Information theory and coding

Bartek Blaszczyszyn
Notes by Antoine Groudiev

18th February 2024

Introduction

This document is Antoine Groudiev's class notes while following the class *Théorie de l'information et codage* (Information theory and coding) at the Computer Science Department of ENS Ulm. It is freely inspired by Bartek Blaszczyszyn's class notes.

This class contains two main parts: information theory, and coding theory. Information theory gives mathematical basis to build a notion of *information quantity*: given a text, how to "weight" the information contained in the sequence of characters? Some languages are more concise, but still provide the same quantity of information. Coding theory aims at finding the most concise way to represent information, with the smallest number of characters. Such theory – *source coding* – has many applications (storage, . . .). Another branch of coding is *canal coding* – allowing "repetition" in a message to avoid the loss of information in a canal.

1 Entropy and source coding

We shall introduce *Shannon's entropy* of a probability distribution on a discrete space and study its basic properties. Our goal is to prove *Shannon's source coding theorem* formulated in 1948. It will allow us to interpret the entropy as a notion of the *amount of information* "carried" by random variables of a given distribution.

1.1 Shannon's entropy

Let $\mathcal{X}$ be a finite or countable set, and $p := \{p(x) \mid x \in \mathcal{X}\}$ be a probability distribution on $\mathcal{X}$.

Definition (Shannon's entropy)**.** We define (Shannon's) entropy $H(p)$ of p to be:

$$H(p) := -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{1.1.1}$$

with the convention that $0\log 0 = 0$, and $a\log 0 = -\infty$ for $a > 0$. We will later on discuss the base of the logarithm.

Remark. *In this mathematical generalisation, $\mathcal{X}$ is the equivalent of a symbol alphabet, and p represents the text. The nature of the elements of $\mathcal{X}$ is not important: the entropy, the average, . . . depend only on $\mathcal{X}$ and on the distribution p. Therefore, we can re-label the elements of $\mathcal{X}$.*

Definition (Entropy of a random variable)**.** Let X be a random variable on $\mathcal{X}$ with distribution p, that is $\mathbb{P}(X = x) = p(x)$, also denoted $X \sim p$. We define:

$$H(X) := H(p) = -\mathbb{E}(\log p(X)) \tag{1.1.2}$$

Observe that $0 \leqslant H(p) \leqslant +\infty$, and that $H(p) = 0$ if and only if X is constant almost surely.

Property. *Entropy is invariant with respect to deterministic injective mapping $f : \mathcal{X} \to \mathcal{Y}$:*

$$H(X) = H(f(X))$$

The entropy $H(p)$ can be interpreted as the *amount of information* carried on average by one realization from the distribution p. Later in this chapter, we shall prove a result supporting this interpretation.

Definition (Entropy units)**.** The unit of the entropy depends on the *base of the logarithm*:

- In binary basis, when $\log = \log_2$, we denote $H(p) = H_2(p)$, and its unit is the $[bit/symbol]$ (per realization of X).
- In arbitrary basis $b > 0$, when $\log = \log_b$, we denote $H(p) = H_b(p)$, and its unit is the $[b - digit/symbol]$ (a b-digit is a digit which can take b values).
- In basis e, when $\log = \ln$, we denote $H(p) = H_e(p)$, and its unit is the $[nat/symbol]$ (nat is the natural unit of information).

The conversion between units can be done by changing the base of the logarithm:

$$H_b(p) = \frac{H_2(p)}{\log_2(b)}$$

Example (Bernoulli distribution)**.** *Let $\mathcal{X} = \{0, 1\}$, and p the Bernoulli distribution such as*

$$\begin{cases} p(0) = p \\ p(1) = 1 - p \end{cases}$$

Therefore, we have $H(p) = -p\log(p) - (1-p)\log(1-p)$. The Bernoulli distribution with the maximum entropy is:

$$\max_{0 \leqslant p \leqslant 1} H_2(p) = H_2(1/2) = 1\ [bit/symbol]$$

Example (Uniform distribution)**.** *Let $\mathcal{X}$ be a finite set, and p the uniform distribution, that is:*

$$\forall x \in \mathcal{X},\ p(x) := \frac{1}{|\mathcal{X}|}$$

Therefore, we have $H(p) = \log(|X|)$.

Example (Geometric distribution)**.** *Let $\mathcal{X} = \mathbb{N}^\star$ and p the geometric distribution of parameter $p > 0$, that is:*

$$\forall n \in \mathbb{N}^\star,\ p(n) = p(1-p)^{n-1}$$

Recall that $\mathbb{E}[X] = \frac{1}{p}$ when X follows a geometric law of parameter p.

Therefore, we have:

$$H(p) = \log\left(\frac{1-p}{p}\right) - \frac{1}{p}\log(1-p)$$

1.2 Gibbs' inequality

Theorem (Gibbs' inequality)**.** *Let p and q be two probability distributions on $\mathcal{X}$. Then:*

$$H(p) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \leqslant -\sum_{x\in\mathcal{X}} p(x)\log q(x) \tag{1.2.1}$$

Moreover, if $H(p) < \infty$, then there is equality in (1.2.1) *if and only if $p = q$.*

The right-hand-side of (1.2.1) is called *cross entropy* between p and q.

Proof. Let $x \sim p$. Gibbs' inequality is equivalent to:

$$\mathbb{E}[\log p(X)] \geqslant \mathbb{E}[\log q(X)]$$

If $\mathbb{E}[\log q(X)] = -\infty$, the inequality is trivial. Otherwise, since we have $\mathbb{E}[\log q(X)] \leqslant 0$:

$$\begin{aligned}\mathbb{E}[\log q(X)] - \mathbb{E}[\log p(X)] &= \mathbb{E}[\log q(X) - \log p(X)]\\ &= \mathbb{E}\left[\log\left(\frac{q(X)}{p(X)}\right)\right]\end{aligned}$$

log being concave, by applying Jensen's inequality, we obtain:

$$\begin{aligned}\mathbb{E}\left[\log\left(\frac{q(X)}{p(X)}\right)\right] &\leqslant \log\mathbb{E}\left[\frac{q(X)}{p(X)}\right]\\ &= \log\sum_{x\in\mathcal{X}}\frac{q(x)}{p(X)}p(x)\\ &= \log\sum_{x\in\mathcal{X}} q(x)\\ &= \log 1 = 0\end{aligned}$$

The equality in Jensen's inequality holds if and only if $\frac{q(X)}{p(X)}$ is almost surely constant, that is $p = \lambda q$ almost surely; furthermore, we must have $\lambda = 1$ since both p and q are distributions, hence $p = q$ almost surely. □

Corollary (Uniform distribution maximizes entropy)**.** *Let p be a probability distribution on some set $\mathcal{X}$ with $|\mathcal{X}| < \infty$. Then:*

$$0 \leqslant H(p) \leqslant \log(|\mathcal{X}|)$$

and the equality holds if and only if p is uniform on $\mathcal{X}$.

