a vector with equal number of Head and Tail is about

$$P(50\,\text{Head}, 50\,\text{Tail}) = 2^{-100}\binom{50}{100} \approx 0.080$$

which is relatively high. To conclude, it is most likely that the outcome of 100 tosses with a fair dice will result in approximately the same numbers of Heads and Tails. This is in fact a consequence of the weak law of large numbers, which we recapitalize here.

Theorem 19 (The weak law of large numbers) *Let $X_1, X_2, \dots, , X_n$ be i.i.d. random variables with mean $E\big[X\big]$. Then,*

$$\frac{1}{n}\sum_i X_i \xrightarrow{p} E\big[X\big]$$

where $\xrightarrow{p}$ denotes convergence in probability. □

The notation that it converges in probability can also be expressed as

$$\lim_{n\to\infty} P\Big(\Big|\frac{1}{n}\sum_i X_i - E[X]\Big| < \varepsilon\Big) = 1$$

for any $\varepsilon > 0$. This means that a arithmetic mean of several i.i.d. random variables approaches the expected value of the variables. For an example on the weak law of large numbers, see Example 3.

Consider instead the logarithmic probability for a vector of length n, consisting of i.i.d. random variables. By the weak law of large numbers we can conclude that

$$-\frac{1}{n}\log p(\boldsymbol{x}) = \frac{1}{n}\sum_i -\log p(x_i) \xrightarrow{p} H(X)$$

or, equivalently, that

$$\lim_{n\to\infty} P\Big(\Big|\frac{1}{n}p(\boldsymbol{x}) - E[X]\Big| < \varepsilon\Big) = 1$$

for any $\varepsilon > 0$. That is, for all vectors that can happen, the mean logarithmic probability for the random variable approaches the entropy. For a finite vector of length n we can have a closer look at those vectors that fulfill this criteria. They are called *typical sequences* and are defined in the next definition.

Definition 16 (AEP) *The set of* ε***-typical sequences*** $A_\varepsilon(X)$ *is the set of all* n*-dimensional vectors* $\boldsymbol{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ *such that*

$$\Big|-\frac{1}{n}\log p(\boldsymbol{x}) - H(X)\Big| \le \varepsilon \tag{5.1}$$

□

It is possible to rewrite (5.1) as follows

$$\begin{aligned} -\varepsilon \le -\frac{1}{n}\log p(\boldsymbol{x}) - H(X) \le \varepsilon \\ H(X) - \varepsilon \le -\frac{1}{n}\log p(\boldsymbol{x}) \le H(X) + \varepsilon \\ -n\big(H(X)+\varepsilon\big) \le -\frac{1}{n}\log p(\boldsymbol{x}) \le -n\big(H(X)-\varepsilon\big) \\ 2^{-n(H(X)+\varepsilon)} \le p(\boldsymbol{x}) \le 2^{-n(H(X)-\varepsilon)} \end{aligned}$$

This is the base for an alternative definition of the AEP.

Definition 17 (AEP, Alternative definition) *The* ε*-typical sequences can definition as the set of vectors* $\boldsymbol{x}$ *such that*

$$2^{-n(H(X)+\varepsilon)} \le p(\boldsymbol{x}) \le 2^{-n(H(X)-\varepsilon)} \tag{5.2}$$

□

Both definitions of the AEP are frequently used in the literature. It differs which one is used as the main definition, but quite often both are presented. In the next example we try to show what it means with ε-typical sequences.

Example 29 Consider a binary 5-dimensional vector with i.i.d. elements where $p_X(0) = \frac{1}{3}$ and $p_X(1) = \frac{2}{3}$. Then the entropy for each symbol is $H(X) = h(1/3) = 0.918$. In the following tabular all vectors are listed together with their respectively probability.

