A short introduction to information theory and Bayesian Model Averaging

Olivier Catoni

March 22, 2012

1. Information theory and lossless codes

1.1. Binary codes. Let us consider a finite alphabet A and a random sequence $(X_n)_{n=1}^{\infty}$ taking its values in A. (The alphabet can be any finite set here.)

Let $\{0,1\}^* = \bigcup_{n=1}^{\infty} \{0,1\}^n$ be the set of finite binary sequences.

Definition 1.1 *Given some block length n, a binary code c is an injective map from A^n to $\{0,1\}^n$. The mean length of c is*

$$\mathbb{E}\big\{\ell\big[c(X^n)\big]\big\},$$

where the expectation is taken with respect to the distribution of the sequence $X^n \stackrel{\text{def}}{=} (X_1, \dots, X_n)$ and where the length function ℓ is defined as $\ell(w) = k$ for any $w \in \{0,1\}^k$.

Minimizing the mean code length under various assumptions on the source and the set of authorized codes is the main subject of lossless coding theory.

1.2. Optimal code length for a known source. In the case when $\mathbb{P}_{X^n}$ (the distribution of $X^n = (X_1, \dots, X_n)$) is known and c is arbitrary, minimizing the code length is achieved by sorting A^n in order of decreasing probabilities. More precisely, let us write $A^n = \{b_i, i = 1, \dots, d^n\}$, where $d = |A|$ is the size of the alphabet and where the blocks of n letters are indexed by order of decreasing probabilities:

$$\mathbb{P}_{X^n}(b_i) \leq \mathbb{P}_{X^n}(b_{i-1}), i = 2, \dots, d^n.$$

Let us sort $\{0,1\}^*$ by order of increasing lengths, writing

$$\{0,1\}^* = \big\{w(j), j \in \mathbb{N} \setminus \{0\}\big\},$$

CNRS – UMR 8553, Département de Mathématiques et Applications, Ecole Normale Supérieure, 45, rue d'Ulm, F75230 Paris cedex 05, and INRIA Paris-Rocquencourt – CLASSIC team.

where $w(1) = 0$, $w(2) = 1$, $w(3) = 00$, $w(4) = 01$, ... and more generally

$$w\Big(2^\ell + \sum_{k=1}^{\ell} d_{\ell-k} 2^k - 1\Big) = \big(d_k\big)_{k=1}^{\ell},$$

where $d_k \in \{0,1\}$ (this means that $w(i)$ is the binary representation of $i+1$ read from left to right, after removing the leftmost bit, which is always equal to one). Let us remark that $\ell\big[w(i)\big] = \lfloor \log_2(i+1) \rfloor$.

Proposition 1.1 *The code $c(b_i) = w(i)$ minimizes the mean code length among all binary codes.*

Proof. Let c' be some other code. Its mean length can be written as

$$\mathbb{E}\Big\{\ell\big[c'(X^n)\big]\Big\} = \sum_{k=1}^{\infty} \mathbb{P}\Big\{\ell\big[c'(X^n)\big] \geq k\Big\}. \tag{1}$$

On the other hand, since $i \mapsto \ell\big[w(i)\big]$ is non-decreasing, for any code length k,

$$\big\{i \in \mathbb{N} \setminus \{0\} : \ell\big[w(i)\big] < k\big\} = [\![1, i_k]\!],$$

where $i_k = 2^k - 2$ is the number of binary codes of length less than k, and $[\![1, i_k]\!] \stackrel{\text{def}}{=} \{1, \ldots, i_k\}$. Thus

$$\mathbb{P}\Big\{\ell\big[c'(X^n)\big] < k\Big\} = \mathbb{P}_{X^n}\Big\{c'^{-1}\big[w\big([\![1, i_k]\!]\big)\big]\Big\}.$$

