Notes on Information Theory

Cristian Di Pietrantonio Michele Laurenti

January 3, 2017

Chapter 1

Introduction

Information Theory is born from one man, Cloude Shannon (1926 - 2001). It has to do with probability, algebra, coding theory, ergodic theory and appears in daily life.

- Tom Cover, Joy Thomas, Information Theory, Wiley;
- IT: cooling theorems for discrete memoryless systems, Korner;
- IT, Robert Ash.

Chapter 2

The Hamming Ball

This Chapter introduces an important concept: the Hamming Ball.

2.1 Hamming space

A *space* is a set that has structure. The set $\{0,1\}^n$ of binary strings of length n can be made into a metric space. To make it a metric space we have define a metric (or distance) on it.

A metric over set $\mathfrak{X}$ is a function $d : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ that has the following properties:

1. it's greater than zero, *i.e.*, $d(x, y) \geq 0 \ \forall (x, y) \in \mathfrak{X} \times \mathfrak{X}$, with equality when its arguments are the same, *i.e.*, $d(x, y) = 0 \iff x = y$;

2. it's symmetric, *i.e.*, $d(x, y) = d(y, x) \ \forall (x, y) \in \mathfrak{X} \times \mathfrak{X}$;

3. it satisfies the triangular inequality, *i.e.*, $d(x, y) \leq d(x, z) + d(z, y)$.

For $\{0,1\}^n$ we define the *Hamming metric.* Consider two strings, $\underline{x}, \underline{y} \in \{0,1\}^n$. Then their Hamming distance is defined as

$$d_H\left(\underline{x}, \underline{y}\right) = \left|\{i : x_i \neq y_i\}\right|. \tag{2.1}$$

The first two properties are trivial to see. For the third property, consider three strings $\underline{x}$, $\underline{y}$, $\underline{z} \in \{0,1\}^n$. Let $D = \{i : x_i \neq y_i\}$ be the set of coordinates where they differ. If $i \in D$, we have $x_i \neq y_i$. What can happen to z_i? Either $z_i \neq x_i$ or $z_i \neq y_i$. If $i \notin D$, z_i is either equal to both x_i and y_i, or it differs from both of them. Thus, $d_H(x_i, y_i) \leq d_H(x_i, z_i) + d_H(z_i, y_i)$ for all i.

Now, since the Hamming metric is additive, we have

$$\begin{aligned} d_H\left(\underline{x}, \underline{y}\right) &= \sum_{i=1}^{n} d_H\left(x_i, y_i\right) \\ &\leq \sum_{i=1}^{n} d_H\left(x_i, z_i\right) + \sum_{i=1}^{n} d_H\left(z_i, y_i\right) = d_H\left(\underline{x}, \underline{z}\right) + d_H\left(\underline{z}, \underline{y}\right) \end{aligned}$$

which gives us the triangular inequality.

In general, a distance is extended to a product space by summing the distances of the single components, as happens for the Hamming metric.

Consider a storage device that has n cells, each containing either 0 or 1. Suppose memory decays with time in some unknown way. After some time the string memorised in the device will be different. The Hamming distance tells us how much different.

If a string $\underline{x}$ has been changed no more than r times to become $\underline{y}$, then $d_H(\underline{x}, \underline{y}) \leq r$. We say that $\underline{y}$ is in a *Hamming Ball* of radius r around $\underline{x}$.

Definition 1 (Hamming Ball)**.** *The Hamming Ball of radius r and centre $\underline{x}$ is the set*

$$B_H\left(\underline{x}, r\right) = \{\underline{y} : d_H\left(\underline{x}, \underline{y}\right) \leq r\}. \tag{2.2}$$

Observation 1. *r needs not to be an integer, but for $r \in \mathbb{R}$ it holds that*

$$B_H\left(\underline{x}, r\right) = B_H\left(\underline{x}, \lfloor r \rfloor\right).$$

Note that $\{0,1\}^n$ is a Hamming Ball of radius n, for any centre.

If you have a ball that is not the whole space, then the centre of this ball is unique.

What we can say about the size of a generic Hamming Ball $B_H\left(\underline{x}, r\right)$, with $r > 0$? For the sake of simplicity, consider $B_H\left(\underline{0}, r\right)$, where $\underline{0}$ is the string of all zeros. Then the size of this Hamming Ball is

$$\left|B_H\left(\underline{0}, r\right)\right| = \sum_{i=0}^{\lfloor r \rfloor} \binom{n}{i} \tag{2.3}$$

i.e., the number of ways in which we can flip to 1 up to r bits.

Definition 2 (Hamming weight)**.** *The Hamming weight of a string is its distance from $\underline{0}$, i.e.,*

$$w_H\left(\underline{x}\right) = d_H\left(\underline{0}, \underline{x}\right). \tag{2.4}$$

If one subtracts $B_H\left(\underline{x}, r\right)$ to $\{0,1\}^n$, then the result is also an Hamming ball.

To see this, consider a string $\underline{y}$ that is not in $B_H\left(\underline{0}, r\right)$. Then it must be that $w_H\left(\underline{y}\right) > r$. What is its distance from the string of all ones, *i.e.*, $d_H\left(\underline{y}, \underline{1}\right)$? One can see that $d_H\left(\underline{y}, \underline{1}\right) = n - w_H\left(\underline{y}\right)$, since $d_H\left(\underline{0}, \underline{x}\right) + d_H\left(\underline{1}, \underline{x}\right) \geq d_H\left(\underline{0}, \underline{1}\right) = n$. More in general we can bound the distance as $d_H\left(\underline{y}, \underline{1}\right) < n - r$, or $d_H\left(\underline{y}, \underline{1}\right) \leq n - (r+1)$, which means that $\underline{y}$ is in a Hamming Ball of radius $n - (r+1)$ with centre $\underline{1}$.

From it we derive the following result:

$$\overline{B_H\left(\underline{0}, n\right)} = \{0,1\}^n \setminus B_H\left(\underline{0}, n\right) = B_H\left(\underline{1}, n - (r+1)\right). \tag{2.5}$$

In other words the Hamming space can be partitioned into two Hamming balls.

If n is odd, the Hamming space can be partitioned in the following way:

$$\{0,1\}^n = B_H\left(0, \left\lfloor\frac{n}{2}\right\rfloor\right) \cup B_H\left(1, \left\lfloor\frac{n}{2}\right\rfloor\right)$$

The Hamming space cannot be partitioned into three balls. In how many balls can $\{0,1\}^n$ be partitioned? It cannot partitioned be partitioned into s balls with $3 \leq s \leq n+1$ (this result has not been demonstrated).

2.2 The volume of a Hamming ball

We have seen that the volume of a generic Hamming ball is described by Equation 2.3. We can use the Pascal triangle to get a feeling about the order of magnitude of the Hamming ball. Consider the nth row in the triangle; since it is symmetric, if we split the row in half the sum of the terms of the first part is equal to the sum of the terms of the second part of the row. The volume of the greatest ball is $|B_H(0,n)| = 2^n$. Instead, the volume of the ball with radius $n/2$ (in the triangle's point of split) is $2^n/2 = 2^n 2^{-1} = 2^{n-1}$. It follows, if $r \geq n/2$, that $2^{n-1} \leq |B_H(0,r)| \leq 2^n$. Can we bound the volume for $r < n/2$?

2.2.1 Entropy

First we introduce the notion of *entropy*. Entropy is a function $h : [0,1] \to [0,1]$ defined as follows:

$$h(t) = t\log_2(\frac{1}{t}) + (1-t)\log_2\left(\frac{1}{1-t}\right) \tag{2.6}$$

The plot of h looks like this:

This function is not defined for the values 0 and 1. However, we have limits defined on these points and they are both 0. So we artificially set $h(0) = 0$ and $h(1) = 0$. It is better to think about it as probability distribution $(t, 1-t)$ and entropy is a number attached to it. Entropy measures symmetry and with $t = 1/2$ we have maximum chaos (the future outcomes are equally likely).

2.2.2 Lower and upper bounds

Theorem 1. *The upper bound on the volume of a Hamming ball, if $r \leq n/2$, is*

$$|B_H(0,r)| \leq 2^{nh(\frac{r}{n})}$$

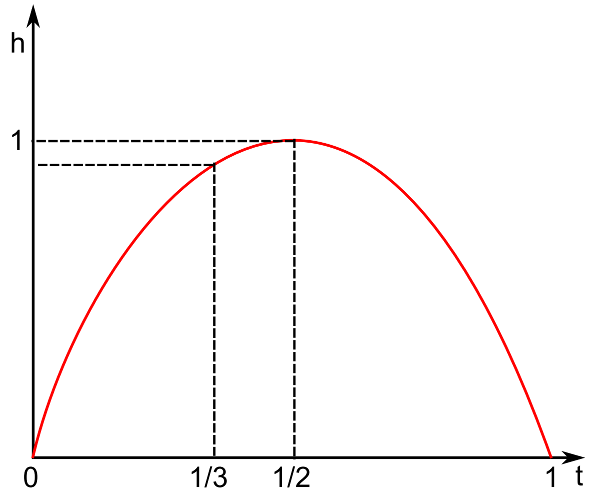

Figure 2.1: The entropy function.

In order to prove this theorem an analogy is introduced. Suppose there is a box with a number of chickens in it. We want to count those animals without withdrawing all of them out of the box. What can be done is to take the lightest one and measure its weight w; then also the weight W of the box is measured. The number of the total chickens in the box can't be greater than w/W.

