Lectures Notes : Algorithmic Information Theory

Itisan Halias
itha.mail0@gmail.com

September 2025

Contents

Prerequisites

In this chapter $\mathcal{A}$ and $\mathcal{A}'$ are arbitrary alphabets [todo: add ref]. We also set $\mathbb{B} = \{0, 1\}$. We further assume $\mathbb{B} \subset \mathcal{A}$. We set $\mathbb{B} = \{0; 1\}^*$.

0.1 Languages

0.1.1 Definitions: Alphabets, Words and Languages

Definition: Alphabet/ word/ empty word/ length/ language

Let $\mathcal{A}$ be a non-empty finite set whose elements are called *symbols*. We define:

1. **Alphabet.** An alphabet is a non-empty finite set $\mathcal{A}$ of symbols.
2. **Word.** A word over $\mathcal{A}$ is a finite sequence of elements of $\mathcal{A}$. More precisely, for any natural integer $l \geq 1$, a word of length l is a mapping
$$w : \{1, 2, \ldots, l\} \mapsto \mathcal{A}$$
which we will denote in the form
$$w = (v_i)_{1 \leq i \leq l} \quad \text{or even} \quad w = v_1 v_2 \cdots v_l$$
The set of non-empty words over $\mathcal{A}$ is denoted by $\mathcal{A}^+$.
3. **Empty word.** The empty word, denoted by ϵ, is the word of zero length. For a given alphabet $\mathcal{A}$, the set of words (including the empty word) is defined by
$$\mathcal{A}^* \stackrel{\text{def}}{=} \mathcal{A}^+ \cup \{\epsilon\}$$
4. **Infinite word.** An infinite word over $\mathcal{A}$ is an infinite sequence of elements of $\mathcal{A}$, that is to say a mapping
$$w : \mathbb{N} \mapsto \mathcal{A}$$
The set of infinite words is denoted by $\mathcal{A}^\infty$.
5. **Set of finite and infinite words.** We define
$$\mathcal{A}^\propto \stackrel{\text{def}}{=} \mathcal{A}^* \cup \mathcal{A}^\infty$$
6. **Length of a word.** The length of a finite word $w \in \mathcal{A}^+$, denoted by $|w|$, is defined by
$$|w| = \sum_{a \in \mathcal{A}} |w|_a$$
where $|w|_a$ denotes the number of occurrences of the symbol a in w. By convention, the length of the empty word is zero, that is $|\epsilon| = 0$.

7. **Language.** A language over the alphabet $\mathcal{A}$ is a subset of $\mathcal{A}^*$, that is

$$L \subseteq \mathcal{A}^*$$

0.1.2 Operations on words and languages

Given two words, it is natural to wish to have an operation allowing them to be concatenated.

Definition: Concatenation of words

Let $\mathcal{A}$ be an alphabet. Concatenation is the internal operation

$$\cdot : \mathcal{A}^* \times \mathcal{A}^* \to \mathcal{A}^*$$

defined by

$$\forall x, y \in \mathcal{A}^*, \quad \begin{cases} x \cdot y = x_1 x_2 \cdots x_{|x|}\, y_1 y_2 \cdots y_{|y|} \\ \epsilon \cdot x = x \cdot \epsilon = x \end{cases}$$

where $|x|$ denotes the length of the word x (and similarly for y). For convenience, we will write xy instead of $x \cdot y$.
For any word $r \in \mathcal{A}^*$ and for any integer $n \in \mathbb{N}$, the power r^n is defined by recurrence as follows:

$$\begin{cases} r^0 = \epsilon, \\ r^{n+1} = r \cdot r^n. \end{cases}$$

Analogously, we define a concatenation operation on languages.

Definition: Concatenation of languages

Let $\mathcal{A}$ and $\mathcal{A}'$ be two alphabets. Concatenation of languages is the operation

$$\cdot : \mathcal{P}(\mathcal{A}^*) \times \mathcal{P}(\mathcal{A}'^*) \to \mathcal{P}((\mathcal{A} \cup \mathcal{A}')^*)$$

defined by

$$\forall L_1 \subset \mathcal{A}^*,\, L_2 \subset \mathcal{A}'^*, \quad \begin{cases} L_1 \cdot L_2 = \{\, w_1 w_2 \mid w_1 \in L_1 \text{ and } w_2 \in L_2 \,\} \\ \{\epsilon\} \cdot L = L \cdot \{\epsilon\} = L \end{cases}$$

We will also denote $L_1 L_2$ instead of $L_1 \cdot L_2$.
The power of a language $L \subset \mathcal{A}^*$ is then defined by recurrence:

$$\begin{cases} L^0 = \{\epsilon\} \\ L^{n+1} = L \cdot L^n \quad \text{for all } n \geq 0. \end{cases}$$

Thus, L^n denotes the set of all words obtained by the concatenation of n words coming from L.

0.2 Prefix-Free Languages

A notion that will be particularly useful throughout this course is that of prefix and prefix-free language.

0.2.1 Definitions

We introduce the notion of prefix, which is essential in the study of formal languages.

Definition: Prefix

We say that $v \in \mathcal{A}^*$ is a prefix of $w \in \mathcal{A}^*$ if there exists $z \in \mathcal{A}^*$ such that $w = vz$. We then denote

$$v \leq_p w$$

Moreover if $z \neq \epsilon$ then we say that v is a strict prefix of w and we denote

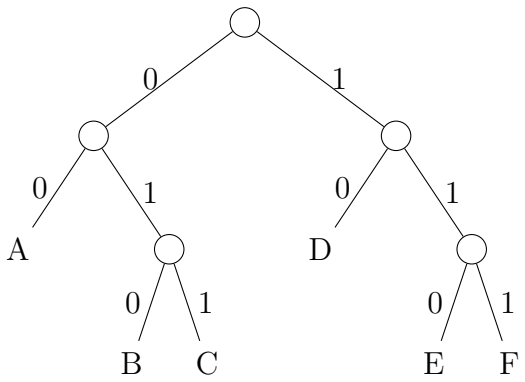
$$v <_p w$$

The notion of prefix allows us to define languages that do not contain words having a common sub-word, which leads to the concept of prefix-free language.

