2. Universal codes and Bayesian model averaging

We have seen that prefix codes were closely related to probability distributions, due to the Kraft inequality. From now on, we will identify the two, and consider any probability measure on A^n as an *ideal code*. In the previous section, we described some ways of coding efficiently a source X^n whose distribution was known. Here, we are going to study the case of an unknown distribution.

Definition 2.1 *The performance of an ideal code $Q \in \mathcal{M}_+^1(A^n)$ applied to a block X^n of length n of a source with probability law $\mathbb{P}_{X^n}$, is measured, in bits, by its redundancy*

$$\mathcal{R}\ \mathbb{P}_{X^n}, Q\ = \log(2)^{-1}\mathcal{K}\ \mathbb{P}_{X^n}, Q\ = \mathbb{E}_{X^n}\ -\log_2\ Q(X^n)\ \ -H\ \mathbb{P}_{X^n}$$

More precisely, the redundancy measures the difference between the mean length of a prefix code built according to Q and the optimal code built using $\mathbb{P}_{X^n}$, up to the discretization errors of 1 (Huffman) or 2 (Shannon-Fano-Elias) bits introduced by the actual coding algorithms.

2.1. Bayes redundancy.

One important approach to choose Q when $\mathbb{P}_{X^n}$ is unknown is provided by the Bayesian approach.

Definition 2.2 *Let π be some probability measure on some measurable set of parameters Θ. Let $P_\theta \in \mathcal{M}_+^1(A^n); \theta \in \Theta$ be a family of measures on A^n, where A is some finite set (extensions to more general measurable sets being possible). Let us assume that $\theta \mapsto P_\theta(x^n)$ is measurable for any $x^n \in A^n$. The mean Bayesian redundancy with respect to π is defined as*

$$\mathcal{R}_\pi(Q) = \quad \mathcal{R}\ P_\theta, Q\ \mathrm{d}\pi(\theta).$$

Thus, the Bayesian redundancy measures the average loss of efficiency of the coding distribution Q, when the source distribution is drawn from $\{P_\theta; \theta \in \Theta\}$ according to the "prior" probability measure π.

Proposition 2.1 *The infimum $\inf_{Q \in \mathcal{M}_+^1(A^n)} \mathcal{R}_\pi(Q)$ is reached for only one coding distribution Q, called the bayesian mixture ideal code, and defined as*

$$Q = \quad P_\theta\, \mathrm{d}\pi(\theta) \stackrel{\mathrm{def}}{=} P_\pi.$$

Proof. The Bayesian redundancy is composed of

$$\begin{aligned}\mathcal{R}_\pi(Q) = & \quad \mathcal{R}\ P_\theta, P_\pi\ \mathrm{d}\pi(\theta) + \mathcal{R}\ P_\pi, Q \\ & = \mathcal{R}_\pi\ P_\pi\ + \mathcal{R}\ P_\pi, Q\ .\end{aligned}$$

The conclusion comes from the fact that $\mathcal{R}(P,Q) > 0$ for $P \neq Q$ and $\mathcal{R}(P,P) = 0$, because it is proportional to the Kullback Leibler divergence. □

Definition 2.3 *The mutual information of the joint distribution of two random variables X and Y defined on the same probability space is*

$$\begin{aligned}\mathcal{I}\ \mathbb{P}_{X,Y}\ & = \log(2)^{-1}\mathcal{K}\ \mathbb{P}_{X,Y}, \mathbb{P}_X \otimes \mathbb{P}_Y \\ & = \log(2)^{-1}\mathbb{E}_X\ \mathcal{K}(\mathbb{P}_{Y|X}, \mathbb{P}_Y)\ .\end{aligned}$$

Proposition 2.2 *The optimal Bayesian redundancy, achieved by the Bayesian ideal code, is equal to the mutual information between the parameter θ and the source $(X_1, \ldots, X_n)$, when their joint distribution is defined as*

$$\mathbb{P}_{(\theta,X^n)}(B) = \quad \mathbb{1}\ (\theta, x^n) \in B\ \mathrm{d}\pi(\theta)\mathrm{d}P_\theta(x^n),$$

for any measurable set B. In other words,

$$\mathcal{R}_\pi(P_\pi) = \mathcal{I}\ \mathbb{P}_{\theta,X^n}\ .$$

Proof. This is a straightforward consequence of the definitions. The mutual information between X and Y tells how many bits can be saved on the transmission of Y when X is known both to the sender and the receiver of the message, as is clear from the second expression of the mutual information. In the case of the Bayesian redundancy, we see that the average increase of code length due to the fact that the parameter of the source has to be learnt, depends on how much information about the unknown parameter θ is contained in the source X^n. □