$\boldsymbol{x}$	$p(\boldsymbol{x})$	
00000	0.0041	
00001	0.0082	
00010	0.0082	
00011	0.0165	
00100	0.0082	
00101	0.0165	
00110	0.0165	
00111	0.0329	$\star$
01000	0.0082	
01001	0.0165	
01010	0.0165	
01011	0.0329	$\star$
01100	0.0165	
01101	0.0329	$\star$
01110	0.0329	$\star$
01111	0.0658	$\star$
10000	0.0082	
10001	0.0165	
10010	0.0165	
10011	0.0329	$\star$
10100	0.0165	
10101	0.0329	$\star$
10110	0.0329	$\star$
10111	0.0658	$\star$
11000	0.0165	
11001	0.0329	$\star$
11010	0.0329	$\star$
11011	0.0658	$\star$
11100	0.0329	$\star$
11101	0.0658	$\star$
11110	0.0658	$\star$
11111	0.1317	

As expected the all zero vector is the least possible vector, while the all one vector is the most likely. However, even the most likely vector is not very likely to happen, with probability 0.1317. Still, if we were to pick one vector of all as a guess for what the outcome would be, this is the one. But since the elements are i.i.d. it can be argued that the order of the elements are unimportant. If we instead would take a guess on the *type* of vector, meaning the number of ones and zeros the answer would be different. The probability for a vector containing k ones and $5-k$ zeros is

$$P(k1s, 5-k0s) = \binom{n}{k}\Big(\frac{2}{3}\Big)^k\Big(\frac{1}{3}\Big)^{n-k} = \binom{n}{k}\frac{2^k}{3^n}$$

which we saw already in Chapter 3. Viewing these numbers in a table we get (in Chapter 3 it is also plotted in a diagram),

k	$P(\#1 \text{ in } \boldsymbol{x}) = \binom{5}{k}\frac{2^k}{3^5}$
0	0.0041
1	0.0412
2	0.1646
3	0.3292
4	0.3292
5	0.1317

Here it is clear that the most likely vector, the all one vector, does not belong to the most likely type of vector. When guessing of the number of ones, it is more likely to get 3 or 4 ones. This is of course due to that there are more vectors that fulfill this criteria then the single all one vector. So, this concludes that vectors with 3 or 4 ones are sort of "typical".

The question is then how this relates to the previous definitions of typical sequences. To see this, chose an ε. Here we use 15% of the entropy, which is $\varepsilon = 0.138$,

$$2^{-n(H(X)+\varepsilon)} = 2^{-5(h(\frac{1}{3})+0.138)} \approx 0.0257$$
$$2^{-n(H(X)+\varepsilon)} = 2^{-5(h(\frac{1}{3})-0.138)} \approx 0.0669$$

Then, the ε-typical sequences are the ones in between those numbers. In the table these are marked with a $\star$. Luckily these are the same vectors as we concluded should be the typical ones.

In the previous example we saw that there are something called typical sequences, and that they are the most likely sequences to occur, seen from the contents. In this example we used a very short vector to be able to list all of them, but for longer sequences it can be seen that the ε-typical sequences are just a fraction of all the sequences. On the other hand, it can also be seen that the probability for a random sequence to belong to the typical sequences is close to one. More formally, we can state the following theorem.

Theorem 20 *Consider sequences of length n of i.i.d. random variables. For each ε there exists an integer n_0 such that, for each $n > n_0$, the set of ε-typical sequences, $A_\varepsilon^{(n)}(X)$, fulfills*

$$P\big(\boldsymbol{x} \in A_\varepsilon^{(n)}(X)\big) \geq 1 - \varepsilon \tag{5.3}$$
$$(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq \big|A_\varepsilon^{(n)}(X)\big| \leq 2^{n(H(X)+\varepsilon)} \tag{5.4}$$

□

The first part of the theorem, (5.3), is a direct consequence of the law of large numbers stating that $-\frac{1}{n}\log p(\boldsymbol{x})$ approaches $H(X)$ as n grows. That means that there exists an n_0, such that for all $n \geq n_0$

$$P\Big(\Big|-\frac{1}{n}\log p(\boldsymbol{x}) - H(X)\Big| < \varepsilon\Big) \geq 1 - \delta$$

for any δ between zero and one. Letting $\delta = \varepsilon$ gives

$$P\Big(\Big|-\frac{1}{n}\log p(\boldsymbol{x}) - H(X)\Big| < \varepsilon\Big) \geq 1 - \varepsilon$$

which is equivalent to (5.3). This shows that the probability for an arbitrary sequence to belong to the typical set approaches one as n grows.

