ECOLE NORMALE SUPERIEURE

Double Mixture and Universal Inference

J.-P. VERT

DMA - 00 - 15

Département de mathématiques et applications

CNRS UMR 8553

Double Mixture and Universal Inference

J.-P. VERT

DMA - 00 - 15

may 2000

Département de mathématiques et applications - École normale supérieure
45 rue d'Ulm 75230 PARIS Cedex 05
Tel : (33)(1) 01 44 32 30 00
E-mail : jean-philippe.vert@ens.fr

DOUBLE MIXTURE AND UNIVERSAL INFERENCE

By Jean-Philippe Vert

Ecole normale supérieure de Paris

Given a family of finite dimensional statistical models and a finite number of observations of a random variable, we show how to build a "double mixture" estimator for the density of the random variable whose risk in terms of Kullback-Leibler divergence has a sharp bound compared to the risk of the best model in the family. This estimator is a mixture of model estimators which are themselves mixtures in the continuous parameter spaces of the corresponding models.

The idea of using double mixtures has been studied for a long time in the field of universal compression in coding theory but we highlight the fundamental differences between our statistical estimator and "twice-universal" coding algorithm, due to the difference in the criteria to optimize.

1. Introduction The problem of estimating the probability distribution P of a random variable X on a space $\mathcal{X}$ from a finite number of i.i.d. observations $X_1, \dots, X_n$ is a central but difficult problem in statistics. As pointed out by Vapnik ([22]) it is generally ill-posed when no assumption is made on P. In the real world, however, the statistician usually knows nothing in advance about P. In that case a natural approach consists in building a family of parametric models $\left\{P_{\theta_m} \in \mathcal{M}_+^1(\mathcal{X})\,; \theta_m \in \Theta_m \subset \mathbb{R}^{D_m}, m \in \mathcal{M}\right\}$ and considering the minimization problem

$$\inf_{m \in \mathcal{M}} \inf_{\theta_m \in \Theta_m} l\left(P, P_{\theta_m}\right), \tag{1.1}$$

where l is a loss function between probability distributions. In other words, rather than trying to guess P, the statistician looks for the most informative projection of P on the most reasonable model Θ_m.

An estimator for this problem of distribution estimation is a measurable mapping $\hat{P}$ from $\mathcal{X}^n$ to $\mathcal{M}_+^1(\mathcal{X})$. The performance of such an estimator with respect to a true distribution P is usually measured in terms of its average loss, also called *risk*:

$$\mathbf{E}_{P^{\otimes n}\left(dX_1^N\right)} l\left(P, \hat{P}(X_1^N)\right).$$

For an estimator with value within a particular model (parametric estimation), this risk can often be expressed or at least upper bounded by the sum of two terms:

- a *bias* term which represents the distance between the actual probability P and its projection P_m on the particular model;

May 2000.

AMS 2000 subject classifications. Primary 62F35; secondary 62B10, 62G07, 94A17.

Key words and phrases. Adaptive Mixture of Models, Double Mixture, Mean Kullback Risk, Gibbs Estimator, Context-Tree Weighting Method, Universal Inference, Universal Prediction.

- a *fluctuation* term which represents the difficulty of estimating P_m.

Usually the larger the model the smaller the bias, but the larger the fluctuation risk. A natural way of solving the estimation problem is to decompose it into two stages : first build good estimators $\hat{P}_m$ for every model $m \in \mathcal{M}$ and then select one model $\hat{m}$ with the lowest total risk. With this approach the final estimator $\hat{P}_{\hat{m}}$ is the estimator associated with the model supposed to realize the best trade-off between bias and fluctuation. This philosophy is the starting point of many techniques in the field of *model selection*, to which a huge amount of literature has been devoted. As we won't further develop this approach let us just mention typical references including works by Akaike ([1]), Mallows ([16]), Schwarz ([21]), Rissanen ([19]), Barron and al. ([4]) and Vapnik ([22]).

Model selection is not the only way to deal with problem (1.1). An other approach is gaining attention in statistical estimation: the idea of *model mixture*. This idea led to remarkable theoretical and experimental results in coding theory for compression purpose where mixture codes ([13]) are known to be universal with respect to a class of encoders, under quite general assumptions.

As far as statistical estimation is concerned, every Bayesian estimator can be considered as a mixture estimator. While these estimators are optimal for the Bayesian risk theoretical results concerning their performance in the worst case setting are difficult to obtain. Barron ([3]) and Barron and Yang ([5]) considered a Cesaro mean of Bayesian estimators to derive minimax density estimators for non parametric density classes. More recently Catoni considered an equivalent estimator together with a half-sample trick to deal with parametric density classes ([8], [10]), and showed that a thermalized version of the Bayesian estimator ([9]) could approach the minimax risk under very general assumption.