As c' is an injective map, $\big|c'^{-1}\big[w\big([\![1, i_k]\!]\big)\big]\big| \leq i_k$, and can thus be written as $\big\{b_{j_1}, \ldots, b_{j_m}\big\}$, where $j_1 < j_2 < \cdots < j_m$ and $m \leq i_k$. Thus

$$\begin{aligned}
\mathbb{P}\Big\{\ell\big[c'(X^n)\big] < k\Big\} &= \mathbb{P}_{X^n}(b_{j_1}) + \cdots + \mathbb{P}_{X^n}(b_{j_m}) \\
&\leq \mathbb{P}_{X^n}(b_1) + \cdots + \mathbb{P}_{X^n}(b_m) \\
&\leq \mathbb{P}_{X^n}(b_1) + \cdots + \mathbb{P}_{X^n}(b_{i_k}) = \mathbb{P}\Big\{\ell\big[c(X^n)\big] < k\Big\},
\end{aligned}$$

since $s \leq j_s$, $i \mapsto \mathbb{P}_{X^n}(b_i)$ is non-increasing and $m \leq i_k$. This proves that

$$\mathbb{P}_{X^n}\Big\{\ell\big[c'(X^n)\big] \geq k\Big\} \geq \mathbb{P}_{X^n}\Big\{\ell\big[c(X^n)\big] \geq k\Big\},$$

and therefore according to equation (1) that the mean code length of c is optimal. □

The mean length of optimal codes is related to the Shannon entropy $H(\mathbb{P}_{X^n})$ of the distribution of X^n.

Definition 1.2 *The Shannon entropy $H(p)$ of a probability measure $p \in \mathcal{M}_+^1(\mathcal{X})$ on a finite set $\mathcal{X}$ is defined as*

$$H(p) = -\sum_{x\in\mathcal{X}} p(x)\log_2\big[p(x)\big],$$

where $\log_2(z) = \log(z)/\log(2)$. The entropy is measured in bits, for reasons that will become clear later.

Another quantity of interest is the Kullback Leibler divergence, or relative entropy, between two probability measures P and Q.

Definition 1.3 *Let P and $Q \in \mathcal{M}_+^1(\mathcal{X})$ be two probability measures defined on some measurable space $\mathcal{X}$ (that needs not be finite). The Kullback Leibler divergence of P with respect to Q is defined as*

$$\mathcal{K}(P,Q) = \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P & \text{when } P \ll Q, \\ +\infty & \text{otherwise.} \end{cases}$$

Let us remark that when $P \ll Q$, $\log\left(\dfrac{\mathrm{d}P}{\mathrm{d}Q}\right)$ has an integrable negative part, because

$$\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)_- \mathrm{d}P = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)_- \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q < \infty,$$

where $z_- = \max\{-z, 0\}$, due to the fact that $z \mapsto z\log(z) : \mathbb{R}_+ \to \mathbb{R}$ is bounded from below (by $-1/e$). Thus the integral $\displaystyle\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P$ is defined as a generalized integral, taking values in $\mathbb{R}\cup\{+\infty\}$.

Exercise 1 *Give an example, where $P \ll Q$ and $\mathcal{K}(P,Q) = +\infty$.*

Proposition 1.2 *The Kullback Leibler divergence is a non-negative function.*

Proof. When $P \ll Q$, we can write the divergence as

$$\mathcal{K}(P,Q) = \int 1 - \frac{\mathrm{d}P}{\mathrm{d}Q} + \frac{\mathrm{d}P}{\mathrm{d}Q}\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q.$$

Since the function $z \mapsto 1 - z + z\log(z) : \mathbb{R}_+ \to \mathbb{R}$ has a positive range, it shows that $\mathcal{K}(P,Q) \in \mathbb{R}_+ \cup \{+\infty\}$. □