2.2. Minimax redundancy. In this approach we want to make sure to do our best in the worst case. In the setting of Definition 2.2 (page 11), let us define the worst case redundancy as

$$\mathcal{R}(Q) = \sup_{\theta\in\Theta} \mathcal{R}\ P_\theta, Q\ .$$

Definition 2.4 *The ideal code $Q \in \mathcal{M}_+^1(A^n)$ is said to be a minimax coding distribution when*

$$\mathcal{R}(Q) = \inf_{Q \in \mathcal{M}_+^1(A^n)} \mathcal{R}(Q)$$

Proposition 2.3 (Least favorable prior) *Under the same hypotheses as in Definition 2.2 (page 11),*

$$\operatorname*{ess\,inf}_{\mathrm{d}\pi(\theta)} \mathcal{R}\ P_\theta, P_\pi \ \leq \quad \mathcal{R}\ P_\theta, P_\pi \ \mathrm{d}\pi(\theta) \leq \inf_Q \sup_\theta \mathcal{R}\ P_\theta, Q \ \leq \sup_{\theta \in \Theta} \mathcal{R}\ P_\theta, P_\pi \ .$$

If moreover, for some $\pi \in \mathcal{M}_+^1(\Theta)$, $\mathcal{R}\ P_\theta, P_{\widehat{\pi}} \ = \sup_\theta \mathcal{R}\ P_\theta, P_{\widehat{\pi}}$, π almost surely, then π is called a least favorable prior. In this case $P_{\widehat{\pi}}$ is the unique minimax coding distribution. Moreover π is solution of

$$\mathcal{R}_{\widehat{\pi}}\ P_{\widehat{\pi}} \ = \sup_{\pi \in \mathcal{M}_+^1(\Theta)} \mathcal{R}_\pi\ P_\pi \ . \tag{2}$$

On the other hand, if π satisfies equation (2), then $P_{\widehat{\pi}}$ is the unique minimax coding distribution.

Remark 2.1 *Equation (2) does have a solution under mild assumptions, we refer to [1][theorem 1.2.1 page 17] for further details on this question.*

Proof. The chain of inequalities at the beginning of the proposition is a consequence of Proposition 3.1, saying that

$$\mathcal{R}(P_\theta, P_\pi)\mathrm{d}\pi(\theta) = \inf_{Q \in \mathcal{M}_+^!(\mathcal{X})} \mathcal{R}(P_\theta, Q)\mathrm{d}\pi(\theta).$$

In the case when $\mathcal{R}(P_\theta, P_{\widehat{\pi}}) = \sup_\theta \mathcal{R}(P_\theta, P_{\widehat{\pi}})$ almost surely, all the inequalities become equalities, showing that $P_{\widehat{\pi}}$ is a minimax coding distribution. To show that it is unique, let us consider another minimax distribution Q. It would satisfy necessarily

$$\sup_{\theta \in \Theta} \mathcal{R}\ P_\theta, Q \ = \quad \mathcal{R}(P_\theta, P_{\widehat{\pi}})\mathrm{d}\pi(\theta),$$

and therefore

$$\mathcal{R}(P_\theta, Q)\,\mathrm{d}\pi(\theta) \leq \quad \mathcal{R}(P_\theta, P_{\widehat{\pi}})\,\mathrm{d}\pi(\theta).$$

On the other hand

$$\mathcal{R}(P_{\widehat{\theta}}, Q)\,\mathrm{d}\pi(\theta) = \quad \mathcal{R}(P_\theta, P_{\widehat{\pi}})\,\mathrm{d}\pi(\theta) + \mathcal{R}(P_{\widehat{\pi}}, Q),$$

implying that $\mathcal{R}(P_{\widehat{\pi}}, Q) = 0$, and consequently that $Q = P_{\widehat{\pi}}$. Finally, equation (2) is a consequence of the equality

$$\mathcal{R}(P_\theta, P_{\widehat{\pi}})\mathrm{d}\pi(\theta) = \inf_Q \sup_\theta \mathcal{R}(P_\theta, Q) \geq \quad \mathcal{R}(P_\theta, P_\pi)\,\mathrm{d}\pi(\theta).$$

On the other hand, let us now assume that π is solution of equation (2). In this case, we can write

$$\mathcal{R}_{\widehat{\pi}}(P_{\widehat{\pi}}) - \mathcal{R}_\pi(P_\pi) = \quad \mathcal{R}(P_\theta, P_{\widehat{\pi}})\,(\mathrm{d}\pi - \mathrm{d}\pi)(\theta) + \mathcal{R}(P_\pi, P_{\widehat{\pi}}).$$

Applying this inequality to $\pi = \lambda\nu + (1-\lambda)\pi$, for all $\lambda \in [0,1]$, and $\nu \in \mathcal{M}_+^1(\Theta)$, we see that the right-hand side should have a positive derivative at $\lambda = 0$, showing that

$$\mathcal{R}(P_\theta, P_{\widehat{\pi}})(\mathrm{d}\pi - \mathrm{d}\nu)(\theta) \geq 0.$$

