Notes on Information Theory

Cristian Di Pietrantonio Michele Laurenti

February 3, 2017

Contents

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

Entropy and the Hamming Space

This Chapter introduces an important concept: the Hamming Ball.

2.1 The 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 to define a metric (or distance) on it.

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

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

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

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 (Radius of the Hamming ball)**.** *r does not need 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:

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

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:

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

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

2.2 The Volume of the 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 n-th 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 $|B_H(\underline{0}, n)| = 2^n$. Instead, the volume of the ball with radius $n/2$ (in the triangle's point of split) is $\frac{2^n}{2} = 2^n 2^{-1} = 2^{n-1}$. It follows, if $r \geq \frac{n}{2}$, that $2^{n-1} \leq |B_H(\underline{0}, r)| \leq 2^n$. Can we bound the volume for $r < \frac{n}{2}$?

Entropy

First we introduce the notion of *entropy*.

Definition 3 (Entropy)**.** *Entropy is a function* $h : [0,1] \to [0,1]$ *defined as follows:*

$$h(t) = t \cdot \log_2\left(\frac{1}{t}\right) + (1-t) \cdot \log_2\left(\frac{1}{1-t}\right). \tag{2.5}$$

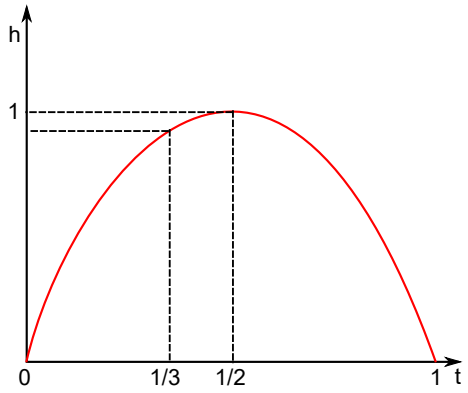

Figure 2.1: The entropy function.

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 a probability distribution $(t, 1-t)$ and entropy is a number attached to it. Entropy measures symmetry and with $t = \frac{1}{2}$ we have maximum chaos (the future outcomes are equally likely).

Lower and Upper Bounds

Theorem 1 (Upper bound of the volume of the Hamming Ball)**.** *The volume of the Hamming ball of radius* r*, for* $r \leq \frac{n}{2}$*, is*

$$|B_H(\underline{0}, r)| \leq 2^{n\, h\left(\frac{r}{n}\right)}.$$

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 $\frac{w}{W}$.

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

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

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

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

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

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

at this point we take out the lightest string and write

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

Which are the lightest strings? Since we assumed that $r \leq \frac{n}{2}$ then

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

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

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

So we have

$$1 \geq |B_H(\underline{0}, r)| \cdot \frac{1}{2^{n\,h\left(\frac{r}{n}\right)}} \implies |B_H(\underline{0}, r)| \leq 2^{n\,h\left(\frac{r}{n}\right)}. \qquad \square$$

Theorem 2 (Lower bound of the volume of the Hamming Ball)**.** *The volume of the Hamming ball of radius r, for $r \leq \frac{n}{2}$, is*

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

Proof.

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

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

$$T_q = \{\underline{x} \ : \ w_H(x) = q\},$$

obtaining $n+1$ classes. We know that $|T_q| = \binom{n}{q}$; we are not interested in how many strings are in that set, but what is the total weight. There is not symmetry in the weight of the partitions. In fact we can prove that

$$\frac{P^n(T_q)}{P^n(T_r)} \leq 1, \ \forall q.$$

In the formula above there are binomials we want to bound. In order to do that, we need the following observation.

Observation 2 (Upper bound to factorial)**.**

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

Verification. If $k \geq l$:

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

If $k < l$:

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

Define $p = \frac{r}{n}$ 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

$$\begin{aligned}
\frac{P^n(T_q)}{P^n(T_r)} &= \frac{p^q(1-p)^{n-q}\,|T_q|}{p^r(1-p)^{n-r}\,|T_r|} \\
&= p^{q-r}(1-p)^{r-q}\,\frac{\frac{n!}{q!(n-q)!}}{\frac{n!}{r!(n-r)!}} \\
&= p^{q-r}(1-p)^{r-q}\,\frac{r!}{q!}\,\frac{(n-r)!}{(n-q)!} \\
&\leq p^{q-r}(1-p)^{r-q} r^{r-q}(n-r)^{q-r} \\
&= p^{q-r}(1-p)^{r-q}(np)^{r-q}[n(1-p)]^{q-r} \qquad \text{(since } r = np\text{)} \\
&= p^{q-r}(1-p)^{r-q} n^{r-q+q-r} p^{r-q}(1-p)^{q-r} \\
&= p^{q-r+r-q}(1-p)^{r-q+q-r} = 1.
\end{aligned}$$

So we can write

$$\begin{aligned} 1 = P^n(\{0,1\}^n) &= P^n\left(\bigcup_{q=0}^{n} T_q\right) \\ &= \sum_{q=0}^{n} P^n(T_q) \\ &\leq (n+1)\max_q P^n(T_q) \\ &= (n+1)P^n(T_r) \\ &= (n+1)\,|T_r| \cdot 2^{-n\,h\left(\frac{r}{n}\right)} \end{aligned}$$

Thus, we know that $|T_r| \geq \frac{1}{n+1}2^{n\,h\left(\frac{r}{n}\right)}$, and since $T_r \subseteq B_H\left(\underline{0}, r\right)$ we have proved that $|T_r| \leq |B_H\left(\underline{0}, r\right)|$. □

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 altered 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, which we call the alphabet. We are interested 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.

Definition 4 (Frequency of an alphabet symbol)**.** *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:*

$$N(a|\underline{x}) = \left|\{i \ : \ x_i = a\}\right|.$$

where $\underline{x} = x_1 x_2 \ldots x_n$.

One can think about "normalized" relative frequencies of symbols

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

Moreover the following holds:

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

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

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

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, $\mathcal{T}_P$, with $P|\mathcal{X}$.

$$\mathcal{T}_P = \mathcal{T}_P^n = \{\underline{x} \ : \ P_{\underline{x}} = P\}.$$

Definition 5 (Generalised Entropy)**.** *We introduce the* generalized entropy $H(P)$*, defined as*

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

Theorem 3 (Cardinality of $\mathcal{T}_P$)**.** *If $\mathcal{T}_P \neq \emptyset$, then*

$$\frac{1}{(n+1)^{|\mathcal{X}|-1}} \cdot 2^{n\,H(P)} \leq |\mathcal{T}_P| \leq 2^{n\,H(P)}.$$

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

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

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

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

Now,

$$\begin{aligned}
\forall \underline{x} \in \mathcal{T}_P^n . P^n(\underline{x}) &= \prod_{a\in\mathcal{X}} P(a)^{nP(a)} \\
&= \prod_{a\in\mathcal{X}} 2^{n\,P(a)\log_2 P(a)} \qquad \text{(since it's independent of } \mathcal{X}) \\
&= 2^{n\left[\sum_{a\in\mathcal{X}} P(a)\log_2 P(a)\right]} \\
&= 2^{-n\,H(p)}.
\end{aligned}$$

So,

$$1 = P^n(\mathcal{X}^n) \geq P^n(\mathcal{T}_P^n) = |\mathcal{T}_P^n| \cdot 2^{-n\,H(p)}. \qquad \square$$

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

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

Observation 3 (Ratio between probability of types)**.** *If $\mathcal{T}_P^n \neq \emptyset$, then*

$$\frac{P^n(\mathcal{T}_Q^n)}{P^n(\mathcal{T}_P^n)} \leq 1.$$

This means that if a distribution is a type, it maximizes its (product) value on the strings of that type.

Proof. We can suppose, without loss of generality, that $\mathcal{T}_Q^n \neq \emptyset$.

$$
\begin{aligned}
P^n(\mathcal{T}_Q^n) &= \prod_{a\in\mathcal{X}} P(a)^{n\,Q(a)}\,\left|\mathcal{T}_Q^n\right| \\
&\Longrightarrow \\
\frac{P^n(\mathcal{T}_Q^n)}{P^n(\mathcal{T}_P^n)} &= \frac{\left|\mathcal{T}_Q^n\right|\prod_{a\in\mathcal{X}}[P(a)]^{nQ(a)}}{|\mathcal{T}_P^n|\prod_{a\in\mathcal{X}}[P(a)]^{nP(a)}} \\
&= \frac{\frac{n!}{\prod_{a\in\mathcal{X}}[nQ(a)]!}\prod_{a\in\mathcal{X}}[P(a)]^{nQ(a)}}{\frac{n!}{\prod_{a\in\mathcal{X}}[nP(a)]!}\prod_{a\in\mathcal{X}}[P(a)]^{nP(a)}} \\
&= \prod_{a\in\mathcal{X}}\frac{[n\,P(a)]!}{[n\,Q(a)]!}\prod_{a\in\mathcal{X}}[P(a)]^{n\,(Q(a)-P(a))} \\
&\leq \prod_{a\in\mathcal{X}}[nP(a)]^{n\,(P(a)-Q(a))}\prod_{a\in\mathcal{X}}[P(a)]^{n\,(Q(a)-P(a))} \\
&= n^{n\left[\sum_{a\in\mathcal{X}}P(a)-Q(a)\right]}\frac{\prod_{a\in\mathcal{X}}[P(a)]^{n(P(a)-Q(a))}}{\prod_{a\in\mathcal{X}}[P(a)]^{n(P(a)-Q(a))}} \\
&= n^{n\left[\sum_{a\in\mathcal{X}}P(a)-\sum_{a\in\mathcal{X}}Q(a)\right]} \\
&= n^{n[1-1]} = 1.
\end{aligned}
$$

So we can write

$$
\max_{Q|\mathcal{X}} P^n(\mathcal{T}_Q^n) = |\mathcal{T}_P^n|\prod_{a\in\mathcal{X}}P(a)^{n\,P(a)} = |\mathcal{T}_P^n|\cdot 2^{-n\,H(P)}
$$

and conclude that

$$
1 \leq (n+1)^{|\mathcal{X}|-1}\,|\mathcal{T}_P^n|\cdot 2^{-n\,H(P)} \implies \frac{1}{(n+1)^{|\mathcal{X}|-1}}\cdot 2^{n\,H(P)} \leq |\mathcal{T}_P^n|\,. \qquad \square
$$

Chapter 3

The log sum inequality

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

Observation 4 (Logarithm cap-convex)**.** *The logarithm function is cap-convex (∩-convex). Remember that* $\ln(t) \leq t - 1$ *(with equality if and only if* $t = 1$*) and* $\log_2(t) = \frac{\ln(t)}{\ln(2)}$.