Proof. In this proof we will use a similar technique. Consider $\{0,1\}^n$. We "sparkle" a substance on the strings in the set; this substance looks like probability, but it doesn't matter. Define the weight of 1 and 0 as

$$P(1) = \frac{r}{n}, \ P(0) = 1 - P(1).$$

Notice that is not an uniform distribution. Define the weight of a string as

$$P^n(\underline{x}) = \prod_{i=1}^{n} P(x_i).$$

If $A \subseteq \{0,1\}^n$ then the weight of the set A is

$$P^n(A) = \sum_{\underline{x} \in A} P^n(\underline{x}).$$

$P^n(\{0,1\}^n)$ is the total weight of the substance sparkled on the strings and it is the probability distribution of binomial. For this reason one can claim that

$$1 = P^n(\{0,1\}^n) \geq P^n(B_H(0,r)) = \sum_{\underline{x} \in B_H(0,r)} P^n(\underline{x})$$

at this point we “take out the lightest chicken” and write

$$\sum_{\underline{x}\in B_H(0,r)} P^n(\underline{x}) \geq |B_H(0,r)| \cdot \min_{\underline{x}\in B_H(0,r)} P^n(\underline{x})$$

Which are the lightest chickens? Because we assume $r \leq n/2$ then

$$r \leq \frac{n}{2} \Rightarrow P(1) = \frac{r}{n} \leq \frac{1}{2} \Rightarrow P(1) \leq P(0).$$

It follows that the the lightest strings are the ones on the border of the ball, with r 1's. We now compute their weight.

$$\begin{aligned}
\min_{\underline{x}\in B_H(0,r)} P^n(\underline{x}) &= [P(1)]^r \cdot [P(0)]^{n-r} = \\
&= \left(\frac{r}{n}\right)^r \left(1-\frac{r}{n}\right)^{n-r} = \\
&= \left(\frac{r}{n}\right)^{n\frac{r}{n}} \left(1-\frac{r}{n}\right)^{n\left(1-\frac{r}{n}\right)} = \\
&= [\left(\frac{r}{n}\right)^{\frac{r}{n}} \left(1-\frac{r}{n}\right)^{\left(1-\frac{r}{n}\right)}]^n = \\
&= 2^{n\log_2[\left(\frac{r}{n}\right)^{\frac{r}{n}}\left(1-\frac{r}{n}\right)^{\left(1-\frac{r}{n}\right)}]} = \\
&= 2^{n[\frac{r}{n}\log_2\frac{r}{n}+\left(1-\frac{r}{n}\right)\log_2\left(1-\frac{r}{n}\right)]} = 2^{-nh\left(\frac{r}{n}\right)}
\end{aligned}$$

So we have

$$1 \geq |B_H(0,r)| \frac{1}{2^{nh\left(\frac{r}{n}\right)}} \Rightarrow |B_H(0,r)| \leq 2^{nh\left(\frac{r}{n}\right)}$$

□

Theorem 2. *The lower bound on the volume of a Hamming ball, if $r \leq n/2$, is*

$$|B_H(0,r)| \geq \frac{1}{n+1} 2^{nh\left(\frac{r}{n}\right)}$$

Proof.

$$P(1) = \frac{r}{n},\ P(0) = 1 - P(1),\ P^n(\{0,1\}^n) = 1.$$

Consider the set of all strings of length n and partition it in the following way.

$$\mathrm{T}_q^n = \{\underline{x} \mid w_H(x) = q\},$$

obtaining $n+1$ classes. We know that $|\mathrm{T}_q^n| = \binom{n}{q}$; we are not interested in how many strings are in that set, but what is the total weight; we want to prove

$$|T_r| \geq \frac{1}{n+1} 2^{nh\left(\frac{r}{n}\right)}.$$

There is not symmetry in the weight of the partitions so there is a weight r so that

$$\frac{P^n(\mathrm{T}_q)}{P^n(\mathrm{T}_r)} \leq 1,\ \forall q$$

In the formula above there are binomials we want to bound.

Observation 2.

$$\frac{k!}{l!} \leq k^{k-l}.$$

Proof. We are going to prove this observation in two steps.

- $k \geq l$.

$$\frac{k!}{l!} = \frac{k(k-1)\cdots l(l-1)\cdots 1}{l(l-1)\cdots 1} \leq k^{k-l}.$$

- $k < l$.

$$\frac{k!}{l!} = \frac{k(k-1)\cdots 1}{l(l-1)\cdots k(k-1)\cdots 1} \leq \left(\frac{1}{k+1}\right)^{l-k} < \left(\frac{1}{k}\right)^{l-k} = k^{k-l}.$$

□

Define $p = r/n$ so that we have a distribution $P(p, 1-p)$ that picks a set and concentrate the weight (probability) on it. We observe that the probability of each string in a class depends only on the number of 1's in it. So we can write

$$\frac{P^n(\mathrm{T}_q)}{P^n(\mathrm{T}_r)} = \frac{p^q(1-p)^{n-q}|\mathrm{T}_q|}{p^r(1-p)^{n-r}|\mathrm{T}_q|} = p^{q-r}(1-p)^{r-q}\frac{\frac{n!}{q!(n-q)!}}{\frac{n!}{r!(n-r)!}} =$$

$$= p^{q-r}(1-p)^{r-q}\frac{r!}{q!}\frac{(n-r)!}{(n-q)!} \leq p^{q-r}(1-p)^{r-q}r^{r-q}(n-r)^{q-r} =$$

considering that $r = np$ it follows that

$$p^{q-r}(1-p)^{r-q}(np)^{r-q}[n(1-p)]^{q-r} =$$
$$= p^{q-r}(1-p)^{r-q}n^{r-q+q-r}p^{r-q}(1-p)^{q-r} =$$

$$= p^{q-r+r-q}(1-p)^{r-q+q-r} = 1.$$

So we can write

$$1 = P^n(\{0,1\}^n) = P^n\left(\bigcup_{q=0}^{n} T_q\right) = \sum_{q=0}^{n} P^n(T_q) \leq (n+1)\max_q P^n(T_q) =$$

$$= (n+1)P^n(T_r) = (n+1)|T_r|2^{-nh\frac{r}{n}}$$

From the previous result we know $|T_r| \geq \dfrac{1}{n+1} 2^{nh(\frac{r}{n})}$ and $T_r \subseteq B_H(0,r)$ so it follows that $|T_r| \leq |B_H(0,r)|$. □

So this proof is important because the cardinality of the Hamming ball comes up in many contexts, such as error correction (a string that has been haltered at most r times is in a certain radius from the original string).

2.3 Generalization to any finite alphabet

Let $\mathcal{X}$ be the (usual) finite set that is an alphabet. The interest lies in sequences of elements of $\mathcal{X}$, called *strings* or *words*. So $\mathcal{X}^n, n \in \mathbb{N}$, is a set of words. $\mathcal{X}^n$ can be partitioned by putting together those sequences that can be transformed one into the other by permutation, i. e. sequences that have the same number of occurrences of elements in the alphabet.

Let $a \in \mathcal{X}$ and $\underline{x} \in \mathcal{X}^n$. We define the frequency of an alphabet symbol a in a string $\underline{x}$ in the following way:

$$N(a|\underline{x}) = |\{i \mid x_i = a\}|, \tag{2.7}$$

where $\underline{x} = x_1 x_2 \dots x_n$. One can think about "normalized" relative frequencies of symbols

$$\frac{1}{n} N(a|\underline{x}).$$

Moreover the following holds:

$$\sum_{a \in \mathcal{X}} N(a|\underline{x}) = n \Rightarrow \sum_{a \in \mathcal{X}} \frac{1}{n} N(a|\underline{x}) = 1$$

so from a string $\underline{x}$ one can obtain a probability distribution over $\mathcal{X}$. We define

$$P_{\underline{x}} = \left\{ \frac{N(a|\underline{x})}{n} \mid a \in \mathcal{X} \right\} \tag{2.8}$$

to be the *type* of $\underline{x}$. There are just that many distributions for a number n; now fix a distribution $P|\mathcal{X}$. $\exists \underline{x} \in \mathcal{X}^n$ such that $P_{\underline{x}} = P$? Yes, if and only if

$$P(a) = \frac{N(a|\underline{x})}{n},\ \forall a \in \mathcal{X}.$$

Consider a product measure over $\mathcal{X}$; strings in the same partition have also the same "length" or measure. Now, given $\mathcal{X}$ and n, how many distributions $P|\mathcal{X}$ are types in $\mathcal{X}^n$? A rough upper bound is $(n+1)^{|\mathcal{X}|}$. The last value is redundant, since the values sum up to 1. So we could do better with $(n+1)^{|\mathcal{X}|-1}$. We can partition $\mathcal{X}^n$ into sets of strings of the same type, T_p, with $P|\mathcal{X}$.

$$\mathrm{T}_p = \mathrm{T}_p^n = \{\underline{x} \mid P_{\underline{x}} = P\}.$$

Theorem 3. *If* $\mathrm{T}_p \neq \emptyset$ *then*

$$\frac{1}{(n+1)^{|\mathcal{X}|-1}} 2^{nH(p)} \leq |\mathrm{T}_p| \leq 2^{nH(p)}$$

Proof. In order to prove the above theorem, we first define the product distribution $P|\mathcal{X} \to P^n|\mathcal{X}^n$ as

$$P^n(\underline{x}) = \prod_{i=1}^{n} P(x_i). \tag{2.9}$$

We can define it additively on subsets of $\mathcal{X}^n$.

$$1 = P^n(\mathcal{X}^n) \geq P^n(\mathrm{T}_p^n)$$

We also introduce the *generalized entropy* $H(P)$, defined as

$$H(P) = -\sum_{a \in \mathcal{X}} P(a) \log_2 P(a).$$

Now,

$$\forall \underline{x} \in \mathrm{T}_p^n P^n(\underline{x}) = \prod_{a \in \mathcal{X}} P(a)^n = nP(a)$$

notice that this is independent from $\mathcal{X}$. So we have

$$= \prod_{a \in \mathcal{X}} 2^{nP(a) \log_2 P(a)} = 2^{n\left[\sum_{a \in \mathcal{X}} P(a) \log_2 P(a)\right]} = 2^{-nH(p)}$$