Considering for any $\theta \in \Theta$, $\nu = \delta_\theta$, the Dirac mass at θ, we deduce that

$$\mathcal{R}(P_\theta, P_{\widehat{\pi}})\,\mathrm{d}\pi(\theta) = \sup_{\theta\in\Theta} \mathcal{R}(P_\theta, P_{\widehat{\pi}}),$$

showing from the first part of the proposition that $P_{\widehat{\pi}}$ is indeed the minimax coding distribution in this case. □

2.3. Oracle inequalities. Another point of view is to consider a family of coding distributions, $Q_\theta : \theta \in \Theta$, instead of considering a familiy of source distributions. Given some prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$, it is easy to compare the mean length of the code mixture

$$Q_\pi \stackrel{\text{def}}{=} \quad Q_\theta\,\mathrm{d}\pi(\theta),$$

with the mean code length of any Q_θ, $\theta \in \Theta$. Indeed,

Proposition 2.4 *For any* $(x_1, \ldots, x_n) \in A^n$,

$$-\log_2 Q_\pi(x^n) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \quad -\log_2 Q_\theta(x^n)\,\mathrm{d}\rho(\theta) + \mathcal{R}(\rho,\pi)$$
$$\leq \inf_{\theta\in\Theta} -\log_2 Q_\theta(x^n) \quad -\log_2[\pi(\{\theta\})].$$

Consequently, for any source distribution $P \in \mathcal{M}_+^1(A^n)$,

$$\mathcal{R}\ P, Q_\pi) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \quad \mathcal{R}(P, Q_\theta)\mathrm{d}\rho(\theta) + \mathcal{R}(\rho, \pi)$$
$$\leq \inf_{\theta \in \Theta} \mathcal{R}\ P, Q_\theta\ - \log_2\ \pi(\{\theta\})\ .$$

Proof. If $\mathcal{R}(\rho, \pi) = +\infty$, there is nothing to prove. Let us assume consequently that $\mathcal{R}(\rho, \pi) < \infty$. Let $\pi = \pi_s + \pi_a$ where $\pi_a \ll \rho$ and π_s and ρ are singular. It is easy to see that $\dfrac{\mathrm{d}\pi_a}{\mathrm{d}\rho}(\theta) = \dfrac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)^{-1}$, ρ almost surely, therefore

$$\begin{aligned} Q_\theta(x^n)\mathrm{d}\pi(\theta)\,\mathrm{d}\pi(\theta) \geq & \quad Q_\theta(x^n)\,\mathrm{d}\pi_a(\theta) \\ = & \quad \exp\ \log\ Q_\theta(x^n)\ + \log\ \frac{\mathrm{d}\pi_a}{\mathrm{d}\rho}(\theta) \quad \mathrm{d}\rho(\theta) \\ = & \quad \exp\ \log\ Q_\theta(x^n)\ - \log\ \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) \quad \mathrm{d}\rho(\theta) \\ & \geq \exp \quad \log\ Q_\theta(x^n)\ \mathrm{d}\rho(\theta) - \mathcal{K}(\rho, \pi)\ , \end{aligned}$$

by Jensen's inequality. □

Corollary 2.5 (Double mixture codes and Bayesian Model Averaging) *Consider a family of models of ideal codes* $Q_\theta \in \mathcal{M}_+^1(A^n); \theta \in \Theta_k\ ,\ k \in \mathbb{N}$, *where* Θ_k *are measurable parameter spaces. Let* $\mu \in \mathcal{M}_+^1(\mathbb{N})$ *and for each* $k \in \mathbb{N}$ *let* $\nu_k \in \mathcal{M}_+^1\ \Theta_k$ *be some prior measure on* Θ_k. *Let us consider* $\Theta = \underset{k \in \mathbb{N}}{\ } \{k\} \times \Theta_k$. *Let the prior distribution* π *on* Θ *be defined for any measurable set* $B \subset \Theta$ *by the formula*

$$\pi(B) = \quad \mathbb{1}\ (k, \theta) \in B\ \ \mathrm{d}\mu(k)\,\mathrm{d}\nu_k(\theta).$$

The coding distribution Q_π *is called a double mixture code. The sequential prediction method that estimates* $\mathbb{P}_{X_i|X^{i-1}}$ *by*

$$Q_\pi\ x_i|x^{i-1}) \stackrel{\text{def}}{=} Q_\pi(x^i)/Q_\pi(x^{i-1})$$

is called Bayesian Model Averaging. It is such that

$$\sum_{i=1}^{n} - \log_2\ Q_\pi\ x_i|x^{i-1})\ \leq \inf_{k \in \mathbb{N}} - \log_2 Q_{\nu_k}\ x_i|x^{i-1}\ - \log_2\ \mu(k)\ .$$