Definition: Prefix-free language

A language $L \subset \mathcal{A}^*$ is prefix-free if

$$\forall x, y \in L : \quad (x \leq_p y \implies x = y)$$



Definition: Minimal prefix languages

Let $P \subseteq \mathbb{B}^*$. The set of minimal prefix representatives of P, denoted $\lfloor P \rfloor$, is defined by:

$$\lfloor P \rfloor := \{p \in P \mid \forall s \in \mathbb{B}^*, (s <_p p \implies s \notin P)\}$$

This set contains exactly the elements p of P such that no proper prefix s of p belongs to P.

Definition: Compatibility

We define the following notions:

- Two words $x, y \in \mathbb{B}^*$ are said to be compatible if either $x \leq_p y$ or $y \leq_p x$

- Two languages $L, L' \subset \mathbb{B}^*$ are said to be compatible if for all $x \in L$ and $y \in L'$ we have x and y compatible.

0.2.2 Cylinders and Binary Intervals

Definition: Cylinder

We call cylinder with respect to $x \in \mathbb{B}^*$ the set Γ_x such that

$$\Gamma_x \underset{def.}{=} \{xw \mid w \in \mathbb{B}^\infty\}$$

Property: Compatibility of cylinders

Let $x, y \in \mathbb{B}^*$, we then have $\Gamma_y \subset \Gamma_x \iff x \leq_p y$

Proof. $\Rightarrow$: Assume $\Gamma_y \subset \Gamma_x$. If $x = y$ then we immediately have $x \leq_p y$, so we will assume in the following $x \neq y$. Since $\Gamma_y \subset \Gamma_x$, there exist $z, z' \in \mathbb{B}^*$ such that $yz = xz'$. We then notice two possible cases: $x \leq_p y$ or $y \leq_p x$. Moreover as we have assumed $x \neq y$ then necessarily $x <_p y$ or $y <_p x$. By contradiction, assume $y <_p x$. We can therefore write $yp = x$ for a $p \in \mathbb{B}^+$. Let $p' \in \mathbb{B}^*$ of the same length as p such that $p' \neq p$ (it suffices to flip any bit of p). We notice two points $yp' \in \Gamma_y$ and $yp' \neq x$. Thus for all $w \in \mathbb{B}^\infty$ we have $yp'w \in \Gamma_y$ and $yp'w \notin \Gamma_x$ which is absurd.

$\Leftarrow$: Assume $x \leq_p y$. There thus exists $z \in \mathbb{B}^*$ such that $xz = y$. Thus for all $w \in \mathbb{B}^\infty$ we have $xzw \in \Gamma_x$ that is $yw \in \Gamma_x$ which means $\Gamma_y \subset \Gamma_x$. ■

One can think of the notation $0.x$, with $x \in \mathbb{B}^\infty$, as a way to represent a real number in the interval $[0, 1[$ from a finite or infinite sequence of binary digits. In other words, each $x \in \mathbb{B}^\infty$ is associated with a real number defined as follows:

Definition: Mapping $\mathbb{B}^*$ to $[0; 1[$

Let φ be a mapping from $\mathbb{B}^*$ to $[0, 1[$ defined by

$$\varphi : \begin{cases} x = x_1 x_2 \cdots x_{|x|} & \mapsto \sum_{i=1}^{|x|} \frac{x_i}{2^i} \\ \mathbb{B}^* & \mapsto [0, 1[\end{cases}$$

We then set the following notation for $x = x_1 x_2 \ldots \in \mathbb{B}^*$:

$$0.x \underset{def.}{=} \varphi(x) = \sum_{i=1}^{|x|} \frac{x_i}{2^i}$$

We seek to understand how the cylinder Γ_x, defined for a word $x \in \mathbb{B}^*$, translates under the numerical representation provided by the mapping φ. In other words, what does $\varphi(\Gamma_x)$ correspond to? This leads us to introduce the notion of binary intervals, which offer a geometric interpretation of the cylinder and will prove practical in the subsequent construction of prefix codes.

Definition: Binary Intervals (incompatible)

Let $x \in \mathbb{B}^*$, we call binary interval of x the interval [a]

$$I_x \underset{def.}{=} [0.x; 0.x + 2^{-|x|}[$$

Moreover if for $x, y \in \mathbb{B}^*$ we have $I_x \cap I_y = \emptyset$ then we say that I_x and I_y are two incompatible binary intervals.

[a] Note that the interval is open on the right, this detail is important.

Remark. We then notice that we have for $x \in \mathbb{B}^*$ that $\varphi(\Gamma_x) = I_x$ ◇

We have a theorem that characterizes prefix-free languages. This theorem is of particular importance, as it will be useful to demonstrate several significant results, notably Kraft's inequality, but especially a fundamental result: Levin's coding theorem.

Theorem: Characterization of prefix languages by binary intervals

Let $x_1, x_2, \dots x_l$ be pairwise distinct in $\mathbb{B}^*$, with l possibly infinite, we then have equivalence between

1. For all $1 \leq i, j \leq l$, such that $i \neq j$, the binary intervals I_{x_i} and I_{x_j} are incompatible (i.e. $I_{x_i} \cap I_{x_j} = \emptyset$).
2. The language $L = \{x_1, x_2, \dots x_l\}$ is prefix-free.

Proof. $\underline{1 \Rightarrow 2}$: By contradiction, assume that there exists $x_i \leq x_j$ with $i \neq j$. Since $x_i \neq x_j$, then there exists $z \in \mathbb{B}^+$ such that $x_i z = x_j$. Let $z = z_1 z_2 \dots z_k$. We have

$$0.x_j = 0.x_i + \sum_{q=1}^{k} z_q \cdot 2^{-(|x_i|+q)} < 0.x_i + \sum_{q=1}^{\infty} 2^{-(|x_i|+q)} = 0.x_i + 2^{-|x_i|}$$

Thus, by trivially noting that $0.x_i \leq 0.x_j$, we have $0.x_j \in I_{x_i}$. Now we also have $0.x_j \in I_{x_j}$. In summary $I_{x_j} \cap I_{x_i} \neq \emptyset$ which is absurd.