So,

$$1 = P^n(\mathcal{X}^n) \geq P^n(\mathrm{T}_p^n) = |\mathrm{T}_p^n| 2^{-nH(p)}$$

□

The lower bound proof is a straightforward generalization of what has been don in the binary case. Entropy is greatest when the distribution is uniform. Now, to prove the lower bound, consider

$$1 = \sum_{P \ | \ \mathrm{T}_p^n \neq \emptyset} P^n(\mathrm{T}_p^n) \leq (n+1)^{|\mathcal{X}|-1} \max_{Q|\mathcal{X}} P^n(\mathrm{T}_q^n).$$

Observation 3. *If* $\mathrm{T}_p \neq \emptyset$ *then*

$$\frac{P^n(\mathrm{T}_q^n)}{P^n(\mathrm{T}_p^n)} \leq 1$$

If a distribution is a type, it maximizes its (product) value on the strings of that type. We can suppose without loss of generality (w.l.o.g) that $\mathrm{T}_q^n \neq \emptyset$.

$$P^n(\mathrm{T}_q^n) = \prod_{a\in\mathcal{X}} P(a)^{nQ(a)} |\mathrm{T}_q^n| \Rightarrow \frac{P^n(\mathrm{T}_q^n)}{P^n(\mathrm{T}_p^n)} = \frac{|\mathrm{T}_q|^n \prod_{a\in\mathcal{X}}[P(a)]^{nQ(a)}}{|\mathrm{T}_p|^n \prod_{a\in\mathcal{X}}[P(a)]^{nP(a)}} =$$

$$= \frac{\frac{n!}{\prod_{a\in\mathcal{X}}[nQ(a)]!} \prod_{a\in\mathcal{X}}[P(a)]^{nQ(a)}}{\frac{n!}{\prod_{a\in\mathcal{X}}[nP(a)]!} \prod_{a\in\mathcal{X}}[P(a)]^{nP(a)}} = \prod_{a\in\mathcal{X}} \frac{[nP(a)]!}{[nQ(a)]!} \prod_{a\in\mathcal{X}} [P(a)]^{n(Q(a)-P(a))} \leq$$

$$\leq \prod_{a\in\mathcal{X}} [nP(a)]^{n(P(a)-Q(a)} \prod_{a\in\mathcal{X}} [P(a)]^{n(Q(a)-P(a))} =$$

$$= n^{n\left[\sum_{a\in\mathcal{X}} P(a)-Q(a)\right]} \frac{\prod_{a\in\mathcal{X}}[P(a)]^{n(P(a)-Q(a)}}{\prod_{a\in\mathcal{X}}[P(a)]^{n(P(a)-Q(a)}} =$$

$$= n^{n\left[\sum_{a\in\mathcal{X}} P(a) - \sum_{a\in\mathcal{X}} Q(a)\right]} = n^{n[1-1]} = 1.$$

Chapter 3

The log sum inequality

We now prove a consequence od the concavity of the logarithm, which will be used to prove some results about entropy.

Observation 4. *The logarithm function is cap-convex ($\cap$-Convex). Remember that* $\ln(t) \leq t-1$ *(with equality iff* $t=1$*) and* $\log_2(t) = \dfrac{\ln(t)}{\ln(2)}$.

Proposition 1 (Log sum inequality)**.** *Let* $a_i \geq 0, i = 1, 2, \dots, t$ *with* $a = \sum_{i=1}^{t} a_i$ *and* $b_i \geq 0, i = 1, 2, \dots, t,$ *with* $b = \sum_{i=1}^{t} b_i$*, then*

$$\sum_{i=1}^{t} a_i \log\left(\frac{a_i}{b_i}\right) \geq a \log\left(\frac{a}{b}\right).$$

We are ignoring for now the cases where a_i or $b_i = 0$. The relation is with equality if and only if the two sets are proportionate, *i.e.* $\exists c \ a_i = cb_i, \ \forall i$. When $a = b = 1$ we have two distributions $P|[t]$ and $Q|[t]$. So:

$$\sum_{i=1}^{t} P(i) log\left(\frac{P(i)}{Q(i)}\right) \geq 0$$

and we have equality iff $P = Q$. We denote this with $D(P||Q)$, called the informational divergence of P from Q. This is not a metric (it lacks of simmetry and triangle inequality), but it can be seen as a "dissimilarity" measure. It's called also Kullback-Leibler divergence or *relative entropy*[1].

Proposition 1 is based on the Observation 4. The Proposition will be proved for the natural logarithm.

Proof. We would like to prove that

$$\sum_{i=1}^{t} a_i \log\left(\frac{a_i}{b_i}\right) \geq a \log\left(\frac{a}{b}\right).$$

[1] From the book Elements of Information Theory, Wiley.

First, there is the need to set some convetions. When $b_i = 0$ and $a_i = 0$, we have

$$0\ln(\frac{0}{0}) = 0$$

by convention.

The reason why? $[t] \subset [w]$, you can think of $\{a_i\}$ as a subset of some other set where the other values are all 0's. Otherwise, if $a_i \geq 0$ and $b_i = 0$, we convene that

$$a_i \log\left(\frac{a_i}{b_i}\right) = +\infty.$$

We accept this convention since $b_i \geq 0$, so we can think of $a_i/0$ as the limit of a_i/f_n, for some $f_n \geq 0$ such that $f_n \to 0$. The third case is

$$\sum_{i=1}^{\hat{t}} a_i \log\left(\frac{a_i}{b_i}\right) + \sum_{i=\hat{t}+1}^{t} 0\log\left(\frac{0}{b_i}\right)$$

with $\hat{t} < t$. Here we convene that

$$0\log\left(\frac{0}{b_i}\right) = 0.$$

Notice that

$$\sum_{i=1}^{\hat{t}} a_i \log\left(\frac{a_i}{b_i}\right) + \sum_{i=\hat{t}+1}^{t} 0\log\left(\frac{0}{b_i}\right) \geq a\log\left(\frac{a}{\hat{b}}\right) + 0 \geq a\log\left(\frac{a}{b}\right),$$

with $\hat{b} < b$. Now the proof. First, suppose $a = b$. Keep in mind that

$$\ln(x) \leq x - 1 \Rightarrow \ln\left(\frac{1}{x}\right) \leq \frac{1}{x} - 1.$$

So

$$\sum_{i=1}^{t} a_i \ln\left(\frac{a_i}{b_i}\right) \geq \sum_{i=1}^{t} a_i \left(1 - \frac{b_i}{a_i}\right) =$$

with equality iff $a = b$ (the case then they are different can be easily reduced to this one.)

$$= \sum_{i=1}^{t} a_i - \sum_{i=1}^{t} a_i \frac{b_i}{a_i} = a - b = 0.$$

Assume $b = ca$, for $c \neq 1$ we introduce

$$b_i = c\hat{b}_i \Rightarrow \hat{b}_i = \frac{b_i}{c}.$$

$$\sum_{i=1}^{t} a_i \ln\left(\frac{a_i}{c\hat{b}_i}\right) =$$

$$= \sum_{i=1}^{t} a_i \ln\left(\frac{1}{c}\right) + \sum_{i=1}^{t} a_i \ln\left(\frac{a_i}{\hat{b}_i}\right) \geq \sum_{i=1}^{t} a_i \ln\left(\frac{1}{c}\right) + a \ln\left(\frac{a}{a}\right) =$$

with equality iff $a_i = \hat{b}_i, \ \forall i$

$$= a \ln\left(\frac{1}{c}\right) + a \ln\left(\frac{a}{a}\right) = a \ln\left(\frac{a}{ca}\right) = a \ln\left(\frac{a}{b}\right).$$

Chapter 4

Variable-length codes

A variable-length binary code is a function

$$f : \mathcal{M} \to \{0,1\}^*,\ |\mathcal{M}| < \infty,\ \{0,1\}^* = \bigcup_{i=1}^{\infty} \{0,1\}^i.$$

that can be extended by concatenation. If $m \in M^* \Rightarrow \exists i\ m \in M^i$. We can write

$$m = m_1 m_2 \dots m_i.$$

It follows that f must be invertible, even after concatenation. However, the function $f^* : \mathcal{M}^* \to \{0,1\}^*$, the extension by concatenation, defined as

$$f^*(m_1 \dots m_i) = f(m_1) \dots f(m_i),$$

is not invertible. What is needed is the prefix-free property for f^*.

Let $\underline{x}, \underline{y} \in \{0,1\}^*$. We say that $\underline{x}$ is prefix of $\underline{y}$ if $\underline{x} = \underline{y}$ or $\exists \underline{z} \in \{0,1\}^*$ such that $\underline{x}\underline{z} = \underline{y}$.

So, f is prefix-free if

$$m' \neq m'' \Rightarrow f(m') \not\hookleftarrow f(m''),$$

where "$\hookleftarrow$" is the "is prefix of" relation. If f is prefix-free, f^* is invertible. $|f(m)| = l \Leftrightarrow f(m) \in \{0,1\}^l$. This proposition tells us that lots of short codewords imply that the set of messages is small.