Proof. Let $X \sim p$ and be q the uniform distribution on $\mathcal{X}$. By Gibbs' inequality:

$$H(p) \leqslant -\sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{|X|}\right) = \log|X|$$

Notice that $\log|X|$ is the entropy of the uniform distribution q. □

Corollary (Geometric distribution maximizes entropy in the set of probability measures on $\mathbb{N}^\star$ having given expectation)**.** *Let p be a probability distribution on $\mathcal{X} = \mathbb{N}^\star$ with mean $\mu = \sum_{n\geqslant 1} np(n) < \infty$. Then:*

$$H(p) \leqslant \mu\log(\mu) - (\mu - 1)\log(\mu - 1)$$

where the right-hand-side is the entropy of the geometric distribution with parameter $1/\mu$.

Proof. Let p be a probability distribution on $\mathcal{X} = \mathbb{N}^\star$ with mean $\mu < \infty$, and q the geometric distribution of parameter $1/\mu$. According to Gibbs' inequality,

$$\begin{aligned}
H(p) &\leqslant -\sum_{n\geqslant 1} p(n)\log q(n) \\
&= -\sum_{n\geqslant 1} p(n)\log\left(\frac{1}{\mu}\left(1-\frac{1}{\mu}\right)^{n-1}\right) \\
&= \sum_{n\geqslant 1} p(n)\log\mu - \sum_{n\geqslant 1}(n-1)p(n)\log\left(1-\frac{1}{\mu}\right) \\
&= \log\mu - \log\left(1-\frac{1}{\mu}\right)(\mu-1) \\
&= \log\mu - (\log(\mu-1)-\log\mu)(\mu-1) \\
&= \log\mu + \mu\log\mu - \mu\log(\mu-1) + \log(\mu-1) - \log\mu \\
&= \mu\log\mu - (\mu-1)\log(\mu-1) = H(q)
\end{aligned}$$

□

1.3 Entropy of random vectors

Definition (Entropy of random vectors)**.** Let $X := (X_1,\dots,X_n)$ be a random vector on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, for some $n \geqslant 1$, with distribution

$$p(x_1^n) = p(x_1,\dots,x_n) = \mathbb{P}(X_1 = x_1,\dots,X_n = x_n)$$

The entropy of X is defined as the entropy of its distribution:

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) = -\mathbb{E}[\log p(X)] \tag{1.3.1}$$

Property (Entropy of independent variables)**.** *Let $X := (X_1,\dots,X_n)$ be a vector of* independent*, random variables. Then:*

$$H(X) = \sum_{i=1}^n H(X_i) \tag{1.3.2}$$

Proof. Let p be the joint distribution of X. By independence, $p(x) = \prod_{i=1}^n p_i(x_i)$. Hence:

$$\begin{aligned}
H(X) &= -\mathbb{E}[\log p(X)] \\
&= -\mathbb{E}\left[\log\prod_{i=1}^n p_i(X_i)\right] \\
&= -\mathbb{E}\left[\sum_{i=1}^n \log p_i(X_i)\right] \\
&= \sum_{i=1}^n -\mathbb{E}[\log p_i(X_i)] \\
&= \sum_{i=1}^n H(X_i)
\end{aligned}$$

□

Property (Independence maximizes entropy)**.** *Let $X := (X_1, \dots, X_n)$ be a vector of (arbitrary) random variables for some $n \geqslant 1$. Then:*

$$H(X) \leqslant \sum_{i=1}^{n} H(X_i) \tag{1.3.3}$$

Moreover, the equality holds if and only if $X_1, \dots, X_n$ are independent.

Proof. By induction. If $n = 1$, the results holds. Let X be an n-vector of random variables and X_{n+1} another random variable. Denote $q(x, y) = p(x)p_{n+1}(y)$, where $X \sim p$ and $X_{n+1} \sim p_{n+1}$, and $(X_1, \dots, X_n, X_{n+1}) \sim p'$. Since:

$$\begin{aligned} H(X) + H(X_{n+1}) &= -\mathbb{E}[\log p(X) + \log p_{n+1}(X_{n+1})] \\ &= -\mathbb{E}[\log q(X, X_{n+1})] \\ &\geqslant -\mathbb{E}[\log p'(X, X_{n+1})] = H(X_1, \dots, X_n, X_{n+1}) \end{aligned}$$

by Gibbs' inequality, the property is hereditary. Furthermore, there is equality in Gibbs' when $p' = q$, hence when X_{n+1} is independent from X, i.e. when $X_1, \dots, X_n, X_{n+1}$ are independent. □

1.4 Typical sequences of random variables

Definition (Typical sequences)**.** Let $\mathcal{X}$ be a set with $D := |\mathcal{X}| < \infty$, and $X = (X_1, \dots, X_n) \in \mathcal{X}^n$ a vector of independent and identically distributed random variables. Let p be the distribution of X on $\mathcal{X}$, with $p(x) = \prod_{i=1}^{n} p(x_i)$. We denote $H_D := H_D(p) = -\mathbb{E}[\log_D p(X)]$, expressed in D-digits/symbol.

For $\varepsilon > 0$, the following subset of realizations of $\mathcal{X}^n$

$$A_\varepsilon^{(n)} := \left\{ x \in \mathcal{X}^n \ : \ \Big| -\frac{1}{n} \sum_{i=1}^{n} \log_D p(x_i) - H_D \Big| \leqslant \varepsilon \right\} \subseteq \mathcal{X}^n \tag{1.4.1}$$

is called the set of ε-typical vectors in $\mathcal{X}^n$ with respect to p.

Intuitively, the typical vectors are the vectors that probabilistically appear a lot, and that we need to represent faithfully.

Remark.

$$\mathbb{E}\left[-\frac{1}{n} \sum_{i=1}^{n} \log_D p(X_i) \right] = \mathbb{E}[-\log_D p(X)] = H_D$$

and, by the Law of Large Numbers (LNN for short):

$$\lim_{n \to +\infty} -\frac{1}{n} \sum_{i=1}^{n} \log_D p(X_i) = \mathbb{E}[-\log_D p(X)] = H_D$$

We shall see that the probability distribution of X concentrates on the set of typical sequences, and, dependending on the entropy H_D, the dimension of this set can be smaller than n (the dimension of the whole space $\mathcal{X}^n$).