Exercise 2 *Show that for any $P_1, P_2, Q \in \mathcal{M}_+^1(\mathcal{X})$, any $\lambda \in [0,1]$,*

$$\mathcal{K}\big(\lambda P_1 + (1-\lambda)P_2, Q\big) \leq \lambda \mathcal{K}(P_1, Q) + (1-\lambda)\mathcal{K}(P_2, Q),$$
$$\mathcal{K}(Q, \lambda P_1 + (1-\lambda)P_2) \leq \lambda \mathcal{K}(Q, P_1) + (1-\lambda)\mathcal{K}(Q, P_2).$$

Proposition 1.3 *The map $p \mapsto H(p) : \mathcal{M}_+^1(\mathcal{X}) \to \mathbb{R}_+$ is concave.*

Proof. The function $z \mapsto z \log_2(z)$ is convex. □

Proposition 1.4 *The Shannon entropy is sub-additive: For any random sequence X^{n+m},*

$$H(\mathbb{P}_{X^{n+m}}) \leq H(\mathbb{P}_{X^n}) + H(\mathbb{P}_{X_{n+1}^{n+m}}),$$

where $X_{n+1}^{n+m} \stackrel{\text{def}}{=} (X_{n+1}, \ldots, X_{n+m})$.

Proof. It is enough to prove that for any couple (X, Y) of random variables,

$$H(\mathbb{P}_{X,Y}) \leq H(\mathbb{P}_X) + H(\mathbb{P}_Y).$$

We can then remark that

$$\begin{aligned} H(\mathbb{P}_{X,Y}) &= -\int \log_2\big[\mathbb{P}_{X,Y}(x,y)\big]\, \mathrm{d}\mathbb{P}_{X,Y}(x,y) \\ &= -\int \log_2\big[\mathbb{P}_X(x)\big]\, \mathrm{d}\mathbb{P}_X(x) - \int \log_2\big[\mathbb{P}_{Y|X=x}(y)\big]\, \mathrm{d}\mathbb{P}_{Y|X=x}(y)\, \mathrm{d}\mathbb{P}_X(x) \\ &= H(\mathbb{P}_X) + \int H\big(\mathbb{P}_{Y|X=x}\big)\, \mathrm{d}\mathbb{P}_X(x) \\ &\leq H(\mathbb{P}_X) + H\left(\int \mathbb{P}_{Y|X=x}\, \mathrm{d}\mathbb{P}_X(x)\right) = H(\mathbb{P}_X) + H(\mathbb{P}_Y), \end{aligned}$$

where we have used the fact that H is concave. □

Definition 1.4 *A random sequence $(X_n)_{n=1}^{+\infty}$ is said to be stationary when, for any $n, m \in \mathbb{N} \setminus \{0\}$, $\mathbb{P}_{X_{n+1}^{n+m}} = P_{X^m}$.*

Proposition 1.5 *If $(X_n)_{n=1}^{+\infty} \stackrel{\text{def}}{=} X^{+\infty}$ is a stationary source, the following limit exists*

$$\lim_{n \to +\infty} \frac{1}{n} H(\mathbb{P}_{X^n}) = \inf_{n=1,\ldots,+\infty} \frac{1}{n} H(\mathbb{P}_{X^n}) \stackrel{\text{def}}{=} \overline{H}(\mathbb{P}_{X^{+\infty}}).$$

It is called the Shannon entropy of $X^{+\infty}$.

Proof. Let $h(n) = n^{-1}H(\mathbb{P}_{X^n})$. From the previous proposition $h(n+m) \leq \frac{nh(n)+mh(m)}{n+m}$. Let $n = pq + r$ where $0 \leq r < p$ be the result of the Euclidean division of n by p. We obtain

$$h(n) \leq \frac{pqh(pq) + rh(r)}{pq+r} \leq \frac{pqh(p) + rh(r)}{pq+r} \leq h(p) + \frac{p-1}{n}\max\bigl\{h(1),\dots,h(p-1)\bigr\}.$$

Thus $\limsup_n h(n) \leq h(p)$, for any $p \in \mathbb{N}\backslash\{0\}$. This implies that $\limsup_n h(n) \leq \inf_n h(n) \leq \liminf_n h(n)$, proving that $\lim_n h(n)$ exists and is equal to $\inf_n h(n)$. □