Consequently for any source $X^n \in A^n$,

$$\sum_{i=1}^{n} \mathbb{E}\ \mathcal{K}\ \mathbb{P}_{X_i|X^{i-1}}, Q_\pi(\cdot|X^{i-1}) \leq \inf_{k\in\mathbb{N}} \sum_{i=1}^{n} \mathbb{E}\ \mathcal{K}\ \mathbb{P}_{X_i|X^{i-1}}, Q_{\nu_k}(\cdot|X^{i-1}) \quad - \log\ \mu(k) \quad .$$

PROOF. It is a consequence of the previous proposition and of the identity

$$\sum_{i=1}^{n} - \log_2\ Q_\pi\ \ x_i \,|\, x^{i-1} \quad = - \log_2\ Q_\pi(x^n)\ .$$

Let us remark that we could have replaced in this proof the family $\{Q_{\nu_k}, k \in \mathbb{N}\}$ by any family of ideal codes $\{Q_k, k \in \mathbb{N}\}$. $\square$

This is a result about sequential prediction. We consider estimators $Q_k(\cdot|X^{i-1})$ of the conditional distributions $\mathbb{P}_{X_i|X^{i-1}}$ and are provided with an aggregated estimator $Q_\pi(\cdot|X^{i-1})$ that performs almost as well as the best choice of k, when performance is measured by the sum of the expected Kullback divergences over all sample sizes in the range $1, \ldots, n$. The sequential nature of the result comes from the fact that the criterion is a sum over all sample sizes : this is an example of cumulated risk function. The following two subsections are devoted to an important example of double mixture codes : the context tree weighting algorithm, where we are going to compute everything explicitly.

2.4. MIXTURES OF I.I.D. DISTRIBUTIONS. Let $\Theta = \mathcal{M}_+^1(A)$ and consider the Lebesgue measure λ on Θ. Let the prior $\nu \ll \lambda$ be defined by its density

$$\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(\theta) = \frac{\Gamma\ \frac{d}{2}}{\sqrt{d}\,\Gamma\ \frac{1}{2}\ ^d} \prod_{x\in A} \theta(x)^{-1/2}, \tag{3}$$

where $\Gamma(z) = \int_{\mathbb{R}_+} t^{z-1} \exp(-t)\,\mathrm{d}t = (z-1)\Gamma(z-1)$ is the usual Γ function.

LEMMA 2.6 *The mixture code based on the prior ν and the family of product measures $Q_\theta(x^n) = \prod_{i=1}^n \theta(x_i)$ is such that*

$$Q_\nu(x^n) = \frac{\Gamma\ \frac{d}{2}}{\Gamma\ \frac{1}{2}\ ^d \Gamma\ n + \frac{d}{2}} \prod_{y\in A} \Gamma\ n\overline{\mathbb{P}}_{x^n}(y) + \frac{1}{2} = \frac{\prod_{y\in A} \prod_{j=1}^{n\overline{\mathbb{P}}_{x^n}(y)} j - \frac{1}{2}}{\prod_{j=1}^{n} j + \frac{d}{2} - 1},$$

where

$$\overline{\mathbb{P}}_{x^n} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

is the empirical probability measure of the sequence $(x_1, \ldots, x_n)$*, so that* $n\overline{\mathbb{P}}_{x^n}(y) = \sum_{i=1}^{n} \mathbb{1}(x_i = y)$.

PROOF. Let us remark that

$$Q_\theta(x^n) = \prod_{y \in A} \theta(y)^{n\overline{\mathbb{P}}_{x^n}(y)},$$

and thus

$$Q_\nu(x^n) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{d}\Gamma\left(\frac{1}{2}\right)^d} \int_\Theta \prod_{y \in A} \theta(y)^{n\overline{\mathbb{P}}_{x^n}(y) - \frac{1}{2}} \mathrm{d}\lambda(\theta). \tag{4}$$

On the other hand, for any $\alpha \in]-1, +\infty[^A$,

$$\int_\Theta \prod_{y \in A} \theta(y)^{\alpha_y} \, \mathrm{d}\lambda(\theta) = \frac{\sqrt{d} \prod_{y \in A} \Gamma(\alpha_y + 1)}{\Gamma\left(\sum_{y \in A} \alpha_y + d\right)}. \tag{5}$$

Indeed, the change of variables $t_y = s\theta(y)$, where $t_y \in \mathbb{R}_+$, $\theta \in \mathcal{M}_+^1(A)$ and $s = \sum_{y \in A} t_y \in \mathbb{R}_+$ shows that

$$\int_{\mathbb{R}_+^d} \prod_{y \in A} t_y^{\alpha_i} \exp(-t_y) \, \mathrm{d}t_y = \int_{\mathbb{R}_+} \int_{\theta \in \Theta} \prod_{y \in A} \theta(y)^{\alpha_y} s^{\alpha_y} \exp(-s) s^{d-1} \mathrm{d}\lambda(\theta) \frac{\mathrm{d}s}{\sqrt{d}}.$$