Property (Typical sequences concentrate probability)**.** *Let $X = (X_1, \dots, X_n)$ be a vector of i.i.d. random variables, with $X_i \sim p$ on $\mathcal{X}$, and $D := |\mathcal{X}| < \infty$. We have:*

$$\lim_{n \to +\infty} \mathbb{P}(X \in A_\varepsilon^{(n)}) = 1 \tag{1.4.2}$$

and

$$|A_\varepsilon^{(n)}| \leqslant D^{n(H_D + \varepsilon)} \tag{1.4.3}$$

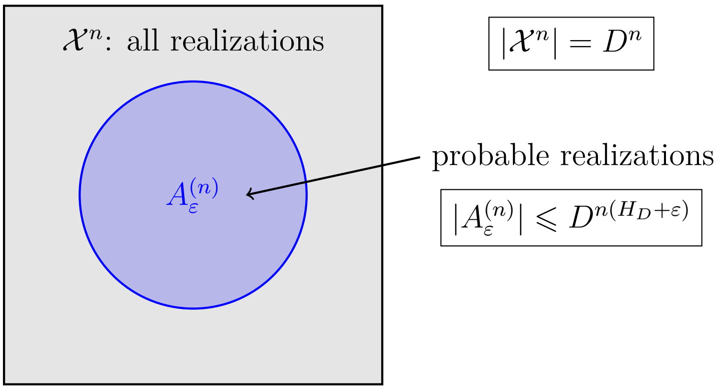
$\mathcal{X}^n$: all realizations
$|\mathcal{X}^n| = D^n$
$A_\varepsilon^{(n)}$
probable realizations
$|A_\varepsilon^{(n)}| \leqslant D^{n(H_D+\varepsilon)}$

Figure 1: Representation of $A_\varepsilon^{(n)}$

Proof. (1.4.2) follows from the LLN. For (1.4.3), observe that:

$$\begin{aligned} x \in A_\varepsilon^{(n)} &\implies -\sum_{i=1}^{n} \log_D p(x_i) \leqslant n(H_D+\varepsilon) \\ &\iff \log_D\left(\prod_{i=1}^{n} p(X_i)\right) \geqslant -n(H_D+\varepsilon) \\ &\iff \log_D p(x) \geqslant -n(H_D+\varepsilon) \\ &\iff p(x) \geqslant D^{-n(H_D+\varepsilon)} \end{aligned}$$

and since:

$$\begin{aligned} 1 \geqslant \mathbb{P}(X \in A_\varepsilon^{(n)}) &= \sum_{x\in A_\varepsilon} p(x) \\ &\geqslant |A_\varepsilon^{(n)}| D^{-n(H_D+\varepsilon)} \end{aligned}$$

which completes the proof. □

Property ($A_\varepsilon^{(n)}$ is the smallest set concentrating probability)**.** *Under the assumptions of Property 1.4, let $B \subseteq \mathcal{X}^n$ and $R > 0$ such that*

$$\lim_{n\to+\infty} \mathbb{P}(X \in B) = 1$$

and

$$|B| \leqslant D^{nR}$$

Then $R \geqslant H_D$, that is that $A_\varepsilon^{(n)}$ is the smallest set concentrating probability.

Proof. Let $\varepsilon > 0$, and assume $D > 1$, otherwise the result is trivial. Observe that:

$$\begin{aligned} x \in A_\varepsilon^{(n)} &\implies -\sum_{i=1}^{n} \log_D p(X_i) \geqslant n(H_D-\varepsilon) \\ &\iff p(x) \leqslant D^{-n(H_D-\varepsilon)} \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{P}(X \in A_\varepsilon^{(n)} \cap B) &\leqslant |B| D^{-n(H_D\varepsilon)} \\ &\leqslant D^{-n(H_D-R-\varepsilon)} \end{aligned}$$

Since $D > 1$ and $\lim_{n\to+\infty} \mathbb{P}(X \in A_\varepsilon^{(n)} \cap B) = 1$, we must have $H_D - R - \varepsilon \leqslant 0$, meaning that $H_D - \varepsilon \leqslant R$. We complete the proof by letting $\varepsilon \to 0$. □

1.5 Entropy and source coding rate – Shannon's first theorem

Definition (Encoding and decoding)**.** Let $X^n = (X_1, \dots, X_n)$ be a vector of i.i.d. random variables, with $X_i \sim p$ on $\mathcal{X}$. We call X_i *source symbols*, and the vector X^n of n source symbols is a *source message* (or *source word*, or *block of source symbols*).

Our goal is to *encode* the source message X^n consisting of n symbols using some (hopefully smaller) number of symbols in $\mathcal{X}$. That is, to represent X^n via a function $Y^m = c^{(n)}(X^n) \in \mathcal{X}^m$ for some $m \leqslant n$, in such a way that one can recover X^n from Y^m via a *decoding function* $d^{(n)}$ at least with high probability.

Definition (Compression rate)**.** The following ratio R is called *(sources) compression rate*:

$$R := \frac{m}{n} \tag{1.5.1}$$

Definition (Error probability)**.** We define the *error probability* $P_e^{(n)}$ to be:

$$P_e^{(n)} := \mathbb{P}\left(d^{(n)}\left(c^n(X^n)\right) \neq X^n\right) \tag{1.5.2}$$

which is nothing but the probability that the decoded message is different from the source message.

Theorem (Source coding theorem – Shannon's first theorem (1948))**.** *Let $X^n \in \mathcal{X}^n$ be a vector of i.i.d. random variables, with $X_i \sim p$ on $\mathcal{X}$. Denote $D := |\mathcal{X}|$ with $1 < D < \infty$. Then:*

$$\forall R > H_D(p), \quad \begin{cases} \exists c^{(n)} : \mathcal{X}^n \to \mathcal{X}^{\lceil nR \rceil} \\ \exists d^{(n)} : \mathcal{X}^{\lceil nR \rceil} \to \mathcal{X}^n \end{cases} \quad \text{such that } \lim_{n \to +\infty} P_e(n) = 0 \tag{1.5.3}$$

Furthermore,

$$\forall R < H_D(p), \quad \begin{cases} \forall c^{(n)} : \mathcal{X}^n \to \mathcal{X}^{\lceil nR \rceil} \\ \forall d^{(n)} : \mathcal{X}^{\lceil nR \rceil} \to \mathcal{X}^n \end{cases} \quad \text{we have } \lim_{n \to +\infty} P_e(n) > 0 \tag{1.5.4}$$

Proof. Let's start by proving (1.5.3). Let $R > H(p)$ and consider $0 < \varepsilon < R - H_D(p)$. For $n \geqslant 1$, let $A_\varepsilon^{(n)}$ be the set of ε-typical sequences for the distribution p, and let $f^{(n)}$ be an injection of $A_\varepsilon^{(n)}$ into $\mathcal{X}^{\lceil nR \rceil}$. Notice that such an injection exists by (1.4.3).