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

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

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.*, $\exists c : a_i = c \cdot b_i$, $\forall i$. When $a = b = 1$ we have two distributions $P|[t]$ and $Q|[t]$. So:

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

and we have equality if and only if $P = Q$. We denote this with $D(P||Q)$, called the informational divergence of P from Q. This is not a metric (it lacks of symmetry 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 observation 4. The Proposition will be proved for the natural logarithm.

Proof. We would like to prove that

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

First, there is the need to set some conventions. When $b_i = 0$ and $a_i = 0$, we say, by convention, that

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

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

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 0s. Otherwise, if $a_i > 0$ and $b_i = 0$, we convene that

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

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

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

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

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

Notice that

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

with $\hat{b} < b$.

Now the proof. First, suppose $a = b$. Keep in mind that

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

So,

$$\begin{aligned} \sum_{i=1}^{t} a_i \ln\left(\frac{a_i}{b_i}\right) &\geq \sum_{i=1}^{t} a_i \left(1 - \frac{b_i}{a_i}\right) && \text{(w. eq. iff } a = b) \\ &= \sum_{i=1}^{t} a_i - \sum_{i=1}^{t} a_i \frac{b_i}{a_i} = a - b = 0. \end{aligned}$$

The case then they are different can be easily reduced to this one.

Assume $b = c \cdot a$, for $c \neq 1$. We introduce

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

Then,

$$\begin{aligned} \sum_{i=1}^{t} a_i \ln\left(\frac{a_i}{c \cdot \hat{b}_i}\right) &= \sum_{i=1}^{t} a_i \ln\left(\frac{1}{c}\right) + \sum_{i=1}^{t} a_i \ln\left(\frac{a_i}{\hat{b}_i}\right) \\ &\geq \sum_{i=1}^{t} a_i \ln\left(\frac{1}{c}\right) + a \ln\left(\frac{a}{a}\right) && \text{(w. eq. iff } a_i = \hat{b}_i,\ \forall i) \\ &= a \ln\left(\frac{1}{c}\right) + a \ln\left(\frac{a}{a}\right) \\ &= a \ln\left(\frac{a}{c \cdot a}\right) = a \ln\left(\frac{a}{b}\right). && \square \end{aligned}$$

Chapter 4

Variable-length codes

Let $\mathcal{M}$ be some message space, with $|M| < \infty$. A variable-length binary code is a function $f^\star : \mathcal{M}^\star \to \{0,1\}^\star$, where $\{0,1\}^* = \bigcup_{i=1}^{\infty} \{0,1\}^i$ denotes the set of all binary strings of any given length, and $\mathcal{M}^\star$ denotes the concatenation of messages, *i.e.*, if $m \in \mathcal{M}^*$, then $\exists i : m \in \mathcal{M}^i$. We can write

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

A variable-length code must be invertible.

If we take a function $f : \mathcal{M} \to \{0,1\}^*$, its extension by concatenation $f^* : \mathcal{M}^* \to \{0,1\}^*$, defined as

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

is not always invertible. For it to be invertible, f must be *prefix-free*.

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 $\underline{xz} = \underline{y}$.

So, f is prefix-free if

$$m' \neq m'' \implies f(m') \not\prec f(m''),$$

where "$\hookleftarrow$" is the "is prefix of" relation. If f is a prefix-free code (or a prefix code for short), f^* is invertible. We denote with $|f(m)|$ the length of the codeword assigned to m by f, *i.e.*, $|f(m)| = l$ if and only if $f(m) \in \{0,1\}^l$.

Proposition 2 tells us that lots of short codewords imply that the set of messages is small.

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

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

Proof. Let $\underline{x}, \underline{v} \in \{0,1\}^*$. We define $Y_L(\underline{x})$ to be the set of all extension strings of $\underline{x}$ of length L, *i.e.*,

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

Notice that if $L < |\underline{x}|$, then $Y_L(\underline{x}) = \emptyset$.

Now, it can be that $Y_L(\underline{x}) \cap Y_L(\underline{v}) \neq \emptyset$, or that $Y_L(\underline{x}) \cap Y_L(\underline{v}) = \emptyset$, or maybe that $Y_L(\underline{x}) \subset Y_L(\underline{v})$ or the other way around.

$$\begin{aligned} \underline{x} \prec\!\!\cdot\, \underline{v} &\implies Y_L(\underline{x}) \supseteq Y_L(\underline{v}), \\ \underline{x} \not\prec\!\!\cdot\, \underline{v} \wedge \underline{v} \not\prec\!\!\cdot\, \underline{x} &\implies Y_L(\underline{x}) \cap Y_L(\underline{v}) = \emptyset. \end{aligned}$$

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 the sets

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

are all non-empty.

For this reason, if f is a prefix code, then for any $m' \neq m''$ we have that

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

Let $L \geq \max_{m \in \mathcal{M}} |f(m)|$. Then it must be that

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

Since $\left|\{0,1\}^L\right| = 2^L$, and since the sets $Y_L(f(m'))$ and $Y_L(f(m''))$ are pairwise disjoint for any two distinct m', m'', we can write

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

Now our thesis follows, *i.e.*,

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

Proposition 3 (Prefix codes and entropy)**.** *If f is a prefix code then, for any distribution $P|\mathcal{M}$,*

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

Proof. The log sum inequality (proposition 1) gives us that

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

with equality if and only if $P(m) = 2^{-|f(m)|}$. Then:

$$\begin{aligned} \sum_{m \in \mathcal{M}} P(m) \log_2 \left(\frac{P(m)}{2^{-|f(m)|}}\right) &= \sum_{m \in \mathcal{M}} P(m) \log_2 (P(m)) \\ &\quad - \sum_{m \in \mathcal{M}} P(m) \log_2 \left(2^{-|f(m)|}\right) \\ &= -H(P) + \sum_{m \in \mathcal{M}} P(m) \cdot |f(m)| \\ &\geq 0 \\ &\quad \implies \\ H(P) &\leq \sum_{m \in \mathcal{M}} P(m) \cdot |f(m)|. \end{aligned}$$

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

Observation 5 (Upper bound to the entropy of a distribution)**.** *Let $P|\mathcal{M}$ be a probability distribution over $\mathcal{M}$, then it holds that $H(P) \leq \log_2(|\mathcal{M}|)$, with equality if and only if P is the equidistribution.*

Proof. Again, the log sum inequality (proposition 2) gives us that

$$\sum_{m\in\mathcal{M}} P(m)\log_2\left(\frac{P(m)}{\frac{1}{|\mathcal{M}|}}\right) \geq 0,$$

with equality if and only if $P(m) = \frac{1}{|\mathcal{M}|}$ for any m.

$$\begin{aligned}\sum_{m\in\mathcal{M}} P(m)\log_2\left(\frac{P(m)}{\frac{1}{|\mathcal{M}|}}\right) &= \sum_{m\in\mathcal{M}} P(m)\log_2(P(m)) \\ &\quad - \sum_{m\in\mathcal{M}} P(m)\log_2\left(\frac{1}{|\mathcal{M}|}\right) \\ &= -H(P) + \log_2(|\mathcal{M}|).\end{aligned}$$

□

Theorem 4 (Kraft theorem)**.** *Let $l : \mathcal{M} \to \mathbb{N}$ be a prescribed codeword length, i.e., $l(m)$ is the length of the codeword we want to assign to m. If l satisfies Kraft's inequality (proposition 2), then $\exists f : \mathcal{M} \to \{0,1\}^*$ prefix code such that $|f(m)| = l(m)$ for all m.*

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

The algorithm works as follows:

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

2. After constructing strings $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{t-1}$, we choose $\hat{\underline{x}}^{(t)}$ from $\{0,1\}^L \setminus \left(Y_L(\underline{x}_1) \cup \cdots \cup Y_L(\underline{x}_{t-1})\right)$. Then we set $f(m_t)$ to the prefix $\underline{x}_t$ of length $l(m_t)$ of $\hat{\underline{x}}^{(t)}$.

We have to prove that the algorithm ends giving to each string in $\mathcal{M}$ an image, and that it builds a prefix code.

Suppose that the algorithm stops at step t. Then it means $\{0,1\}^L = \bigcup_{i=1}^{t-1} Y_L(\underline{x}_i)$. 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, since we have at least t messages.

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

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

Since two sets $Y_L(\underline{x})$ and $Y_L(\underline{v})$ for distinct $\underline{x}, \underline{v}$ cannot be in general position, we just have to show that they do not contain one another.

Note that, for $i < t$,

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

since $l(m_i) \leq l(m_t)$ means that $|Y_L(\underline{x}_i)| \geq |Y_L(\underline{x}_t)|$. The sets get smaller and smaller.

On the other hand, for $i < t$,

$$Y_L(\underline{x}_t) \not\subset Y_L(\underline{x}_i).$$

Recall that $\hat{\underline{x}}^{(t)}$ was chosen in such a way that $\hat{\underline{x}}^{(t)} \in Y_L(\underline{x}_t)$, and that $\hat{\underline{x}}^{(t)} \notin Y_L(\underline{x}_i)$ for all $i < t$. So $Y_L(\underline{x}_t)$ has an element not in $Y_L(\underline{x}_i)$ for all $i < t$, so it can't be included in any of them. □

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

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

then $\exists \lambda \geq L, \lambda \in \mathbb{N}$ such that

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

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 cannot add more codewords (and still have a prefix-code).