Proposition 1.6 *For any source, any optimal code c,*

$$(1-n^{-1})\bigl[H(\mathbb{P}_{X^n}) - \log_2(n-1)\bigr] - 1 \leq \sup_{\alpha>1}\alpha^{-1}\bigl[H(\mathbb{P}_{X^n}) + \log_2(\alpha-1)\bigr] - 1 \leq \mathbb{E}_{X^n}\bigl[\ell\bigl(c(X^n)\bigr)\bigr] \leq H(\mathbb{P}_{X^n}) + 1.$$

Consequently, if the source is stationary,

$$\lim_{n\to+\infty}\frac{1}{n}\mathbb{E}\bigl[\ell\bigl(c(X^n)\bigr)\bigr] = \overline{H}(\mathbb{P}_{X^{+\infty}}).$$

Proof. Let $\mathbb{P}_{X^n}(b_i) \stackrel{\text{def}}{=} p(b_i)$. As

$$ip(b_i) \leq \sum_{j=1}^{i} p(b_j) \leq 1,$$

$p(b_i) \leq i^{-1}$. Since all optimal codes have the same mean length, we can choose the optimal code of Proposition 1.1 (page 2). We get

$$\mathbb{E}\bigl[\ell\bigl(c(X^n)\bigr)\bigr] = \sum_{i=1}^{d^n} p(b_i)\lfloor\log_2(i+1)\rfloor \leq \sum_i p(b_i)\log_2\bigl(p(b_i)^{-1}+1\bigr) \leq 1 - \sum_i p(b_i)\log_2\bigl(p(b_i)\bigr) = H(p) + 1.$$

On the other hand,

$$\mathbb{E}\big[\ell\big(c(X^n)\big)\big] \geq \sum_i p(b_i)\big[\log_2(i+1)-1\big]$$
$$\geq -\alpha^{-1}\sum_i p(b_i)\log_2\left(\frac{\alpha-1}{(i+1)^\alpha}\right) - 1 + \frac{\log_2(\alpha-1)}{\alpha}.$$

Since $\displaystyle\sum_i \frac{\alpha-1}{(i+1)^\alpha} \leq (\alpha-1)\int_1^{+\infty}\frac{dz}{z^\alpha} = 1$, there is a probability measure $q \in \mathcal{M}_+^1(D^n)$ such that $q(b_i) \geq \dfrac{\alpha-1}{(i+1)^\alpha}$. We can then remark that

$$-\sum_i p(b_i)\log_2\left(\frac{\alpha-1}{(i+1)^\alpha}\right) - H(p) \geq \log(2)^{-1}\mathcal{K}(p,q) \geq 0.$$

This proves the second inequality of the proposition. The first one is obtained by choosing $\alpha = \big(1-n^{-1}\big)^{-1}$. □

1.3. Instantaneous codes. When transmitting $c(X^n)$ through a channel, it is useful for the receiver to know without delay when the message ends. More precisely, if $c(A^n) = W$, the receiver would like to dispose of a function $f : \{0,1\}^* \to \{0,1\}$ such that $\mathbb{1}\big(i = \ell(w)\big) = f\big(w(1),\dots,w(i)\big)$. If such a function exists, it is necessarily equal to $\mathbb{1}\big(w(1,\dots,i) \in W\big)$, so that in this case

$$\mathbb{1}\big(i = \ell(w)\big) = \mathbb{1}\big(w(1,\dots,i) \in W\big), \quad w \in W, 1 \leq i \leq \ell(w),$$

which can also be written as

$$w(1,\dots,i) \notin W, \qquad w \in W, 1 \leq i < \ell(w).$$

A code having this property is called an instantaneous code, and the previous reasoning proves

Proposition 1.7 *The binary code $c : A^n \to \{0,1\}^*$ is an instantaneous code if and only if it is a prefix code, which means that no codeword is the prefix of another codeword. In this case, the range $c(A^n) = W \subset \{0,1\}^*$ of c is called a prefix set.*