In a recent paper ([23]) we applied Catoni's estimators in the framework of regression. We showed how to build a mixture estimator $\hat{P}_w = \sum_{i \in \mathcal{I}} w(i) \hat{P}_{m_i}$ where the weights w as well as the estimators $\hat{P}_m$ are built from the observation, and obtain a universal risk bound. This approach involved a split of the observations into an estimation set used to build the estimators $\hat{P}_m$ for every $m \in \mathcal{M}$ and a validation set used to compute the weights $w(i)$ of each estimator.

In this paper we go one step further in this mixing approach. After observing that the estimators $\hat{P}_m$ for every model can be mixtures on the continuous parameter set themselves (think of the Laplace estimator for a Bernoulli distribution for instance), we show how it is possible to carry out a *double mixture* in one stage by considering the larger parameter set $\{(m, \theta_m), m \in \mathcal{M}, \theta_m \in \Theta_m\}$. This means in particular that the observations are not split into two sets any more, and that the same observations are used to estimate continuous parameters and model structure in the same time. The idea of a double mixture finds its roots in coding theory where double mixture codes have given very interesting results ([14], [20]). An important source of inspiration was the work of Willems, Shtarkov and Tjalkens concerning the context tree weighting algorithm ([24] , [25]), together with Catoni's Gibbs estimator ([9]) which can be used to mix discrete as well as continuous parameters.

This paper is organized as follows. After setting up the general regression framework which will be used afterward in section 2 we state the main result of this paper in section 3 (theorem 1) whose proof is postponed to section 6 because of its length. Section 4 is a comparison between the estimator we propose for statistical estimation and "universal" estimators used in coding theory, and section 5 presents a particular application of our estimator for string analysis, with an efficient implementation. We refer to a previous works ([23]) for suggestions on how to use such models for natural language processing applications.

2. Notations and framework In this section we present the regression framework together with general notations which will be used within the paper.

Let $(\mathcal{X}, \mathcal{B}_1)$ be a measurable space and $(\mathcal{Y}, \mathcal{B}_2)$ be a *finite* measurable space endowed with the discrete σ-algebra. We note α the size of the set $\mathcal{Y}$. The goal of statistical modeling is to predict the value of a variable $Y \in \mathcal{Y}$ from an observation $X \in \mathcal{X}$. The set $\mathcal{Y}$ being finite this covers in particular the problem of categorization. However we focus on estimating the conditional law of Y knowing X, and not on the design of a classifier. In particular the criterion we will use is a measure of the difference between laws and not the number of categorization errors. Note that the variable X can be almost anything.

To model the random nature of X and Y we suppose that a family of unknown exchangeable probability distributions is given :

$$\forall N \in \mathbb{N} \qquad P_N \in \mathcal{M}_+^1\left((\mathcal{X} \times \mathcal{Y})^N, (\mathcal{B}_1 \otimes \mathcal{B}_2)^{\otimes N}\right),$$

and we let $\{(X_i, Y_i) = Z_i; i = 1, \dots, N\}$ be the canonical process.

One can for instance think of P_N as a product measure $P^{\otimes N}$ with P being a probability on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_1 \otimes \mathcal{B}_2)$, if the observations are supposed to be i.i.d. However we will only use the weaker assumption that P_N is *exchangeable*, i.e. that for any permutation σ of $\{1, \dots, N\}$ and any $A \in (\mathcal{B}_1 \otimes \mathcal{B}_2)^N$,

$$P_N\left(Z_1^N \in A\right) = P_N\left((\sigma Z)_1^N \in A\right),$$

where σZ is the exchanged process

$$(\sigma Z)_i = Z_{\sigma(i)}, \qquad i = 1, \dots, N.$$

The property of being exchangeable is more general than the property of being a product measure, and it covers more situations which can happen in the real world, e.g. random splitting of the observations into different sets, or sampling from a finite set without replacement.

Within this framework the observation is Z_1^{N-1} and the goal is to estimate the unknown conditional distribution $P_N\left(dY_N | X_N; Z_1^{N-1}\right)$.

2.1. *Finite context model* Without any further restriction on P_N the problem of density estimation based on empirical data can be ill-posed. Therefore we suppose a family of models is given. It will be used to approximate the unknown distribution.

DEFINITION 1. *A model $m = (\mathcal{S}_m, s_m)$ consists of:*

- *a finite set $\mathcal{S}_m = \{s_1, \ldots, s_{D_m}\}$. D_m is called the dimension of the model.*
- *A measurable mapping $s_m : (\mathcal{X}, \mathcal{B}_1) \to \mathcal{S}_m$, which describes how the space $\mathcal{X}$ is partitioned according to the model.*

For any model $m \in \mathcal{M}$ and $x \in \mathcal{X}$, $s_m(x)$ is called the context of x with respect to the model m.

The goal of any model is to partition the space $\mathcal{X}$ into D_m categories through the mapping s_m, and to build a conditional distribution for Y which only depends on the category of X. Such a class of models includes in particular regression based on histograms, CART models ([7]) or representation of complex objects (e.g. images) using filtering of features extraction.