Let $x_\star \in X^{\lceil nR \rceil} \setminus f^{(n)}(A_\varepsilon^{(n)})$ arbitrary. Such an $x_\star$ exists since the inequality is strict in (1.4.3) (we have $H_D + \varepsilon < R$). We define the following coding function:

$$c^{(n)} := x \to \begin{cases} f^{(n)}(x) & \text{for } x \in A_\varepsilon^{(n)} \\ x_\star & \text{otherwise} \end{cases}$$

As a decoding function, consider the inverse of $f^{(n)}$ on its image $f^{(n)}(A_\varepsilon^{(n)})$, completed arbitrarily on the whole domain $\mathcal{X}^{\lceil nD \rceil}$, that is:

$$d^{(n)} := x \to \begin{cases} \left[f^{(n)}\right]^{-1}(x) & \text{if } x \in f^{(n)}(A_\varepsilon^{(n)}) \\ x_0 & \text{otherwise} \end{cases}$$

for some arbitrary $x_0 \in \mathcal{X}^n$. Finally, note that:

$$P_e^{(n)} \leqslant \mathbb{P}(X \notin A_\varepsilon^{(n)})$$

Therefore, (1.5.3) follows from (1.4.2).

The second statement, (1.5.4), follows from Property 1.4 considering the set:

$$B := \{x \mid d^{(n)}\left(c^n(x)\right) = x\}$$

□

Remark (Achievable compression rates)**.** *The source coding theorem – Theorem 1.5 – says that independent D-symbols emitted by a source with distribution p can be encoded asymptotically without errors using $H_D(p)$ encoding symbols per sources symbol. Note that*

$$H_D(p) \leqslant H_D(u) = \log_D(D) = 1$$

where u is the uniform distribution. The zero-error probability is approached asymptotically when increasing the length of the encoding blocks of source symbols.

Remark (Source coding rates)**.** *In general, one may use different sets of coding symbols $\mathcal{Y}$, having arbitrary number $b := |\mathcal{Y}| > 1$ of elements (set of b-digits) together with some coding and decoding functions:*

$$\begin{cases} c^{(n)} : \mathcal{X}^n \mapsto \mathcal{Y}^{n''} \\ d^{(n)} : \mathcal{Y}^{n''} \mapsto \mathcal{X}^n \end{cases}$$

In this more general scheme, the ratio:

$$R_s := \frac{\text{number of } b\text{-digits used to encode one source message}}{\text{number of source symbols in one source message}} = \frac{n''}{n} \tag{1.5.5}$$

is called (source) coding rate. *It is expressed in b-digits/(source) symbol. Considering a bijective mapping $\mathcal{X}^{n'} \mapsto \mathcal{Y}^{n''}$ with n', n'' such that $D^{n'} = b^{n''}$ in conjunction with Shannon's first theorem, it is straightforward to see that*

$$H_D(p) \log_b D = H_b(p) \text{ [}b\text{-digit/(source) symbol]}$$

is the infimum of coding rates over asymptotically error-free source coding.

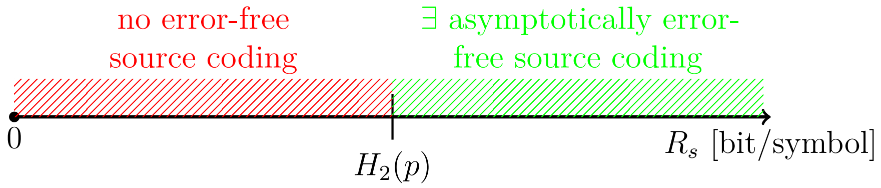
no error-free source coding
∃ asymptotically error-free source coding
0
$H_2(p)$
R_s [bit/symbol]

Figure 2: Source coding "phase transition"

2 Uniquely decipherable codes

We shall introduce some error-free source coding schemes and prove that in this class the entropy still indicates the infimum of achievable coding rates. The price to pay for error-free coding is *variable code-word length.* The main tool is Kraft's code proposed in 1949.

2.1 Uniquely decipherable and prefix-free codes

In the following, we let $\mathcal{X}$ be a finite set of *source symbols*, and $\mathcal{A}$ a set of coding symbols, called *alphabet.* We define $D := |\mathcal{A}| \in \mathbb{N}^\star$. Note that in general, $D \neq |\mathcal{X}|$. We denote $\mathcal{A}^\star$ the set of *words*, i.e. the finite (including null) sequences of character over alphabet $\mathcal{A}$. We denote $x * y$ or simply xy the concatenation between words x and y.

Definition (Code)**.** A *code* c encoding source symbols $\mathcal{X}$ in alphabet $\mathcal{A}$ is a function

$$c : \mathcal{X} \to \mathcal{A}^\star \setminus \{\varepsilon\}$$

$c(x)$ is called *code-word* of the source symbol x, and $l(x) := l(c(x))$ is the *code-word length.* The image of $\mathcal{X}$ by c, i.e. the set of all possible code-words, is called *codebook.*

In general, efficient codes will have code-words with variable length, with more probable source symbols having short code-words. In order to be able to uniquely decode a sequence of source symbols, we require more than just the injectivity of c.

Definition (Codes uniquely decipherable (UD))**.** A code $c : \mathcal{X} \to \mathcal{A}^\star$ is *uniquely decipherable* (UD) if $\forall k, l \geqslant 1, \forall x_1 \dots x_k \in \mathcal{X}, \forall y_1, \dots, y_k \in \mathcal{X}$,

$$c(x_1) * \dots * c(x_k) = c(y_1) * \dots * c(y_k) \implies k = l \wedge \forall i, x_i = y_i$$

Remark. *Note that c is UD if the following mapping is injective:*

$$\begin{gathered} X^\star \to \mathcal{A}^\star \\ (x_1, \dots, x_n) \mapsto c(x_1) * \dots * c(x_n) \end{gathered}$$

Definition (Prefix-free (PF) codes)**.** A code $c : \mathcal{X} \to \mathcal{A}^\star$ is *prefix-free (PF)* if

$$\nexists x \neq y \in \mathcal{X}, \exists a \in \mathcal{A}^\star,\ c(x) * a = c(y)$$

When such an $a \in \mathcal{A}^\star$ exists, $c(x)$ is said to be a *prefix* of $c(y)$.

Lemma (PF $\implies$ UD)**.** *A prefix-free code is uniquely decipherable.*

Proof. Let c be a PF code. Suppose by contradiction that $x_1, \dots, x_k, y_1, \dots, y_l$ violate the UD condition. Let i be the smallest index such that $x_i \neq y_i$. Then, $c(x_i)$ is a prefix of $c(y_i)$ or the other way around, depending on the code length. □

Remark (Decoding v. sequential decoding of block messages)**.** *There is a difference between the decoding of UD and PF codes:*

- *A UD code allows one to uniquely decode, that is to find, given a concatenation $(a_1, \dots, a_n)$ of some code-words, the sequence of symbols x_i for $1 \leqslant i \leqslant n$ such that*

$$c(x_1) * \dots * c(x_n) = (a_1, \dots, a_n)$$

- *A PF code allows one to decode the symbols x_i sequentially, by decoding with c the successive shortest prefixes of the encoded sequence found in the codebook. Sequential decogin simplifies decoding of blocks of symbols. As we shall see, it does not restrict the achievable performance of source coding.*

Example. *Let's consider codes $c : \mathcal{X} = \{1, 2, 3, 4\} \to \{0, 1\}^\star$. We define:*

- *A PF hence UD code of constant length, such that*

$$c(1) = (0, 0), c(2) = (0, 1), c(3) = (1, 0), c(4) = (1, 1)$$

- *A PF hence UD code of variable length*

$$c(1) = (0), c(2) = (1, 0), c(3) = (1, 10), c(4) = (1, 1, 1)$$

- *A not UD hence not PF code:*

$$c(1) = (0), c(2) = (1), c(3) = (1, 0), c(4) = (1, 1)$$

We have for instance that $c(2) * c(1) = c(3)$ and $c(2) * c(2) = c(4)$.