To show the second part, that the number of ε-typical sequences is bounded by (5.4) we need to split the equation in two parts. Starting with the left hand inequality we have, for large enough n_0

$$\begin{aligned}1-\varepsilon \leq P\big(\boldsymbol{x} \in A_\varepsilon^{(n)}(X)\big) &= \sum_{\boldsymbol{x}\in A_\varepsilon^{(n)}(X)} p(\boldsymbol{x}) \\ &\leq \sum_{\boldsymbol{x}\in A_\varepsilon^{(n)}(X)} 2^{-n(H(X)-\varepsilon)} = \big|A_\varepsilon^{(n)}(X)\big|2^{-n(H(X)-\varepsilon)}\end{aligned}$$

where we in the second inequality used the alternative definition of the AEP. The right hand side can be shown by

$$\begin{aligned} 1 &= \sum_{\boldsymbol{x}\in\mathcal{X}^n} p(\boldsymbol{x}) \geq \sum_{\boldsymbol{x}\in A_\varepsilon^{(n)}(X)} p(\boldsymbol{x}) \\ &\geq \sum_{\boldsymbol{x}\in A_\varepsilon^{(n)}(X)} 2^{-n(H(X)+\varepsilon)} = \big|A_\varepsilon^{(n)}(X)\big|2^{-n(H(X)+\varepsilon)} \end{aligned}$$

which completes the theorem. To see what the theorem means we consider longer sequences than in the previous example.

Example 30

Let $\mathcal{X}^n = \{\boldsymbol{x}\}$ be the set of length n vector of i.i.d. binary random variables with $p(0) = \frac{1}{3}$ and $p(1) = \frac{2}{3}$. Let ε be 5% of the entropy $H(X) = h(\frac{1}{3})$, $\varepsilon = 0.046$. Then the number of *eps*-typical sequences adn their bounding functions are given in the next table for $n = 100$, $n = 500$ and $n = 1000$. As a comparison, the fraction of ε-typical sequences compared to the total number of sequences is also shown.

n	$(1-\varepsilon)2^{n(H(X)-\varepsilon)}$	$\lvert A_\varepsilon^{(n)}(X)\rvert$	$2^{n(H(X)+\varepsilon)}$	$\lvert A_\varepsilon^{(n)}(X)\rvert/\lvert\mathcal{X}^n\rvert$
100	$1.17\cdot 10^{26}$	$7.51\cdot 10^{27}$	$1.05\cdot 10^{29}$	$5.9\cdot 10^{-3}$
500	$1.90\cdot 10^{131}$	$9.10\cdot 10^{142}$	$1.34\cdot 10^{145}$	$2.78\cdot 10^{-8}$
1000	$4.16\cdot 10^{262}$	$1.00\cdot 10^{287}$	$1.79\cdot 10^{290}$	$9.38\cdot 10^{-15}$

This table shows that the ε-typical sequences only constitute a fraction of the total number of sequences.

Next, the probability for the ε-typical sequences are given together with the probability for the most probable sequence, the all one sequence. Here it can be clearly seen that the most likely sequence has a very low probability, and is in fact very unlikely to happen. Instead, the most likely event is that a random sequence is taken from the typical sequences.

n	$P(A_\varepsilon^{(n)}(X))$	$P(\boldsymbol{x} = 11\dots1)$
100	0.660	$2.4597\cdot 10^{-18}$
500	0.971	$9.0027\cdot 10^{-89}$
1000	0.998	$8.1048\cdot 10^{-177}$

5.2 Source Coding Theorem

In Figure 5.2 a block diagram of a communication system is shown. One of the results from information theory is that the sequence generated from a source often contains redundancy. The source encoder is intended to remove the redundancy and use a minimum size representation of the information. The problem is that as the redundancy is

19=>5(c),21=>5(a) Q. With suitable figure explain Shannon’s model for a communication system removed the information is very vulnerable to disturbances in the transmission. To circumvent this new redundant data is added in a controlled way, by the channel encoder. If errors occur during transmission, the channel decoder will be able to detect and/or correct most of them. Finally, the source decoder decompress the received sequence back to the original. If the channel decoder manage to correct all errors occurred on the channel, i.e. $\widehat{Y} = Y$, the source decoder should implement the inverse of the compression function. Shannon showed that the the source coding and the channel coding can be performed independent of each other. Therefore, in this section only the source encoding and decoding will be treated.