Finally we suppose given a countable family of such models $\mathcal{M} = \{m_i\}_{i \in \mathcal{I}}$ with $\mathcal{I}$ being a countable index set, as well as a prior probability distribution π on $\mathcal{I}$. The role of the prior distribution π which influences the performance of the final estimator will become clearer in the sequel.

The variable Y being discrete its distribution is a Bernoulli distribution characterized by a parameter of the α-dimensional simplex $\Sigma = \{\theta \in [0,1]^\alpha / \sum_{i=1}^{\alpha} \theta^i = 1\}$. Therefore any model m is associated with a parameter space $\Theta_m = \Sigma^{D_m}$ to define a family of conditional probability distributions with the following density:

$$\forall m \in \mathcal{M}, \forall \theta_m = \left(\theta_{s_1}, \ldots, \theta_{s_{D_m}}\right) \in \Theta_m, \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \qquad p_{m,\theta_m}(y|x) = \theta^y_{s_m(x)}.$$

2.2. *Problem* As we want to compare estimators based on different models we can not use a distance defined on the parameter space. In order to measure directly the distance between the true sample conditional distribution and the estimated one we use the conditional Kullback Leibler divergence (also called *conditional relative entropy*, see e.g. [12, p. 22]) which is an intrinsic and fundamental measure of risk defined for two probabilities P_1 and P_2 with densities p_1 and p_2 by :

$$\mathcal{K}\left(P_1(dY|X), P_2(dY|X)\right) = \mathbf{E}_{P_1(dX,dY)} \log \frac{p_1(y|x)}{p_2(y|x)}.$$

The model selection problem for the average Kullback risk is to solve approximately, knowing the sample Z_1^{N-1}, the minimization problem:

$$\inf_{m \in \mathcal{M}, \theta_m \in \Theta_m} \mathbf{E}_{P_N\left(dZ_1^{N-1}\right)} \mathcal{K}\left(P_N\left(dY_N|X_N; Z_1^{N-1}\right), P_{m,\theta_m}\left(dY_N|X_N\right)\right), \tag{2.2}$$

where $\mathcal{K}(.,.)$ is the conditional Kullback Leibler divergence.

3. The double mixture estimator The continuous parameter set associated with a model $m \in \mathcal{M}$ is $\Theta_m = \Sigma^{D_m}$. Let us define a probability distribution on this set as a product measure $\mu_m = \mu^{\otimes D_m}$ where μ is the Dirichlet distribution with parameter $1/2$ on Σ, i.e. the measure with the following density with respect to Lebesgue's measure $\lambda(d\theta)$:

$$\mu(d\theta) = \frac{1}{\sqrt{\alpha}} \frac{\Gamma\left(\frac{\alpha}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{\alpha}} \cdot \prod_{i=1}^{\alpha} \frac{1}{\sqrt{\theta^i}} \lambda(d\theta).$$

This prior, also known as Jeffrey's prior, arises naturally in coding theory for compression purpose because it asymptotically maximizes Shannon's mutual information between an i.i.d. sample drawn according to a Bernoulli law of parameter θ and the parameter ([15], [11]). The reason why we use it here will appear in the computation of the performance of our double mixture estimator. Let us recall a formula that will be used frequently in the sequel:

$$\forall \lambda \in \left(\mathbb{R}^+\right)^{\alpha} \qquad \int_{\Sigma} \left(\theta^1\right)^{\lambda^1} \ldots \left(\theta^{\alpha}\right)^{\lambda^{\alpha}} \mu\left(d\theta\right) = \frac{\Gamma\left(\frac{\alpha}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{\alpha}} \frac{\prod_{i=1}^{\alpha} \Gamma\left(\lambda^i + \frac{1}{2}\right)}{\Gamma\left(\sum_{i=1}^{\alpha} \lambda^i + \frac{\alpha}{2}\right)}. \tag{3.3}$$

Based on these priors for the continuous parameters and on an arbitrary prior π on the model set $\mathcal{M}$, we can construct a double mixture estimator which takes the form of a Gibbs mixture, as defined by Catoni in [9]. In order to clarify the definition of this estimator it is convenient to introduce notations for the entropy of a Bernoulli model and the Kullback-Leibler divergence between two such models, respectively as:

$$\forall \theta \in \Sigma \qquad h(\theta) = -\sum_{y \in \mathcal{Y}} \theta^y \log \theta^y,$$

$$\forall \left(\theta_1, \theta_2\right) \in \Sigma^2 \qquad d\left(\theta_1 || \theta_2\right) = \sum_{y \in \mathcal{Y}} \theta_1^y \log \frac{\theta_1^y}{\theta_2^y}.$$