Proposition 4 (Prefix codes and entropy +1)**.** *For all $P|\mathcal{M}$ probability distributions over $\mathcal{M}$, $\exists f : \mathcal{M} \to \{0,1\}^*$ prefix code such that*

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

Proof. We'll give a prescription satisfying both Kraft's inequality and this inequality. Recall that entropy is defined as

$$H(P) = \sum_{m \in \mathcal{M}} P(m) \log_2 \left(\frac{1}{P(m)} \right).$$

We could choose $l(m) = \frac{1}{P(m)}$, but this is not always an integer, and we could not build a prescription satisfying Kraft's inequality this way. Thus, we choose

$$l(m) = \left\lceil \log_2 \left(\frac{1}{P(m)} \right) \right\rceil.$$

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

$$\begin{aligned}\sum_{m \in \mathcal{M}} P(m) \left\lceil \log_2 \left(\frac{1}{P(m)} \right) \right\rceil &< \sum_{m \in \mathcal{M}} P(m) \log_2 \left(\frac{1}{P(m)} \right) \\ &\quad + \sum_{m \in \mathcal{M}} P(m) \\ &= H(P) + 1.\end{aligned}$$

We now check that l satisfies Kraft's inequality.

$$\begin{aligned}\sum_{m \in \mathcal{M}} 2^{-l(m)} &= \sum_{m \in \mathcal{M}} 2^{-\left\lceil \log_2 \left(\frac{1}{P(m)} \right) \right\rceil} \\ &\leq \sum_{m \in \mathcal{M}} 2^{-\log_2 \left(\frac{1}{P(m)} \right)} && \text{(since } \lceil t \rceil \geq t\text{)} \\ &= \sum_{m \in \mathcal{M}} 2^{\log_2 (P(m))} \\ &= \sum_{m \in \mathcal{M}} P(m) = 1. && \square\end{aligned}$$

Chapter 5

Entropy of Random Variables

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

5.1 Joint Entropy, Conditional Entropy and Mutual Information

Given a RV 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 RVs? If $X = Y$ then $H(X,Y) = H(X)$.

Definition 6 (Joint Entropy)**.** *In general, the entropy of a pair is the entropy of the RV 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 RV. $H(X,Y)$ is defined a the joint entropy *of X and Y.*

Proposition 5 (Upper bound to the entropy of a distribution)**.** *Let $P|\mathcal{X}$, then*

$$H(P) \leq \log_2(|\mathcal{X}|),$$

with equality if and only if P is the equidistribution, i.e., $P(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$.

Proof. Call $\mathcal{U}$ the uniform distribution. $D(P||\mathcal{U}) \geq 0$, with equality if and only if $P = \mathcal{U}$.

$$\begin{aligned} D(P||\mathcal{U}) &= \sum_{x\in\mathcal{X}} P(x)\log_2\left(\frac{P(x)}{\frac{1}{|\mathcal{X}|}}\right) \\ &= \sum_{x\in\mathcal{X}} P(x)\log_2(P(x)) + \sum_{x\in\mathcal{X}} P(x)\log_2(|\mathcal{X}|) \\ &= -H(P) + \log_2(|\mathcal{X}|). \end{aligned}$$

This quantity is non-negative with equality to zero if and only if $P = \mathcal{U}$. □

We write $X \in \mathcal{X}$ since X is like an unknown element of $\mathcal{X}$. RVs don't always have an expected value. We can think of $H(X)$ as the information content of X. If we consider entropy as the amount of information in a RV, 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 (Lower bound to joint entropy)**.**

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

Proof. Consider the following quantities:

$$H(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr[X = x, Y = y] \log_2 \left(\frac{1}{\Pr[X = x, Y = y]} \right),$$

and

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} \Pr[X = x] \log_2 \left(\frac{1}{\Pr[X = x]} \right) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr[X = x, Y = y] \log_2 \left(\frac{1}{\Pr[X = x]} \right). \end{aligned}$$

We take the difference between the two:

$$H(X,Y) - H(X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr[X = x, Y = y] \log_2 \left(\frac{\Pr[X = x]}{\Pr[X = x, Y = y]} \right).$$

The definition of conditional probability gives us that $\Pr[X = x, Y = y] = \Pr[X = x] \cdot \Pr[Y = y | X = x]$, and as such

$$\begin{aligned} &\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr[X = x] \cdot \Pr[Y = y | X = x] \log_2 \left(\frac{1}{\Pr[Y = y | X = x]} \right) \\ &= \sum_{x \in \mathcal{X}} \Pr[X = x] \sum_{y \in \mathcal{Y}} \Pr[Y = y | X = x] \log_2 \left(\frac{1}{\Pr[Y = y | X = x]} \right). \end{aligned}$$

We would like to say that this quantity is non-negative. Since $H(\cdot)$ is non-negative, this difference is actually greater than (or equal to) zero. □

Definition 7 (Conditional Entropy)**.** *We call*

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

the conditional entropy *of* Y *given* X*.*

$H(X,Y) - H(X)$ is the convex combination of the entropies of the conditional distribution of Y given the various values of X. It's like the expected value of the entropy of the conditional distribution $\Pr[Y = y|X = x]$. It can be seen as the residual information when X is known.

Proposition 7 (Upper bound to conditional entropy)**.**

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

Proof.

$$\begin{aligned} H(Y) - H(Y|X) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr[X = x, Y = y] \log_2\left(\frac{1}{\Pr[Y = y]}\right) \\ &\quad - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr[X = x, Y = y] \log_2\left(\frac{\Pr[X = x]}{\Pr[X = x, Y = y]}\right) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr[X = x, Y = y] \log_2\left(\frac{\Pr[X = x, Y = y]}{\Pr[X = x]\cdot\Pr[Y = y]}\right). \end{aligned}$$

This is symmetrical in X and Y. Notice that $H(Y) - H(Y|X)$ is not. 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 independence of X and Y. □

Definition 8 (Mutual Information)**.** *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 non-negative.

Notice that

$$\begin{aligned} I(X \wedge Y) &= H(Y) - H(Y|X) \\ &= H(Y) + H(X) - H(X,Y) \end{aligned}$$

thus

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

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

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

Proposition 8 (Chain Rule)**.** *The* chain rule *states that*

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

We have said that $I(X \wedge Y) \sim \mu(A \cap B)$, and since $H(Y|X,Z) \leq H(Y|Z)$ we can state that

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

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.

$$\begin{aligned} I(X \wedge Y) - I(X \wedge Y|Z) &\sim \mu(A \cap B) - \mu((A \cap B) \setminus C) \\ &= \mu((A \cap B) \setminus ((A \cap B) \setminus C)) \\ &= \mu(A \cap B \cap C). \end{aligned}$$

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

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

If $X \equiv Y \equiv Z$ then

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

so eq. (5.1) is possible. But we can also have it the other way around:

$$I(X \wedge Y|Z) - I(X \wedge Y) \geq 0.$$

It follows that the inequality of eq. (5.1) does not hold always.

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 if and only if they are pairwise disjoint, but RV independence does not satisfy this property. The analogy fails on assuming that RVs independence is similar to set disjointness. n-way independence is binary, but n-way independence is unrelated to m-way independence if $n \neq m$.

5.2 Information Source and Speed of Information

An *information source* is an infinite sequence $X_1, X_2, \ldots, X_n, \ldots$ of RVs 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^∞, and with $X^n = X_1, \ldots, X_n$ a sequence of n RVs.

The "speed" of information from a sequence of RVs is

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

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

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

Consider $\{X_i\}_{i=1}^{\infty}$, an infinite sequence of independent identically distributed (i.i.d.) RVs, and denote by P the common distribution of $\{X_i\}_{i=1}^{\infty}$, so that $H(X_i) = H(P)$ for all i.

In this case,

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

so

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

What is the sufficient condition for this to hold (*i.e.*, for the limit to exists)? We need stationary RVs: impredictable but "stationary".

Definition 9 (Stationary Information Source)**.** *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 entropy of a sequence

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

decreases.

Definition 10 (Memoryless Information Source)**.** *We call an information source* memoryless *if* X_i*, for all* 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 RVs, does its entropy rate exists?

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

When does this quantity diverge? Take $H(X_i) \in \{0, 1\}$, what we want to know is if it exists the following limit, for $\varepsilon_i \in \{0, 1\}$:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{\varepsilon_i}{n}.$$

Consider the following information source:

$$\overbrace{0101\ldots0101}^{n \text{ bits}}\overbrace{0101\ldots0101}^{n \text{ bits}}\overbrace{1111\ldots1111}^{2n \text{ bits}}.$$

The first $2n$ bits, a sequence of alternating bits, have average $\frac{1}{2}$. Add the second $2n$ and you get $\frac{3}{4}$. Repeat this and the sequence oscillates between $\frac{3}{4}$ and $\frac{1}{2}$.

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

Proof. It's reasonable to think that

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

goes to 0. What we are going to prove is that this is in fact true, and thus the information rate of a stationary information source exists (and it is 0).

To prove it, we just have to show that the sequence is decreasing, since $H(\cdot)$ is always positive, *i.e.*, we show that

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

which we could also write as

$$(n-1) \cdot H(X^n) \leq n \cdot H\left(X^{n-1}\right),$$

that is equal to saying that

$$(n-1) \cdot \left[H\left(X^{n-1}\right) + H\left(X_n | X^{n-1}\right)\right] \leq n \cdot H\left(X^{n-1}\right).$$

Thus,

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

We apply the definition of joint entropy (proposition 8):

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

Since the source is stationary, we have that $H(X_1) = H(X_2)$, and that $H(X_2|X_1) = H(X_n|X_{n-1})$, and so on. In general, $H(X_i|X_1, \ldots, X_{i-1}) = H(X_n|X_{1+n-i}, \ldots, X_{n-1})$. Thus our thesis becomes

$$(n-1) \cdot H\left(X_n | X^{n-1}\right) \leq H(X_n) + \sum_{i=2}^{n-1} H(X_n | X_{1+n-i}, \ldots, X_{n-1}).$$

This is a weaker statement than the following $n-1$ statements:

$$\begin{aligned}
H\left(X_n | X^{n-1}\right) &\leq H(X_n), \\
&\vdots \\
H\left(X_n | X^{n-1}\right) &\leq H(X_n | X_{1+n-i}, \ldots, X_{n-1}), \\
&\vdots \\
H\left(X_n | X^{n-1}\right) &\leq H\left(X_n | X^{n-1}\right).
\end{aligned}$$

We have coupled the $n-1$ terms in the Left Hand Side (LHS) with one of the $n-1$ terms in the Right Hand Side (RHS).