Property. *There exists codes that are UD but not PF.*

2.2 Codes on trees

Definition (k-ary tree)**.** For $k \in \mathbb{N}^\star$, a *k-ary tree* is a rooted tree in which each node (vertex) has no more than k children. A node of a k-ary tree having no children is called a *leaf*; otherwise, it is called an *intermediate node*. A tree in which each node has exactly k-children is called an *entire k-ary tree*, and is hence an infinite tree with no leaf.

Remark. *There is a natural bijection between the set of words $\mathcal{A}^\star$ expressed in the D-elements alphabet $\mathcal{A}$, and the vertices of the entire D-ary tree. Bearing in mind this bijection, we shall often identify this set of vertices with $A\star$. In particular, the empty word ε is identified with the root of the tree.*

Using the above bijection, a code $c : \mathcal{X} \to \mathcal{A}^\star$ can be seen as a mapping from $\mathcal{X}$ to the nodes of the entire D-ary tre.

Definition (Coding tree)**.** The minimal subtree of the entire D-ary tree containing the root and the code-words of the code c is called the *coding tree of c*.

Note that the coding tree is finite since $|\mathcal{X}| < \infty$, and that the leaves of the coding tree are necessarily the code-words of c, but some intermediate nodes might also be code-words.

Lemma (PF coding trees)**.** *Code c is PF if and only if it is injective, and it does not have code-words at the intermediate nodes in its coding tree.*

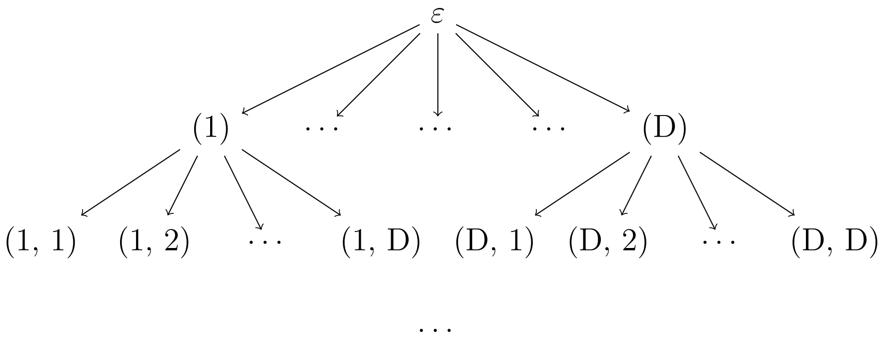

Figure 3: The entire D-ary tree

A useful equivalent characterization of PF codes involve subtrees of the entire regular tree. For any $w \in \mathcal{A}^\star$, denote by $\mathcal{T}(w)$ the subset of elements of $\mathcal{A}^\star$ for which w is a prefix. In graph representation, $\mathcal{T}(w)$ corresponds to the subtree rooted at w and consisting of all its descendants in the entire $|\mathcal{A}|$-regular tree.

Lemma (PF and subtrees)**.** *A code c is PF if and only if the subtrees of the entire regular tree rooted at the code-words corresponding to distinct source symbols are mutually disjoint:*

$$\forall x \neq y \in \mathcal{X}, \mathcal{T}(c(x)) \cap \mathcal{T}(c(y)) = \varnothing$$

Proof. $\Longrightarrow$ Assume $x \neq y \in \mathcal{X}$ and $a \in \mathcal{T}(c(x)) \cap \mathcal{T}(c(y)) \neq \varnothing$. Then both $c(x)$ and $c(y)$ are prefixes of a, and hence the shorter code-word of the two is a prefix of the longer one.

$\Longleftarrow$ If $c(x)$ is a prefix of $c(y)$ then $\mathcal{T}(c(y)) \subseteq \mathcal{T}(c(x))$, hence the intersection is not empty. □

Example (Uniform coding on the full binary tree)**.** *Assume $|\mathcal{X}| = 2^m$ for some $1 \leqslant m < \infty$, $\mathcal{A} = \{0, 1\}$. All elements of $\mathcal{X}$ can be encoded by the vertices of the m-th generation of the binary tree, or, equivalently, by the binary sequences of length m.*

Since the mapping is bijective and all code-words are equal, this is a PF and hence UD code.

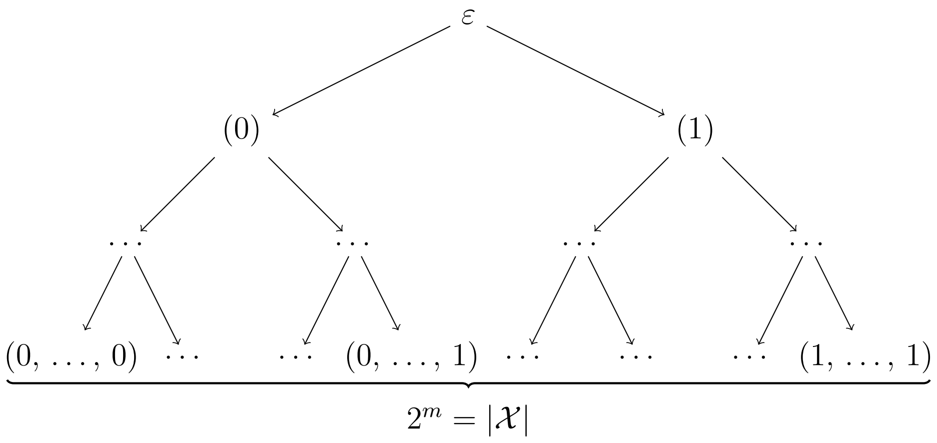

Figure 4: Binary sequences

The source coding rate is

$$R_s = \frac{\text{number of bits}}{\text{one source symbol}} = \frac{m}{1} = \log_2 |\mathcal{X}| = H_2(\text{uniform}(\mathcal{X}))$$

which corresponds to the infimum achievable coding rate when the source symbols are independently, uniformly sampled in $\mathcal{X}$. By Shannon's source coding theorem we cannot do better even accepting only asymptotically error-less coding.

A remaining question is that in case of independent source symbols sampled from some general distribution p on $\mathcal{X}$, how to achieve rates $R_S > H_2(p)$ with error-free (UD, PF) codes? The idea is to encode elemnts of $\mathcal{X}$ using binary sequences of *variable length*, with more likely source symbols having shorter code-words.