For any model $m \in \mathcal{M}$ let us also introduce the following random variables which are expressed in terms of Z_1^N (the dependency w.r.t. m is not indicated in order to simplify the notations, and because no ambiguity about which m these variables refer to should arise in the sequel):

$$\forall (y, s) \in \mathcal{Y} \times \mathcal{S}_m \qquad a_s^y = \sum_{j=1}^{N-1} \mathbf{1}(s_m(X_j) = s \text{ and } Y_j = y),$$

$$\forall (y, s) \in \mathcal{Y} \times \mathcal{S}_m \qquad b_s^y = \mathbf{1}(s_m(X_N) = s \text{ and } Y_N = y),$$

$$\forall (y, s) \in \mathcal{Y} \times \mathcal{S}_m, \forall (\beta, \xi) \in \mathbb{R}^2 \qquad \eta_s^y(\beta, \xi) = \beta a_s^y + \xi b_s^y,$$

$$\forall s \in \mathcal{S}_m, \forall (\beta, \xi) \in \mathbb{R}^2 \qquad n_s(\beta, \xi) = \sum_{y \in \mathcal{Y}} \eta_s^y,$$

$$\forall s \in \mathcal{S}_m \qquad \bar{\theta}_s(\beta,\xi) = \left(\frac{\eta_s^1(\beta,\xi)}{n_s(\beta,\xi)}, \dots, \frac{\eta_s^\alpha(\beta,\xi)}{n_s(\beta,\xi)} \right).$$

The reason for using these notations basically comes from the following equality used to express the thermalized likelihood of a sequence Z_1^N with respect to a particular model (m, θ_m):

$$\begin{aligned} &\left(\prod_{i=1}^{N-1} p_{m,\theta_m}\left(Y_i|X_i\right) \right)^\beta p_{m,\theta_m}\left(Y_N|X_N\right)^\xi \\ &\qquad = \prod_{s\in\mathcal{S}_m} \prod_{y\in\mathcal{Y}} \left(\theta_s^y\right)^{\eta_s^y(\beta,\xi)} \\ &\qquad = \prod_{s\in\mathcal{S}_m} \exp\left[n_s(\beta,\xi) \sum_{y\in\mathcal{Y}} \bar{\theta}_s^y(\beta,\xi) \log \theta_s^y \right] \\ &\qquad = \prod_{s\in\mathcal{S}_m} \exp\left[-n_s(\beta,\xi) \left(h\left(\bar{\theta}_s(\beta,\xi)\right) + d\left(\bar{\theta}_s(\beta,\xi)||\theta_s\right)\right)\right]. \end{aligned} \tag{3.4}$$

Following these preliminaries we can now state the main theorem of this paper which contains the definition of the double mixture estimator as well as a universal bound for its risk:

THEOREM 1. *Let*

$$\bar{\chi} = 24 + 8 \log \left(N + \frac{\alpha}{2} + 1 \right),$$

and $\beta > 0$ *such that*

$$\beta < \frac{1}{\bar{\chi}-1} \left(\sqrt{1 + (\bar{\chi}-1)\left(2 - \frac{\log \bar{\chi}}{\bar{\chi}}\right) \frac{\log \bar{\chi}}{\bar{\chi}}} - 1 \right) \underset{N\to\infty}{\sim} \frac{\sqrt{2 \log\log(N)}}{8 \log(N)}.$$

For any exchangeable distribution P_N *on* $\left(\mathcal{X}\times\mathcal{Y}\right)^N$ *, for any choice of prior probability distribution* π *on* $\mathcal{M}$*, the posterior Gibbs distribution* ρ *defined on the set*

$$\left\{ \left(m, \theta_m\right), m \in \mathcal{M}, \theta_m \in \Theta_m \right\}$$

by

$$\rho\left(d\theta_m|m\right) \sim \prod_{s\in\mathcal{S}_m} \exp\left[-n_s(\beta,0) d\left(\bar{\theta}_s(\beta,0)||\theta_s\right)\right] \mu\left(d\theta_s\right),$$

and

$$\begin{aligned}\rho\left(m\right) \sim \pi\left(m\right) \prod_{s\in\mathcal{S}_m} \Bigg\{ & \frac{\pi^{\alpha/2}}{\Gamma\left(\frac{\alpha}{2}\right)} \left(\frac{n_s\left(\beta,\beta\right)}{2\pi e}\right)^{\frac{\alpha-1}{2}} \\ & \times \int_{\Sigma} \exp\left\{-n_s(\beta,0)\left[h\left(\bar{\theta}_s(\beta,0)\right) + d\left(\bar{\theta}_s(\beta,0)||\theta_s\right)\right]\right\} \mu\left(d\theta_s\right)\Bigg\},\end{aligned}$$

can be used to form the double mixture estimator

$$G_\beta^N\left(dY_N|X_N;Z_1^{N-1}\right) = \mathbf{E}_{\rho(m,d\theta_m)} P_{m,\theta_m}\left(dY_N|X_N\right)$$