Look at the generic term: for all k,

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

By definition of mutual information, we have that

$$\begin{aligned}
&H(X_n | X_{n-k}, \ldots, X_{n-1}) - H(X_n | X_1, \ldots, X_{n-1}) = \\
&\quad I(X_n \wedge X_1 \wedge \cdots \wedge X_{n-k-1} | X_{n-k}, \ldots, X_{n-1}) \geq 0.
\end{aligned}$$ □

This reminds us that

$$I(A \wedge B | C) \geq 0 \implies H(A|C) \geq H(A|B, C).$$

5.3 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 a stationary information source $\{X_i\}_{i=1}^{\infty}$. Given f, a variable length prefix code, we consider $|f(X_i)|$ as a RV. We can talk about the expected value $E(|f(X_i)|)$ of this RV. Since $E(|f(X_i)|) = \sum_{x\in\mathcal{X}} \Pr[X_i = x]\,|f(x)|$, by proposition 3 and proposition 4 we know that

$$H(X_i) \leq E(|f(X_i)|) \leq H(X_i) + 1. \tag{5.2}$$

Note that the length of the string obtained by applying f to n RVs is

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

and that, by eq. (5.2) above,

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

If the source is stationary all RVs have the same probability distribution P, and thus

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

but this does not tell us much.

Instead, we join RVs together. Let f_n be an optimal (in the sense of length of output) prefix code for $X_1, \ldots, X_n$.

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

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

Chapter 6

Error Correcting Codes

Error Correcting Codes (ECCs) 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 radius 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. The first property that we want from $\mathcal{C}$ is that $|\mathcal{C}|$ should be large.

Next, we extend the definition of Hamming Distance to sets, *i.e.*, we define

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

where

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

So $d_H(\mathcal{C})$ tells how close the closest elements of $\mathcal{C}$ are. $d_H(\mathcal{C})$ should be large too.

What is the best possible trade-off?

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

This says that we are concentrating on codes for which $d_H(\mathcal{C}) \geq d$ for some d.

We can think of the graph for which $V = \{0,1\}^n$ and $\underline{x}, \underline{y} \in \binom{V}{2}$ share an edge if and only if $d_H\left(\underline{x}, \underline{y}\right) \geq d$. $M(n,d)$ is the size of the largest clique in this graph.

We try to make this simpler by looking at the problem from an asymptotic point of view. We consider $M(n, n\delta)$ for $\delta \in [0,1]$. This grows to infinity exponentially in n, while

$$\frac{1}{n} \log_2 (M(n, n\delta))$$

does not grow exponentially. So we look at the superior limit of

$$R(\delta) = \limsup_{n \to \infty} \frac{1}{n} \log_2 (M(n, n\delta)).$$

R here stands for *rate*.

$M(n, n\delta)$ is the size of largest set of strings one can put inside n bits of memory, with minimum distance $n\delta$. Its logarithm is the number of bits of true information that we can store. Dividing by n we get the amount of information bit by bit. We can easily see that $R(0) = 1$ and that $R(1) = 0$.

Theorem 6 (Gilbert-Varshamov bound)**.** *If* $\delta \in \left[0, \frac{1}{2}\right]$*, then*

$$R(\delta) \geq 1 - h(\delta).$$

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 later improved to $n2^{n\,[1-h(\delta)]}$.

For now, fix $n, d \in \mathbb{N}$. Take an arbitrary string $\underline{x} \in \{0,1\}^n$, and exclude $B_H(\underline{x}, d-1)$ from our future choices.

Then, after having found $\underline{x}_1, \ldots, \underline{x}_{t-1}$, we choose arbitrarily

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

so, after picking each string, we exclude the Hamming Ball of radius $d-1$ around it from our future choices.

How long can this go on? Maybe after m steps we have that

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

How large m can be? We stop when

$$\begin{aligned}
\{0,1\}^n &\subseteq \bigcup_{i=1}^{m} B_H(\underline{x}_i, d-1) \\
&\Longrightarrow \\
2^n &\leq \left| \bigcup_{i=1}^{m} B_H(\underline{x}_i, d-1) \right| \\
&\leq \sum_{i=1}^{m} |B_H(\underline{x}_i, d-1)| \\
&\leq \sum_{i=1}^{m} |B_H(\underline{x}_i, d)| \\
&\leq m 2^{n\,h\left(\frac{d}{n}\right)},
\end{aligned}$$

where the last inequality holds since $d \leq \frac{n}{2}$. We have proven the thesis, since now

$$M(n,d) \geq m \geq \frac{2^n}{2^{n\,h\left(\frac{n}{d}\right)}} = 2^{n\left[1-h\left(\frac{n}{d}\right)\right]},$$

where $\frac{n}{d} = \delta$. □

Theorem 7 (Hamming bound)**.** *For* $\delta \in [0,1]$*,*

$$R(\delta) \leq 1 - h\left(\frac{\delta}{2}\right).$$

Proof. Fix $n, d \in \mathbb{N}$, $d_H(\mathcal{C}) \geq d$, arbitrarily.

Consider the centres $\underline{x}$ and $\underline{y}$ of two Hamming balls, with $d_H(\underline{x}, \underline{y}) = d$. For the two Hamming balls to be disjoint, we must choose

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

Assume the two Hamming Balls are not disjoint, *i.e.*,

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

but this means that

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

and that, since $\underline{z}$ is in both Hamming Balls,

$$d_H(\underline{x}, \underline{z}) \leq \frac{d-1}{2} \qquad d_H(\underline{y}, \underline{z}) \leq \frac{d-1}{2}.$$

But then we have that

$$d = d_H(\underline{x}, \underline{y}) \leq d_H(\underline{x}, \underline{z}) + d_H(\underline{y}, \underline{z}) \leq 2 \cdot \frac{d-1}{2} = d - 1 < d,$$

in contradiction with the assumption that $d_H(\underline{x}, \underline{y}) = d$.

If all strings have a distance of at least d, we can correct up to $\frac{d-1}{2}$ errors.

Assume $\mathcal{C}$ is built using disjoint balls, and that $M(n,d) = |\mathcal{C}|$. Then

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

Now, since the Hamming Balls are disjoint, we have that

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

which in turn gives us that $M(n,d) = |\mathcal{C}|$ can be upper bounded as

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

We now take the logarithm and normalise by n, and obtain

$$\frac{1}{n}\log_2(M(n,d)) \leq \frac{1}{n}\log_2(n+1) + 1 - h\left(\frac{d-1}{2n}\right),$$

and, by taking the limit superior of that quantity, we get

$$\begin{aligned} R(\delta) &= \limsup_{n\to\infty} \frac{1}{n} \log_2 \left(M(n, n\delta)\right) \\ &= \limsup_{n\to\infty} \cancel{\frac{1}{n} \log_2 (n+1)}^{\to 0} + 1 - h\left(\frac{\cancel{n}\delta}{2\cancel{n}} - \cancel{\frac{1}{2n}}^{\to 0}\right), \\ &\leq 1 - h\left(\frac{\delta}{2}\right). \end{aligned}$$

□

6.1 Uniquely Decodable codes

Definition 11 (Uniquely Decodable code)**.** *Let $f : \mathcal{M} \to \{0,1\}^\star$, and define $f^\star : \mathcal{M}^\star \to \{0,1\}^\star$ as*

$$f^\star(\underline{m}) = f(m_1)f(m_2)\dots f(m_t),$$

for $\underline{m} \in \mathcal{M}^\star$ with $\underline{m} = m_1 m_2 \dots m_t$ for some t. We say that f is Uniquely Decodable (UD) if $f^\star$ is injective.

If f is a prefix code, f is UD.

Theorem 8 (Kraft - McMillan theorem on Uniquely Decodable codes)**.** *If f is UD then Kraft's inequality holds, i.e.,*

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

Proof. Let

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

We would like to prove that $q \leq 1$. To do so, we consider q^n, which will involve the length of concatenations of code words. We will show that q^n "grows slowly", *i.e.*, it does not grow exponentially, and thus is less than or equal to 1.

$$\begin{aligned} q^n &= \left[\sum_{m\in\mathcal{M}} 2^{-|f(m)|}\right]^n \\ &= \prod_{i=1}^{n} \left[\sum_{m\in\mathcal{M}} 2^{-|f(m)|}\right] \\ &= \sum_{\underline{m}\in\mathcal{M}^n} \left[\prod_{i=1}^{n} 2^{-|f(m_i)|}\right] \\ &= \sum_{\underline{m}\in\mathcal{M}^n} 2^{-|f^\star(\underline{m})|}. \end{aligned}$$

$f^\star(\underline{m})$ is just a binary string. We can break up the summation over the length of these strings, as follows:

$$\sum_{\underline{m}\in\mathcal{M}^n} 2^{-|f^\star(\underline{m})|} = \sum_{t=n}^{nL} \sum_{\substack{\underline{m}\in\mathcal{M}^n: \\ |f^\star(\underline{m})|=t}} 2^{-|f^\star(\underline{m})|},$$

with $L = \max_{m \in \mathcal{M}} |f(m)|$. Now we use the fact that $f^\star(\cdot)$ is injective: each binary string of length t appears just once in the sum. We can't have two strings of messages encoded by the same binary string.

$$\sum_{t=n}^{nL} \sum_{\substack{\underline{m} \in \mathcal{M}^n: \\ |f^\star(\underline{m})|=t}} 2^{-|f^\star(\underline{m})|} \leq \sum_{t=n}^{nL} 2^t \cdot 2^{-t} = nL.$$

This follows from the fact that we can have at most 2^t messages encoded by binary strings of length t.