$\underline{2 \Rightarrow 1}$: Let $1 \leq i, j \leq l$ be two indices with $i \neq j$. Without loss of generality, assume $|x_i| \leq |x_j|$. Let $r \in I_{x_i} \cap I_{x_j}$. By definition, r starts with x_i, and x_j which gives

$$r = 0.x_i \dots \quad \text{and} \quad r = 0.x_j \dots$$

Thus, the first $|x_i|$ bits of r are also those of x_j. This implies that x_i is a prefix of x_j. However, L is prefix-free, so the uniqueness of prefixes forces $x_i = x_j$, which contradicts the hypothesis that x_i and x_j are two distinct elements. ■

0.2.3 Kraft's Inequality

Lemma:

Let $(l_i)_{i\in[\![1,k]\!]}$ be a finite increasing sequence in $\mathbb{N}^*$ such that $\sum_{i=1}^{k-1} 2^{-l_i} < 1$, then there exists a unique $x := b_1 b_2 \dots b_{l_k}$ with $b_i \in \mathbb{B}$ such that

$$\sum_{i=1}^{k-1} 2^{-l_i} = \sum_{i=1}^{l_k} b_i \cdot 2^{-i} := \varphi(x)$$

Proof. assumed for now ■

We will now prove a result that will be notably necessary to establish the definition of an a priori probability later in this course. Kraft's inequality also allows characterizing a prefix-free language. In reality, it is a corollary of the theorem characterizing prefix-free languages by binary intervals.

Theorem: Kraft's Inequality

$\Rightarrow$: Let $L \subset \mathbb{B}^*$ be a non-empty prefix-free language, then

$$\sum_{x\in L} 2^{-|x|} \leq 1$$

$\Leftarrow$: Let $(l_n)_{n\in[\![1;l]\!]}$, with l potentially infinite, be a sequence of natural numbers in $\mathbb{N}^*$ such that

$$S := \sum_{i=1}^{l} 2^{-l_i} \leq 1.$$

Then, there exists a prefix-free language $L = \{x_1, \dots, x_l\}$ of words in $\mathbb{B}^*$ such that, for all $1 \leq i \leq l$, we have $|x_i| = l_i$.

Proof. $\underline{\Rightarrow}$: Let $L = \{x_1, x_2, \dots, x_l\}$ be a prefix-free language, with l possibly infinite. Let us set the binary intervals for all $1 \leq i \leq l$ such that $I_{x_i} = [0.x_i + 2^{-|x_i|}[$. By the characterization of prefix-free languages, we have that the I_{x_i} are disjoint. Moreover, we evidently have that $I_{x_i} \subset [0; 1[$. Thus, denoting by $\delta(I_{x_i})$ the length of the interval, we have

$$1 = \delta([0;1[) \geq \delta\left(\bigcup_{i=1}^{l} I_{x_i}\right) = \sum_{i=1}^{l} \delta(I_{x_i}) = \sum_{i=1}^{l} \delta([0.x_i; 0.x_i + 2^{-|x_i|}[) = \sum_{i=1}^{l} 2^{-|x_i|}$$

$\underline{\Leftarrow}$: Let $(l_k)_{k\in[\![1;l]\!]}$, with l potentially infinite, be defined as in the statement of the theorem. Without loss of generality, assume that $(l_k)_{k\in[\![1;l]\!]}$ is increasing. We then notice, knowing that for all $1 \leq k \leq l$ we have $S_{k-1} := \sum_{i=1}^{k-1} 2^{-l_i} < 1$, (the inequality is strict because $2^{-l_{k+1}} > 0$ and $S \leq 1$) and $S_0 = 0$. Thus, by the previous lemma, there exists a unique $x_k := b_1 \dots b_{l_k}$ such that $S_{k-1} = \sum_{i=1}^{l_k} b_i \cdot 2^{-i}$.

Let us set the binary intervals for all $1 \leq k \leq l$ such that

$$I_{x_k} := [0.x_k; 0.x_k + 2^{-l_k}[$$

We observe then that the I_{x_k} are pairwise disjoint. Indeed, let $1 \leq k < k' \leq l$, we have

$$I_{x_k} \cap I_{x_{k'}} = [0.x_k; 0.x_k + 2^{-l_k}[\cap [0.x_{k'}; 0.x_{k'} + 2^{-l_{k'}}[$$

$$= \left[\sum_{i=1}^{k-1} 2^{-l_i}; \left(\sum_{i=1}^{k-1} 2^{-l_i}\right) + 2^{-l_k}\right[\cap \left[\sum_{i=1}^{k'-1} 2^{-l_i}; \left(\sum_{i=1}^{k'-1} 2^{-l_i}\right) + 2^{-l_{k'}}\right[$$

Now $\sum_{i=1}^{k-1} 2^{-l_i} + 2^{-l_k} \leq \sum_{i=1}^{k'-1} 2^{-l_i} + 2^{-l_{k'}}$ which yields $I_{x_k} \cap I_{x_{k'}} = \emptyset$. Thus, by the characterization of prefix-free languages by binary intervals, $L = \{x_1, \ldots, x_l\}$ is prefix-free. ■

0.2.4 Convergence of monotone sequences for $\leq_p$

We will establish a property that will be useful later, which we designate as the monotone convergence theorem for $\leq_p$, by analogy with the monotone convergence theorem for bounded sequences in $\mathbb{R}$.

Definition-Theorem: Convergence of monotone sequences for $\leq_p$

Let $(x_i)_{i\in\mathbb{N}}$ be a sequence of words in $\mathbb{B}^*$ that is increasing with respect to $\leq_p$ we have

1. $\exists!\, l \in \mathbb{N} \cup \infty, \quad \lim_{i\to\infty} |x_i| = l.$
2. $\exists!\, s \in \mathbb{B}^{=l}, \quad \forall i \in \mathbb{N}, \quad x_i \leq_p s$

we then set

$$\lim_{i\to\infty} x_i \underset{def.}{=} x$$

Proof. Let $(x_i)_{i\in\mathbb{N}}$ be a sequence of words in $\mathbb{B}^*$ that is increasing with respect to $\leq_p$. Let us prove each of the points sequentially.

1) For all $k \in \mathbb{N}$ we have $x_k \leq_p x_{k+1}$ therefore $|x_k| \leq |x_{k+1}|$, that is to say that the sequence $(|x_i|)_{i\in\mathbb{N}}$ is increasing. By the monotone sequence convergence theorem its limit exists and $\lim_{i\to\infty} |x_i| = l := \sup\{|x_i| \mid i \in \mathbb{N}\} \in \mathbb{N} \cup \infty$.

2) Existence: Let $0 \leq i \leq l$. Since $\lim_{k\to\infty} |x_i| = l$, there exists a unique minimum

$$N_i = \min\{N_i \in \mathbb{N} \text{ such that } |x_{N_i}| \geq i\}$$

We then set $s[i]$ equal to the i-th letter of x_{N_i}, which we will denote $s[i] = x_{N_i}[i]$. As this holds for all $1 \leq i \leq l$ we can set $s = s[1] \ldots s[l]$.