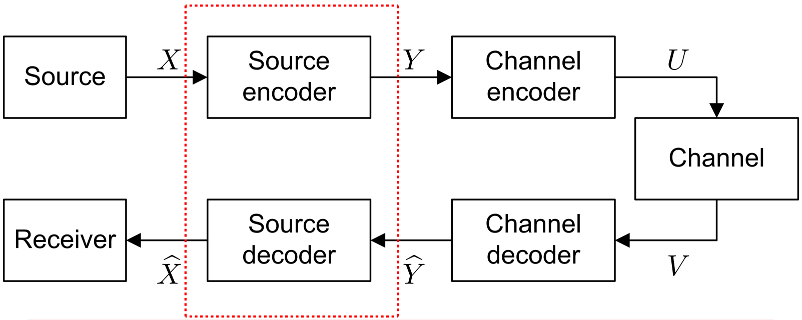


Figure 5.1: Shannon's model for a communication system.

One of the questions that information theory tries to answer is how much information is stored in a set of data. A typical example is text, that contain redundancy for us to be able correct spelling errors or misprints. In English written language it is often possible to cross out every third letter and still be able to read the text. The possible compression that can be done for a sequence show how much information it contains. Earlier, we considered the entropy as a measure of the uncertainty, which is the amount of information needed to determine the value. We will see in this section that the entropy is in fact a measure of the possible compression of a source.

In Figure 5.2 the model for source encoding and decoding used is shown. The sequence from the source is fed to the source encoder where the redundancy is removed. The source decoder is the inverse function and reconstruct the source data. In the figure $\boldsymbol{X}$ is an n-dimensional vector from the alphabet $\mathcal{X}$. The length n of the vector is considered to be a random variable and the average length is $E[n]$. The corresponding compressed sequence, $\boldsymbol{Y}$, is an ℓ-dimensional vector from the alphabet $\mathcal{Y}$. Also ℓ is a random variable so the average length of the codeword is $E[k]$. The decoder estimates the source vector as $\widehat{\boldsymbol{X}}$. When working with lossless source coding as in this section it is required that the original message is perfectly reconstructed, $\widehat{X} = X$. This is the case for compression algorithms working on e.g. texts and also some image processing. However, in many cases for media coding, such as image, video and speech coding, there are lossy algorithms. The compression is much better in those, but the prize paid is that he original message cannot be reconstructed exactly, i.e. $\widehat{X} \approx X$.

The compression rate is defined as

$$R = \frac{E[k]}{E[\ell]}$$

If the alphabet sizes are equal this means there is a compression if R is less than one. The reason to view n and ℓ as random variables is that, to have compression either the length of the source vector, the code vector or both must vary. If both the lengths would be fixed for all messages there would not be any compression. In the next definition the above reasoning is formalized.

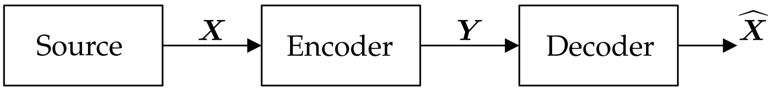


Figure 5.2: Block model for source coding system..

Definition 18 (Not too general) *A **source code** is a mapping from a finite source vector $\boldsymbol{x} \in \mathcal{X}^n$ to a finite vector of variable length ℓ, $\boldsymbol{y} = (y_1 y_2 \dots y_\ell)$, where $y_i \in \mathcal{D} = \mathbb{Z}_D$ are drawn from a D-ary alphabet.*
The length of the codeword corresponding to $\boldsymbol{x}$ is denote $\ell(x)$. □

Naturally the codeword length is a good measure of the efficiency of a chosen code. This is defined as

$$L = E[\ell(\boldsymbol{x})] = \sum_{\boldsymbol{x} \in \mathcal{X}^n} p(\boldsymbol{x}) \ell(\boldsymbol{x})$$