Putting together equations (4) and (5) gives the desired result. □

PROPOSITION 2.7 (KRICHEVSKI, TROFIMOV) *For any* $x^n \in A^n$,

$$Q_\nu(x^n) \geq \frac{2^{-nH(\overline{\mathbb{P}}_{x^n})}}{dn^{(d-1)/2}} = d^{-1} n^{-(d-1)/2} \sup_{\theta \in \Theta} Q_\theta(x^n).$$

PROOF. Let us put for any $a = (a_i)_{i=1}^d \in \mathbb{N}^d$

$$\Delta(a) = \frac{\Gamma\left(\frac{d}{2}\right) \prod_{i=1}^d \Gamma\left(a_i + \frac{1}{2}\right) \prod_{i=1}^d a_i^{\sum_{i=1}^d a_i + (d-1)/2}}{\Gamma\left(\frac{1}{2}\right)^d \Gamma\left(\sum_{i=1}^d a_i + \frac{d}{2}\right) \prod_{i=1}^d a_i^{a_i}}$$

We have to show that $\Delta(a) \geq d^{-1}$. Let us notice first that $\Delta(1, 0, \ldots, 0) = d^{-1}$, and that Δ is invariant under any permutation of the vector $(a_i)_{i=1}^d$. It

is therefore enough to check that $\Delta(a) \geq \Delta\ (a_1 - 1, a_2, \ldots, a_d)$. Let us put $s = \sum_{i=1}^{d} a_i$ and $t = a_1$. With these notations,

$$\Delta(a) = \Delta(a_1 - 1, a_2^d) \frac{t - \frac{1}{2}\quad t - 1\ ^{t-1} s^{s+\frac{d-1}{2}}}{s + \frac{d}{2} - 1\ t^t\ s - 1\ ^{s-1+\frac{d-1}{2}}}.$$

To end the proof, we have to check that

$$\frac{t - \frac{1}{2}\quad t - 1\ ^{t-1} s^{s+\frac{d-1}{2}}}{s + \frac{d}{2} - 1\ t^t\ s - 1\ ^{s-1+\frac{d-1}{2}}} \geq 1, \qquad t \geq 1, s \geq 2.$$

This can also be written as $g(s) \geq f(t)$, where

$$g(s) = -\ s + \tfrac{d-3}{2}\ \log\ \frac{s-1}{s}\ - \log\ \frac{s + \frac{d-3}{2}}{s}\ ,$$
$$f(t) = t \log(t) - (t-1)\log(t-1) - \log\ t - \tfrac{1}{2}\ .$$

Using the fact that $\log(1+z) \leq z$, we see that $g(s) \geq\ s + \frac{d-3}{2}\ s^{-1} - \frac{d-3}{2s} = 1$. On the other hand, making a Taylor expansion of $z \mapsto z\log(z)$, we see that

$$(z+u)\log(z+u) = z\log(z) + u\ \log(z) + 1\ +\ \int_0^u (u-v)\frac{\mathrm{d}v}{z+v}.$$

Applying this to $z = t - \frac{1}{2}$ and $u \in\ -\frac{1}{2}, \frac{1}{2}\ $, gives

$$t\log(t) - (t-1)\log(t-1) - \log\ t - \tfrac{1}{2}\ - 1$$
$$= \int_0^{\frac{1}{2}} \tfrac{1}{2} - v\quad \frac{1}{t - \frac{1}{2} + v} - \frac{1}{t - \frac{1}{2} - v}\quad \mathrm{d}v \leq 0,$$

proving that $f(t) \leq 1$. □

Corollary 2.8 *The redundancy of the KT ideal code is such that*

$$\mathcal{R}(P, Q_\nu) \leq \inf_{\theta \in \Theta} \mathcal{R}(P, Q_\theta) + \frac{d-1}{2}\log_2(n) + \log_2(d).$$

2.5. Universal codes and context tree weighting. Let $\mathcal{D}$ be the set of complete suffix sets of A^* and $\mathcal{D}_L$ the subset of complete suffix sets of length not greater than L. More explicitly, let us introduce the notations $A^*w = \{w'w, w' \in A^*\}$ and

$$\overline{D} = D \cup\ \ w \in A^*; A^*w \cap D \neq \varnothing\ \ ,$$

the set of suffixes of D. We can define $\mathcal{D}_L$ formely as

$$\mathcal{D}_L = \quad D \subset A^*; D \cap A^*D = \varnothing, A^*\overline{D} = A^*D, \text{ and } \max_{w \in D} \ell(w) \leq L \quad .$$