Proposition 1.8 (Kraft inequality) *For any prefix set $W \subset \{0,1\}^*$,*

$$\sum_{w\in W} 2^{-\ell(w)} \leq 1.$$

Proof. Consider $a(w) = \sum_{i=1}^{\ell(w)} w(i)\, 2^{-i}$ and $b(w) = a(w) + 2^{-\ell(w)}$. Let us write $w \prec w'$ when $w(1,\dots,\ell(w)) = w'(1,\dots,\ell(w))$, that is when w is a prefix of w'. To each finite binary word $w \in \{0,1\}^*$ corresponds an interval $I(w) = [a(w), b(w)[\subset [0,1[$. It is easy to see that $I(w') \subset I(w)$ when $w \prec w'$ and that $I(w') \cap I(w) = \varnothing$ otherwise. Thus W is a prefix set if and only if the intervals $I(w)$ are disjoint. In this case, considering the Lebesgue measure λ, we see that $1 = \lambda([0,1[) \geq \sum_{w\in W} \lambda\big(I(w)\big) = \sum_{w\in W} 2^{-\ell(w)}$. □

Proposition 1.9 (Inverse Kraft inequality) *If*

$$\sum_{i=1}^{T} 2^{-r_i} \leq 1, \quad (r_i, i = 1,\dots,T) \in \mathbb{N}^T,$$

then there is a prefix set $\{w_1,\dots,w_T\}$ *such that* $\ell(w_i) = r_i$.

Proof. Let us sort r_i so that $r_{i-1} \leq r_i$, $i = 2,\dots,T$. Let

$$\alpha_i = \sum_{j=1}^{i-1} 2^{-r_j} = \sum_{k=1}^{r_i} w_i(k)\, 2^{-k}, \text{ where } w_i \in \{0,1\}^{r_i}$$

is the binary representation of α_i up to precision 2^{-r_i}. As $a(w_i) = \alpha_i$ and $b(w_i) = \alpha_{i+1}$, it is clear that $I(w_i) \cap I(w_j) = \varnothing$, for any $i \neq j$, proving that $\{w_i, i = 1,\dots,T\}$ is indeed a prefix set. □

Definition 1.5 (Complete prefix code and prefix set) *A prefix set W is said to be complete if no other prefix set W' is such that $W \subsetneq W'$. A prefix code is complete if it uses a complete prefix set of codewords.*

Proposition 1.10 (Kraft equality) *A prefix set $W \subset \{0,1\}^*$ is complete if and only if*

$$\sum_{w\in W} 2^{-\ell(w)} = 1.$$

Proof. We are going to show that W is *not* complete if and only if $\sum_{w\in W} 2^{-\ell(w)} < 1$. If W is not complete, $W \subsetneq W'$, where W' is another prefix set, so that $\sum_{w\in W} 2^{-\ell(w)} < \sum_{w\in W'} 2^{-\ell(w)} \leq 1$, and $\sum_{w\in W} 2^{-\ell(w)} < 1$. On the other hand, if

$$\sum_{w\in W} 2^{-\ell(w)} < 1, \text{ then } \bigcup_{w\in W} I(w) \subsetneq [0,1[.$$

Let $L = \max\{\ell(w), w \in W\}$ and consider the partition of $[0,1[$ made by the intervals $\big\{J(k) = \big[k2^{-L}, (k+1)2^{-L}\big[,\ k = 0, \dots, 2^L - 1\big\}$. Since for any $w \in W$, $I(w) = \bigcup_{k \in K(w)} J(k)$, we see that $\bigcup_{w \in W} K(w) \subsetneq \{0, \dots, 2^L - 1\}$. Therefore, there is $k \in \{0, \dots, 2^L - 1\}$ such that $J(k) \notin \bigcup_{w \in W} I(w)$. Define $w' \in \{0,1\}^L$ by the formula $k2^{-L} = \sum_{i=1}^{L} w'(i)2^{-i}$. Since $I(w') = J(k)$, $W \cup \{w'\}$ forms a prefix set, hence W is not complete. □