What we have shown is that $q^n \leq nL$, and thus

$$q \leq \sqrt[n]{nL} = \sqrt[n]{n}\sqrt[n]{L} \to 1,$$

and so $q \leq 1$. □

Theorem 9 (Plotkin bound)**.**

$$\delta \geq \frac{1}{2} \implies R(\delta) = 0.$$

We already know that, if $\delta \leq \frac{1}{2}$,

$$1 - h(\delta) \leq R(\delta) \leq 1 - h\left(\frac{\delta}{2}\right).$$

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

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

where $\underline{x} = x_1 \dots x_n$.

Now, picture a matrix with n columns and $M = |\mathcal{C}|$ rows. Each row is an element of $\mathcal{C}$, each column is a coordinate of $\underline{x} \in \mathcal{C}$. One can think of $\sum_{\{\underline{x},\underline{y}\} \in \binom{\mathcal{C}}{2}} d_H(x_i, y_i)$ as the number of 1s times the number of 0s in the i-th column of the matrix: we add 1 to the summation each time two strings $\underline{x}$ and $\underline{y}$ differ on column i. Thus, calling M_i the number of 1s in column i, we have

$$\begin{aligned} \frac{2}{M(M-1)} \sum_{i=1}^{n} \sum_{\{\underline{x},\underline{y}\} \in \binom{\mathcal{C}}{2}} d_H(x_i, y_i) &= \frac{2}{M(M-1)} \sum_{i=1}^{n} M_i \cdot (M - M_i) \\ &\leq \frac{2}{M(M-1)} \left(\frac{M}{2}\right)^2 \\ &= \frac{nM}{2(M-1)}. \end{aligned}$$

So the minimum distance d is

$$d = d_H(\mathcal{C}) \leq \frac{nM}{2(M-1)}. \tag{6.1}$$

To complete the proof we just need some manipulation of eq. (6.1)

$$\begin{aligned}
&2(M-1)d \leq nM \\
&\Longrightarrow \\
&2Md - 2d \leq nM \\
&\Longrightarrow \\
&M(2d-n) \leq 2d \\
&\Longrightarrow \\
&M(2n\delta - n) \leq 2n\delta && \text{(since } d = n\delta\text{)} \\
&\Longrightarrow \\
&M(2\delta - 1) \leq 2\delta && \text{(if } \delta > \tfrac{1}{2} \text{ we can divide)} \\
&\Longrightarrow \\
&M \leq \frac{2\delta}{2\delta - 1}.
\end{aligned}$$

So, when $\delta \geq \frac{1}{2}$, we have that $M(n, n\delta)$ is a constant. Thus $R(\delta) = 0$, since it is defined as the limit superior of $\frac{1}{n}M(n, n\delta)$. □

6.2 Parity Check Codes

Consider the set

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

that is, the set of binary strings which have an even number of 1s. $|\mathcal{C}_n| = 2^{n-1}$. This set will not help us 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 $GF(2)$, the Galois field over $\{0,1\}$.

Define $\langle \underline{x}, \underline{y} \rangle$ to be the scalar product between $\underline{x}$ and $\underline{y}$, *i.e.*,

$$\langle \underline{x}, \underline{y} \rangle = \sum_{i=1}^{n} x_i \cdot y_i \mod 2.$$

$\mathcal{C}_n$ can be defined also as

$$\mathcal{C}_n = \{ \underline{x} : \langle \underline{x}, \underline{1} \rangle = 0 \}.$$

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

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

Consider the set $S \subseteq [n]$ of indices of $\underline{s}$ which are 1. If we take $\mathcal{C}_n(\underline{s})$, we are only looking for errors on x_i for $i \in S$. $\mathcal{C}_n(\underline{s})$ is a linearly closed space, with addition $\oplus \mod 2$. Furthermore, these spaces are closed under intersection.

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 we can construct codes that not only detect errors, but also correct them.

A set $\mathcal{L} \subseteq \{0,1\}^n$ 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 its orthogonal complement

$$\mathcal{L}^\perp = \{\underline{z} : \underline{z} \in \{0,1\}^n, \forall \underline{x} \in \mathcal{L}.\, \langle \underline{z}, \underline{x} \rangle = 0\} = \bigcap_{\underline{x}\in\mathcal{L}} \mathcal{C}_n(\underline{x}).$$

Consider a basis of $\mathcal{L}^\perp$. Write the vectors from this basis as column vectors of a matrix A with n rows and m columns.

For all $\underline{x} \in \mathcal{L}$, we have that $\underline{x} \times A = \underline{0}_m$. So $\mathcal{L} = \{\underline{x} : \underline{x} \times A = \underline{0}_m\}$. A linear code can be specified by the matrix A, which is called the parity check matrix of $\mathcal{L}$. Given a matrix A we have that

$$\begin{aligned} \ker A &= \{\underline{x} : \underline{x} \times A = \underline{0}_m\}, \\ \operatorname{Im} A &= \{\underline{z} : \underline{z} \in \{0,1\}^m, \exists \underline{x} \in \{0,1\}^n . \underline{x} \times A = \underline{z}\}. \end{aligned}$$

So the set $\operatorname{Im} A$ is the linear combination of the rows of A. Consider the i-th canonical vector $e_i \in \{0,1\}^m$, we have that

$$e_i \times A = A_i^T$$

that is, the i-th row of A.

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

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

Observation 6 (Hamming distance of a linear code)**.** *If $\mathcal{L} \subseteq \{0,1\}^n$ 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 $\underline{x} \in \mathcal{L}$ different from $\underline{0}$ and such $\underline{x}$ minimises $w_H(\underline{x})$. Then

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

Now we prove that

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

To prove this we rely on the intuitive idea that distance is translation invariant. Note that for any $\underline{z} \in \mathcal{L}$ we have that

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

since $\mathcal{L}$ is linearly closed.

Consider any $\underline{x}$ and $\underline{y}$. Since $\underline{x} \oplus \underline{y} \in \mathcal{L}$, we have that

$$d_H\left(\underline{x}, \underline{y}\right) = d_H\left(\underline{0}, \underline{x} \oplus \underline{y}\right) = w_H\left(\underline{x} \oplus \underline{y}\right) \geq \min_{\underline{z} \in \mathcal{L}: \underline{z} \neq \underline{0}} w_H\left(\underline{z}\right). \qquad \square$$

$M(n, 3)$ is the largest cardinality of a code correcting 1 error. In the proof of the Hamming bound (theorem 7) we have seen that we can build such a code using disjoint balls of radius 1.

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

with equality if the Hamming Balls cover the entire space. So we have that

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

Theorem 10 (Hamming theorem on Error Correcting Codes for 1 error)**.** *If* $\exists m$ *such that* $n = 2^m - 1$*, then*

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

Proof. 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). The code we are looking for is $\mathcal{C} = \ker A$.

We have to prove that $d_H\left(\mathcal{C}\right) \geq 3$, and that consequently

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

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

- If $\underline{x} \in \mathcal{C}$ and $\underline{x} \neq \underline{0}$, then $w_H\left(\underline{x}\right) \neq 1$. This is true since if $w_H\left(\underline{x}\right) = 1$ then $\underline{x} = e_i$ for some i, and $e_i \times A$ is the i-th row of A, which is different from $\underline{0}_m$;
- if $w_H\left(\underline{x}\right) = 2$ then $\underline{x} \notin \mathcal{C}$. If $w_H\left(\underline{x}\right) = 2$ then $\underline{x} = e_i \oplus e_j$ for two distinct i, j. But then $\underline{x} \times A$ is a linear combination of two rows of A, and since all rows of A are different $e_i \times A \oplus e_j \times A \neq \underline{0}_m$.

What is left to show is that the Hamming Balls of radius 1 around the elements of $\mathcal{C}$ fill up $\{0,1\}^n$, *i.e.*,

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

- If $\underline{z} \in \mathcal{C}$, we are ok;
- if $\underline{z} \notin \mathcal{C}$, then $\underline{z} \times A \neq \underline{0}_m$. We have that $\underline{z} \times A \in \{0,1\}^m$, and since $\underline{z} \times A \neq \underline{0}_m$, it must be that it is a row of A. It follows that $\underline{z} \times A = e_i \times A$ for some i. So $(\underline{z} \oplus e_i) \times A = \underline{0}_m$, thus $\underline{z} \oplus e_i \underline{x} \in \ker A$. So we have found that $\underline{z} \in B_H\left(\underline{x}, 1\right)$. $\square$

Chapter 7

The communication model

Shannon developed a mathematical model of the communication among parties. In particular, we will 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*.

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, and we could think of a communication that evolves over time at the same place, like data storage and retrieval.

The noisy channel is both the vehicle 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 trade-off between data integrity and fast communication.

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 cannot 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.

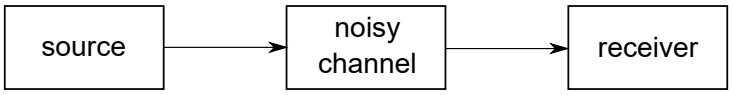

Figure 7.1: Graphic model of a simple communication.

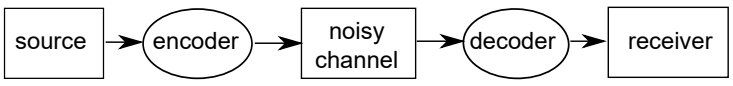

Figure 7.2: The role of encoder and decoder.