Proposition 2 (Kraft's inequality)**.** *If $f : \mathcal{M} \to \{0,1\}^*$ is a prefix code, then*

$$\sum_{m \in \mathcal{M}} 2^{-|f(m)|} \leq 1.$$

Proof. Let $\underline{x}, \underline{v} \in \{0,1\}^*$. Define Y to be the set of all extension strings of $\underline{x}$ of length L

$$Y_L(\underline{x}) = \{\underline{y} \mid \underline{y} \in \{0,1\}^L \wedge \underline{x} \hookleftarrow \underline{y}\}.$$

Notice that $L < |\underline{x}| \Rightarrow Y_L = \emptyset$. Now, either $Y_L(\underline{x}) \cap Y_L(\underline{v}) \neq \emptyset$, or $Y_L(\underline{x}) \cap Y_L(\underline{v}) = \emptyset$, and maybe $Y_L(\underline{x}) \subset Y_L(\underline{v})$ or the other way.

$$\underline{x} \hookleftarrow \underline{v} \Rightarrow Y_L(\underline{x}) \subseteq Y_L(\underline{v}).$$

$$\underline{x} \not\hookleftarrow \underline{v} \wedge \underline{v} \not\hookleftarrow \underline{x} \Rightarrow Y_L(\underline{x}) \cap Y_L(\underline{v}) = \emptyset.$$

We say that $Y_L(\underline{x})$ and $Y_L(\underline{v})$ can never be in *general position.*

Let A and B be two sets. They are in general position if

$$A \cap B,\ A \setminus B,\ B \setminus A,\ \overline{A \cup B}$$

are all non-empty. For a prefix code, given $m' \neq m''$, then

$$Y_L(f(m')) \cap Y_L(f(m'')) = \emptyset.$$

So consider

$$\{0,1\}^L \supseteq \bigcup_{m \in \mathcal{M}} Y_L(f(m)),$$

and since $|\{0,1\}^L| = 2^L$ and

$$|\{0,1\}^L| \geq \left| \bigcup_{m \in \mathcal{M}} Y_L(f(m)) \right| = \sum_{m \in \mathcal{M}} |Y_L(f(m))| = \sum_{m \in \mathcal{M}} 2^{L-|f(m)|}.$$

Of course, $L \geq \max_{m \in \mathcal{M}} |f(m)|$. Now, we have

$$2^L \geq \sum_{m \in \mathcal{M}} 2^{L-|f(m)|} \Rightarrow 1 \geq \sum_{m \in \mathcal{M}} 2^{-|f(m)|}.$$

□

Proposition 3. *If f is a prefix code then, for any distribution $P|\mathcal{M}$,*

$$\sum_{m \in \mathcal{M}} |f(m)| P(m) \geq H(P).$$

Proof.

$$\sum_{m \in \mathcal{M}} P(m) \log_2 \left(\frac{P(m)}{2^{-|f(m)|}} \right) \geq 0$$

with equality iff $P(m) = 2^{-|f(m)|}$.

$$\sum_{m \in \mathcal{M}} P(m) \log_2(P(m)) - \sum_{m \in \mathcal{M}} P(m) \log_2(2^{-|f(m)|}) =$$

$$= -H(P) + \sum_{m \in \mathcal{M}} P(m)|f(m)| \geq 0 \Rightarrow H(P) \leq \sum_{m \in \mathcal{M}} P(m)|f(m)|.$$

We have equality when $P(m) = 2^{-|f(m)|}$. □

Observation 5. *Given $P|\mathcal{M}$ it is true that $H(P) < \log(|\mathcal{M}|)$, with equality iff P is the equidistribution.*

Proof.

$$\sum_{m\in\mathcal{M}} P(m)\log\left(\frac{P(m)}{\frac{1}{|\mathcal{M}|}}\right) \geq 0$$

with equality iff $P(m) = 1/|\mathcal{M}|$

$$\sum_{m\in\mathcal{M}} P(m)\log\left(\frac{P(m)}{\frac{1}{|\mathcal{M}|}}\right) = -H(P) + \log(|\mathcal{M}|).$$

□

Theorem 4 (Kraft)**.** *Suppose $l : \mathcal{M} \to \mathbb{N}$, a prescribed codeword length, satisfies Kraft's inequality (2). Then $\exists f : \mathcal{M} \to \{0,1\}^*$ prefix code such that $|f(m)| = P(m)$, $\forall m$.*

Proof. We prove this with a greedy algorithm. We'll find an ordering of $\mathcal{M}$, which helps us with being greedy. We order $\mathcal{M}$ so that $l(m_1) \leq l(m_2) \leq \ldots \leq l(m_{|\mathcal{M}|})$.

First step. Set $L = l(m_{|\mathcal{M}|}) = \max_m l(m)$. We work with strings of length L and then we shorten them. Choose arbitrary $\underline{\hat{x}}^{(1)} \in \{0,1\}^L$ and let $f(m_1)$ be the prefix of $\underline{\hat{x}}^{(1)}$ of length $l(m_1)$. We then exclude the set of $2^{L-l(m_1)}$ extensions of $f(m_1)$.

General step. After constructing strings $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{t-1}$, choose $\underline{\hat{x}}^{(t)}$ from $\{0,1\}^L \setminus (Y_L(\underline{x}_1) \cup \cdots \cup Y_L(\underline{x}_{t-1}))$. Then $\underline{x}_t$ = the prefix of length $l(m_t)$ of $\underline{\hat{x}}^{(t)}$. We have to prove that the algorithm does end, and that it builds a prefix code.

The algorithm stops at step t iff $\{0,1\}^L = \bigcup_{i=1}^{t-1} Y_L(\underline{x}_i)$. We have seen that $|Y_L(\underline{x}_i)| = 2^{L-l(m_i)}$. This means that

$$2^L = \left|\bigcup_{i=1}^{t-1} Y_L(\underline{x}_i)\right| \leq \sum_{i=1}^{t-1} |Y_L(\underline{x}_i)| = \sum_{i=1}^{t-1} 2^{L-l(m_i)}$$

dividing by 2^L we obtain

$$1 \leq \sum_{i=1}^{t-1} 2^{-l(m_i)}$$

in contradiction with Kraft's inequality (Proposition 2), since we have at least t messages. If $t < |\mathcal{M}|$ then

$$\sum_{i=1}^{t} 2^{L-l(m_i)} < 2^L$$

so the procedure terminates.

Correctness. f is a prefix-code. We have to show that

$$i \neq j \Rightarrow Y_L(\underline{x}_i) \cap Y_L(\underline{x}_j) = \emptyset.$$

Since they are not in general position, we just have to show that they don't contain one another.

$$Y_L(\underline{x}_t) \not\supset Y_L(\underline{x}_i),\ i < t,$$

since $l(m_i) \leq l(m_t)$, $|Y_L(\underline{x}_i) \geq |Y_L(\underline{x}_t)|$. The sets get smaller and smaller. On the other hand

$$Y_L(\underline{x}_t) \not\subset Y_L(\underline{x}_i),\ i < t$$

recall that $\underline{\hat{x}}^{(t)}$ was chosen in such a way that $\underline{\hat{x}}^{(t)} \in Y_L(\underline{x}_t)$, and that $\underline{\hat{x}}^{(t)} \notin Y_L(\underline{x}_i)$, $\forall i < t$. So $Y_L(\underline{x}_t)$ has an element not in $Y_L(\underline{x}_i)$, so i can't be included.

If our code does not satisfy Kraft's inequality to equality, *i.e.*

$$\sum_{m\in\mathcal{M}} 2^{|f(m)|} < 1,$$

then $\exists \lambda \geq L, \lambda \in \mathbb{N}$ t.c.

$$\sum_{m\in\mathcal{M}} 2^{-|f(m)|} + 2^{-\lambda} \leq 1$$

with $l(m_1) \leq \cdots \leq l(m_{|\mathcal{M}|}) \leq \lambda$ so we could add some more words to our code. A maximal prefix code is a prefix code to which you can't add more codewords (and still get a prefix-code). □

Proposition 4. $\forall P|\mathcal{M}\ \exists f : \mathcal{M} \to \{0,1\}^*$ *prefix code such that*

$$\sum_{m\in\mathcal{M}} P(m)|f(m)| < H(P) + 1$$

Proof. We'll give a prescription satisfying Kraft's inequality and this inequality too.

$$H(P) = \sum_{m\in\mathcal{M}} P(m) \log\left(\frac{1}{P(m)}\right)$$

$$l(m) = \frac{1}{P(m)}$$

is not always an integer and could not satisfy Kraft's inequality.
we could choose

$$l(m) = \left\lceil \log\left(\frac{1}{P(m)}\right)\right\rceil$$

since $\lceil t \rceil < t+1$ we can easily see that

$$\sum_{m\in M} P(m) \left\lceil \log\left(\frac{1}{P(m)}\right)\right\rceil < \sum_{m\in\mathcal{M}} P(m)\log\left(\frac{1}{P(m)}\right) + \sum_{m\in\mathcal{M}} P(m) =$$

$$= H(P) + 1.$$

We now check that l satisfies Kraft's inequality.

$$\sum_{m\in\mathcal{M}} 2^{-l(m)} = \sum_{m\in\mathcal{M}} 2^{-\left\lceil \log\left(\frac{1}{P(m)}\right)\right\rceil} \leq$$

(here we are using $\lceil t \rceil \geq t$)

$$\sum_{m\in\mathcal{M}} 2^{-\log\left(\dfrac{1}{P(m)}\right)} = \sum_{m\in\mathcal{M}} 2^{\log(P(m))} =$$

$$\sum_{m\in\mathcal{M}} P(M) = 1.$$

$\square$

Chapter 5

Entropy of random variables

Let's reason on the probability of a message. You can consider an infinite sequence of random variables X_i which take values in $\mathcal{M}$. We implicitly assumed that X_i are independent random variables. If we group random variables before encoding we can asymptotically reach entropy.

Given a random variable X, the entropy of X is defined as

$$H(X) = H(P_X),$$

where P_X is the distribution of X. How is entropy defined for two random variables? If $X = Y$ then $H(X,Y) = H(X)$. In general, two random variables can be considered as components of a vector random variable:

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

Proposition 5.

$$H(P) \leq log(|\mathcal{X}|),$$

with $P|\mathcal{X}$ and with equality iff $P(x) = 1/|\mathcal{X}|, \forall x \in \mathcal{X}$.