To derive the source coding theorem, stating there exists a source code where the average codeword length approaches the entropy of the source, we first need to set up an encoding rule. We saw in the previous section that the typical sequences are typically a small fraction of all sequences, but the are the most likely to happen. Assuming then that the non-typical sequences almost never happen, we can concentrate on the typical. Starting with a list of all sequences, it can be partitioned in two parts, one with the typical and one with the non-typical. To construct the codewords we use a binary prefix stating which set it belongs to. Lets say we use 0 for the typical sequences and 1 for the non typical. Each of the sets are listed and indexed by binary vectors. For the set of typical sequences the index vector will be of length $\lceil \log |A_\varepsilon^{(n)}(X)| \rceil$. The non-typical sequences only happens rarely so we can use n bits. The following two tables show the idea of the look-up tables.

Typical
Prefix $= 0$

$\boldsymbol{x}$	Index vec
$\boldsymbol{x}_0$	$0 \dots 00$
$\boldsymbol{x}_1$	$0 \dots 0$
$\vdots$	$\vdots$

Non-typical
Prefix $= 1$

$\boldsymbol{x}$	Index vec
$\boldsymbol{x}_a$	$0 \dots \dots 00$
$\boldsymbol{x}_b$	$0 \dots \dots 0$
$\vdots$	$\vdots$
$\vdots$	$\vdots$

The same table look-up can be used by the decoding algorithm. The number of typical sequences is

$$\big|A_\varepsilon^{(n)}(X)\big| \leq 2^{n(H(X)+\varepsilon)}$$

and the length is

$$\ell(\boldsymbol{x}) = \Big\lceil \log\big|A_\varepsilon^{(n)}(X)\big|\Big\rceil + 1 \leq \log\big|A_\varepsilon^{(n)}(X)\big| + 2 \leq n\big(H(X)+\varepsilon\big) + 2$$

Similarly, we know there are at most $|\mathcal{X}|^n$ non-typical sequences and that their length is

$$\ell(\boldsymbol{x}) = \Big\lceil \log\big|\mathcal{X}\big|^n\Big\rceil + 1 \leq \log\big|\mathcal{X}\big|^n + 2 \leq n\log|\mathcal{X}| + 2$$

In Figure 5.2 the procedure is shown graphically.

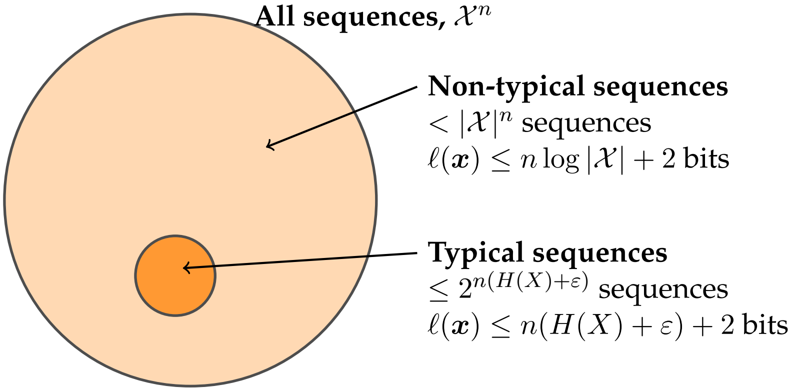

All sequences, $\mathcal{X}^n$
Non-typical sequences
$< |\mathcal{X}|^n$ sequences
$\ell(\boldsymbol{x}) \leq n\log|\mathcal{X}| + 2$ bits
Typical sequences
$\leq 2^{n(H(X)+\varepsilon)}$ sequences
$\ell(\boldsymbol{x}) \leq n(H(X)+\varepsilon) + 2$ bits

Figure 5.3: Principle of Shannon's source coding algorithm.