Let us now show that for all $k \in \mathbb{N}$ we have $x_k \leq_p s[1] \ldots s[l]$. Let $k \in \mathbb{N}$ and $1 \leq \gamma \leq |x_k|$. Since $|x_k| \geq \gamma$ we have by the minimality of N_γ that $k \geq N_\gamma$, which by the increasing nature of $(x_i)_{i\in\mathbb{N}}$ with respect to $\leq_p$ implies $x_{N_\gamma} \leq_p x_k$. Now by definition $s[\gamma] = x_{N_\gamma}[\gamma]$, which gives $x_{N_\gamma}[\gamma] = x_k[\gamma]$. Thus we have shown that for all $1 \leq \gamma \leq |x_k|$ we have $x_k[\gamma] = s[\gamma]$. Knowing moreover that $|s| = l$ and $|x_k| \leq l$ this immediately provides that $x_k \leq_p s$.

Uniqueness: Suppose there exist s and s' satisfying for all $i \in \mathbb{N}$ that $x_i \leq_p s$ and $x_i \leq_p s'$. By the uniqueness of l we know that $|s| = |s'| = l$. Let $1 \leq i \leq l$. By the growth of $(|x_k|)_{k\in\mathbb{N}}$ and the definition of N_i we have that for all $k \geq N_i$, $|x_k| \geq i$. Thus x_k is of size at least i and moreover $x_k \leq s$ and $x_k \leq_p s'$. In summary we have $x_k[i] = s[i]$ and $x_k[i] = s'[i]$ therefore $s[i] = s'[i]$. Knowing that this holds for all $0 \leq i \leq |s| = |s'| = l$ this yields $s = s'$. ■

0.2.5 Dyadic Intervals

Lemma: Prefix language associated with an interval with dyadic endpoints

Let $[a, b[\subset [0, 1[$ be an interval with dyadic endpoints, that is to say a and b are dyadic, then there exists a finite prefix-free language $Q \subset \mathbb{B}^*$ such that

$$\bigcup_{p \in Q} \varphi(\Gamma_p) = [a, b[$$

Proof. Let a and b be two dyadic numbers. By definition, there exist an integer $N \in \mathbb{N}$ and integers $k_a, k_b \in \mathbb{Z}$ such that $a = \frac{k_a}{2^N}$ and $b = \frac{k_b}{2^N}$.

For all $0 \leq j < k_b - k_a$ there exists a unique word $p \in \mathbb{B}^{=N}$ such that $0.p = \frac{k_a + j}{2^N}$. We then define the language Q as follows:

$$Q := \{p_j \mid j = 0, 1, \ldots, k_b - k_a - 1\}$$

We then remark that for each word $p_j \in \Xi$, the associated binary interval is

$$I_{p_j} = \varphi(\Gamma_{p_j}) = \left[0.p_j,\, 0.p_j + 2^{-N}\right[= \left[\frac{k_a + j}{2^N},\, \frac{k_a + j + 1}{2^N}\right[$$

By performing the union over all words in Ξ, we obtain a union of contiguous semi-open intervals:

$$\bigcup_{p \in \Xi} \varphi(\Gamma_p) = \bigcup_{j=0}^{k_b - k_a - 1} \left[\frac{k_a + j}{2^N},\, \frac{k_a + j + 1}{2^N}\right[$$

This "telescoping" union is equal to the interval extending from the beginning of the first to the end of the last:

$$\bigcup_{p \in \Xi} \varphi(\Gamma_p) = \left[\frac{k_a}{2^N},\, \frac{k_b}{2^N}\right[= [a, b[$$

The equality is thus proven. Moreover, since the I_j are pairwise disjoint, the words of Ξ are incomparable, which means the set is prefix-free. ■

Lemma: Measure of the union of dyadic intervals

Let $P \subseteq \mathbb{B}^*$ be any language. The Lebesgue measure of the union of the intervals associated with P is given by:

$$\lambda\left(\bigcup_{p \in P} I_p\right) = \sum_{p' \in \lfloor P \rfloor} 2^{-\ell(p')}$$

Proof. Let us first prove, by double inclusion, that $\bigcup_{p \in P} I_p = \bigcup_{p' \in \lfloor P \rfloor} I_{p'}$.

- $\subseteq$: Let $z \in \bigcup_{p \in P} I_p$. By definition, there exists a $p \in P$ such that $z \in I_p$. By the definition of $\lfloor P \rfloor$, there exists a unique $p' \in \lfloor P \rfloor$ such that $p' \leq_p p$. The characterization of prefix words by binary intervals gives us $I_p \subseteq I_{p'}$. Consequently, $z \in I_{p'}$. Since $p' \in \lfloor P \rfloor$, we have $z \in \bigcup_{q \in \lfloor P \rfloor} I_q$. This proves the inclusion.

- $\supseteq$: Let $z \in \bigcup_{p' \in \lfloor P \rfloor} I_{p'}$. There exists therefore a $p' \in \lfloor P \rfloor$ such that $z \in I_{p'}$. By definition, every element of $\lfloor P \rfloor$ is also an element of P. Thus, $p' \in P$. Consequently, z belongs to an interval $I_{p'}$ where $p' \in P$, which means that $z \in \bigcup_{p \in P} I_p$. This proves the inclusion.

By taking the measure, we therefore have $\lambda\left(\bigcup_{p\in P} I_p\right) = \lambda\left(\bigcup_{p'\in\lfloor P\rfloor} I_{p'}\right)$. By definition, $\lfloor P\rfloor$ is prefix-free, so for all p'_1 and p'_2 in $\lfloor P\rfloor$, we have $I_{p'_1} \cap I_{p'_2} = \emptyset$. Thus, the measure of the disjoint union is the sum of the measures:

$$\lambda\left(\bigcup_{p'\in\lfloor P\rfloor} I_{p'}\right) = \sum_{p'\in\lfloor P\rfloor} \lambda(I_{p'})$$

The measure of each dyadic interval $I_{p'}$ is its length, that is to say $\lambda(I_{p'}) = 2^{-\ell(p')}$. Thus by substituting we obtain

$$\lambda\left(\bigcup_{p\in P} I_p\right) = \sum_{p'\in\lfloor P\rfloor} 2^{-\ell(p')}$$

■

0.3 Code (prefix)

As indicated in the introduction of this section, a coding links a word to its representation, which is called a code. If a code were not an injection, there would be ambiguity regarding the origin of the encoded word. Thus, we can define a code as an injective map that associates a word with its code, which gives formally

0.3.1 Definition

Definition: Code

We call a code an injective map

$$C : \Theta \subset \mathbb{B}^* \mapsto \mathbb{B}^*$$

We will now term a prefix code a code whose $C(\mathbb{B}^*)$ is a prefix-free language. This type of code will be very useful to us, notably due to a property demonstrated in the non-ambiguity theorem for prefix codes.