Proof. Call U the uniform distribution. $D(P||U) \geq 0$ with equality iff $P = U$.

$$D(P||U) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{\frac{1}{|\mathcal{X}|}} \right) =$$

$$= \sum_{x \in \mathcal{X}} P(x) \log(x) + \sum_{x \in \mathcal{X}} P(x) \log(|\mathcal{X}|) =$$

$$= -H(P) + \log(|\mathcal{X}|).$$

this quantity is nonnegative with equality to zero iff $P = U$ □

We write $X \in \mathcal{X}$ since X is like an unknown element of $\mathcal{X}$. Random variables don't always have an expected value. We can think of $H(X)$ as the information content of X. The entropy of a pair is the entropt of the random variable derived by the pair, *i.e.* $H(X,Y) = H((X,Y))$, since (X,Y) has a probability distribution P_{XY} and we can think of the pair as just a random variable. $H(X,Y)$ is defined a the *joint entropy* of X and Y. If we consider entropy as the amount of information in a random variable, then it should be that $H(X,Y) \geq X$. We can think of entropy as measure over a set.

$$H(X) \sim \mu(A),\ A \subseteq U$$
$$H(X,Y) \sim \mu(A \cup B),\ \mu(A \cup B) \geq \mu(A)$$

Proposition 6.

$$H(X,Y) \geq H(X).$$

Proof.

$$H(X,Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} Pr\{X = x, Y = y\} \log\left(\frac{1}{Pr\{X = x, Y = y\}}\right)$$

$$H(X) = \sum_{x\in\mathcal{X}} Pr\{X = x\} \log\left(\frac{1}{Pr\{X = x\}}\right) =$$

$$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} Pr\{X = x, Y = y\} \log\left(\frac{1}{Pr\{X = x\}}\right)$$

We take the difference between the two quantities:

$$H(X,Y) - H(X) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} Pr\{X = x, Y = y\} \log\left(\frac{Pr\{X = x\}}{Pr\{X = x, Y = y\}}\right).$$

We can think of $Pr\{X = x, Y = y\} = Pr\{X = x\}Pr\{Y = y|X = x\}$, then

$$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} Pr\{X = x\}Pr\{Y = y|X = x\} \log\left(\frac{1}{Pr\{Y = y|X = x\}}\right) =$$

$$\sum_{x\in\mathcal{X}} Pr\{X = x\} \sum_{y\in\mathcal{Y}} Pr\{Y = y|X = x\} \log\left(\frac{1}{Pr\{Y = y|X = x\}}\right).$$

We would like to say that this quantity is nonnegative. Since $H(\cdots)$ is nonnegative, this difference is actually greater than (or equal to) zero.

$H(X,Y) - H(X)$ is the convex combination of the entropies of the conditional distribution if Y given the various values of X. It's like the expected value of the entropy of the conditional distribution $Pr\{Y = y|X = x\}$. We call it *conditional entropy* of Y given X,

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

It can be seen as the residual information when X is known.

Proposition 7.

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

Proof.

$$H(Y) - H(Y|X) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} Pr\{X=x, Y=y\} \log\left(\frac{1}{Pr\{Y=y\}}\right) -$$

$$-\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} Pr\{X=x, Y=y\} \log\left(\frac{Pr\{X=x\}}{Pr\{X=x, Y=y\}}\right) =$$

$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} Pr\{X=x, Y=y\} \log\left(\frac{Pr\{X=x, Y=y\}}{Pr\{X=x\}Pr\{Y=y\}}\right).$$

This is symmetrical in X and Y. Notice that $H(Y)-H(Y|X)$ is [?]. This also looks as a log sum inequality. In Particular, it is similar to information divergence of the distribution P_{XY} from the distribution $P_X \times P_Y$. So,

$$H(Y) - H(Y|X) = D(P_{XY}||P_X \times P_Y) \geq 0$$

and we have equality when $P_{XY} = P_X \times P_Y$. It is a measure of indipendence of X and Y. □

We define the amount of information that X and Y share as follows:

$$I(X \wedge Y) = H(Y) - H(Y|X),$$

and it is called *mutual information.* It is symmetric and nonnegative. Notice that

$$I(X \wedge Y) = H(Y) - H(Y|X) = H(Y) + H(X) - H(X,Y)$$

thus

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

with equality iff X and Y are independent. Finally, we state (without proof) that

$$H(Y|Z) \geq H(Y|Z,X).$$

The *chain rule* states that

$$H(X_1,\dots,X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_1,\dots,X_{i-1})$$

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

$$I(X \wedge Y|Z) \sim \mu(A \cap B \setminus C)$$

that is the conditional mutual information. We now wonder about which quantity between $I(X \wedge Y|Z)$ and $I(X \wedge Y)$ is greater.

$$I(X \wedge Y) - I(X \wedge Y|Z) \sim \mu(A \cap B) - \mu((A \cap B) \setminus C) =$$

$$= \mu((A \cap B) \setminus ((A \cap B) \setminus C)) = \mu(A \cap B \cap C)$$

this is what the analogy suggests, but we need a proof to believe that this is true.

$$I(X \wedge Y) - I(X \wedge Y|Z) \geq 0 \tag{5.1}$$

If $X \equiv Y \equiv Z$ then

$$I(X \wedge X) = H(X) - H(X|X) = H(X)$$

so Equation 5.1 is possible. We can also have it the other way around:

$$I(X \wedge Y|Z) \geq I(X \wedge Y).$$

It follows that equality does not hold. Suppose $X, Y, Z \in \{0, 1\}$, and also they are uniformly distributed. Moreover, X and Y are independent so $I(X \wedge Y) = 0$. We then define $Z = X \oplus Y$, but then $X = Y \oplus Z$ and $Y = X \oplus Z$. X, Y, Z are pairwise independent, but they are not three-way independent, since every couple of them determines the third.

$$I(X \wedge Y|Z) = H(X|Z) - H(X|Y, Z) = H(X) - 0 = 1.$$

Any number $m \geq 2$ of sets are disjoint iff they are pairwise disjoint, but random variable independence does not satisfy this property. The analogy fails on assuming that random variables independence is similar to set disjointness. n-way independence is binary, but n-way independence is unrelated to m-way independence if $n \neq m$.

An *information source* is an infinite sequence $X_1, X_2, \ldots, X_n$ of random variables with $X_i \in \mathcal{X}$, *i.e.* they take values from the same set. How can we measure the information content in an information source? We denote the information source as X^∞, commonly $X^n = X_1, \ldots, X_n$.

The "speed" of information from a sequence of random variables is

$$\frac{1}{n} H(X_1, \ldots, X_n).$$

The information rate of H^∞, if it exists, is

$$\lim_{n \to \infty} \frac{1}{n} H(X^n).$$

Consider $\{X_i\}_{i=1}^\infty$, i.i.d., and denote by P the common distribution of $\{X_i\}_{i=1}^\infty$, so that $H(X_i) = H(P)$, $\forall i$.

In this case,

$$H(X_1,\ldots,X_n)=\sum_{i=1}^{n}H(X_i)=nH(P)$$

so

$$\lim_{n\to\infty}\frac{1}{n}H(X^n)=H(P).$$

What is the sufficient condition for this to hold (i.e. for the limit to exists)? We need stationary random variables: impredictable but "stationary". We call an information source $\{X_i\}_{i=1}^{\infty}$ *stationary* if

$$\forall n,k\in\mathbb{N},\ \forall\underline{x}\in\mathcal{X}^n\ Pr\{X^n=\underline{x}\}=Pr\{X_{k+1},\ldots,X_{k+n}=\underline{x}\}.$$

We will see that if an information source is stationary then it has an information rate. In particular, with a stationary source the sequence

$$H(X_i|X_1,\ldots,X_{i-1})$$

decreases.

We call an information source *memoryless* if $X_i,\forall i$, are totally independent. We will show that lack of memory is not a sufficient condition for the existence of the information rate.

Consider $\{X_i\}_{i=1}^{\infty}$, a sequence of totally independent random variables, does its entropy rate exists?

$$\frac{1}{n}H(X_1,\ldots,X_n)=\frac{1}{n}\left(\sum_{i=1}^{n}H(X_i)\right)$$

when does this quantity diverge. Take $H(X_i)\in\{0,1\}$, the question is

$$\exists?\lim_{n\to\infty}\sum_{i=1}^{n}\frac{\epsilon_i}{n},\ \epsilon_i\in\{0,1\}$$

[SOMETHING HERE IS MISSING, ASK MICHELE]

Theorem 5. *If $\{X_i\}_{i=1}^{\infty}$ is stationary, then its entropy rate exists.*

Proof. It's reasonable to think that

$$\frac{1}{n}H(X_1,\ldots,X_n)$$

goes to 0. It's sufficient to prove this to show that the rate exists (and it's 0). We just show it's decreasing:

$$\frac{1}{n}H(X_1,\ldots,X_n)\leq\frac{1}{n-1}H(X_1,\ldots,X_{n-1})$$

i.e. we prove

$$(n-1)H(X^n) \leq nH(X^{n-1})$$

$$(n-1)[H(X^{n-1}) + H(X_n|X^{n-1})] \leq nH(X^{n-1})$$

we thus have to prove that

$$(n-1)H(X_n|X^{n-1}) \leq H(X^{n-1})$$

we apply the definition of joint entropy

$$(n-1)H(X_n|X^{n-1}) \leq H(X_1) + \sum_{i=2}^{n-1} H(X_i|X^{i-1})$$

since the source is stationary, we rewrite the right hand side in

$$H(X_n) + \sum_{i=2}^{n-1} H(X_n|X_{1+n-i} \dots X_{n-1})$$

the proof then is

$$(n-1)H(X_n|X^{n-1}) \leq H(X_n) + \sum_{i=2}^{n-1} H(X_n|X_{1+n-i} \dots X_{n-1})$$

this means something like

$$H(X_n|X^{n-1}) \leq H(X_n)$$
$$\vdots$$
$$H(X_n|X^{n-1}) \leq H(X_n|X_{1+n-i} \dots X_{n-1})$$
$$\vdots$$
$$H(X_n|X^{n-1}) \leq H(X_n|X^{n-1})$$

for $n+1$ times for $i \in [2, n-1]$. Look at the generic term

$$H(X_n|X_1 \dots X_{n-1}) \leq H(X_n|X_{n-k} \dots X_{n-1}), \ \forall k$$

this is

$$H(X_n|X_{n-k} \dots X_{n-1}) - H(X_n|X_1 \dots X_{n-1}) =$$
$$I(X_n \wedge X_1 \dots X_{n-k-1}|X_{n-k} \dots X_{n-1}) \geq 0$$

that remind us

$$I(A \cap B|C) \geq 0 \Rightarrow H(A|C) \geq H(A|B,C).$$

□

5.1 Universal compression

Lempel and Ziv worked on universal compression algorithms. Universal compression is possible with stationary source. How would you compress a stationary source to its entropy?