2.3 Kraft's inequality

Kraft's inequality will be our main tool in studying the performance of UD and PF codes.

Definition (Aggregated block codes)**.** For a code c and $n \in \mathbb{N}^\star$, the corresponding *aggregated block code* $c^{(n)} : \mathcal{X}^n \to \mathcal{A}^\star$ of block-length n is defined as follows:

$$\forall x \in \mathcal{X}^n, c^{(n)}(x) := c(x_1) * \cdots * c(x_n)$$

Lemma. *If c is UD then the aggregated block code $c^{(n)}$ is UD.*

Theorem (Kraft's inequality)**.** *Let c be a code encoding source symbols $\mathcal{X}$ with alphabet $\mathcal{A}$. Assume $D := |\mathcal{A}| < \infty$.*

1. *If c is UD then*

$$\sum_{x \in \mathcal{X}} D^{-l(c(x))} \leqslant 1 \tag{2.3.1}$$

2. *Let $\{l(x) \mid x \in \mathcal{X}\}$ be a collection of non-negative integers. If*

$$\sum_{x \in \mathcal{X}} D^{-l(c(x))} \leqslant 1$$

then there exists a PF (hence UD) code c with code-word lengths $l(x) = l(c(x))$.

Proof of Case 1. Assuming $|X| < \infty$ consider UD code c and denote by $l_{\max}$ its maximal code-word length:

$$l_{\max} = \max_{x \in \mathcal{X}} l(c(x)) < \infty$$

Let $n \in \mathbb{N}^\star$. Consider the aggregated block code $c^{(n)}$ corresponding to c. Recall from the preceding lemma that $c^{(n)}$ is UD. Denote by $\alpha(k)$ the number of code-words of $c^{(n)}$ of length k:

$$\alpha(k) := |\{x \in \mathcal{X} \mid l(c^{(n)}(x)) = k\}|$$

Note that $\alpha(k) = 0$ for $k = 0$ and for $k > nl_{\max}$. In general $\alpha(k) \leqslant D^k$, since $c^{(n)}$ is UD and there are at most D^k differen code-words of length k.

We have

$$\begin{aligned}\left(\sum_{x \in \mathcal{X}} D^{-l(c(x))}\right)^n &= \sum_{x \in \mathcal{X}} D^{-l(c(x_1)+\cdots+c(x_n))} \\ &= \sum_{x \in \mathcal{X}} D^{-l(c^{(n)}(x))} \\ &= \sum_k \alpha(k) D^{-k} \\ &\leqslant \sum_{k=1}^{nl_{\max}} D^k D^{-k} \\ &= nl_{\max}\end{aligned}$$

Consequently,

$$\sum_{x \in \mathcal{X}} D^{-l(c(x))} \leqslant (nl_{\max})^{1/n} \to_{n \to \infty} 1$$

which completes the proof of the first statement for $|\mathcal{X}| < \infty$.

If $|\mathcal{X}| = \infty$, consider an increasing sequence of subsets $\mathcal{X}_m$ such that $\bigcup_m \mathcal{X}_m = \mathcal{X}$ and codes truncated to $\mathcal{X}_m$. □

Proof of Case 2. We assume that $|\mathcal{X}| < \infty$. WLOG, let $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$, chosen such that $l(1) \leqslant \ldots \leqslant l(|\mathcal{X}|)$. Define $m := l(|\mathcal{X}|) = \max_{x \in \mathcal{X}} l(x)$. We construct a code c with code-word lengths $l(c(i)) = l(i)$ by encoding messages with some vertices of the entire D-ary tree as follows (recall $D := |\mathcal{A}|$):

- The code-word of $1 \in \mathcal{X}$ is corresponds to the most recent common ancestor of the lexicographically first group of $D^{m-l(1)}$ on the m-th generation.

- By the induction, the code-word of $i \in \mathcal{X}$ corresponds to the most recent common ancestor of the lexicographically i-th group of $D^{m-l(i)}$ nodes of the m-th generation.

Note that by the assumption of the second statement,

$$\sum_{i=1}^{|\mathcal{X}|} D^{m-l(i)} \leqslant D^m$$

and thus the above construction can be completed. By the construction, subtrees rooted at the selected code-words are disjoint, $\mathcal{T}(c(i) \cap \mathcal{T}(c(j))) = \varnothing$ for $1 \leqslant i \neq j \leqslant |\mathcal{X}|$, hence by lemma, the constructed code is PF.

We can similarly extend this to the case where $|\mathcal{X}| = \infty$. □

2.4 Kraft's code

We have suggested that an efficient UD code has in general variable length of the code-words. Its coding rate can be defined asymptotically on a sequence of source symbols $X_1, X_2, \ldots$

$$R_s := \lim_{n\to\infty} \frac{l(c(X_1)) + \cdots + l(c(X_n))}{n} \quad \left[\frac{D - \text{digit}}{\text{source symbol}}\right] \tag{2.4.1}$$

assuming that the limit exists, thus extending the previous definition of R (1.5.1) for fixed length codes.

For $X_1, X_2, \ldots$ i.i.d. with $X \sim p$, by the LLN, the limit in (2.4.1) exists a.s. and

$$R_s = \mathbb{E}[l(c(X))]$$

Therefore, looking for a rate-optimal UD code consists in minimizing the mean code-word $\mathbb{E}[l(c(X))]$. In particular, we want to know whether there exist UD codes for all compression rates $R > H_D(p)$, as it is the case for asymptotically error-free codes, according to Shannon's source coding theorem (1.5.3).

Definition (Kraft's code)**.** For a probability distribution p on $\mathcal{X}$, Kraft's code with alphabet $\mathcal{A}$ of cardinality $D := |\mathcal{A}|$ is any PF code c_K having code-words of length

$$\forall x \in \mathcal{X}, \quad l(c_K(x)) = \lceil -\log_D(p(x)) \rceil$$

Remark. *Note that*

$$\sum_{x\in\mathcal{X}} D^{-\lceil -\log_D p(x) \rceil} \leqslant \sum_{x\in\mathcal{X}} D^{\log_D p(x)} = \sum_{x\in\mathcal{X}} p(x) = 1$$

Property (Mean length of Kraft's code)**.** *The mean code-word length of any Kraft's code satisfies*

$$H_D(p) \leqslant \mathbb{E}[l(c_K(X))] \leqslant H_D(p) + 1 \tag{2.4.2}$$

Furthermore, there is no UD code c with mean code-word length

$$\mathbb{E}[l(c(X))] \leqslant H_D(p)$$

Proof. For the upper bound, note that

$$\begin{aligned}\mathbb{E}[l(c_K(X))] &= \sum_{x\in\mathcal{X}} p(x) l(c(x)) \\ &\leqslant \sum_{x\in\mathcal{X}} p(x)(-\log_D p(x) + 1) \\ &= H_D(p) + 1\end{aligned}$$