Definition: Prefix code

We call a prefix code a function $C : \Theta \subset \mathbb{B}^* \mapsto \mathbb{B}^*$ satisfying

1. C is a code (i.e., injective).
2. $\{C(x) \mid x \in \Theta\}$ is prefix-free.

The prefix-free nature of a prefix code immediately entails the following property.

Property: Non-existence of a preimage for ε

Let $C : \Theta \subset \mathbb{B}^* \mapsto \mathbb{B}^*$ be a prefix code. Then, for all $x \in \theta$, we have $C(x) \neq \varepsilon$.

Proof. By contradiction, assume that there exists $x \in \mathbb{B}^*$ such that $C(x) = \varepsilon$. Then, for all $y \neq x$, $C(x) \leq_p C(y)$ implies $C(x) = C(y)$. By injectivity, $x = y$, which is a contradiction. Therefore, $C(x) \neq \varepsilon$ for all $x \in \mathbb{B}^*$. ■

0.3.2 Non-ambiguity theorem for prefix codes

Theorem: Non-ambiguity of prefix codes

Let $C : \Theta \subset \mathbb{B}^* \mapsto \mathbb{B}^*$ be a prefix code and $x_1, x_2, \ldots, x_l$ and $y_1, y_2, \ldots, y_{l'}$ all in Θ with $l' \geq l$ [to do: this hypothesis is not necessary]. We then have

$$C(x_1)C(x_2)\ldots C(x_l) \leq_p C(y_1)C(y_2)\ldots C(y_{l'}) \implies \forall i \in \{1,\ldots,l\}, \begin{cases} C(x_i) = C(y_i) \\ x_i = y_i \end{cases}$$

Proof. **Base case**: for $l = 1$
Assume that $C(x_1) \leq_p C(y_1)C(y_2)\ldots C(y_{l'})$. Then, either

$$C(x_1) \leq_p C(y_1), \quad \text{or} \quad C(y_1) \leq_p C(x_1).$$

In each of these cases, since C is a prefix code, we have $C(x_1) = C(y_1)$, and by injectivity of C, this implies $x_1 = y_1$.

Inductive step: assume the result holds for a certain $l \geq 1$ and let us show it for $l + 1$. Assume that

$$C(x_1)C(x_2)\ldots C(x_{l+1}) \leq_p C(y_1)C(y_2)\ldots C(y_{l'})$$

In particular, we have

$$C(x_1)C(x_2)\ldots C(x_l) \leq_p C(y_1)C(y_2)\ldots C(y_{l'})$$

By the induction hypothesis, for all $i \in \{1,\ldots,l\}$, we have $C(x_i) = C(y_i)$. We can then write

$$p = C(x_1)C(x_2)\ldots C(x_l) = C(y_1)C(y_2)\ldots C(y_l)$$

The relation then translates to

$$p\,C(x_{l+1}) \leq_p p\,C(y_{l+1})\ldots C(y_{l'})$$

By removing the common prefix p, we obtain

$$C(x_{l+1}) \leq_p C(y_{l+1})\ldots C(y_{l'})$$

This reduces to the case $l = 1$, hence $C(x_{l+1}) = C(y_{l+1})$. By injectivity of C, we deduce that $x_{l+1} = y_{l+1}$. ■

Corollary: Non-ambiguity of prefix codes

Let $C : \Theta \subset \mathbb{B}^* \mapsto \mathbb{B}^*$ be a prefix code and two sequences $x_1, x_2, \ldots x_l$ and $y_1, y_2, \ldots y_{l'}$ both in Θ. We then have

$$C(x_1)C(x_2)\ldots C(x_l) = C(y_1)C(y_2)\ldots C(y_{l'}) \implies \begin{cases} l = l' \\ \forall i \in [\![1; l]\!], \begin{cases} x_i = y_i \\ C(x_i) = C(y_i) \end{cases} \end{cases}$$

Proof. Suppose $C(x_1)C(x_2)\ldots C(x_l) = C(y_1)C(y_2)\ldots C(y_{l'})$. Then, we have

$$\begin{cases} C(x_1)C(x_2)\ldots C(x_l) \geq_p C(y_1)C(y_2)\ldots C(y_{l'}) \\ C(x_1)C(x_2)\ldots C(x_l) \leq_p C(y_1)C(y_2)\ldots C(y_{l'}) \end{cases}$$

By applying the theorem to each inequality (in the prefix sense), we immediately obtain the result. ■

0.4 Information Theory

Throughout this section, we assume a probability space $(\mathcal{X}, \mathcal{F}, \mathbb{P})$.

0.4.1 Definition: Shannon Entropy

Definition: Shannon Entropy

We call Shannon entropy the function H that associates to any discrete random variable $X : \Omega \mapsto \Xi$ (that is to say Ξ is any finite or countable set) the value

$$H(X) = \sum_{x \in X(\Omega)} \mathbb{P}[X = x] \log \frac{1}{\mathbb{P}[X = x]}$$

taking the convention $0 \log(0) = \lim_{p \to 0} p \log(\frac{1}{p}) = 0$

Remark. We then have, in the case where $X(\Omega) = \{x_1, x_2, \ldots, x_n\}$ is finite (and letting $p_i = \mathbb{P}[X = x_i]$), that

$$H(X) = \sum_{x \in \{x_1, x_2, \ldots, x_n\}} p_i \log \frac{1}{p_i}$$

◇

0.4.2 Combinatorial approach to Shannon entropy

In this theory we ignore the meaning of a message; we are interested only in the problem of communicating a message between a sender and a receiver under the assumption that the universe of possible messages is known to both the sender and the receiver.

This notion of information is a measure of one's freedom of choice when one selects a message. Given the choice of transmitting a message consisting of the contents of this entire book, and the message "let's get a beer," the information concerned is precisely one bit. Obviously this does not capture the information content of the individual object itself. Kolmogorov's intention for introducing algorithmic complexity is as a measure of the information content of individual objects.

