Notes on Information Theory

Cristian Di Pietrantonio, Michele Laurenti

November 2, 2016

Chapter 1

Introduction

1.1 Lesson 1

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. In communication you want to be understood, but you also want to avoid redundancy and words tend to be lengthy; there is a conflict of interest. Why some words are short and others are long? Short words encode common abstract pictures whereas long words must carry the whole information not available in common knowledge [nessuna fonte].

Suggested readings:

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

1.2 Lesson 2

Define $\mathcal{M}$ as the finite set of all messages over an alphabet Σ. A message m is an element of $\mathcal{M}$. Hartley, who introduced this concept before Shannon, stated that the maximum amount of information is equal to $\log_2 |\mathcal{M}|$, since one can identify elements of $\mathcal{M}$ with binary strings of length $\log_2 |\mathcal{M}|$. Formally

$$\mathcal{M} \sim \{0,1\}^{\lceil \log_2 |\mathcal{M}| \rceil}$$

and

$$m \sim \Delta \in \{0,1\}^{\lceil \log_2 |\mathcal{M}| \rceil}.$$

One can think of Δ as a sequence of $\log_2 |\mathcal{M}|$ binary questions that identify an element in $\mathcal{M}$; a single unit of information is a 0 or 1 (Binary

Information Theory). In the *20 question game*, where you need to minimize the number of questions to guess what a person thinks, is given a probability distribution P over $\mathcal{M}$ s.t. $P(m) \geq 0,\ \forall m \in \mathcal{M}$ and $\sum_{m\in\mathcal{M}} P(m) = 1$. The questions can be designed to make the game as short as possible if P is known.

Must be noted that $|\underline{x}| = i \Leftrightarrow \underline{x} \in \{0,1\}^i$. The number of questions asked on average must be minimized. Let $f : \mathcal{M} \to \{0,1\}^*$, in other words a mapping from the message set to the sequences of answers to find the message then we want to minimize

$$\sum_{m\in\mathcal{M}} P(m)|f(m)|. \tag{1.1}$$

This equation is tied only to the probability distribution, not to the set or the questions. If P is uniform then 1.1 is equal to $\lceil \log_2 |\mathcal{M}| \rceil$. The function f is called **code**. The code is not always invertible.

Communication happens at a distance, otherwise it would be trivial. There is some device that makes communication possible; the channel is not perfect. The noise it is random, modelled through a probability distribution. The information source is not predictable either and will have its own probability distribution too (transmission error ?). Encoding and decoding must be optimized to make communications short to reduce "lost", and to reliably transmit, i.e. the receiver is reasonably sure that it gets what was sent. The two aspects are in contrast, you need and optimal trade-off. Usually partners in communication are separated in space and communication takes place in time. Sometimes in computer science the opposite thing happens: communication is storing information in space to retrieve it later in time. Sometimes you want to shorten the communication time, sometimes the space communication takes. The model is the same. Entropy comes from thermodynamics, introduced by Boltzmann.

1.3 Lesson 3 - Hamming Space

A space is a set that has structure. Consider the set of all binary strings $\{0,1\}^*$; we can define a space with the distance metric. The **distance** is defined as a function $d : X \times X \to \mathbb{R}_0^+$. A metric has the following properties:

1. (a) $d(x,y) \geq 0,\ \forall (x,y) \in X \times X$
   (b) $d(x,y) = 0 \Leftrightarrow x = y$
2. $d(x,y) = d(y,x), \forall (x,y) \in X \times X$
3. $d(x,y) \leq d(x,z) + d(z,y)$

What is the Hamming metric? Consider $\underline{x}, \underline{y} \in \{0,1\}^n$. Then the Hamming distance is defined as

$$d_H(\underline{x}, \underline{y}) = |\{i \mid x_i \neq y_i\}| \tag{1.2}$$

Why is it a metric? Consider $\underline{x}$, $\underline{y}$, $\underline{z} \in \{0,1\}^n$. Then, with $D = \{i \mid x_i \neq y_i\}$ is true that

$$i \in D \Rightarrow x_i \neq y_i \Rightarrow \neg(z_i = x_i \wedge z_i = y_i) \Rightarrow d_H(x_i, y_i) \leq d_H(z_i, x_i) + d_H(z_i, y_i).$$

By the additive property we can write

$$d_H(x, y) = \sum_{i=1}^{n} d_H(x_i, y_i).$$

Summing up gives the triangle inequality.