Consider $\{X_i\}_{i=1}^{\infty}$, that is stationary. Given f, a variable length prefix code, we consider $|f(X_i)|$ as a random variable. We can talk about the expected value $E|f(X_i)|$. We know that

$$H(X_i) \leq E|f(X_i)| \leq H(X_i) + 1$$

now,

$$|f(X_1)\dots f(X_n)| = \sum_{i=1}^{n} |f(X_i)|.$$

$$\frac{1}{n}\sum_{i=1}^{n} H(X_i) \leq \frac{1}{n}\sum_{i=1}^{n} |f(X_i)|$$

if the source is stationary all variables have distribution P.

$$H(P) \leq \frac{1}{n}\sum_{i=1}^{n} |f(X_i)| \leq H(P) + 1$$

but we didn't get far. Instead, join random variables together. Let f_n be an optiman (in the sense of length of output) prefix code for $X_1 \dots X_n$.

$$\frac{1}{k}H(X^k) \leq \frac{1}{k}E(|f_n(X^k)|) \leq \frac{1}{k}[H(X^k) + 1]$$

if $k \to \infty$ (and we enlarge the conding window), both sides go to the entropy rate. Thus the entropy is the "limit".

Chapter 6

Error correcting codes

Error correcting codes (ECC) arise in probabilistic contexts. Consider a memory device with n cells. We can model the content as a string $\underline{x} \in \{0,1\}^n$. The device decays; each cell can flip its value independently from the others with a certain probability (uniform). The probability p is usually $< 1/2$. We can expect no more than np flips. Can we guarantee recovery from so many errors? Yes, thanks to algebra. The string $\underline{x}$ can be converted into anything inside a ball of radious np. We want that two strings $\underline{x}$ and $\underline{y}$ are distant, so that their balls do not intersect. The pairwise distance of the words should be $\geq 2np + 1$.

Let $\mathcal{C} \subseteq \{0,1\}^n$ be a block code. This should have some properties, for example $|C|$ should be large. Define

$$d(\mathcal{C}) = \min_{\{\underline{x},\underline{y}\} \in \binom{\mathcal{C}}{2}} d_H(\underline{x}, \underline{y}),$$

where

$$\binom{\mathcal{C}}{2} = \{\{\underline{x}, \underline{y}\} \mid \underline{x}, \underline{y} \in \mathcal{C} \wedge \underline{x} \neq \underline{y}\}.$$

So d tells how close the closest elements are. $d(\mathcal{C})$ should be large too. What is the best possible tradeoff?

$$M(n,d) = \max_{C \mid d(\mathcal{C}) \geq d} |\mathcal{C}|$$

This says that we are concentrating on codes for which $d(\mathcal{C}) \geq d$. We can think of the graph for which $V = \{0,1\}^n$ and $X, Y \in \binom{V}{2}$ have an edge iff $d_H(\underline{x}, \underline{y}) \geq d$. We are looking for the largest clique. We try to make this simpler by looking at the problem from an asymptotic point of view. $M(n, n\delta)$ grows to infinity exponentially in n.

$$\frac{1}{n} \log(M(n, n\delta))$$

does not (grow exponentially ?). We look at the superior limit

$$R(\delta) = \overline{\lim_{n\to\infty}} \frac{1}{n} \log(M(n, n\delta)),$$

where R denotes the rate. $M(n, n\delta)$ is the largest set you can put inside n bits of memory. Its log is the number of bits of true information that we can store. Dividing by n we get the amount of information bit by bit. $\delta \in [0, 1]$, we know $R(0) = 1$ and $R(1) = 0$.

Theorem 6 (Gilbert-Varshamov bound)**.**

$$R(\delta) \geq 1 - h(\delta),\ \ \delta \in [0, \frac{1}{2}].$$

If $\delta > \frac{1}{2}$ *then* $R(\delta) = 0$.

Proof. This means that

$$M(n, n\delta) \geq 2^{n[1-h(\delta)]}$$

(this bound was actually improved to $n2^n[1 - h(\delta)]$).

Fix $n, d \in \mathbb{N}$. Take an arbitrary string $\underline{x} \in \{0, 1\}^n$, exclude $B_H(x, d - 1)$. Then, after having found $\underline{x}_1, \ldots, \underline{x}_{t-1}$, we choose arbitrarily

$$\underline{x}_t \in \{0, 1\}^n \setminus \bigcup_{i=1}^{t-1} B_H(x_i, d - 1).$$

How long can this go on? Maybe after m steps

$$\{0, 1\}^n \setminus \bigcup_{i=1}^{m} B_H(x_i, d - 1) = \emptyset$$

How large m can be? We stop when

$$\{0, 1\}^n \subseteq \bigcup_{i=1}^{m} B_H(x_i, d - 1) \Rightarrow 2^n \leq \left| \bigcup_{i=1}^{m} B_H(x_i, d - 1) \right| \leq$$

$$\leq \sum_{i=1}^{m} |B_H(x_i, d - 1)| \leq \sum_{i=1}^{m} |B_H(x_i, d)| \leq m2^{nh(\frac{d}{n})},$$

where the last inequality derives from the fact that $d \leq n/2$. That's all, since now

$$M(n, d) \geq m \geq \frac{2^n}{2^{nh(\frac{n}{d})}} = 2^{n(1-h(\frac{n}{d}))},$$

where $n/d = \delta$. □

Theorem 7 (Hamming bound)**.**

$$R(\delta) \leq 1 - h(\frac{\delta}{2}),\ \forall \delta \in [0,1].$$

Proof. Fix $n, d \in \mathcal{N}, d_H(\mathcal{C}) \geq d$, arbitrarily. For two Hamming Balls to be disjoint, given the centers $\underline{x}$ and $\underline{y}$, such that $d_H(\underline{x}, \underline{y}) = d$, you must take

$$B_H\left(x, \frac{d-1}{2}\right) \text{ and } B_H\left(y, \frac{d-1}{2}\right).$$

Assume

$$B_H\left(x, \frac{d-1}{2}\right) \neq \emptyset\ \wedge\ B_H\left(y, \frac{d-1}{2}\right) \neq \emptyset$$

But then

$$\exists z \in B_H\left(x, \frac{d-1}{2}\right) \cap B_H\left(y, \frac{d-1}{2}\right)$$

with

$$d_H\left(x\right) z \leq \frac{d-1}{2} \wedge d_H\left(y\right) z \leq \frac{d-1}{2}.$$

It follows that

$$d = d_H\left(x\right) y \leq d_H\left(x\right) z + d_H\left(y\right) z \leq 2\frac{d-1}{2} = d - 1 \leq d$$

(contradiction.)

So we can correct up to $d - 1/2$ errors. Assume $\mathcal{C}$ is built using disjoint balls, and $M(n,d) = |\mathcal{C}|$. Then

$$\left|\bigcup_{\underline{x}\in\mathcal{C}} B_H\left(x, \frac{d-1}{2}\right)\right| = |\mathcal{C}| \cdot \left|B_H\left(0, \frac{d-1}{2}\right)\right| \geq \frac{|\mathcal{C}|}{n+1} 2^{nh\left(\frac{d-1}{2}\right)}$$

since

$$\left|\bigcup_{\underline{x}\in\mathcal{C}} B_H\left(x, \frac{d-1}{2}\right)\right| \leq 2^n,$$

$$M(n,d) = |\mathcal{C}| \leq n + 12^{n\left(1 - h\left(\frac{d-1}{2}\right)\right)}$$

but then

$$\frac{1}{n}\log(M(n,d)) \leq \frac{1}{n}\log(n+1) + 1 - h(\frac{d-1}{2n}$$

then

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\log(M(n,d)) \leq 1 - h\left(\frac{\delta}{2}\right)$$

□

6.1 Uniquely decodable codes

$f : \mathcal{M} \to \{0,1\}^*$ is *uniquely decodable* (UD) id $f^* : M* \to \{0,1\}^*$ is injective. Let $\underline{m} \in M^*$, with $\underline{m} = m_1 m_2 \ldots m_t$.

$$f^*(\underline{m}) = f(m_1) f(m_2) \ldots f(m_t).$$

If f is a prefix code, f is UD.