Theorem: Combinatorial approach to Shannon entropy

Consider an alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_s\}$ composed of s distinct symbols. We define the following elements:

1. We define for each $i = 1, 2, \ldots, s$ a sequence of strictly positive integers $(m_{i,n})_{n \geq 0}$ such that

$$\forall n \geq 0, \sum_{k=1}^{s} m_{k,n} = n \quad \text{and} \quad \forall k \in [\![1; s]\!], \lim_{n \to \infty} \frac{m_{k,n}}{n} = p_i \in]0; 1[$$

2. Let us introduce a random variable X defined on the alphabet $\mathcal{A}$, following the probability distribution:

$$\forall a_i \in \mathcal{A}, \quad \mathbb{P}[X = a_i] = p_i$$

3. For all $n \geq 0$ we then define the language Λ_n such that [a]

$$\Lambda_n = \{x \in \mathcal{A}^* \mid \forall i \in [\![1; s]\!], |x|_{a_i} = m_{i,n}\}$$

We then have

$$\log(\Lambda_n) \underset{n\to\infty}{\sim} nH(X)$$

[a] i.e., the words of the alphabet $\mathcal{A}$ in which there are exactly $m_1, \ldots, m_s$ occurrences of the letters $a_1, \ldots, a_s$.

Proof. Let $n \geq 0$. To simplify the notation, we will denote in the proof m_i instead of $m_{i,n}$. The number of words of length n in Λ_n is given by the multinomial coefficient

$$|\Lambda_n| = \frac{n!}{m_1!\, m_2! \cdots m_s!}$$

We use Stirling's formula, which ensures that for any integer k tending to infinity,

$$k! \underset{k\to\infty}{\sim} \sqrt{2\pi k}\, \left(\frac{k}{e}\right)^k$$

By taking the logarithm of this formula and since the term $\log\left(\sqrt{2\pi k}\right)$ is negligible compared to $k \log k$ and k when $k \to \infty$, we obtain

$$\log k! \underset{k\to\infty}{\sim} \log\left(\sqrt{2\pi k}\right) + k \log k - k \underset{k\to\infty}{\sim} k \log k - k$$

Let us apply this equivalence to $n!$ and to each of the $m_i!$ (for $i = 1, \ldots, s$):

$$\begin{cases} \log(n!) & \underset{n\to\infty}{\sim} n \log n - n, \\ \log(m_i!) & \underset{n\to\infty}{\sim} m_i \log m_i - m_i. \end{cases}$$

We then deduce

$$\log|\Lambda_n| = \log(n!) - \sum_{i=1}^{s} \log(m_i!)$$

$$\underset{n\to\infty}{\sim} \Big[n \log n - n\Big] - \sum_{i=1}^{s} \Big[m_i \log m_i - m_i\Big].$$

Now we have $\sum_{i=1}^{s} m_i = n$ which implies that the terms in $-n$ cancel out with $-\sum_{i=1}^{s} m_i$, and therefore,

$$\log|\Lambda_n| \underset{n\to\infty}{\sim} n \log n - \sum_{i=1}^{s} m_i \log m_i$$

To relate the obtained expression to probabilities, recall that, by hypothesis, for each $i \in \{1, \ldots, s\}$ we have

$$\lim_{n\to\infty} \frac{m_i}{n} = p_i \quad \text{which is equivalent to} \quad m_i \underset{n\to\infty}{\sim} n\, p_i$$

It follows that

$$m_i \log m_i \underset{n\to\infty}{\sim} n\,p_i \log(n\,p_i) = n\,p_i\Big(\log n + \log p_i\Big)$$

By summing over i from 1 to s and using $\sum_{i=1}^{s} p_i = 1$, we obtain:

$$\sum_{i=1}^{s} m_i \log m_i \underset{n\to\infty}{\sim} n\log n + n\sum_{i=1}^{s} p_i \log p_i$$

By substituting this expression into that of $\log|\Lambda_n|$, we have:

$$\log|\Lambda_n| \underset{n\to\infty}{\sim} n\log n - \Big[n\log n + n\sum_{i=1}^{s} p_i \log p_i\Big] = -n\sum_{i=1}^{s} p_i \log p_i$$

Note finally that the entropy of the random variable X (defined by $\mathbb{P}[X = a_i] = p_i$) is given by

$$H(X) = -\sum_{i=1}^{s} p_i \log p_i$$

We can therefore conclude that, when $n \to \infty$,

$$\log\big(|\Lambda_n|\big) \underset{n\to\infty}{\sim} n\,H(X)$$

■

0.4.3 Source Coding Theorem

Lemma: Gibbs' Inequality

Let $\{p_1, p_2, \ldots, p_N\}$ and $\{q_1, q_2, \ldots, q_N\}$ be all in $\mathbb{R}^+$ such that $\sum_{i=1}^{N} p_i = 1$ and $\sum_{i=1}^{N} q_i \leq 1$. We then have

$$\sum_{i=1}^{N} p_i \log p_i \geq \sum_{i=1}^{N} p_i \log q_i$$

With equality if and only if $\forall i \in [\![1, N]\!], p_i = q_i$. We adopt here the convention $0\log(0) = 0$ and $-\log(0) = \infty$.

Proof. We have $\forall x > 0, \ln(x) \leq 1 - x$ with equality if and only if $x = 1$. Multiplying by -1 gives $\forall x > 0, \ln(\frac{1}{x}) \geq x - 1$ which yields. Let I be the set of i such that $p_i > 0$

$$-\sum_{i\in I} p_i \ln\frac{q_i}{p_i} \geq -\sum_{i\in I} p_i\left(\frac{q_i}{p_i} - 1\right) \geq -\sum_{i\in I}(q_i - p_i) = \sum_{i\in I} p_i - \sum_{i\in I} q_i \geq 0$$

We therefore have, by dividing by $\ln(2)$ and rearranging, that

$$-\sum_{i\in I} p_i \ln\frac{p_i}{q_i} \geq 0 \iff -\sum_{i\in I} p_i \log p_i \geq -\sum_{i\in I} p_i \log q_i$$