Proposition 1.11 *For any prefix code c, and any random source $X^n \in A^n$,*

$$\mathbb{E}\big[\ell\big(c(X^n)\big)\big] \geq H(\mathbb{P}_{X^n}).$$

Moreover, there exists a prefix code c such that

$$\mathbb{E}\big[\ell\big(c(X^n)\big)\big] < H(\mathbb{P}_{X^n}) + 1.$$

Proof. Due to the Kraft inequality, we can find a probability measure $Q \in \mathcal{M}_+^1(A^n)$ such that $Q(x^n) \geq 2^{-\ell(c(x^n))}$, for any $x^n \in A^n$. We can then remark that

$$\mathbb{E}\big[\ell\big(c(X^n)\big)\big] - H(\mathbb{P}_{X^n}) \geq \mathcal{K}(P, Q)/\log(2) \geq 0,$$

as proved in Proposition 1.2 (page 3). As for the second part of the proposition, let us put

$$r(x^n) = \Big\lceil -\log_2\big[\mathbb{P}_{X^n}(x^n)\big] \Big\rceil.$$

Since $\sum_{x^n \in A^n} 2^{-r(x^n)} \leq 1$, there is, according to Proposition 1.9 (page 7), a prefix code c such that $\ell\big(c(x^n)\big) = r(x^n) < -\log_2\big(\mathbb{P}_{X^n}(x^n)\big) + 1$, for any $x^n \in A^n$, and consequently such that $\mathbb{E}\big[\ell\big(c(X^n)\big)\big] < H(\mathbb{P}_{X^n}) + 1$. □

Proposition 1.12 (Huffman code) *Let $p \in \mathcal{M}_+^1\big(\{1, \dots, d\}\big)$ be a probability measure on the d first integers. Let i and $j \leq d$ be such that $p(i)$ and $p(j)$ are the smallest, meaning that*

$$\max\{p(i), p(j)\} \leq \min\{p(k), k \neq i, k \neq j\}.$$

Let c be some optimal prefix code for the probability vector

$$\big(p(k), k \notin \{i, j\}, p(i) + p(j)\big)$$

of length $d-1$, that is some prefix code with minimum mean code length. Let $c' : \{1, \dots, d\} \to \{0,1\}^*$ be the prefix binary code defined as

$$c'(k) = \begin{cases} c(k), & k \notin \{i, j\}, \\ \big(c(i,j), 0\big), & k = i, \\ \big(c(i,j), 1\big), & k = j. \end{cases}$$

Then, c' is an optimal code for the source with probability p.

Proof. Let $c'' : \{1,\dots,d\} \to \{0,1\}^*$ be some optimal prefix code and let $W = c''(\{1,\dots,d\})$. Let us choose $w \in W$ such that $\ell(w) = \max\big\{\ell(v), v \in W\big\}$. Let $w' = \big(w(1,\dots,\ell(w)-1), 1 - w(\ell(w))\big)$ be its brother. If $w' \notin W$, then $w(1,\dots,\ell(w)-1)$ would be the prefix of w only, so that it would be possible to replace w with it and to decrease the mean code length of c'', leading to a contradiction. Thus necessarily $w' \in W$ also. We can assume that $c''(i) = w$ and $c''(j) = w'$, indeed, if this is not the case, we can exchange w and $c''(i)$ and w' and $c''(j)$ without increasing the mean length of c''. We can then consider the prefix code c''' on $\{k \notin \{i,j\}, 1 \le k \le d\} \cup \{(i,j)\}$ defined as

$$c'''(k) = \begin{cases} c''(k), & k \notin \{i,j\}, \\ w\big(1,\dots,\ell(w)-1\big), & k = (i,j). \end{cases}$$