which satisfies

$$\begin{aligned}&\mathbf{E}_{P_N\left(dZ_1^{N-1}\right)}\mathcal{K}\left(P_N\left(dY_N|X_N;Z_1^{N-1}\right), G_\beta^N\left(dY_N|X_N;Z_1^{N-1}\right)\right) \\ &\quad \le \inf_{m\in\mathcal{M},\theta_m\in\Theta_m}\Bigg\{\mathbf{E}_{P_N\left(dZ_1^{N-1}\right)}\mathcal{K}\left(P_N\left(dY_N|X_N;Z_1^{N-1}\right), P_{m,\theta_m}\left(dY_N|X_N\right)\right) \\ &\qquad\qquad + \frac{1}{\beta N}\left(D_m\frac{\alpha-1}{2} + \log\frac{1}{\pi\left(m\right)} + C_N\left(m\right)\right)\Bigg\},\end{aligned}$$

with

$$C_N(m) = \mathbf{E}_{P_N}\sum_{s\in\mathcal{S}_m}\left(\frac{\alpha^2}{4n_s(\beta,\beta)} + \frac{\alpha}{4n_s\min_i(\bar{\theta}^i(\beta,\beta)) + 2}\right).$$

This theorem is proved in section 6.

Remark 1.
The double mixture estimator can be expressed in the following way:

$$\begin{aligned}G_\beta^N\left(dY_N|X_N;Z_1^{N-1}\right) &= \mathbf{E}_{\rho(m,d\theta_m)} P_{m,\theta_m}\left(dY_N|X_N\right) \\ &= \sum_{m\in\mathcal{M}} \rho(m)\mathbf{E}_{\rho(d\theta_m|m)} P_{m,\theta_m}\left(dY_N|X_N\right).\end{aligned}$$

Therefore it is a mixture under $\rho(m)$ of the estimator $\mathbf{E}_{\rho(\theta_m|m)} P_{m,\theta_m}\left(dY_N|X_N\right)$ which is, for any given model m, a thermalized version (at inverse temperature β) of a Bayesian estimator for the continuous parameters with respect to Jeffrey's prior on every simplex. This Bayesian estimator for θ_m has been studied in particular by Krichevsky and Trofimov ([15]). One interesting feature is that it can easily be computed using the following formula:

$$\mathbf{E}_{\rho(d\theta_m|m)} p_{m,\theta_m}(y_N|x_N) = \frac{\beta a^{y_N}_{s_m(x_N)} + \frac{1}{2}}{\beta \sum_{y\in\mathcal{Y}} a^{y}_{s_m(x_N)} + \frac{\alpha}{2}} \tag{3.5}$$

Remark 2. *For a given model m the additional term $C_N(m)$ decreases to zero and becomes negligible compared to the other terms as soon as the projections $P(dY|s_m(X) = s)$ are in the interior of the simplex for all $s \in \mathcal{S}_m$. Every node $s \in \mathcal{S}_m$ for which this projection is on the vertex of the simplex adds a risk of order $\frac{\alpha}{2}$ which is not negligible anymore compared to $D_m(\alpha - 1)/2 + \log 1/\pi(m)$. This is due to the fact that Jeffrey's prior is asymptotically minimax in the interior of the simplex, but only maximin on the whole simplex (see [26]).*

Remark 3. *The upper bound on the inverse temperature β is of order $(\log\log N)^{\frac{1}{2}}/\log N$. However this bound might be conservative, and it is reasonable to think from the computations in [9] and the experiments in [23] that $\beta = 1/2$ will work in many cases.*

4. Twice-universal coding and statistical estimation Let us have a look at the differences between the double mixture estimator we introduce for statistical regression purpose and "twice-universal" estimators used for compression in coding theory. In the compression framework the variables $Y_1^i = Y_1 \dots Y_i$ play the role of the variables X_i and the goal is to design a family of conditional probabilities $\left\{\hat{P}^i(dY_i|X_i)\right\}_{i\in\mathbb{N}}$ such that their per-sample redundancies be small compared to the per-sample redundancy of the best model when the number of observation goes to infinity, i.e. that

$$\frac{1}{N}\mathbf{E}_{P^N} \log \frac{1}{\prod_{i=1}^N \hat{P}^i(Y_i|X_i)} \leq \inf_{m\in\mathcal{M},\theta_m\in\Theta_m} \frac{1}{N}\mathbf{E}_{P^N} \log \frac{1}{\prod_{i=1}^N P_{m,\theta_m}(Y_i|X_i)} + \epsilon_N(m,\theta_m)$$

with

$$\forall m \in \mathcal{M}, \forall \theta_m \in \Theta_m \qquad \lim_{N\to\infty} \epsilon_N(m,\theta_m) = 0.$$

More precisely such a family of estimators is called a "twice-universal" code ([20]) if it is strongly universal (in the sense of [13]) with respect to every model $m \in \mathcal{M}$, i.e. if $\sup_{\theta_m} \epsilon_N(m,\theta_m)$ goes to zero as N goes to infinity with the minimax rate of convergence for the model m. In other words a twice-universal code is minimax up to a vanishing term in the convergence rate with respect to every model m.