Let $\bar{I}$ denote the i in $\{1,\dots,N\}$ such that $i \notin I$. We then have that for all $p \in \bar{I}$, $p = 0$ therefore $p\log(p) = 0$ and $-\log(p) = \infty$. Thus we can add these terms to the sum without modifying the inequality

$$-\sum_{i=1}^{N} p_i \log p_i \geq -\sum_{i=1}^{N} p_i \log q_i$$

Moreover, if for all i in $\{1,\dots,N\}$ we have $\frac{q_i}{p_i} = 1$, that is to say $p_i = q_i$, then the inequalities all transform into equality, which proves the equality case. ■

Definition: Optimal Prefix Code

Let X be a random variable taking values in $\mathcal{A} = \{a_1, a_2, \dots, a_s\}$ with, for all i,

$$p_i = \mathbb{P}[X = a_i]$$

An optimal prefix code for X is a prefix code $C^* : \mathcal{A} \to \mathbb{B}^*$ such that

$$\mathbb{E}[|C^*(X)|] = \min\{\mathbb{E}[|C(X)|] \mid C : \mathcal{A} \to \mathbb{B}^* \text{ is a prefix code}\}$$

We will now introduce Shannon's source coding theorem which defines a fundamental limit to lossless data compression. Roughly speaking, it tells us that for a given information source (represented by a random variable X), there exists a minimum average number of bits per symbol necessary to represent the information optimally. This minimal number is the entropy H(X) of the source! In summary, Shannon's source coding theorem teaches us that, despite all our efforts to compress information, there is an intrinsic limit dictated by the entropy of the source itself.

Theorem: Source coding theorem for symbol codes

Let $\mathcal{A} = \{a_1, a_2, \dots, a_s\}$ be an alphabet and X a random variable taking values in $\mathcal{A}$, with distribution given by

$$p_i = \mathbb{P}[X = a_i], \quad i = 1, \dots, s.$$

For any optimal prefix code $C^* : \mathcal{A} \to \{0,1\}^*$, we then have

$$H(X) \leq \mathbb{E}[|C^*(X)|] \leq H(X) + 1,$$

where $H(X) = -\sum_{i=1}^{s} p_i \log_2 p_i$ denotes the entropy of X.

Proof. Let $C : \mathcal{A} \to \mathbb{B}^*$ be any prefix code and, for each $1 \leq i \leq s$, let us denote $\ell_i = |C(a_i)|$ the length of $C(a_i)$. The proof is divided into two parts.

(1) Lower bound.
For all i, let us define

$$q_i = \frac{2^{-\ell_i}}{S}, \quad \text{with } S = \sum_{i=1}^{s} 2^{-\ell_i}$$

Kraft's inequality guarantees that $S \leq 1$, therefore $\log_2 S \leq 0$. Using Gibbs' inequality, we obtain

$$H(X) = -\sum_{i=1}^{s} p_i \log_2 p_i \leq -\sum_{i=1}^{s} p_i \log_2 q_i$$

Since

$$-\log_2 q_i = -\log_2\left(2^{-\ell_i}\right) + \log_2 S = \ell_i + \log_2 S$$

we have

$$-\sum_{i=1}^{s} p_i \log_2 q_i = \sum_{i=1}^{s} p_i\, \ell_i + \log_2 S \sum_{i=1}^{s} p_i = \mathbb{E}\left[|C(X)|\right] + \log_2 S$$

Given that $\log_2 S \leq 0$, it follows that

$$H(X) \leq \mathbb{E}\left[|C(X)|\right]$$

Since this inequality is valid for any code C, it holds in particular for the optimal code C^*.

(2) Upper bound.
We explicitly construct a prefix code approaching the entropy. For each $i \in \{1, \ldots, s\}$, let us set

$$\ell_i = \lceil -\log_2 p_i \rceil$$

Then,

$$-\log_2 p_i \leq \ell_i < -\log_2 p_i + 1$$

By exponentiating, this yields

$$2^{-\ell_i} \leq p_i$$

Thus,

$$\sum_{i=1}^{s} 2^{-\ell_i} \leq \sum_{i=1}^{s} p_i = 1$$

By the converse of Kraft's inequality, there exists a prefix code $C : \mathcal{A} \to \mathbb{B}^*$ such that $|C(a_i)| = \ell_i$ for all i. For this code, the average length is given by

$$\mathbb{E}\left[|C(X)|\right] = \sum_{i=1}^{s} p_i\, \ell_i < \sum_{i=1}^{s} p_i\Big(-\log_2 p_i + 1\Big) = H(X) + 1$$

As C^* is optimal, we then have

$$\mathbb{E}\left[|C^*(X)|\right] \leq \mathbb{E}\left[|C(X)|\right] < H(X) + 1$$

By combining the two inequalities, we obtain

$$H(X) \leq \mathbb{E}\left[|C^*(X)|\right] < H(X) + 1,$$

which concludes the proof. ■

0.4.4 Shannon-Fano code

Definition-Property: Proper binary expansion

Let $b : [0,1[\mapsto \mathbb{B}^{\mathbb{N}}$ be the function that, for all $r \in [0,1[$, associates the sequence $b(r) = (b_k(r))_{k \in \mathbb{N}^*}$ such that

$$\forall n \in \mathbb{N}^* : \quad S_n = \sum_{i=1}^{n} b_i(r) 2^{-i} \quad \text{satisfies} \quad S_n \leq r < S_n + 2^{-n} \qquad [*]$$

Thus we have b which is defined [a] and injective (see proof).

[a] For each $r \in [0,1[$ we have a unique $b(r)$ satisfying [*]

Proof. **Let us prove that b is defined:** Assume for the sake of contradiction that there exists $r \in [0,1[$ such that $(b_k(r))_{k\in\mathbb{N}^*}$ and $(b'_k(r))_{k\in\mathbb{N}^*}$ are different and satisfy [$*$]. There exists a smallest $k_m > 0$ such that $b_{k_m}(r) \neq b'_{k_m}(r)$, that is to say

$$k_m = \min\{k \geq 1 \mid b_k(r) \neq b'_k(r)\}$$

Moreover, without loss of generality, assume $b_{k_m} = 0$ and $b'_{k_m} = 1$. We then have

$$\begin{cases} r = S'_{k_m-1} + \underbrace{b'_{k_m}2^{-k_m}}_{=1} + \sum_{i=k_m+1}^{\infty} b'_i(r)2^{-i} \\ r = S_{k_m-1} + \underbrace{b_{k_m}2^{-k_m}}_{=0} + \sum_{i=k_m+1}^{\infty} b_i(r)2^{-i} \end{cases}$$

