# Information theory and coding

Bartek Blaszczyszyn
Notes by Antoine Groudiev

5th 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.

## 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:

$$H(p) := -\sum_{x\in\mathcal{X}} p(x) \log p(x) \tag{1}$$

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.

**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:

$$H(X) := H(p) = -\mathbb{E}(\log p(X)) \tag{2}$$

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:*

$$H(p) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \leqslant -\sum_{x\in\mathcal{X}} p(x)\log q(x) \tag{3}$$

*Moreover, if $H(p) < \infty$, then there is equality in (3) if and only if $p = q$.*

The right-hand-side of (3) 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:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\mathbb{E}[\log p(X)] \tag{4}$$

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

$$H(X) = \sum_{i=1}^{n} H(X_i) \tag{5}$$

*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:*

$$H(X) \leqslant \sum_{i=1}^{n} H(X_i) \tag{6}$$

*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 $\mathcal{X}^n$

$$A_\varepsilon := \left\{ x \in \mathcal{X}^n \ : \ \Big| -\frac{1}{n} \sum_{i=1}^n \log_D p(x_i) - H_D \Big| \leqslant \varepsilon \right\} \subseteq \mathcal{X}^n \tag{7}$$

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

**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 $\mathcal{X}$, with $D := |\mathcal{X}| < \infty$. We have:*

$$\lim_{n \to +\infty} \mathbb{P}(X \in A_\varepsilon) = 1 \tag{8}$$

*and*

$$|A_\varepsilon| \leqslant D^{n(H_D + \varepsilon)} \tag{9}$$

*Proof.* (8) follows from the LLN. For (9), observe that:

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

and since:

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

which completes the proof. □

**Property** ($A_\varepsilon$ 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 $A_\varepsilon$ is the smallest set concentrating probability.*

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

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

Therefore,

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

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