It is easy to see that

$$\begin{aligned} \mathbb{E}\big[\ell(c'')\big] &= \mathbb{E}\big[\ell(c''')\big] + p(i) + p(j), \\ \mathbb{E}\big[\ell(c')\big] &= \mathbb{E}\big[\ell(c)\big] + p(i) + p(j). \end{aligned}$$

Since c is optimal, $\mathbb{E}\big[\ell(c)\big] \le \mathbb{E}\big[\ell(c''')\big]$, showing that $\mathbb{E}\big[\ell(c')\big] \le \mathbb{E}\big[\ell(c'')\big]$, so that c' is also optimal. □

Exercise 3 *Find an optimal code for the probability vector* $(1/3, 1/3, 1/6, 1/6)$.

1.4. Arithmetic codes. The previous section shows that for prefix codes as well, finding an optimal code requires to sort A^n according to decreasing probabilities. If n is large, this may induce intensive computations. Arithmetic codes are almost optimal and do not share this weakness.

Definition 1.6 *A non ordered vector of probabilities* $p(1,\dots,d)$ *being given, let us define*

$$\xi_i = \sum_{j=1}^{i-1} p(j), \quad i = 1,\dots,d+1.$$

Let us then consider the intervals $J(i) = \big[\xi_i, \xi_{i+1}\big[$ *and define*

$$w_i \in \arg\max_w \big\{\lambda\big(I(w)\big); w \in \{0,1\}^*, I(w) \subset J(i)\big\}.$$

The code $c(i) = w_i$ *is called an arithmetic, or Shannon-Fano-Elias code, and satisfies* $\mathbb{E}[\ell(c)] < H(p) + 2$.

Proof. The code c is prefix because the intervals $J(i)$ being non overlapping, this is also the case for $I(w)$, $w \in c\big(\{1,\dots,d\}\big)$. Let us define $L = \big\lceil 1 - \log_2\big[p(i)\big]\big\rceil$ and let

$$w = \arg\min_w \big\{a(w), w \in \{0,1\}^*, \ell(w) = L, a(w) \geq \xi_i\big\}.$$

Since

$$b(w) = a(w) + 2^{-L} < \xi_i + 2^{-(L-1)} \leq \xi_i + p(i) = \xi_{i+1},$$

$I(w) \subset J(i)$ and therefore $\ell\big[c(i)\big] \leq \ell(w) < 2 - \log_2\big(p(i)\big)$, implying that $\mathbb{E}[\ell(c)] < H(p) + 2$. $\square$ Arithmetic coding of $x^n \in A^n$ is fast when A^n is sorted in lexicographic order. Indeed,

$$\begin{aligned}\xi(x^n) &= \sum_{k=1}^{n} \sum_{y<x_k} p(x^{k-1}, y) \\ &= \sum_{k=1}^{n} \sum_{y<x_k} \prod_{j=1}^{k-1} p(x_j \,|\, x^{j-1}) p(y \,|\, x^{k-1}),\end{aligned}$$

which can be computed from at most dn conditional probabilities performing at most dn additions and dn multiplications.

Exercise 4 *Write a program that draws a probability vector $p \in \mathcal{M}_+^1(\{1,\dots,d\})$ at random according to the uniform probability measure on the simplex. (Hint: draw $(X_1,\dots,X_{d-1})$ i.i.d. according to the uniform measure on the interval $[0,1]$, set $X_0 = 0$, $X_d = 1$, consider the order statistics $X_{(i)}$ such that $\{X_i\} = \{X_{(i)}\}$ and $X_{(0)} \leq \cdots \leq X_{(d)}$ and put $p_i = X_{(i)} - X_{(i-1)}$.) Compute an optimal prefix code for p (following Huffman's algorithm). Compute also an arithmetic prefix code. Print the mean code lengths of the two codes, as well as the Shannon entropy of p, for various draws of p and various values of d.*

References

[1] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. Pages 1–269.

[2] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley and Sons, New York, second edition, 2006.