Since for $i < k_m$ we have $b_i(r) = b'_i(r)$, the partial sums up to index $k_m - 1$ coincide, so we have $S'_{k_m-1} = S_{k_m-1}$. By subtracting the expressions of r from one another

$$S'_{k_m} - S_{k_m} = 2^{-k_m} + \sum_{i=k_m+1}^{\infty} (b'_i(r) - b_i(r))2^{-i} = 0 \implies \sum_{i=k_m+1}^{\infty} (b'_i(r) - b_i(r))2^{-i} = 2^{k_m} \quad [\dagger]$$

Moreover, according to [$*$] for $n = k_m$ we have $S_n \leq r < S_{k_m} + 2^{-k_m}$ and $S'_{k_m} \leq r < S'_{k_m} + 2^{-k_m}$ which implies

$$0 \leq r - S_{k_m} < 2^{k_m} \quad \text{and} \quad 0 \leq r - S'_{k_m} < 2^{k_m}$$

Which, by making it explicit, gives

$$0 \leq \sum_{i=k_m+1}^{\infty} b_i(r)2^{-i} < 2^{-k_m} \quad \text{and} \quad 0 \leq \sum_{i=k_m+1}^{\infty} b'_i(r)2^{-i} < 2^{-k_m}$$

Thus we have

$$\sum_{i=k_m+1}^{\infty} (b'_i(r) - b_i(r))2^{-i} < 2^{-k_m}$$

which contradicts [$\dagger$].

Let us prove that b is injective: Let $r, r' \in [0,1[$ be such that

$$b(r) = (b_k(r))_{k\geq 1} = (b'_k(r))_{k\geq 1} = b(r')$$

For all $n \geq 1$, the associated partial sums are then identical:

$$S_n(r) = \sum_{i=1}^{n} b_i(r)2^{-i} = \sum_{i=1}^{n} b'_i(r)2^{-i} = S_n(r')$$

By property $(*)$ applied to r and to r', we have

$$S_n(r) \leq r < S_n(r) + 2^{-n} \quad \text{and} \quad S_n(r) \leq r' < S_n(r) + 2^{-n}$$

In particular,

$$|r - r'| < 2^{-n} \quad \text{for all } n \geq 1$$

Now $\lim_{n\to\infty} 2^{-n} = 0$, therefore $r = r'$ which demonstrates the injectivity of b. ■

Definition: Shannon-Fano Code

Consider $\mathcal{A} = \{a_1, a_2, \ldots, a_s\}$ an alphabet and $p_1 \geq p_2 \geq \ldots \geq p_s > 0$ such that $p_1 + \ldots + p_s = 1$. Let X be a random variable on $\mathcal{A}$ such that

$$\forall i \in \{1, \ldots, s\}, \quad \mathbb{P}[X = a_i] = p_i > 0 \quad \text{and let us set} \quad \forall r \in \{1, \ldots, s\}, \quad P_r = \sum_{k=1}^{r-1} p_k$$

We then call a Shannon–Fano code for X the code $C : \mathcal{A} \to \mathbb{B}^*$ such that

$$\forall i \in \{1, \ldots, s\} : \qquad C(a_i) := b_1 b_2 \ldots b_{l_i} \qquad \text{where} \begin{cases} l_i = \lceil \log \frac{1}{p_i} \rceil \\ b(P_i) = (b_k)_{k \in \mathbb{N}^*} \end{cases}$$

That is to say that $C(a_i)$ corresponds to the first l_i bits of the proper binary expansion of $b(P_i)$.

Theorem: Shannon-Fano code is a prefix code / Shannon-Fano source code

Let us reconsider the Shannon-Fano code C for the random variable X as defined previously, we then have

1. C is a prefix code.
2. $H(X) \leq \mathbb{E}[|C(X)|] \leq H(X) + 1$ [Shannon-Fano source code]

Proof. Let C be a Shannon-Fano code for a random variable X as defined in the previous definition.

1. C is a prefix code: Let a_i, a_j in $\mathcal{A}$ be such that $i \neq j$. Without loss of generality assume $j > i$. Let us set

$$\begin{cases} C(a_i) = b_1 b_2 \ldots b_{l_i} \\ C(a_j) = b'_1 b'_2 \ldots b'_{l_j} \end{cases} \qquad \text{where} \quad l_i = \lceil \log \frac{1}{p_i} \rceil \text{ and } l_j = \lceil \log \frac{1}{p_j} \rceil$$

We notice that

$$\begin{aligned} l_i = \lceil \log \frac{1}{p_i} \rceil \implies & \log \frac{1}{p_i} \leq l_i \leq \log \frac{1}{p_i} + 1 \qquad [\diamond] \\ \implies & -l_i \leq -\log \frac{1}{p_i} \leq -l_i + 1 \implies 2^{-l_i} \leq p_i \leq 2^{-l_i+1} \end{aligned}$$

We can do the same with $l_j = \lceil \log \frac{1}{p_j} \rceil$ which gives

$$2^{-l_j} \leq p_j \leq 2^{-l_j+1}$$

Let us show that there exists a $1 \leq k \leq l_i$ such that $b_k \neq b'_k$, which will show that $C(a_i) \not\preceq_p C(a_j)$. For this, let us remark that:

$$P_j - P_i = \sum_{k=i}^{j-1} p_k \geq (j - i)\, p_i \geq 2^{l_i} \qquad [\sharp]$$

Now we have

$$P_j - P_i = \sum_{k=1}^{\infty} (b'_k - b_k)\, 2^{-k} = \sum_{k=1}^{l_i} (b'_k - b_k)\, 2^{-k} + \sum_{k=l_i+1}^{\infty} (b'_k - b_k)\, 2^{-k}$$

By definition of the proper binary expansion we have $\sum_{k=l_i+1}^{\infty}(b'_k - b_k)2^{-k} < 2^{-l_i}$. Thus we necessarily have by [♯] that

$$\sum_{k=1}^{l_i}(b'_k - b_k)\,2^{-k} > 0$$

We necessarily have that there exists a k such that $b'_k - b_k \neq 0$, which means we have $C(a_i) \not\preceq_p C(a_j)$. Thus C is a prefix code.

2. Shannon–Fano source code: This is immediate using [⋄]. Indeed

$$H(X) = \sum_i p_i \log_2 \frac{1}{p_i} \leq 1 + \sum_i p_i \log_2 \frac{1}{p_i} = H(X) + 1$$

■