Let $D \in \mathcal{D}_L$ be a complete suffix set and let $f_D : A^{\mathbb{Z}_-} \to D$ be defined by

$$f_D \; x_{-\infty}^0 \; = x_{-k}^0,$$

where k is the only index such that $x_{-k}^0 \in D$. Given a past context $x_{-\infty}^0$, we want to define a probability measure on $x_1^{+\infty}$. A stationary context tree distribution is a probability measure defined by

$$\mathbb{P} \; X_n = x_n \,|\, X_{-\infty}^{n-1} = x_{-\infty}^{n-1} \; = \mathbb{P} \; X_1 = x_n \,|\, f_D(X_{-\infty}^0) = f_D(x_{-\infty}^{n-1}) \; .$$

Its parameter set is D, Θ_D where $\Theta_D = \quad \theta_w \in \mathcal{M}_+^1(A); w \in D \quad = \mathcal{M}_+^1(A)^D$. Consider the parameter space

$$\Theta = \qquad \{D\} \times \Theta_D.$$
$$D \in \mathcal{D}_L$$

Let us define on Θ the prior probability measure π by

$$\pi(B) = \quad \mathbb{1} \; (D, \theta) \in B \; \mathrm{d}\mu(D) \qquad \mathrm{d}\nu(\theta_w),$$
$$w \in D$$

where ν is the Krichevski Trofimov prior on $\mathcal{M}_+^1(A)$ and where for some real parameter $\alpha \in]0, 1[$

$$\mu(D) = \alpha^{(|D|-1)/(d-1)}(1-\alpha)^{|D \setminus A^L|}, \qquad D \in \mathcal{D}_L, \tag{6}$$

is the measure of a Galton Watson process on complete suffix sets with offspring probability α. More precisely, the distribution of $\overline{D}$ under μ is given by the formula

$$\begin{aligned}
\mu(\overline{D} = \overline{B}) &= \prod_{j=1}^{L} \mu \; \overline{D} \cap A^j = \overline{B} \cap A^j \,|\, \overline{D} \cap \overline{A^{j-1}} = \overline{B} \cap \overline{A^{j-1}} \\
&\stackrel{\text{def}}{=} \prod_{j=1}^{L} \prod_{w \in \overline{B} \cap A^{j-1}} \alpha \mathbb{1}(A^*w \subset \overline{B} \; + (1-\alpha)\mathbb{1} \; A^*w \cap \overline{B} = \varnothing \\
&= \alpha^{|\overline{B} \setminus B|+1}(1-\alpha)^{|B \setminus A^L|}, \qquad B \in \mathcal{D}_L.
\end{aligned}$$

Equation (6) is a consequence of the fact that

$$|\overline{B} \setminus B| = \; |B| - 1 \; /(d-1) - 1,$$

for any $B \in \mathcal{D}_L$, a fact that can easily be checked by induction.

PROPOSITION 2.9 *Let us be given some sequence $x_{1-L}^n \in A^{n+L}$. Let us consider the counters*

$$\begin{aligned} a_w^y(n) &= \sum_{k=1}^{n} \mathbb{1}\left\{ x_{k-\ell(w)}^{k-1} = w, x_k = y \right\}, \qquad w \in A^* \cup \{\varnothing\}, y \in A, \\ b_w(n) &= \sum_{y \in A} a_w^y(n), \\ K_w(n) &= \frac{\Gamma\left(\frac{d}{2}\right)^d \prod_{y \in A} \Gamma\left(a_w^y(n) + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^d \Gamma\left(b_w(n) + \frac{d}{2}\right)}, \end{aligned}$$

The mixture code Q_π can be computed as

$$Q_\pi\left(x_1^n \,|\, x_{1-L}^0\right) = \sum_{D \in \mathcal{D}_L} \mu(D) \prod_{w \in D} K_w(n).$$

This computation can be made by induction according to the following scheme

$$\begin{aligned} p_w(n) &= K_w(n), && w \in A^L, \\ p_w(n) &= (1-\alpha) K_w(n) + \alpha \prod_{y \in A} p_{yw}(n), && w \in A^* \cup \{\varnothing\}, \ell(w) < L, \\ p_\varnothing(n) &= Q_\pi\left(x_1^n \,|\, x_{1-L}^0\right). \end{aligned}$$

PROOF. For any $D \in \mathcal{D}_L$,

$$\begin{aligned} &\int_{\Theta_D} Q_{D,\theta}\left(x_1^n \,|\, x_{1-L}^0\right) \prod_{w \in D} \mathrm{d}\nu(\theta_w) \\ &\qquad = \int_{\Theta_D} \prod_{w \in D} \prod_{y \in A} \theta_w(y)^{a_w^y(n)} \prod_{w \in D} \mathrm{d}\nu(\theta_w) = \prod_{w \in D} K_w(n), \end{aligned}$$

so that

$$Q_\pi\left(x_1^n \,|\, x_{1-L}^0\right) = \int_{\mathcal{D}_L} \prod_{w \in D} K_w(n) \, \mathrm{d}\mu(D).$$