The second statement follows from the second statement of Kraft's inequality combined with Gibbs' inequality. Indeed, for UD code c, by Kraft's inequality (2.3.1), $\bar{q}(x) := D^{-l(c(x))}$ can be seen as a sub-probability measure on $\mathcal{X}$. Consider a probability measure q on $\mathcal{X}$ such that $\bar{q}(x) \leqslant q(x)$. We have

$$\begin{aligned}\mathbb{E}[l(c(X))] &= \sum_{x\in\mathcal{X}} p(x) l(c(x)) \\ &= -\sum_{x\in\mathcal{X}} p(x) \log_D(D^{-l(c(x))}) \\ &\geqslant -\sum_{x\in\mathcal{X}} p(x) \log_D q(x) \\ &\geqslant -\sum_{x\in\mathcal{X}} p(x) \log_D p(x) && \text{(Gibbs' inequality)} \\ &= H_D(p)\end{aligned}$$

□

Remark. *The second statement of the previous property can also be deduced from Shannon's first theorem* (1.5.4)*. Indeed, UD codes have zero-error probability and thus necessarily compression rate no smaller than* $H_D(p)$*. Some work is required to make this argument rigorous.*

Property. *The not necessarily integer-valued function*

$$l^\star(x) := -\log_D p(x)$$

minimizes $\sum_{x\in\mathcal{X}} p(x)l(x)$ *within the class of functions satisfying Kraft's inequality constraint*

$$\sum_{x\in\mathcal{X}} D^{-l(x)} \leqslant 1$$

Proof. We use Gibbs' inequality, noticing that $\sum_{x\in\mathcal{X}} p(x)l^\star(x) = H_D(p)$. □

Corollary (Kraft's code par block achieve all coding rates above the entropy)**.** *Let* $X_1, X_2, \dots$ *be i.i.d. source symbols on* $\mathcal{X}$ *with* $X_i \sim p$*. For given* $n, k \in \mathbb{N}^\star$*, denote by*

$$X_k^{(n)} := (X_{(k-1)n+1}, \dots, X_{kn})$$

the source messages consisting of blocks of n *source symbols. Let* $c_K^{(n)} : \mathcal{X}^n \to \mathcal{A}^\star$ *be a Kraft's code related to the* n*-th product probability measure* p^n*, where*

$$p^n(x) := p(x_1)\dots p(x_n)$$

(Note that in general, it is not the aggregated block-code related to a "one-dimensional" Kraft's code $c_k : \mathcal{X} \to \mathcal{A}^\star$*, the former being used in the proof of Kraft's inequality.) The coding rate of this block-code* $c^{(n)}$ *is*

$$\begin{aligned} R_s &= \lim_{k\to\infty} \frac{l(c_K^{(n)}(X_1^{(n)})) + \dots + l(c_K^{(n)}(X_k^{(n)}))}{kn} \qquad \left[\frac{D-\text{digit}}{\text{source symbol}}\right] \\ &= \frac{\mathbb{E}[l(c_K^{(n)}(X_1^{(n)}))]}{n} \qquad \text{(a.s. by the LLN)} \\ &\leqslant \frac{H_D(p^n)+1}{n} \qquad \text{(by (2.4.2))} \\ &= \frac{nH_D(p)+1}{n} \qquad \text{(the variables are independent)} \\ &= H_D(p) + \frac{1}{n} \to_{n\to\infty} H_D(p) \end{aligned}$$

Remark (Entropy is the infimum of coding rates achievable over UD codes)**.** *The previous corollary shows that for any* $\varepsilon > 0$ *there exists a PF code (for example, Kraft's block-code) offering coding rate* $R_s \leqslant H_D(p) + \varepsilon \left[\frac{D-\text{digit}}{\text{source symbol}}\right]$*. Combining this with the second statement of* (2.4.2) *we conclude that* $H_D(p)$ *is the infimum of coding rates achievable over UD codes.*

3 Optimal codes

We shall present *Huffman's code*, introduced in 1952, which is optimal in the sens of minimizing the mean code-word length for a given source distribution.

3.1 Optimality

Let $\mathcal{X}$ be a finite set of source symbols, p a probability distribution on $\mathcal{X}$ and $\mathcal{A}$ a finite alphabet with $D := |\mathcal{A}| < \infty$ characters.

Definition (Optimal codes)**.** A code $c : \mathcal{X} \to \mathcal{A}^\star$ is called an *optimal D-ary code* for p if it is a UD code which minimizes the mean code-word length $\mathbb{E}[l(c(X))]$ with $X \sim p$, in the class of UD codes:

$$\mathbb{E}[l(c(X))] = \min_{\hat{c} \text{ UD code}} \mathbb{E}[l(\hat{c}(X))]$$

Remark. *For any optimal D-ary code c for p, by 2.4.2, we have:*

$$H_D(p) \leqslant \mathbb{E}[l(c(X))] \leqslant \mathbb{E}[l(c_k(X))] \leqslant H_D(p) + 1$$

where c_k is a Kraft code (a PF code having code-words of lengths $l(x) = \lceil 1/\log p(x) \rceil$).

Property. *There exists an optimal code, which moreover can be taken PF.*

Proof. Assume that $p_{\min} := \min_{x \in \mathcal{X}} p(x) > 0$. The elements of $\mathcal{X}$ with zero probability can have arbitrary code-words as they do not impact the mean code length. Observe first than an optimal code exists as a solution of an optimization problem in the finite domain of the non-negative integer-valued functions $l(\cdot)$ on $\mathcal{X}$ bounded by $(H_D(p) + 1)/p_{\min} < \infty$. Indeed, $\mathbb{E}[l(c(X))] \geqslant \max_{x \in \mathcal{X}} l(c(x)) p_{\min}$, and by (3.1), no code with the mean code-word length larger than $H_D(p) + 1$ is optimal.

By Kraft's inequality (2.3.1), the code-word lengths of a given optimal code (which is UD) satisfy

$$\sum_{x \in \mathcal{X}} D^{-l(c(x))} \leqslant 1$$

By the second statement of Kraft's inequality, there exists a PF code having the same code-word lengths. It is hence optimal. □

Remark. *Combining the two statements of Kraft's inequality, a UD code c is an optimal D-ary code for p if and only if its code-word lengths $l(x) := l(c(x))$ solve the following optimization problem:*

$$\begin{cases} \text{minimize } \sum_{x \in \mathcal{X}} p(x) l(x) \\ \text{subject to } \sum_{x \in \mathcal{X}} D^{-l(x)} \leqslant 1 \end{cases} \tag{3.1.1}$$

in the set of non-negative integer valued functions l. Recall that:

- *The function $l^\star(x) := -\log p(x)$ solves* (3.1.1) *in the set of non-negative, real-valued functions yielding $\sum_{x \in \mathcal{X}} p(x) l^\star(x) = H_D(p)$.*

- *Kraft's codes assume $l(x) = \lceil -\log p(x) \rceil$, which is not necessarily a solution to* (3.1.1) *in integer-valued functions.*

3.2 Huffman's code

Let p be a probability distribution on $\mathcal{X}$, with $|\mathcal{X}| < \infty$, and consider the binary alphabet $\mathcal{A} = \{0, 1\}$.