A good solution to this apparently difficult problem is to take for $\hat{P}$ a so-called "two-stage" mixture or double mixture ([14]), that is a discrete mixture of estimators for every model which are themselves mixtures of the probability distributions in the model class w.r.t. to a least favorable prior on the continuous parameter set. An impressive implementation of this idea has been carried out for a binary alphabet ($\alpha = 2$) in the so-called context tree weighting algorithm ([24]), where Jeffrey's prior $\mu(d\theta)$ is used on every simplex together with an arbitrary prior π on the model set to build the estimator:

$$p_w^i(y_i|x_i) = \frac{\sum_{m\in\mathcal{M}} \pi(m) \int_{\theta_m\in\Theta_m} \prod_{j=1}^{i} p_{m,\theta_m}(y_j|x_j)\mu(d\theta_m)}{\sum_{m\in\mathcal{M}} \pi(m) \int_{\theta_m\in\Theta_m} \prod_{j=1}^{i-1} p_{m,\theta_m}(y_j|x_j)\mu(d\theta_m)}, \tag{4.6}$$

which satisfies :

$$\frac{1}{N}\mathbf{E}_{P^N} \log \frac{1}{\prod_{i=1}^N P_w^i(Y_i|X_i)} \leq \inf_{m\in\mathcal{M},\theta_m\in\Theta_m} \frac{1}{N}\mathbf{E}_{P^N} \log \frac{1}{\prod_{i=1}^N P_{m,\theta_m}(Y_i|X_i)} + \frac{1}{N}\left[\log\frac{1}{\pi(m)} + \sum_{s\in\mathcal{S}_m}\left(\frac{\log n_s}{2}+1\right)\right].$$

This expression shows that this family of estimators has a per-sample redundancy which decreases at the minimax rate $D_m \log N/(2N)$.

In the case of statistical regression the criterion we are interested in is slightly different from the redundancy used for compression purpose. Indeed we are only interested in the estimation of the conditional law of Y_N knowing X_N and the observations Z_1^N, while the redundancy is only an average of this criterion for $i = 1, \ldots, N$. The relationship between the redundancy and the statistical risk is more precisely expressed in the following equality :

$$\frac{1}{N}\mathbf{E}_{P^N} \log \frac{1}{\prod_{i=1}^N Q(Y_i|X_i)} = \frac{1}{N}\sum_{i=1}^N \mathbf{E}_{P^i} \log \frac{1}{Q(Y_i|X_i)}.$$

In other words the estimates of the conditional law of Y_N knowing X_N obtained from a universal coding procedure have good performances in terms of *cumulative* risk for an increasing number of observations.

In order to compare the twice universal coding algorithm (4.6) and our double mixture statistical estimator defined in theorem 1 we need to rewrite (4.6) as:

$$p_w^i(y_N|x_N) = \sum_{m\in\mathcal{M}} \rho_c(m) \frac{\int_{\theta_m\in\Theta_m} \prod_{j=1}^{N} p_{m,\theta_m}(y_j|x_j)\mu(d\theta_m)}{\int_{\theta_m\in\Theta_m} \prod_{j=1}^{N-1} p_{m,\theta_m}(y_j|x_j)\mu(d\theta_m)},$$

with

$$\forall m \in \mathcal{M} \qquad \rho_c(m) = \frac{\pi(m) \displaystyle\int_{\theta_m \in \Theta_m} \prod_{j=1}^{N-1} p_{m,\theta_m}(y_j|x_j)\mu(d\theta_m)}{\displaystyle\sum_{m' \in \mathcal{M}} \pi(m') \int_{\theta'_m \in \Theta'_m} \prod_{j=1}^{N-1} p_{m',\theta'_m}(y_j|x_j)\mu(d\theta_m)}. \tag{4.7}$$

If one forgets one second about the inverse temperature β (think of $\beta = 1$), it is possible to compare this posterior with the one expressed in theorem 1 to point out one important difference:

$$\forall m \in \mathcal{M} \qquad \rho(m) \sim \rho_c(m) \times \prod_{s \in \mathcal{S}_m} \left[\left(\frac{n_s(1,1)}{2\pi e} \right)^{\frac{\alpha-1}{2}} . \frac{\pi^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)} \right]. \tag{4.8}$$

This shows that in order to get a small statistical risk instead of a small cumulative risk one needs to modify the prior π on the model set in order to take into account the differences in the difficulty of estimating the continuous parameters. Besides a constant "penalization" term which does not depend on N, one sees in expression (4.8) that as N increases, models with a larger number of parameters should be given more and more weight because the term $\prod_{s \in \mathcal{S}_m} n_s^{\frac{\alpha-1}{2}}$ behaves like $N^{\frac{\alpha-1}{2} D_m}$.