Let

$$p_w(n) = \int_{\mathcal{D}_L} \prod_{w' \in D \cap (A^* w \cup \{w\})} K_{w'}(n) \, \mathrm{d}\mu\left(D \,|\, w \in \overline{D} \cup \{\varnothing\}\right),$$

so that $p_\varnothing(n) = Q_\pi\left(x_1^n \,|\, x_{1-L}^0\right)$. We see from the definition of μ that

$$p_w(n) = \mu\left(w \in D \,|\, w \in \overline{D}\right) K_w(n)$$

$$+\mu\ A^*w \subset \overline{D} \,|\, w \in \overline{D} \qquad\qquad K_{w'}(n)\mathrm{d}\mu\ D \,|\, yw \in \overline{D}$$
$$\sum_{y\in A} \int_{\mathcal{D}_L} \sum_{w'\in D\cap(A^*yw\cup\{yw\})}$$
$$= (1-\alpha)K_w(n) + \alpha \sum_{y\in A} p_{yw}(n), \qquad w \in A^* \cup \{\varnothing\}, \ell(w) < L,$$

and that $p_w(n) = K_w(n)$ for any $w \in A^L$. $\square$

Exercise 5 *Show that the computation of the context tree weighting ideal code can be updated online according to the following rules :*

$$\begin{aligned}
a_w^y(0) &= b_w(y) = 0,\\
K_w(0) &= 1,\\
p_w(0) &= 1,\\
&\vdots\\
a_w^y(n) &= \begin{cases} a_w^y(n-1)+1, & \text{when } w = x_{n-\ell(w)}^{n-1} \text{ and } y = x_n,\\ a_w^y(n-1), & \text{otherwise,}\end{cases}\\
b_w(n) &= \begin{cases} b_w(n-1)+1, & \text{when } w = x_{n-\ell(m)}^{n-1},\\ b_w(n-1), & \text{otherwise,}\end{cases}\\
K_w(n) &= \begin{cases} K_w(n-1)\dfrac{a_w^{x_n}(n)-\frac{1}{2}}{b_w(n)+\frac{d}{2}-1}, & \text{when } w = x_{n-\ell(m)}^{n-1},\\ K_w(n-1), & \text{otherwise,}\end{cases}\\
p_w(n) &= \begin{cases} p_w(n-1), & \text{when } w \neq x_{n-\ell(m)}^{n-1},\\ K_w(n), & \text{when } w = x_{n-L}^{n-1},\\ (1-\alpha)K_w(n) + \alpha \sum_{y\in A} p_{yw}(n), & \text{when } w = x_{n-\ell}^{n-1}, \text{ with } \ell < L.\end{cases}
\end{aligned}$$

Conclude that the number of operations needed to compute $Q_\pi(x_1^n \,|\, x_{1-L}^0)$ for a given value of $x_{1-L}^n \in A^{n+L}$ grows only linearly with n.

PROPOSITION 2.10 *For any $x_{1-L}^n \in A^{n+L}$,*

$$\begin{aligned}
-\log_2 Q_\pi\ x_1^n \,|\, x_{1-L}^0 \quad &\leq \inf_{D\in\mathcal{D}_L} \inf_{\theta\in\Theta_D} -\log_2 Q_{D,\theta}\ x_1^n \,|\, x_{1-L}^0\\
&\quad + \tfrac{|D|(d-1)}{2}\log\ \tfrac{n}{|D|}\ + |D|\log_2(d)\\
&\quad - \tfrac{|D|-1}{d-1}\log_2(\alpha) - |D|\log_2(1-\alpha)\ .
\end{aligned}$$

As a consequence for any process X_{1-L}^n, and any $x_{1-L}^0 \in A^L$,

$$\mathcal{R}\ \mathbb{P}_{X_1^n|X_{1-L}^0=x_{1-L}^0}, Q_\pi(\cdot|x_{1-L}^0)$$
$$\leq \inf_{D\in\mathcal{D}_L}\inf_{\theta\in\Theta_D}\ \mathcal{R}\ \mathbb{P}_{X_1^n|X_{1-L}^0=x_{1-L}^0}, Q_{D,\theta}(\cdot|x_{1-L}^0)$$
$$+\frac{|D|(d-1)}{2}\log\ \frac{n}{|D|}\ +|D|\log_2(d)$$
$$-\frac{|D|-1}{d-1}\log_2(\alpha)-|D|\log_2(1-\alpha)\ .$$