7.1 Discrete Memoryless Channel

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. 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 cannot launch any dice he likes; 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 we 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 $x_1, x_2, \ldots, x_n \in \mathcal{X}^n$ and launch a sequence of dices. There is a probability distribution for $y_1, y_2, \ldots, y_n \in Y^n$. Consider the function

$$W^n : \mathcal{X}^n \to \mathcal{Y}^n,$$

defined as

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

We assume that every symbol is independently modified (lack of memory of the channel). Our Discrete Memoryless Channel (DMC) (over time instants) is defined by the set $\{W^n\}_{n=1}^{\infty}$.

7.2 Shannon's Noisy Channel Theorem

The encoder takes the messages of the source and maps it to an arbitrary subset $\mathcal{C} \subseteq \mathcal{X}^n$. So we have $|\mathcal{C}|$ many messages that can 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 $\log_2(|\mathcal{C}|)$. We define the speed of the channel as

$$\frac{1}{n} \cdot \log_2\left(|\mathcal{C}|\right),$$

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 $\varphi : \mathcal{Y}^n \to \mathcal{C}$. An error event happens when $\underline{x} \in \mathcal{X}^n$ is transmitted and $\varphi(\underline{y}) \neq \underline{x}$ is received. Define

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

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

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

where

$$\varphi^{-1}(\underline{x}) = \left\{\underline{y} : \varphi(\underline{y}) = \underline{x}\right\}.$$

Define the (maximum) error probability of the code $(\mathcal{C}_n, \varphi_n)$ (W^n is fixed) as

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

Proposition 9 (Achievable rate)**.** *$R \geq 0$ is an* achievable rate *of transmission over the DMC $\{W\}$ if exists $\{(\mathcal{C}_n, \varphi_n)\}_{n=1}^{\infty}$ such that:*

- $\lim_{n\to\infty} \frac{1}{n} \log_2\left(|\mathcal{C}_n|\right) \geq R$;
- $\lim_{n\to\infty} e_m(W, \mathcal{C}_n, \varphi_n) = 0$.

What is the highest achievable rate?

$$\mathcal{C} \subseteq \mathcal{X}^n \implies \frac{1}{n}\log_2\left(|C|\right) \leq \log_2\left(|\mathcal{X}^n|\right).$$

So the maximum achievable rate is $\log_2\left(|\mathcal{X}^n|\right)$. Also, if R_t is a series of achievable rates and R_t tends to R, 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 co-workers.

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 a RV $Y \in \mathcal{Y}$ representing 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

$$\Pr\left[X = x, Y = y\right] = P(x)W(y|x).$$

Define $I\left(X \wedge Y\right)$ 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 $P|\mathcal{X}$. With some abuse of notation:

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

What is the formula for I?

$$\begin{aligned} I(X \wedge Y) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \Pr[X=x, Y=y] \cdot \log_2 \left(\frac{\Pr[X=x, Y=y]}{\Pr[X=x]\cdot\Pr[Y=y]} \right) \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x)\cdot W(y|x) \cdot \log_2 \left(\frac{P(x)\cdot W(y|x)}{\sum_{z\in X} P(x)\cdot W(y|z)} \right) \\ &= I(P,W). \end{aligned}$$

We used the fact that $\Pr[Y=y] = \sum_{x\in\mathcal{X}} \Pr[Y=y|X=x] = \sum_{x\in\mathcal{X}} W(y|x)$.

Theorem 11 (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, *i.e.*, all rows in W are the same.

7.3 Binary Symmetric Channel

A Binary Symmetric Channel (BSC) is used to transmit binary strings and the DMC has the form

$$W = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix},$$

with $\mathcal{X} = \mathcal{Y} = \{0,1\}$.

It is called symmetric because W is symmetric. In this context $P|\mathcal{X}$ is called *crossover* probability. The capacity of this channel according to theorem 11 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 equation holds:

$$H(Y|X) = \sum_{x\in\mathcal{X}} \Pr[X=x] \cdot \underbrace{H(Y|X=x)}_{h(p)} = h(p),$$

where

$$\begin{aligned} H(Y|X=x) &= \sum_{y\in\mathcal{Y}} \overbrace{\Pr[Y=y|X=x]}^{W(y|x)} \cdot \log_2 \left(\frac{1}{\Pr[Y=y|X=x]} \right) \\ &= \sum_{y\in\mathcal{Y}} W(y|x) \cdot \log_2 \left(\frac{1}{W(y|x)} \right) \\ &= h(p). \end{aligned}$$

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.*, $\Pr[X = 0] = \frac{1}{2}$.

Observation 7 (Achievable rate in Binary Symmetric Channel)**.** $1 - h(2p)$ *is an achievable rate if* $p < \frac{1}{4}$.

We will do this using ECCs.

Proof. Choose arbitrary $\underline{x} \in \{0,1\}^n$, but think of $\underline{0}$ (it's easier). Consider the Hamming ball $B_H(\underline{0}, n(p+\varepsilon))$; we want that

$$W^n\left(\varphi_n^{-1}(\underline{0})|\underline{0}\right) \to 1.$$

Consider a code for which

$$\forall \underline{x} \in \mathcal{C}_n . \varphi_n^{-1}(\underline{x}) \supseteq B_H(\underline{x}, n(p+\varepsilon));$$

these Hamming balls should be disjoint. We wonder if it is true that

$$W^n\left(B_H(\underline{0}, n(p+\varepsilon))|\underline{0}\right) \to 1.$$

This is what we should achieve, and prove that np is a good choice. What we receive is $\underline{x} \oplus Z^n$, for some RV Z^n, defined as $Z^n = \underline{x} \oplus Y^n$.

$$Y^n \in B_H(\underline{0}, n(p+\varepsilon)) \iff Z^n \text{ has at most } n(p+\varepsilon) \text{ 1s.}$$

The number of 1s in Z^n is equal to $\sum_{i=1}^n Z_i$, where $Z_i \sim (1-p, p)$ is a RV.

$$E(Z_i) = \Pr[Z_i = 0] \cdot 0 + \Pr[Z_i = 1] \cdot 1 = p.$$

$\{Z_i\}_{i=1}^n$ is an i.i.d. sequence of RVs.

$$E(Z^n) = E\left(\sum_{i=1}^n Z_i\right) = \sum_{i=1}^n E(Z_i) = n \cdot p.$$

By the law of large numbers,

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n Z_i - p\right| > \varepsilon\right] \to 0.$$

The complement of this event has a probability that converges to 1. How can we guarantee that these Hamming balls of radius $n(p+\varepsilon)$ are disjoint? This is guaranteed if $d_H(\underline{x}', \underline{x}'') \geq 2n(p+\varepsilon)$. Thus $d_H(\mathcal{C}_n) \geq 2n(p+\varepsilon)$. The Gilbert-Varhsamov bound (theorem 6) says that, if $\delta < \frac{1}{2}$,

$$\exists \mathcal{C}_n : |\mathcal{C}_n| \gtrsim 2^{n(1-h(\delta))},$$

and that

$$d_H(\mathcal{C}_n) \gtrsim n\delta.$$

We want that $d_H(\mathcal{C}_n) \geq 2n(p+\varepsilon)$, thus what we want is achievable if and only if

$$2p + \varepsilon < \frac{1}{2} \sim p < \frac{1}{4}.$$

□

Shannon stated something stronger, namely that the capacity is $1 - h(p)$. Shannon uses "maximum likelihood". He chooses $2^{n\,R}$ strings at random from T_p^n, given $P|\mathcal{X}$.

Theorem 12 (Converse part of Shannon's Noisy Channel Theorem)**.** *If R is such that $\exists\{\mathcal{C}_n, \varphi_n\}_{n=1}^{\infty}$ with*

$$\limsup_{n\to\infty} \frac{1}{n} \cdot \log_2(|\mathcal{C}_n|) \geq R,$$

and

$$\lim_{n\to\infty} e_n(W^n, \mathcal{C}_n, \varphi) = 0,$$

then

$$R \leq \max_{\substack{X\in\mathcal{X}:\\ P_{Y|X}=W}} I(X \wedge Y).$$

Proof. Let M_n be a RV uniformly distributed over $\mathcal{C}_n$. Take $\frac{1}{n} \cdot H(M_n)$, the average entropy.

$$\frac{1}{n} \cdot H(M_n) = \frac{1}{n} \cdot \log_2(|\mathcal{C}_n|).$$

To get the upper bound, we manipulate the LHS.

$$\frac{1}{n} \cdot H(M_n) = \frac{1}{n} \cdot [H(M_n) - H(M_n|\varphi_n(Y^n))] + \frac{1}{n} \cdot H(M_n|\varphi_n(Y^n)),$$

where Y^n is the RV defined by $P_{Y^n|M_n} = W^n$. In the square brackets we have a mutual information, so

$$\frac{1}{n} \cdot H(M_n) = \frac{1}{n} \cdot I(M_n \wedge \varphi_n(Y^n)) + \frac{1}{n} \cdot H(M_n|\varphi_n(Y^n)).$$

We should be able to prove that this is not more than the upper bound. We expect the other therm to be negligible; with high probability the result of decoding the received codeword is the codeword that was sent.

$$I(M_n \wedge \varphi_n(Y^n)) \leq I(M_n \wedge Y^n). \tag{7.1}$$

This is from the fact that information cannot be gained by processing the RVs.

To prove this, first we look at

$$I(M_n \wedge \varphi_n(Y^n)) \leq I(M_n \wedge \varphi_n(Y^n), Y^n). \tag{7.2}$$

By definition of mutual information, eq. (7.2) can be written as

$$H(M_n) - H(M_n|\varphi_n(Y^n)) \leq H(M_n) - H(M_n|\varphi_n(Y^n), Y^n),$$

and since $H(A|B) \geq H(A|B, C)$ we have that

$$H(M_n|\varphi_n(Y^n)) \geq H(M_n|\varphi_n(Y^n), Y^n).$$

Then eq. (7.2) follows.