[three pages missing 22/11/2016, uniquely decodable codes]

$$M(n,d) = \max_{\mathcal{C} \subseteq \{0,1\}^n | d_H(\mathcal{C}) \geq d} |\mathcal{C}|,$$

with

$$d_H(\mathcal{C}) = \min_{\{\underline{x},\underline{y}\} \in \binom{\mathcal{C}}{2}} d_H(\underline{x},\underline{y}).$$

$$R(\delta) = \overline{\lim_{n\to\infty}} \frac{M(n, n\delta}{n}$$

we know that, if $\delta \leq 1/2$,

$$R(\delta) \geq 1 - h(\delta) \ \wedge \ R(\delta) \leq 1 - h(\frac{\delta}{2})$$

Proof. Let $\mathcal{C} \subseteq \{0,1\}^n$, chosen arbitrary.

$$d_H(\mathcal{C}) \leq \sum_{\{\underline{x},\underline{y}\} \in \binom{\mathcal{C}}{2}} \frac{d_H\,(x)\,1}{\frac{1}{\binom{M}{2}}} =$$

with $M = |\mathcal{C}|$,

$$= \frac{2}{M(M-1)} \sum_{\{\underline{x},\underline{y}\} \in \binom{\mathcal{C}}{2}} d_H(\underline{x}, \underline{1}) = \frac{2}{M(M-1)} \sum_{\{\underline{x},\underline{y}\} \in \binom{\mathcal{C}}{2}} \sum_{i=1}^{n} d(x_i, y_i) =$$

where $\underline{x} = x_1 \ldots x_n$. Think of the last expression as a matrix that has n columns and $|\mathcal{C}|$ rows. The contribution at the i-th column is the number of 1's multiplied by the number of 0's.

$$= \frac{2}{M(M-1)} \left(\frac{M}{2}\right)^2 = \frac{nM}{2(M-1)}$$

so, the minimum distance is

$$d_H(\mathcal{C}) = d \leq \frac{nM}{2(M-1)}$$

$$2(M-1)d \leq nM \Rightarrow 2Md - 2d \leq nM \Rightarrow M(2d-n) \leq 2d$$

consider $\delta = d/n$

$$M(2\delta - 1 \leq 2\delta$$

now consider $\delta > 1/2$

$$M \leq \frac{2\delta}{2\delta - 1}$$

So, if $M(n, n\delta) = M_n$ when $\delta \geq 1/2$ we have that $M(n, n\delta)$ is a constant. Thus, $R(\delta) = 0$ for $\delta \geq 1/2$ □

6.2 Parity Check Codes

$$\mathcal{C}_n = \{\underline{x} \mid \underline{x} \in \{0,1\}^n, 2|\sum_{i=1}^{n} x_i\}$$

is the set of strings which have an even number of 1's. $|\mathcal{C}| = 2^{n-1}$. This will not correct errors, but it will detect a single error (or in fact an odd number of errors). Consider $\{0,1\}^n$ as a n-dimensional vector space over GF(2).

Define $(\underline{x}, 1)$ to be the scalar product, i.e.

$$(\underline{x}, 1) = \left(\sum_{i=1}^{n} x_i\right) \mod 2.$$

Let

$$\mathcal{C}_n = \{\underline{x} \mid (\underline{x}, 1) = 0\}.$$

It is a hyperplane made of vectors orthogonal to 1. This can be made more general, and fix an arbitrary vector $\underline{s}$,

$$\mathcal{C}_n(\underline{s}) = \{\underline{x} \mid (\underline{x}, \underline{s}) = 0\}.$$

Consider the set $S \subseteq [n]$ of indices of $\underline{s}$ which are 1. One only care about errors on x_i for $i \in S$. This is a linearly closed space, with addition $\oplus$ (mod 2). Also, their intersection is closed. We can take a bunch of vectors, and the hyperplanes orthogonal to them, and their intersection

$$\bigcap_{\underline{s} \in S} \mathcal{C}_n(\underline{s})$$

In this way I can construct codes that not only detect errors, but also correct them.

$$\mathcal{C}_n = \{\underline{x} \mid \underline{x} \in \{0,1\}^n, (\sum_{i=1}^{n} x_i) \mod 2 = 0\}$$

This is linearly closed, and can also be defined as $\mathcal{C}_n = \{\underline{x} \mid (\underline{x}, 1) = 0\}$. We will consider scalar product with other vectors, which will give us other linearly closed spaces; their intersection will also be a linearly closed space.

Remind that GF(2) means "Galois Field over $\{0,1\}$", and the only linear combination of vectors is the sum. $\mathcal{L} \in \{0,1\}^n$ is a linear space if

- $\mathcal{L} \neq \emptyset$,
- is closed under linear combination, i.e. $\forall \underline{x}, \underline{y} \in \mathcal{L}\ \underline{x} \oplus \underline{y} \in \mathcal{L}$.

Take a linear space and consider the orthogonal complement

$$\mathcal{L}^{\perp} = \{\underline{z} \mid \underline{z} \in \{0,1\}^n, (\underline{z}, \underline{x}) = 0\ \forall \underline{x} \in \mathcal{L}\}$$

We have $\mathcal{L}$. Consider $\underline{x} \in \mathcal{L}$, and a basis for $\mathcal{L}^{\perp}$.

$$\mathcal{L}^{\perp} = \bigcap_{\underline{x} \in \mathcal{L}} \mathcal{C}_n(\underline{x})$$

Let us write all the vectors from the basis of $\mathcal{L}^{\perp}$ as column vectors of a matrix A with n rows and m columns.

$\underline{x}A = \underline{0}$. So $\mathcal{L} = \{\underline{x} \mid \underline{x}A = \underline{0}\}$. A linear code is specified by the A matrix, which is called the parity check matrix of $\mathcal{L}$. Given a matrix A we have

$$\ker A = \{\underline{x} \mid \underline{x}A = \underline{0}\}$$

$$\mathrm{Im} A = \{\underline{z} \mid \exists \underline{x} \in \{0,1\}^n,\ \underline{x}A = \underline{z},\ \underline{z} \in \{0,1\}^m\}$$

So ImA is the linear combination of its rows. Pick the $e_i \in \{0,1\}^m$ vector,

$$e_i A = A_i^T$$

that is, the i-th row of A. Remind that given a code $\mathcal{C}$, if $d_H(\mathcal{C}) = d$ we can correct up to $(d-1)/2$. Also, recall the definition of Hamming weight:

$$w_H(\underline{x}) = \sum_{i=1}^{n} x_i,\ \ \underline{x} \in \{0,1\}^n.$$

Observation 6. *If $\mathcal{L} \in \{0,1\}^n$ is a linear code then*

$$d_H(\mathcal{L}) = \min_{\underline{x} \in \mathcal{L},\ \underline{x} \neq \underline{0}} w_H(\underline{x})$$

Proof.Consider $\underline{x} \neq \underline{0}$ such that it minimizes $w_H(\underline{x})$ over $\mathcal{L}$. Then

$$d_H(\mathcal{L}) \leq \min_{\underline{x} \in \mathcal{L},\ \underline{x} \neq \underline{0}} w_H(\underline{x}),$$

since $d_H(\underline{0}, \underline{x}) = w_H(\underline{x})$ for

$$d_H(\mathcal{L}) \geq \min_{\underline{x} \in \mathcal{L},\ \underline{x} \neq \underline{0}} w_H(\underline{x}).$$

"Distance is translation invariant". Consider $\underline{z} \in \mathcal{L}$, and

$$\underline{z} \oplus \mathcal{L} = \{\underline{x} \oplus \underline{x} \mid \underline{x} \in \mathcal{L}\} = \mathcal{L},$$

since it's linearly closed.

$$d_H(\underline{x}, \underline{y}) = d_H(\underline{0}, \underline{x} \oplus \underline{y}) = w_H(\underline{x} \oplus \underline{y}) \geq \min_{\underline{x} \in \mathcal{L},\ \underline{x} \neq \underline{0}} w_H(\underline{x})$$

□

$$M(n,d) = \max_{\mathcal{C} \subseteq \{0,1\}^n,\ d_H(\mathcal{C}) \geq d}$$

$M(n,3)$ is the largest cardinality of a code correcting 1 error. By the Hamming bound we could build a code using disjoint balls of radius 1, non-intersecting.

$$2^n \geq \left| \bigcup_{\underline{x} \in \mathcal{C}} B_H(x,1) \right| = \sum_{\underline{x} \in \mathcal{C}} |B_H(x,1) = |\mathcal{C}| n + 1 \Rightarrow |\mathcal{C}| \leq \frac{2^n}{n+1}$$

where there is equality in the fist inequality if the balls cover the entire space. So

$$M(n,3) \leq \frac{2^n}{n+1}.$$

Theorem 8 (Hamming). *$M(n,3) = \dfrac{2^n}{n+1}$ if $\exists m\ n = 2^m - 1$*

$$M(2^M - 1, 3)) = \frac{2^{2m}}{22^m} = 2^{m-1}$$

6.3 BCH Codes

Consider all the non-zero vectors in $\{0,1\}^m$. Let A be a matrix having all these strings as its rows ($2^m - 1 = n$ rows, m columns). $\mathcal{C} = \ker A$ is the code we are looking for. We have to prove that $d_H(\mathcal{C}) \geq 3$, thus

$$\min_{\underline{x} \in \mathcal{C},\ \underline{x} \neq 0} w_H(\underline{x}) = 3,$$

or equivalently that $\forall \underline{x} \in \mathcal{C},\ \underline{x} \neq \underline{0},\ w_H(\underline{x}) \geq 3$:

- $\underline{x} \in \mathcal{C}, \underline{x} \neq 0 \Rightarrow w_H(\underline{x}) \neq 1$. This is true since if $w_H(\underline{x}) = 1$ then $\underline{x} = e_i$, for some i ($\underline{x}A$ is a row of A).
- $w_H(\underline{x}) = 2 \Rightarrow \underline{x} \notin \mathcal{C}$. $\underline{x}A$ would be a linear combination of two rows of A, but they are different, so say $\underline{x} = e_i \oplus e_j$. Then

  $$\underline{x}A = e_iA \oplus e_jA$$

  but since $e_j \neq e_i$, also $e_iA \neq e_jA$ and $e_iA \oplus e_jA \neq \underline{0}$.

Observation 7. *The Hamming Balls of radius 1 around the elements of $\mathcal{C}$ fill up the space of $\{0,1\}^n$, thus*

$$\forall \underline{z} \in \{0,1\}^n \ \exists \underline{x} \in \mathcal{C} \ s.t. \ \underline{z} \in B_H\left(x,1\right).$$

Ver. (?).