If a string $\underline{x}$ has been changed r times to become $\underline{y}$ then $d_H(\underline{x}, \underline{y}) \leq r$. Define a **Hamming ball** of radius r around $\underline{x}$ as

$$B_H = B_H(\underline{x}, r) = \{\underline{y} \mid d_H(\underline{x}, \underline{y}) \leq r\}.$$

Obviously r needn't to be an integer (use the *floor* function). Here are some properties of Hamming balls:

- if one subtracts $B_H(\underline{x}, r)$ to $\{0,1\}^n$ then the result is also an Hamming ball;
- $\{0,1\}^n = B_H(\underline{x}, n)$, $\forall \underline{x} \in \{0,1\}^n$, but if $r < n$ then the center is unique;

What we can say about $B_H(\underline{x}, r)$? For simplicity sake, consider $B_H(\underline{0}, r)$. Then the following holds:

$$|B_H(\underline{0}, r)| = \sum_{i=0}^{\lfloor r \rfloor} \binom{n}{i} \tag{1.3}$$

That is the number of ways in which we can flip to 1 the bits. From this follows that $d_H(\underline{0}, \underline{y}) + d_H(\underline{1}, \underline{y}) = n$.

We define the Hamming weight as $w_H(\underline{x}) = d_H(\underline{0}, \underline{x})$. The following holds:

$$\underline{y} \in B_H(\underline{0}, r) \Rightarrow w_H(\underline{y}) > r \Rightarrow d_H(\underline{y}, \underline{1}) < n - r \leq n - r - 1.$$

So it is true that

$$\overline{B_H(\underline{0}, n)} = B_H(\underline{1}, r') = B_H(\underline{1}, n - r - 1).$$

In other words the Hamming space can be partitioned into two Hamming balls. If n is odd an Hamming space can be partitioned in the following way:

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

The Hamming space cannot be partitioned into three balls. In how many balls can $\{0,1\}^n$ be partitioned? It cannot partitioned be partititoned into s balls with $3 \leq s \leq n+1$.

1.3.1 The volume of a Hamming ball

We have seen that the volume of a generic Hamming ball is described by equation 1.3. We can use the Pascal triangle to get a feeling about the order of magnitude of the Hamming ball. Consider the nth row in the triangle; since it is symmetric, if we split the row in half the sum of the terms of the first part is equal to the sum of the terms of the second part of the row. The volume of the greatest ball is $|B_H(\underline{0}, n)| = 2^n$; instead, the volume of the ball with radius $n/2$ (in the triangle's point of split) is $2^n/2 = 2^n 2^{-1} = 2^{n-1}$. So, if $r \geq n/2$ then $2^{n-1} \leq |B_H(\underline{0}, r)| \leq 2^n$. Can we bound the volume for $r < n/2$?

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

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

The plot of h looks like this:

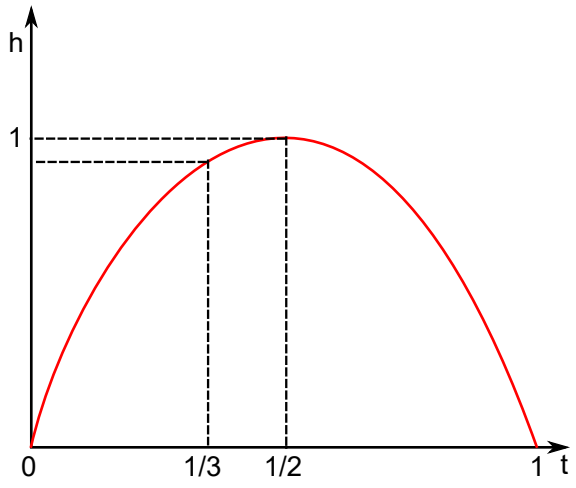

Figure 1.1: The entropy function.

This function is not defined for the values 0 and 1 but we have limits defined on these points, and they are both 0. So we artificially set $h(0) = 0$

and $h(1) = 0$. It is better to think about it as probability distribution $(t, 1-t)$ and entropy is a number attached to it. Entropy measures symmetry and with $t = 1/2$ we have maximum chaos (the future outcomes are equally likely).

1.4 Lesson 4

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

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

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 w/W.

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

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

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

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

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

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

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

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

at this point we "take out the lightest chicken" and write

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

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

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

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

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

So we have

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

□

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

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

Proof.

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

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

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

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

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

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

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

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

Observation 1

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

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

- $k \geq l$.

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

- $k < l$.

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

□

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

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

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

considering that $r = np$ it follows that

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

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

So we can write

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

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

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

So this proof is important because of two reasons:

1. something related to entropy. [LISTEN THE REGISTRATION];
2. Cardinality of the Hamming ball comes up in error correction (a string that has been haltered at most r times is in a certain radius from the original string).

1.5 Lesson 13/10/2016

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

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

$$N(a|\underline{x}) = |\{i \mid x_i = a\}|, \tag{1.4}$$

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

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

Moreover the following holds:

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

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

$$P_{\underline{x}} = \left\{ \frac{N(a|\underline{x})}{n} \mid a \in \mathcal{X} \right\} \tag{1.5}$$

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

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

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

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

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

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

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

$$P^n(\underline{x}) = \prod_{i=1}^{n} P(x_i). \tag{1.6}$$

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

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

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

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

Now,

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

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

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

So,

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

□

The lower bound proof is a straightforward generalization of what has been don in the binary case.