The RHS of eq. (7.2) can be written as

$$\begin{aligned} I\left(M_n \wedge \varphi_n(Y^n), Y^n\right) &= I\left(M_n \wedge Y^n\right) + \underbrace{I\left(M_n \wedge \varphi_n(Y^n) | Y^n\right)}_{0} \\ &= I\left(M_n \wedge Y^n\right). \end{aligned}$$

We used the fact that

$$I\left(M_n \wedge \varphi_n(Y^n) | Y^n\right) \leq H\left(\varphi_n(Y^n) | Y^n\right) = 0.$$

Combining these two things, we have obtained eq. (7.1).

What we wrote as M_n is just a vector of RVs, *i.e.*, something like $X_1 \dots X_n = X^n$. Thus,

$$I\left(M_n \wedge Y^n\right) = I\left(X^n \wedge Y^n\right).$$

This is what we want to upper bound with the conjectured value of capacity.

By definition of information gain, we write

$$I\left(X^n \wedge Y^n\right) = H\left(Y^n\right) - H\left(X^n | Y^n\right) \leq \sum_{i=1}^{n} H\left(Y_i\right) - H\left(Y^n | X^n\right).$$

Now we rewrite $H\left(Y^n | X^n\right)$ using the chain rule (proposition 8):

$$H\left(Y^n | X^n\right) = \sum_{i=1}^{n} H\left(Y_i | X^n, Y_1, \dots, Y_{i-1}\right).$$

Since the channel we are using is a DMC (and thus it is memoryless), we have that

$$H\left(Y_i | X^n, Y_1, \dots, Y_{i-1}\right) = H\left(Y_i | X_i\right),$$

so we can rewrite what we have broken up with the chain rule as

$$\begin{aligned} H\left(Y^n | X^n\right) &= \sum_{i=1}^{n} H\left(Y_i | X^n, Y_1, \dots, Y_{i-1}\right) \\ &= \sum_{i=1}^{n} H\left(Y_i | X_i\right). \end{aligned}$$

Using the fact that we are using a DMC, we have obtained that

$$\begin{aligned} I\left(X^n \wedge Y^n\right) &= \sum_{i=1}^{n} I\left(X_i \wedge Y_i\right). \\ &\leq n \cdot \max_{P_{Y|X}=W} I\left(X \wedge Y\right), \end{aligned}$$

which is n times capacity, so we can write this as

$$\frac{1}{n} \cdot I\left(M_n \wedge \varphi_n(Y^n)\right) \leq C(W).$$

Now we are left to prove that

$$\frac{1}{n} \cdot H\left(M_n | \varphi_n(Y^n)\right) \to 0.$$

We will do this using Fano's inequality.

We introduce the RV Z_n, defined as

$$Z_n = \begin{cases} 1 \text{ if } \varphi_n(Y^n) \neq M_n, \\ 0 \text{ otherwise.} \end{cases}$$

We can say that

$$H\left(M_n|\varphi_n(Y^n)\right) \leq H\left(M_n, Z_n|\varphi_n(Y^n)\right)$$

since we "added" something to a RV.

Now we find an upper bound for it:

$$\begin{aligned} H\left(M_n, Z_n|\varphi_n(Y^n)\right) &= H\left(Z_n|\varphi_n(Y^n)\right) + H\left(M_n|\varphi_n(Y^n), Z_n\right) \\ &\leq H\left(Z_n\right) + H\left(M_n|\varphi_n(Y^n), Z_n\right) \\ &\leq 1 + H\left(M_n|\varphi_n(Y^n), Z_n\right). \end{aligned}$$

This means that

$$\frac{1}{n} \cdot H\left(M_n|\varphi_n(Y^n)\right) \leq \frac{1}{n} + \frac{1}{n} \cdot H\left(M_n|\varphi_n(Y_n), Z_n\right).$$

Since $\frac{1}{n} \to 0$, we are left to show that

$$\frac{1}{n} \cdot H\left(M_n|\varphi_n(Y_n), Z_n\right)$$

is small, *i.e.*, it goes to 0 too.

We can break this up:

$$\begin{aligned} \frac{1}{n} \cdot H\left(M_n|\varphi_n(Y_n), Z_n\right) = &\Pr\left[Z_n = 1\right] \cdot H\left(M_n|\varphi_n(Y^n), Z_n = 1\right) \\ &+ \Pr\left[Z_n = 0\right] \cdot H\left(M_n|\varphi_n(Y^n), Z_n = 0\right). \end{aligned} \tag{7.3}$$

For the case in which $Z_n = 1$, we bring in the error probability:

$$\begin{aligned} \varepsilon_n = \Pr\left[Z_n = 1\right] &= \Pr\left[M_n \neq \varphi_n(Y^n)\right] \\ &\leq e_n(W^n, \mathcal{C}_n, \varphi_n) \to 0. \end{aligned}$$

So $\varepsilon_n \to 0$, since the average is not more than the maximum.

$$\begin{aligned} &H\left(M_n|\varphi_n(Y^n), Z_n = 1\right) = \\ &\qquad \sum_{y \in \mathcal{Y}^n} \Pr\left[Y^n = y\right] \cdot H\left(M_n|\varphi_n(Y^n) = \varphi_n(y), Z_n = 1\right). \end{aligned}$$

We have that $M_n \in \mathcal{X}^n$, so its entropy is no more than $\log_2\left(|\mathcal{X}^n|\right)$. The first term of eq. (7.3) can thus be bounded by

$$\varepsilon_n \cdot \frac{1}{n} \cdot H\left(M_n|\varphi_n(Y^n), Z_n = 1\right) \leq \varepsilon_n \cdot \frac{1}{n} \cdot \log_2\left(|\mathcal{X}^n|\right) = \varepsilon_n \cdot \log_2\left(|\mathcal{X}|\right) \to 0.$$

The second term of eq. (7.3) is equal to 0.

$$\frac{1}{n} \cdot \Pr\left[Z_n = 0\right] \cdot H\left(M_n|\varphi_n(Y^n), Z_n = 0\right).$$

To see why, note that there is no error, so $M_n = \varphi_n(Y^n)$, and thus $H\left(M_n|\varphi_n(Y^n), Z_n\right) = 0$. □

7.4 Zero Error Capacity

Zero stands for zero probability, so the probability of error is bound to be equal to 0. We translate this problem in graph theory: we will look at product of graphs and powers of graphs. We have a DMC $\{W\}$, $W : \mathcal{X} \to \mathcal{Y}$. $W^n : \mathcal{X}^n \to \mathcal{Y}^n$, the mathematical model of a code is the very same. A code is $(\mathcal{C}_n, \varphi)$ with $\mathcal{C}_n \subseteq \mathcal{X}^n$ and $\varphi : \mathcal{Y}^n \to \mathcal{C}_n$. A code has two parameters:

- $\frac{1}{n} \cdot |\mathcal{C}_n|$ is the size of the code compared to n. It is the number of bits transmitted over the channel per use of the channel.
- $e_n(W^n, \mathcal{C}_n, \varphi) = \max_{\underline{x} \in \mathcal{C}_n} W^n \left(\overline{\varphi^{-1}(\underline{x})} | \underline{x} \right)$ is the probability of error.

$(\mathcal{C}_n, \varphi)$ is a zero error code if $e_n(W^n, \mathcal{C}_n, \varphi) = 0$. $R \geq 0$ is an achievable rate for error probability 0 if $\exists (\mathcal{C}_n, \varphi_n)$, $\mathcal{C}_n \subseteq \mathcal{X}^n$ with

$$\limsup_{n \to \infty} \frac{1}{n} \cdot \log_2 \left(|\mathcal{C}_n| \right) \geq R,$$

and

$$e_n(W^n, \mathcal{C}_n, \varphi_n) = 0.$$

If $W(y|x) > 0$ for all x, y then this is not possible. If you have 0s in the matrix, we are only interested in the patterns of the 0s. To have 0 error, we should have that, for all $\underline{x} \in \mathcal{C}_n$,

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

The decoding function must be fixed if we want 0 error. If $W^n(\underline{y}|\underline{x}) \geq 0$, we must have that $\underline{y} \in \varphi^{-1}(\underline{x})$. Given $\underline{x}$, the set $\left\{ \underline{y} | W^n \left(\underline{y} | \underline{x} \right) > 0 \right\}$ is uniquely defined. This is the support of conditional distribution:

$$\mathrm{Supp}_W(\underline{x}) = \left\{ \underline{y} | W^n \left(\underline{y} | \underline{x} \right) > 0 \right\}.$$

We should have that, for all $\underline{x} \in \mathcal{C}_n$,

$$\mathrm{Supp}_W(\underline{x}) \subseteq \varphi^{-1}(\underline{x}).$$

So what we want to achieve, *i.e.*, how to get the 0 error probability, only depends on the set $\mathcal{C}_n$. Support sets have to be disjoint. $\mathcal{C}_n$ is a 0 error code if

$$\forall \{\underline{x}', \underline{x}''\} \in \binom{\mathcal{C}_n}{2} . \mathrm{Supp}_W(\underline{x}') \cap \mathrm{Supp}_W(\underline{x}'') = \emptyset$$

Observation 8 (Support set in $\mathcal{X}^n$)**.** *Let $\underline{x} \in \mathcal{X}^n$. Then*

$$\underline{y} \in \mathrm{Supp}_W(\underline{x}) \iff W^n \left(\underline{y} | \underline{x} \right) > 0,$$

but W^n is a product of probabilities; thus, since

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

it must be that, for all i,

$$W(y_i, x_i) > 0.$$

So the support set is

$$\mathrm{Supp}_W(\underline{x}) = \times_{i=1}^{n} \mathrm{Supp}_W(x_i).$$

This is a combinatorial condition. The support is the set of positive elements of the row of x. You need at least two orthogonal rows in the matrix.

Observation 9 (Disjoint support sets)**.**

$$\mathrm{Supp}_W(\underline{x}') \cap \mathrm{Supp}_W(\underline{x}'') = \emptyset \iff \mathrm{Supp}_W(x_i') \cap \mathrm{Supp}_W(x_i'') = \emptyset,$$

for some i.