- $\underline{z} \in \mathcal{C}$, we are ok.
- $\underline{z} \notin \mathcal{C}$, then $\underline{z}A \neq \underline{0}$. $|\underline{z}A| = m$. $\underline{z}A \in \{0,1\}^m$ but $\underline{z}A \neq \underline{0}$ so it is a row of A. It follows that $\underline{z}A = e_iA$ for some i. So $(\underline{z} \oplus e_i)A = \underline{0}$, thus $\underline{z} \oplus e_i \in \ker A$, so $\underline{z} \oplus e_i = \underline{x}$ is at distance 1 from $\underline{z}$ and $\underline{z} \in B_H\left(x,1\right)$.

□

Chapter 7

The communication model

Shannon developed a mathematical model of the communication among parties. In particular, we will not consider the simplest one, in which the communication is one-way between two entities. In this model, one entity is the *source* that sends information to the second entity, called the *receiver*, through a *noisy channel*.

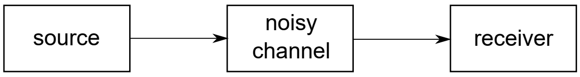

Figure 7.1: Graphic model of a simple communication.

Source and receiver are usually separated in space and communicate (approximately) at the same time. Indeed, the roles of space and time can be swapped (i.e. a communication that evolves over time at the same place, like data storage and retrieval).

The noisy channel is both the veicle and the obstacle. One cannot transmit a signal without it being modified to some extent; moreover the communication must be quick because time is expensive. The alterations to the message are errors and therefore must be corrected. In order to do this we need to remove "bad" redundancy and add "good" redundancy. So we have a tradeoff between data integrity and fast communication.

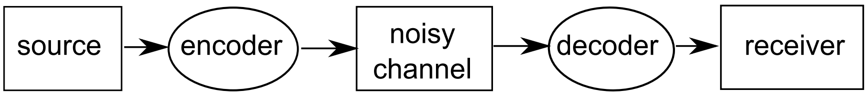

Figure 7.2: The role of encoder and decoder.

The balance is found by the *encoder*. This entity takes the message from the source, modify it in some way, and transmit the result over the channel. A *decoder* is the entity that apply some transformation to the incoming flow of information before handling it to the receiver. The decoder can't perform the inverse of the encoder's function, because the channel applies noise to the information. We will concentrate on the noisy channel, without considering encoder and decoder's interfaces to source and receiver.

7.1 Discrete Memoryless Channel (DMC)

For now, the channel is a black box. Let $\mathcal{X}$ be the input alphabet and $\mathcal{Y}$ be the output alphabet for the strings that respectively enter and exit the channel. Select an input symbol $x \in \mathcal{X}$; a random symbol $y \in \mathcal{Y}$ comes out of the channel.

We can think that there is a "devil" inside the black box (yes, you've read correctly, lol). The devil has dices that have $|\mathcal{Y}|$ faces. Whenever an input signal enters the channel, the devil launches a dice and outputs the symbol on the facet that comes up. The devil can't launch any dice; instead, there is a mapping between dices and input symbols (you can think of x as the name of the dice that the devil has to use). So we can influence the devil in using some specific dice (out of the $|\mathcal{X}|$ dices), but the outcome is still stochastic. The probability distribution of each dice is arbitrary. Note that the channel is usable if e have different dices; otherwise, the output is completely uncontrollable.

Let $W(y|x)$ be the probability that the dice named x will fall onto its facet y, with the following conditions:

- $W(y|x) \geq 0$,
- $\forall x \sum_{y \in \mathcal{Y}} w(y|x) = 1$.

We can think of W as a matrix whose columns are indexed by y and rows are indexed by x.

To communicate iteratively pick $x_1, x_2, \dots, x_n \in \mathcal{X}^n$ and launch a sequence of dices. There is a probability distribution for $y_1, y_2, \dots, y_n \in Y^n$. Consider the function

$$W^n(\underline{x}, \underline{y}) : \mathcal{X}^n \to \mathcal{Y}^n$$

defined as

$$W^n(\underline{x}, \underline{y}) = \prod_{i=1}^{n} W(y_i|x_i)$$

We assume that every symbol is independently modified (lack of memory of the channel). We can now define the DMC (over time instants) as

$$DMC = \{W^n\}_{n=1}^{\infty}$$

7.2 Shannon's Noisy Channel Theorem

The encoder takes the messages of the source and maps it to an arbitrary subset $C \subseteq \mathcal{X}^n$. So we have $|C|$ many messages that flow in the channel; they can be non-binary strings, but to have an approximation of their length we can assume they are binary. Then, the length of these messages is $\log_2 |C|$. We define the speed of the channel as

$$\frac{1}{n} \log_2 |C|,$$

where n is the length of transmission in some unit of time. We can call this *rate* of the code.

Now we are interested in the quality of transmission, that is how well data transmitted can be recovered. The decoder is a function $\varphi : \mathcal{Y}^n \to C$. An error event happens when $\underline{x} \in \mathcal{X}^n$ is transmitted and $\varphi(\underline{y}) \neq \underline{x}$. Define

$$W^n(T|\underline{x}) = \sum_{\underline{y} \in T} W^n(\underline{y}|\underline{x}),$$

where $T \subseteq \mathcal{Y}$. The error probability of a string $\underline{x} \in \mathcal{X}^n$ is

$$1 - W^n(\varphi^{-1}(\underline{x})|\underline{x}) = W^n(\overline{\varphi^{-1}(\underline{x})}|\underline{x})$$

where

$$\varphi^{-1}(\underline{x}) = \{\underline{y} \mid \varphi(\underline{y}) = \underline{x}\}$$

Define the (maximum) error probability of the code (C_n, φ_n) (W^n is fixed) as

$$e_m(W^n, C_n, \varphi_n) = \max_{\underline{x} \in C} W^n(\overline{\varphi^{-1}(\underline{x})}, \underline{x}).$$

Proposition 8. *$R \geq 0$ is an* achievable rate *of transmission over the DMC $\{W\}$ if exists $\{(C_n, \varphi_n)\}_{n=1}^{\infty}$ such that:*

- $\lim_{n\to\infty} \dfrac{1}{n} \log_2(|C_n|) \geq R$*, and*
- $\lim_{n\to\infty} e_m(W, C_n, \varphi_n) = 0$.

What is the highest achievable rate?

$$C \subseteq \mathcal{X}^n \Rightarrow \frac{1}{n} \log_2 |C| \leq \log_2 |\mathcal{X}^n|$$

So the maximum achievable rate is $\log_2 |\mathcal{X}^n|$. Also, if $R_t \to R$ and R_t is a series of values representing achievable rates, then R is an achievable rate too (simple analysis, will not be demonstrated).

Shannon wondered about the highest achievable rate $C(W)$ of a given DMC $\{W\}$. He called this number *capacity of the channel.* The answer came out of his intuition but it has been proved true by his students and coworkers.

Let $x \in \mathcal{X}$ be a symbol chosen following a probability distribution $P|\mathcal{X}$. The input symbol will be sent through the channel W and the output is conditioned by W. Let Y be the random variable (RV), $Y \in \mathcal{Y}$, be the output of W with respect to input X. Then, we have a joint distribution $P, W|\mathcal{X} \times \mathcal{Y}$. The probability that a specific input x corresponds to a received symbol y is

$$Pr\{X = x, Y = y\} = P(x)W(y|x).$$

Define $I(X \wedge Y)$ to be the number of bits that one can transmit over a channel in a unit of time. Intuitively, it is what y retains of input x (mutual information). We can maximize I by controlling the probability distribution $P|\mathcal{X}$ (pardon the abuse of notation):

$$\max_{P|\mathcal{X}} I(P, W)$$

So, what is the formula for I?

$$I(X \wedge Y) = \sum_{(x,y) \in X \times Y} Pr\{X = x, Y = y\} \log_2 \left(\frac{Pr\{X = x, Y = y\}}{Pr\{X = x\} Pr\{Y = y\}} \right) =$$

$$= \sum_{x \in X, y \in Y} Pr\{X = x\} W(y|x) \log_2 \left(\frac{Pr\{X = x\} W(y|x)}{\sum_{x \in X} Pr\{X = x\} W(y|x)} \right) = I(P, W)$$

Theorem 9 (Shannon's Noisy Channel Theorem)**.** *The highest achievable rate, given a DMC W, is*

$$C(w) = \max_{P|X} I(P, W).$$

Notice that $C(W)$ is 0 if X and Y are independent $\Leftrightarrow$ all rows in W are the same.

7.3 Binary Symmetric Channel

A binary symmetric channel is used to transmit binary strings and the DMC has the form

$$W = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}, \mathcal{X} = \mathcal{Y} = \{0, 1\}$$

It is called symmetric because W is. In this context $P|\mathcal{X}$ is called *crossover* probability. The (hypothetical, since we proved nothing) capacity of this channel according to Theorem 9 is:

$$I(X \wedge Y) = H(Y) - H(Y|X) \leq 1 - h(p)$$

This is true because:

- the maximum value for $H(Y)$ is 1;
- the following equations hold:

$$H(Y|X) = \sum_{x \in X} Pr\{X = x\} H(Y|X = x) = h(p),$$

where

$$H(Y|X = x) = \sum_{y \in Y} Pr\{Y = y|X = x\} \log_2 \left(\frac{1}{Pr\{Y = y|X = x\}} \right).$$

Is this upper bound achievable? We must make sure that

$$\exists X \ H(Y) = 1$$

We can impose the uniform distribution when picking the input values, i.e. $Pr\{X = 0\} = 1/2$.

We will show that $1 - h(2p)$ is an achievable rate if $p < 1/4$. We will to this using error correcting codes.