An other way to look at the particularity of our double mixture estimator is to observe that the posterior weight $\rho(m)$ of a model m is essentially proportional to the maximum likelihood of the observed sequence in the model class. Indeed one can notice that if $\bar{\theta}$ is in the interior of the simplex,

$$\int_\Sigma \exp\left(-nd(\bar{\theta}||\theta)\right) \underset{n \to \infty}{\sim} \frac{\Gamma\left(\frac{\alpha}{2}\right)}{\pi^{\frac{\alpha}{2}}} \times \left(\frac{n}{2\pi}\right)^{-\frac{\alpha-1}{2}},$$

and therefore, for any model m in $\mathcal{M}$,

$$\begin{aligned} \rho(m) &\underset{N \to \infty}{\sim} \frac{1}{Z} \pi(m) \exp\left(-D_m(\alpha-1)/2\right) \times \prod_{s \in \mathcal{S}_m} \int_\Sigma \exp\left(-n_s(\beta,0) h(\bar{\theta}_s(\beta,0))\right) \mu(d\theta_s) \\ &\underset{N \to \infty}{\sim} \frac{1}{Z} \pi(m) \exp\left(-D_m(\alpha-1)/2\right) \times \sup_{\theta_m \in \Theta_m} \prod_{i=1}^{N-1} p_{m,\theta_m}(Y_i|X_i)^\beta. \end{aligned} \tag{4.9}$$

Compared with (4.7) one sees that instead of doing a double mixture one should:

- replace the mixture estimator for continuous parameters by the maximum likelihood in every model;
- penalize the likelihood by a factor $\exp\left(-D_m(\alpha-1)/2\right)$.

As far as performance is concerned the main difference between our double mixture estimator and a twice universal code is that the bound of the statistical estimator is not on the cumulated risk.

REMARK 4. *An fundamental link exists between minimax estimators and mixture estimators : in can be proved under quite general assumptions that a mixture estimator with respect to a "least favorable" prior can be a minimax estimator (see a survey and references in [17]). Our formula for ρ gives an idea of what such a least favorable distribution could look like in the problem considered. Two points are of interest:*

- *The factor $exp\left(-D_m(\alpha-1)/2\right)$ in the expression of $\rho\,(m)$ can be regarded as a penalty term for the dimension of the continuous parameter in each model.*
- *The prior on the simplex is Jeffrey's prior. This suggests a penalty term for the continuous parameters inversely proportional to the variance of the corresponding Bernoulli models.*

This can also be related to penalized maximum likelihood estimators ([4]) in which a penalization of models proportional to their dimension arises for other reasons.

5. Double mixture on context trees

5.1. *The estimator* In this section we present a particular form of the double mixture estimator defined in theorem 1 when the variable X is a string and the models considered are context trees. In other words we consider the case $\mathcal{X}=\mathcal{Y}^D$. We will basically use the same models as described in [23] where an application in natural language processing is proposed.

Let D be a fixed integer. We define a model m by a non-empty set $\mathcal{S}_m \subset \bigcup_{i=0}^{D}\mathcal{Y}^i$ of finite strings of length not larger than D such that *any suffix of any string of* m *be also in* m (by definition a suffix of a string $x_1\ldots x_i$ is of the form $x_j\ldots x_i$ for some $j\le i$). This definition implies in particular that the empty string λ belongs to every model.

The projection s_m associated with a model m is simply the transformation from any string $x\in\mathcal{X}$ into its longest suffix that is in $\mathcal{S}_m$.

Finally we can define a natural probability on $\mathcal{M}$ as follows:

$$\forall m\in\mathcal{M}\qquad \pi(m)=C^{D_m}, \tag{5.10}$$

where the constant C is adjusted so that $\sum_{m\in\mathcal{M}}\pi(m)=1$.

It is shown in [23] that in that case,

$$\log\frac{1}{C}\le 1+\log\alpha.$$

Therefore the "model risk" is controlled as follows:

$$\forall m\in\mathcal{M}\qquad \log\frac{1}{\pi(m)}\le\left(1+\log\alpha\right)D_m.$$

As a result we can apply theorem 1 to this particular setting to obtain:

Theorem 2. *Let G_β^N be the double mixture estimator as defined in theorem 1. For any exchangeable distribution P_N on $(\mathcal{X} \times \mathcal{Y})^N$ it satisfies :*

$$\begin{aligned}
&\mathbf{E}_{P_N\left(dZ_1^{N-1}\right)}\mathcal{K}\left(P_N\left(dY_N|X_N;Z_1^{N-1}\right),G_\beta^N\left(dY_N|X_N;Z_1^{N-1}\right)\right) \\
&\quad\leq \inf_{m\in\mathcal{M},\theta_m\in\Theta_m}\left\{\mathbf{E}_{P_N\left(dZ_1^{N-1}\right)}\mathcal{K}\left(P_N\left(dY_N|X_N;Z_1^{N-1}\right),P_{m,\theta_m}\left(dY_N|X_N\right)\right)\right. \\
&\qquad\qquad\left.+\frac{D_m}{\beta N}\left(\frac{\alpha-1}{2}+1+\log\alpha+C_N\left(m\right)\right)\right\},
\end{aligned}$$

with

$$C_N(m) = \frac{1}{D_m}\mathbf{E}_{P_N}\sum_{s\in\mathcal{S}_m}\left(\frac{\alpha^2}{4n_s(\beta,\beta)}+\frac{\alpha}{4n_s\min_i(\bar{\theta^i}(\beta,\beta))+2}\right).$$

Note that the larger the alphabet size the smaller the model risk term $1+\log\alpha$ compared to the parameter risk $(\alpha-1)/2$.