1.6 IDK what he's talking about yet (always 13/10)

Entropy is greatest when the distribution is uniform. Now, to the lower bound.

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

Observation 2 *If* $\mathrm{T}_p \neq \emptyset$ *then*

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

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

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

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

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

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

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

1.7 The log-sum inequality

Observation 3 *The logarithm function is cap-convex (∩-Convex). Remember that* $\ln(t) \leq t-1$ *(with equality iff* $t=1$*) and* $\log_2(t) = \dfrac{\ln(t)}{\ln(2)}$.

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

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

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

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

and we have equality iff $P = Q$. We call this $D(P||Q)$, called the informational divergence of P from Q. This is not a measure, but a "dissimilarity" measure. It's called also Kullback-Leibler divergence.

Proposition 1 is based on the Observation 3. The Proposition will be proved for the natural logarithm.

Proof. So

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

when $b_i = 0$ and $a_i = 0$, we have

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

by convention.

The reason why? $[t] \subset [w]$, you can think of $\{a_i\}$ as a subset of some other set where the other values are all 0s. Otherwise, if $a_i \geq 0$ and $b_i = 0$, we convene that

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

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

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

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

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

Notice that

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

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

$$\ln\left(\frac{1}{x}\right) \geq x - 1 \text{ and } \ln\left(\frac{1}{x}\right) \leq \frac{1}{x} - 1.$$

So

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

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

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

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

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

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

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

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

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

1.8 Variable-length codes

A variable-length binary code is a function $f : \mathcal{M} \to \{0,1\}^*$ for $\mathcal{M} < \infty$ and $\{0,1\}^* = \bigcup_{i=1}^{\infty} \{0,1\}^i$. Consider

$$f : [m] \to \{0,1\}^*$$

defined as $f(i) = 0^i$ is injective. However, consider $f^* : \mathcal{M}^* \to \{0,1\}^*$, the extension by concatenation, defined as

$$f^*(m_1, \ldots, m_e) = f(m_1) \ldots f(m_e).$$

The function f^* is not injective!

When f is prefix-free (or just prefix), *i.e.* let $\underline{x}, \underline{y} \in \{0,1\}^*$, $\underline{x}$ is prefix of $\underline{y}$ id $\underline{x} = \underline{y}$ or $\exists \underline{z} \in \{0,1\}^*$ such that $\underline{xz} = \underline{y}$. So, f is prefix-free if

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

where "$\hookleftarrow$" is the "is prefix of" relation. If f id prefix-free, f^* is injective.

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

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

$|f(m)| = l \Leftrightarrow f(m) \in \{0,1\}^l$. This proposition tells us that lots of short codewords imply that the set of message is small.

Proof. Let $\underline{x}, \underline{y} \in \{0,1\}^*$. Define

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

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

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

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

We say that $Y_L(\underline{x})$ and $Y_L(\underline{y})$ (???) can never be in "general position". Let A, B be two sets. They are in general position if $A \cap B$, $A \setminus B$, $B \setminus A$, $\overline{A \cup B}$ are all non-empty. For a prefix code, fiven $m' \neq m''$, then

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

So consider

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

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

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

Of course $L \geq \max_m |f(m)|$. Now, we have

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

□

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

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

Proof.

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

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

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

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

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

Observation 4 *Given $P|\mathcal{M}$ it is true that $H(P) < \log(|\mathcal{M}|)$, with equality iff P is the equidistribution.*

Proof.

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

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

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

□

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

Proof. We prove this with a greedy algorithm. We'll find an ordering of $\mathcal{M}$, which helps us with being greedy. We order $\mathcal{M}$ so that $l(m_1) \leq l(m_2) \leq \ldots \leq l(m_{|\mathcal{M}|})$.
First step. Set $L = l(m_{|\mathcal{M}|}) = \max_m l(m)$. We work with strings of length L and then we shorten them. Choose arbitrary $\hat{\underline{x}}^{(1)} \in \{0,1\}^L$ and let $f(m_1)$ be the prefix of $\hat{\underline{x}}^{(1)}$ of length $l(m_1)$. We then exclude the set of $L - l(m_1)$ extensions of $f(m_1)$.
General step. After constructing strings $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{t-1}$. We choose $\hat{\underline{x}}^{(t)}$ from $\{0,1\}^L \setminus (Y_L(\underline{x}_1) \cup \cdots \cup Y_L(\underline{x}_{t-1}))$; then $\underline{x}_t =$ the prefix of length $l(m_t)$ of $\hat{\underline{x}}^{(t)}$. We have to prove that the algorithm does end, and that it builds a prefix code.
The algorithm stops at step t iff $\{0,1\}^L = \bigcup_{i=1}^{t-1} Y_L(\underline{x}_i)$. We have seen that $|Y_L(\underline{x}_i)| = 2^{L-l(m_i)}$. This means that

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

dividing by 2^L we obtain

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

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

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

so the procedure terminates.
Correctness. f is a prefix-code. We have to show that

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

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

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

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

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

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

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

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

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

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

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