PROOF. Let $\gamma(z)=\frac{d-1}{2}\log_2(z)+\log_2(d)$. According to Proposition 2.7 (page 17),

$$-\log_2\ \int_{\Theta_D} Q_{D,\theta}\ x_1^n\,|\,x_{1-L}^0\ \prod_{w\in D}\mathrm{d}\nu(\theta_w)$$
$$\leq \inf_{\theta\in\Theta_D}\ -\log_2\ Q_{D,\theta}\ x_1^n\,|\,x_{1-L}^0\ \ +\sum_{w\in D}\gamma\ b_w(n)$$
$$\leq \inf_{\theta\in\Theta_D}\ -\log_2\ Q_{D,\theta}\ x_1^n\,|\,x_{1-L}^0\ \ +|D|\gamma\ \ \frac{1}{|D|}\sum_{w\in D}b_w(n)\ ,$$

where we have used the fact that γ is concave. As $\sum_{w\in D} b_w(n)=n$, we get

$$-\log_2\ Q_\pi\ x_1^n\,|\,x_{1-L}^0\ \ \leq \inf_{D\in\mathcal{D}_L}\inf_{\theta\in\Theta_D}\ -\log_2\ Q_{D,\theta}\ x_1^n\,|\,x_{1-L}^0$$
$$+|D|\gamma\ \ \frac{n}{|D|}\ \ -\log_2\ \mu(D)\ ,$$

as stated in the proposition. □

Exercise 6 (A simulation study of a switch code) *Let us consider the alphabet* $A=\{1,\dots,4\}^2$ *of size* 16. *Let us define* $q\in\mathcal{M}_+^1(A)$ *as*

$$q(i,j)=\frac{4(i-1)+j}{8\times 17}$$

and let $q_1(i)=\sum_{j=1}^4 q(i,j)$ *and* $q_2(j)=\sum_{i=1}^4 q(i,j)$ *be its two marginal distributions. Let*

$$p(i,j)=\frac{4}{5}q(i,j)+\frac{1}{5}q_1(i)q_2(j),\qquad i,j\in\{1,\dots,4\}.$$

For any $\theta_1,\theta_2\in\Theta_1\stackrel{\text{def}}{=}\mathcal{M}_+^1(\{1,\dots,4\})$, *let*

$$Q_{\theta_1,\theta_2}(x^n)=\prod_{i=1}^n\theta_1(x_{i,1})\theta_2(x_{i,2}),\qquad x^n\in A^n,$$

where $x_i = (x_{i,1}, x_{i,2}) \in \{1, \dots 4\}^2$. *In the same way, for any* $\theta \in \Theta_2 \stackrel{\text{def}}{=} \mathcal{M}_+^1(A)$, *let*

$$Q_\theta(x^n) = \prod_{i=1}^{n} \theta(x_i).$$

Let ν_1 *be the Krichevski Trofimov prior distribution on* Θ_1, *given by equation* (3, page 16), *and* ν_2 *the Krichevski Trofimov prior distribution on* Θ_2. *Let us consider the following ideal codes*

$$\begin{aligned}
Q_1(x^n) &= \int_{\Theta_1 \times \Theta_1} Q_{\theta_1,\theta_2}(x^n)\, \mathrm{d}\nu_1(\theta_1)\mathrm{d}\nu_1(\theta_2),\\
Q_2(x^n) &= \int_{\Theta_2} Q_\theta(x^n)\, \mathrm{d}\nu_2(\theta),\\
Q_m(x^n) &= \frac{1}{2}Q_1(x^n) + \frac{1}{2}Q_2(x^n),\\
Q_s(x^n) &= \begin{cases} Q_1(x^n), & \text{when } n < m,\\ Q_1(x^m)Q_2(x^n)/Q_2(x^m), & \text{when } n \geq m. \end{cases}
\end{aligned}$$

Take $m = 2000$, *and write a program to compute for* $n \in 200\mathbb{N}$, $n \leq 10000$ *the empirical mean ideal code length*

$$L(Q, n) = -\frac{1}{100} \sum_{j=1}^{100} \log_2 \; Q \; X_1^n(j) \quad ,$$

where $X_1^n(j)$ *are independent trials of the i.i.d. random process distributed according to* $p^{\otimes n}$, *and where* $Q \in \{Q_1, Q_2, Q_m, Q_s\}$. *Plot the curves*

$$\begin{aligned}
&n \mapsto L(Q_2, n) - L(Q_1, n),\\
&n \mapsto L(Q_m, n) - L(Q_1, n),\\
&n \mapsto L(Q_s, n) - L(Q_1, n).
\end{aligned}$$

Make your comments (about the weak points and possible improvements of mixture codes).

Programming hint : all you need is to simulate the empirical measure of X_1^n, *something you can do in Octave, using the function*

`tabulate(rand(1,n), df)`,

where `df` *is the distribution function of your probability vector (to be computed with* `cumsum`*).*