Proof. Suppose

$$\begin{gathered}\mathrm{Supp}_W(\underline{x}') \cap \mathrm{Supp}_W(\underline{x}'') \neq \emptyset \\ \iff \\ \exists \underline{z} \in \mathrm{Supp}_W(\underline{x}') \cap \mathrm{Supp}_W(\underline{x}'') \\ \iff \\ \forall i. z_i \in \mathrm{Supp}_W(x_i') \cap \mathrm{Supp}_W(x_i'').\end{gathered}$$

□

Consider the graph $G = G(W)$ such that $V(G) = \mathcal{X}$ and $\{x', x''\} \in E(G)$ if $\mathrm{Supp}_W(x') \cap \mathrm{Supp}_W(x'') = \emptyset$. We are looking for a clique in this graph, since its size $\omega(G)$ is the size of the best code one can have.

What about when we use the channel more than once? We look at $G^n = G(W^n)$, where $V(G^n) = \mathcal{X}^n$ and $\{\underline{x}', \underline{x}''\} \in E(G^n)$ if $\exists i : \mathrm{Supp}_W(x_i') \cap \mathrm{Supp}_W(x_i'') = \emptyset$. This sequence of graphs can be built using just the first graph, and forgetting about the matrix.

Take G. If $\{a, b\} \in E(G)$ they are indistinguishable. Now take G^n. $\{\underline{a}, \underline{b}\} \in E(G^n)$ are indistinguishable if $\exists i : \{a_i, b_i\} \in E(G)$ with $a = a_1 \dots a_n$ and $b = b_1 \dots b_n$. Every clique of maximal length is a 0 error code.

We say that

$$\limsup_{n\to\infty} \frac{1}{n} \cdot \log_2 \left(\omega(G^n)\right)$$

is the zero error capacity of G. This limit exists for every graph.

Proposition 10 (Noisy channels and graphs)**.** *For every graph G, $\exists W : G = G(W)$.*

Proof. Consider a matrix W where the rows are the vertexes of the graphs and columns are indexed by the non-edges, *i.e.*, the pairs of $\binom{V(G)}{2}$ that are not in $E(G)$.

$$W\left(\{a, b\} | x\right) = \begin{cases} 0 \text{ if } x \notin \{a, b\}, \\ 1 \text{ if } x \in \{a, b\}. \end{cases} \tag{7.4}$$

If the rows of x and y are not orthogonal, $\{x, y\}$ is not an edge. There are two problems:

- rows could not sum to 1;
- rows could sum up to 0.

The first problem is avoided by dividing the rows. The second is a avoided by doing

$$W = (WI)$$

i.e., by appending the identity matrix to W. Now W can be normalized, and $G = G(W)$. □

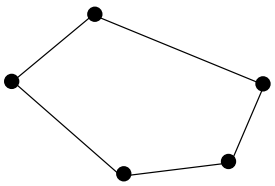

Figure 7.3: C_5.

$\sqrt[n]{\omega(G^n)}$ is how much larger is $\omega(G^n)$ with respect to $\omega(G)$. For graphs of 5 vertices there is only one graph for which

$$\sqrt[n]{\omega(G^n)} \neq \omega(G).$$

The graph is shown in fig. 7.3.

Lovàsz gave a function for $\sqrt[n]{\omega(G^n)}$ in '79. In fact it works for graphs that are self complementary and vertex transitive.

The Shannon capacity of G is defined as

$$C(G) = \limsup_{n\to\infty} \sqrt[n]{\omega(G^n)}.$$

Shannon proved that

$$\sqrt{5} \leq C(C_5) \leq \frac{5}{2}.$$

Lovàsz proved that in fact $C(C_5) = \sqrt{5}$. For the upper bound we have that $C_5 = G(W)$ with

$$W = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Then he said that 0 error capacity is less than or equal to the capacity. Shannon noticed, for the lower bound, that $C(G) > \omega(G)$, but $\omega(G) = 2$. If you look at G^2, *i.e.*, C_5^2, then $\omega(C_5^2) = 5$.

$$\omega(G^n) \geq [\omega(G)]^n \implies \sqrt[n]{\omega(G^n)} \geq \omega(G).$$

Using the product set you only get 4. But Shannon noticed that these 5 strings are distinguishable:

$$\{00, 11, 24, 31, 43\}.$$

Lemma 1 (Fekete)**.** *Take a sequence $a_n \in \mathbb{R}$. If the sequence is super additive, i.e., for all $m, n \in \mathbb{N}$ we have that $a_{m+n} \geq a_m + a_n$, and if for all $n \in \mathbb{N}$*

$$\frac{a_n}{n} \leq M,$$

then

$$\exists \lim_{n\to\infty} \frac{a_n}{n}$$

and

$$\lim_{n\to\infty} \frac{a_n}{n} = \sup_n \frac{a_n}{n}.$$

To apply this to $\omega(G^n)$ we have to take the logarithm, since $\omega(G^n)$ is super multiplicative (and thus $\log_2(\omega(G^n))$ is super additive). We have that $\omega(G^n) \leq |V(G)|^n$, and thus

$$\frac{\log_2(\omega(G^n))}{n} \leq \log_2(|V(G)|).$$

Since $\frac{a_n}{n} \leq M$, $\exists \hat{M} = \sup_n \frac{a_n}{n}$ with $\hat{M} \leq M$. So, for all $\varepsilon > 0$, $\exists n_0$ such that

$$\frac{a_{n_0}}{n_0} > \hat{M} - \varepsilon.$$

Take arbitrary $n > n_0$ with $n = q_n \cdot n_0 + r_n$, with $0 \leq r_n < n_0$.

$$\begin{aligned}\frac{a_n}{n} &= \frac{a_{q_n \cdot n_0 + r_n}}{q_n \cdot n_0 + r_n} \\ &\geq \frac{q_n \cdot a_{n_0} + a_{r_n}}{(q_n + 1)\, n_0} \\ &= \frac{q_n}{q_{n+1}} \cdot \frac{a_{n_0}}{n_0} + \frac{a_{r_n}}{(q_n + 1)\, n_0}.\end{aligned}$$

So the limit can be bounded from below:

$$\liminf_{n\to\infty} \frac{a_n}{n} \geq \liminf_{n\to\infty} \left(\frac{q_n}{q_{n+1}} \cdot \frac{a_{n_0}}{n_0} + \frac{a_{r_n}}{(q_n + 1)\, n_0} \right) \geq \hat{M} - \varepsilon.$$

Since $a_{r_n} \geq \min\{a_0, \ldots, a_{n_0 - 1}\}$, the last fraction goes to 0.

Proposition 11 (Graph capacity)**.**

$$\omega(G) \leq C(G) \leq \chi(G),$$

where $\chi(G)$ is the chromatic number of G.

A colouring of G is a function $f : V(G) \to C$ such that if $\{a, b\} \in E(G)$ then $f(a) \neq f(b)$. $\chi(G)$ is the minimum cardinality of C such that such a function exists.

Proof.

$$[\omega(G)]^n \leq \omega(G^n) \leq \chi(G^n).$$

The first part comes from super-multiplicativity of $\omega(G)$, the second from the fact that the chromatic number of a graph is greater than the largest clique. Since $\chi(G)$ is sub-multiplicative, we have

$$\chi(G^n) \leq [\chi(G)]^n.$$

To show that $\chi(G)$ is sub-multiplicative, consider an optimal colouring of G:

$$f : V(G) \to C \text{ with } |C| = \chi(G).$$

Take $\underline{x} \in V(G^n)$, we colour it separately, for x_i in $x_1 \ldots x_n$. The function $f^n(\underline{x})$, *i.e.*, its extension by concatenation, is a correct colouring.

To see why, note that if $\{\underline{x}, \underline{y}\} \in E(G^n)$, it must be that $\exists i : \{x_i, y_i\} \in E(G)$, but then, since x_i and y_i are adjacent in G, their colourings are different, *i.e.*,

$f(x_i) \neq f(y_i)$, and consequently also the colourings of $\underline{x}$ and $\underline{y}$, *i.e.*, $f^n(\underline{x}) \neq f^n((y))$.

Furthermore, the number of colours used by f^n is no more than $|C^n| = [\chi(G)]^n$, so

$$\omega(G) \leq \sqrt[n]{\omega(G^n)} \leq \chi(G). \qquad \square$$

There is a simple upper bound that is better than this. We use the "fractional" chromatic number, by expressing colouring as a Integer Linear Programming (ILP) problem and by dropping the integer constraint.

Corollary 1 (Graph capacity of perfect graphs)**.**

$$\omega(G) = \chi(G) \implies C(G) = \omega(G).$$

The corollary doesn't give us much. You can make any graph like this. Schutzenberger + Berge showed some classes of graphs for which this holds. One example is graphs for which vertices are intervals, and intersecting intervals share and edge. Furthermore, for interval graphs, $\forall G' \subseteq G$ induced subgraph we have that $\omega(G') = \chi(()G')$.

Definition 12 (Perfect graph)**.** *G is perfect if*

$$\omega(G') = \chi(G') \ \forall G' \subseteq G,$$

where G' is an induced subgraph.

Definition 13 (Minimally imperfect graph)**.** *A graph G is minimally imperfect if*

1. *G is not perfect, i.e., $\omega(G) < \chi(G)$,*
2. *$\forall G' \subseteq G$ induced subgraph, G' is perfect, i.e., $\omega(G') = \chi(G')$.*

What makes definition 12 beautiful are two conjectures, by Berge. By now they are both theorems.

- **Weak**: G is perfect if and only if $\overline{G}$ is perfect.
- **Strong**: minimally imperfect graphs are either odd cycles or their complements.

List of Definitions

List of Main Results

List of Secondary Results

Acronyms

BSC Binary Symmetric Channel

DMC Discrete Memoryless Channel

ECC Error Correcting Code

i.i.d. independent identically distributed

ILP Integer Linear Programming

LHS Left Hand Side

RHS Right Hand Side

RV Random Variable

UD Uniquely Decodable