Next, we derive a bound on the average length of the codewords.

$$\begin{aligned}
L = E\big[\ell(\boldsymbol{x})\big] &= \sum_{\boldsymbol{x}\in\mathcal{X}^n} p(\boldsymbol{x})\ell(\boldsymbol{x}) = \sum_{\boldsymbol{x}\in A_\varepsilon^{(n)}(X)} p(\boldsymbol{x})\ell(\boldsymbol{x}) + \sum_{\boldsymbol{x}\notin A_\varepsilon^{(n)}(X)} p(\boldsymbol{x})\ell(\boldsymbol{x}) \\
&\leq \sum_{\boldsymbol{x}\in A_\varepsilon^{(n)}(X)} p(\boldsymbol{x})\big(n(H(X)+\varepsilon)+2\big) + \sum_{\boldsymbol{x}\notin A_\varepsilon^{(n)}(X)} p(\boldsymbol{x})\big(n\log|\mathcal{X}|+2\big) \\
&= P\big(\boldsymbol{x}\in A_\varepsilon^{(n)}(X)\big)n(H(X)+\varepsilon) + P\big(\boldsymbol{x}\notin A_\varepsilon^{(n)}(X)\big)n\log|\mathcal{X}| + 2 \\
&\leq n(H(X)+\varepsilon) + \varepsilon n\log|\mathcal{X}| + 2 = n\Big(H(X) + \underbrace{\varepsilon + \varepsilon\log|+\frac{2}{n}}_{\varepsilon'}\Big) \\
&= n\big(H(X)+\varepsilon'\big)
\end{aligned}$$

where ε' can be arbitrary small for sufficiently large n. Summarizing, we state the source coding theorem.

Theorem 21 *Let $\boldsymbol{x}\in\mathcal{X}^n$ be length n vectors of i.i.d. random variables with probability function $p(\boldsymbol{x})$. Then there exists a code which maps sequences $\boldsymbol{x}$ of length n into binary sequences such that the mapping is invertible and*

$$\frac{1}{n}E\Big[\ell(\boldsymbol{x})\Big] \leq H(X) + \varepsilon'$$

for sufficiently large n, where ε' can be made arbitrarily small. □

For random processes we do not have i.i.d. variables. However, it is possible to generalize the theorem for this case, using the entropy rate. For a formal proof we refer to e.g. [2].

Theorem 22 *Let X^n be an stationary ergodic process. Then there exists a code which maps sequences $\boldsymbol{x}$ of length n into binary sequences such that the mapping is invertible and*

$$\frac{1}{n}E\Big[\ell(\boldsymbol{x})\Big] \leq H_\infty(X) + \varepsilon'$$

for sufficiently large n, where ε' can be made arbitrarily small. □

5.3 Kraft Inequality

To be done.

5.4 Shannon-Fano Coding

To be done.

5.5 Huffman Coding

To be done.

Bibliography

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition edition, 2006.

[3] Robert G. Gallager. *Information Theory and Reliable Communication*. Wiley, 1968.

[4] Allan Gut. *An Intermediate Course in Probability*. Springer-Verlag, 1995.

[5] Richard Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, Vol. 29:pp. 147–160, 1950.

[6] Ralph V. L. Hartley. Transmission of information. *Bell System Technical Journal*, pages p. 535–563, July 1928.

[7] http://www.libpng.org/pub/png. PNG (Portable Graphics Format) Home site. maintained by Greg Roelofs.

[8] David A. Huffman. A method for the construction of minimum redundancy codes. *Proceedings of the Institute of Radio Engineers*, Vol. 40:pp. 1098–1101, 1952.

[9] Rolf Johannesson. *Informationsteori – grundvalen för (tele-) kommunikation*. Studentlitteratur, 1988. (in Swedish).

[10] Rolf Johannesson and Kamil Zigangirov. *Fundamentals of Convolutional Codes*. IEEE Press, 1999.

[11] Boris Kudryashov. *thori Informatii*. Piter, 2009. (in Russian).

[12] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, Vol. 22, No 1:79–86, March 1951.

[13] Robert McEliece. *The Theory of Information and Coding*. CambridgeUniversity Press, 2004. Student Edition.

[14] Khalid Sayood. *Introduction to Data Compression*. Elsevier Inc., 2006.

[15] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27:pp. 379–423, 623–656, July, October 1948.

[16] Terry Welch. A technique for high-performance data compression. *IEEE Computer*, vol. 17:pp. 8–19, 1984.

[17] John F. Young. *Information Trheory.* Butterworth & Co, 1971.

[18] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, Vol. 23:pp. 337–343, 1977.

[19] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, Vol. 24:pp. 530–536, 1978.