Definition (Huffman's coding tree)**.** We construct a binary tree $\mathcal{T}_H = T_H(p)$ with vertices marked by some probabilities (real values in $[0, 1]$) executing the following algorithm:

1. Assign distinct vertices (of a binary tree to be constructed) to all elements $x \in \mathcal{X}$, *activate* and mark them by the corresponding probability $p(x)$. At this stage, all the vertices are isolated, there are no edges in the graph.

2. Take two different active vertices minimizing the sum of their probabilities. Deactivate these vertices and create a new one, being the direct common ancestor of them. Activate and mark this new vertex by the sum of the probaibilities of the two vertices deactivated in this step.

3. Repeat step 2 until there is only one active vertex. Declare this vertex to be the root $\varnothing$ of the constructed graph, which is a tree. The root is marked with probability 1.

Definition (Huffman's code)**.** Huffman's code is a function $c_H : \mathcal{X} \to \{0,1\}^\star$ which assignes to $x \in \mathcal{X}$ the binary representation of the vertex corresponding to x in Huffman's binary tree $\mathcal{T}_H$.

Example (Huffman's tree and code construction)**.** *Let $\mathcal{X} = \{1, \dots, 9\}$ and*

$$p = (0.01, 0.02, 0.03, 0.1, 0.12, 0.2, 0.2, 0.3)$$

We obtain the following Huffman's coding tree:

Note that for the Huffman's code $c = c_H$, the mean code length is

$$\sum_{i=1}^{9} p(i)l(c(i)) = 2.64$$

and the entropy is

$$\sum_{i=1}^{9} p(i)p(i) \simeq 2.593$$

3.3 Optimality of Huffman's code

Lemma (Optimal binary code – Base condition)**.** *Let $\mathcal{X} = \{1, \dots, n\}$, with $3 \leqslant n < \infty$ and p be a probability distribution on $\mathcal{X}$ such that $p(i) > 0$. There exists binary PF codes*

$$c : \{1, \dots, n\} \to \{0,1\}^\star$$

optimal for p, satisfying

$$\begin{cases} c(n) = \omega * 0 \\ c(n-1) = \omega * 1 \text{ for } \max\limits_{i=1,\dots,n-2} p(i) \geqslant p(n-1) \geqslant p(n) > 0 \end{cases} \tag{3.3.1}$$

for some $\omega \in \{0,1\}^\star$.

Proof. Let $c : \mathcal{X} \to \{0,1\}^\star$ be an optimal binary code for p, which is PF. It exists by Property 3.1. Enumerate $p(1) \geqslant \dots p(n) > 0$. Then the code-word lengths of this optimal code necessarily satisfy $l(1) \leqslant \dots \leqslant l(n)$ (*The proof is left as an exercise.*)

Note that we have $l(n-1) = l(n)$. Indeed, if $l(n-1) < l(n)$, then replacing $c(n)$ by its prefix $c'(n)$ of length $l(n-1)$, we obtain a PF code that has a strictly smaller mean code-word length, thus contradicting the optimality of c.

One can also assume that $c(n) = \omega * (1)$ and $c(n-1) = \omega * (0)$. This might require exchanging some code words of length $l(n)$, preserving the PF property and the mean code-word length. □

Lemma (Optimal binary code – Recursive step)**.** *Under the assumptions of Lemma 3.3 let a code c be optimal for p satisfying the base condition* (3.3.1)*. Then, any optimal code c' for the distribution p' defined by:*

$$\begin{cases} p'(i) = p(i), \quad \forall 1 \leqslant i \leqslant n-1 \\ p'(n-1) = p(n-1) + p(n) \end{cases}$$

makes the following code $\hat{c}'$ optimal for p:

$$\hat{c}'(i) := \begin{cases} c'(i) & \textit{for } 1 \leqslant i \leqslant n-2 \\ c'(n-1) * (0) & \textit{for } i = n-1 \\ c'(n) * (1) & \textit{for } i = n \end{cases}$$

Proof. Denote by L' and $\hat{L}'$ the mean code-word length of c' and $\hat{c}'$ respectively. We have

$$L' + p(n-1) + p(n) = \hat{L}'$$

Indeed,

$$L' = \sum_{i=1}^{n-2} p(i) l(c'(i)) + (p(n-1) + p(n)) l(c'(n-1))$$
$$\hat{L}' = \sum_{i=1}^{n-2} p(i) l(c'(i)) + (p(n-1) + p(n))(l(c'(n-1)) + 1)$$

Consider also a code c'' shortening to $\{1, \ldots, n-1\}$ the original optimal code c for p satisfying the initialization property (3.3.1):

$$c''(i) := \begin{cases} c(i) & \text{for } i = 1, \ldots, n-2 \\ \omega & \text{for } i = n-1 \end{cases}$$

Denote by L'' the mean code of c''. We have a similar relation:

$$L'' + p(n-1) + p(n) = L$$

Therefore,

$$\begin{aligned} L &= L'' + p(n-1) + p(n) \\ &\geqslant L' + p(n-1) + p(n) \\ &= \hat{L}' \\ &\geqslant L \end{aligned}$$

thus proving that $L = \hat{L}'$ and hence that $\hat{c}'$ is optimal for p, which completes the proof. □

Corollary. *Huffman's code is optimal.*

Proof. Indeed, Huffman's code is constructed by recursion:

- It ensures the base condition (3.3.1) (when the second step of the algorithm of Huffman's binary tree is performed).
- Making the recursive step of Lemma 3.3 reduced the cardinality of the set of messages iteratively down to $|\mathcal{X}| = 2$. At the final step any of the two existing bijections of $\mathcal{X}$ to $\{0, 1\}$ is an optimal PF code.

Lemma 3.3 guarantees the optimality by the induction. □

Lemma. *Huffman's code minimizes the mean code-word length amongst* binary *UD codes.*

Remark. *Huffman's code minimizes the mean code-word length amongst* binary *UD codes. Comparing it to Kraft's code c_K introduced previously, using* (2.4.2) *with binary alphabet $D = 2$ we have, for $X \sim p$:*

$$H_2(p) \leqslant \mathbb{E}[l(c_H(X))] \leqslant \mathbb{E}[l(c_K(X))] \leqslant H_2(p) + 1 \tag{3.3.2}$$

Consequently, when applied per block, Huffman's code achieves asymptotically optimal coding rate converging to the entropy.

Proof of Lemma 3.3. Let $c : \mathcal{X} = \{1, \ldots, n\} \to \{0, 1\}^\star$ be an optimal binary code for p, which is PF. It exists by Property 3.1. The code-word lengths of this optimal code necessarily satisfy $l(1) \leqslant \ldots \leqslant l(n)$. □