5.2. *Implementation* The exact Gibbs estimator as a double mixture is difficult to compute efficiently because the number of models is very large as D and α increase. However it is possible to imagine a suboptimal implementation which computes an estimator which might be not so different from the double mixture estimators in many concrete cases.

We propose to replace the double mixture procedure by the selection of the model with the largest posterior distribution. Our hope is that in many cases the Gibbs posterior is unimodal and that the selection of a particular model with a large posterior probability is representative of the mixture in terms of probability law.

Following (4.9) and (5.10) we see that a good candidate for the quantity to maximize is:

$$\gamma(m) = \log\sup_{\theta_m\in\Theta_m}\prod_{i=1}^{N-1}p_{m,\theta_m}(Y_i|X_i) - \frac{D_m}{\beta}\left(\log\frac{1}{C}+\frac{\alpha-1}{2}\right).$$

This equation shows that the model selection we propose takes the form of a penalized maximum likelihood with a penalization proportional to the size of the model D_m. In order to get an efficient implementation of this model selection procedure we can use a context tree (see [24], [10], [23]), i.e. a suffix tree representing all possible strings of length smaller than D hierarchically, the root of the tree being the empty string λ. Let us attach the following counters to every node s of the context tree (we use the equivalence between a node and the associated string):

$$\forall i\in\mathcal{Y} \qquad a_s^i = \sum_{n=1}^{N-1}\mathbf{1}\left(s \text{ is a suffix of } x_n \text{ and } y_n = i\right),$$

$$n_s = \sum_{n=1}^{K} \mathbf{1}\left(s \text{ is a suffix of } x_n\right).$$

If we now note $\delta = \left(\log 1/C + (\alpha - 1)/2\right)/\beta$ and define the recursive function w on the context tree :

$$\begin{cases} \text{If } l(s) = D & w(s) = \displaystyle\sum_{y\in\mathcal{Y}} \frac{a_s^y}{n_s} \log \frac{a_s^y}{n_s} - \delta, \\ \text{If } l(s) < D & w(s) = \displaystyle\max_{\mathcal{N}\subset\mathcal{Y}} \left[\sum_{i\in\mathcal{N}} w(is) + \sum_{y\in\mathcal{Y}} \frac{a_s^y - \sum_{i\in\mathcal{N}} a_{is}^y}{n_s - \sum_{i\in\mathcal{N}} n_{is}} \log \frac{a_s^y - \sum_{i\in\mathcal{N}} a_{is}^y}{n_s - \sum_{i\in\mathcal{N}} n_{is}} \right] - \delta, \end{cases}$$

then it is easy to see that $\max_{m\in\mathcal{M}} \gamma(m) = w(\lambda)$ and that the model m which realizes this maximum is the connected component of λ in the set of nodes that are selected in $\mathcal{N}$ at every node in the definition of w.

For a given node s the problem remains to compute the corresponding subset $\mathcal{N}$ and to mark the selected node. This can be approximated using an iterative procedure to build $\mathcal{N}$, starting with $\mathcal{N} = \emptyset$ and adding nodes one by one until the function to maximize locally stops increasing.

The complexity of such an optimization procedure is linear in the number of nodes of the context tree, because at most α tests are performed at every node to test the children nodes to select. It is also not more than linear in N because only the visited nodes are concerned, and the size of memory required to store the context tree is also not more than linear in the number of observations and of course bounded by the size of the context tree. In [23] we show results of experiments using an implementation of an algorithm very similar to the one described in this paper (a "two-stage double mixture algorithm").

6. Proof of theorem 1

6.1. *The Gibbs estimator* Let us first recall some facts about the so-called Gibbs estimator introduced by Catoni in [9]. For a given class of conditional probability densities $\{p_\theta\}_{\theta\in\Theta}$ indexed by a parameter θ living in a measurable space Θ endowed with a prior probability measure $\pi(d\theta)$, the Gibbs estimator at inverse temperature $\beta \in \mathbb{R}^+$ has a density

$$g_\beta^N \left(y_N | x_N, z_1^{N-1}\right) = \mathbf{E}_{\rho_{\beta,0}(d\theta)} p_\theta \left(y_N | x_N\right)$$

where $\rho_{\beta,\xi}$ is the following Gibbs posterior:

$$\rho_{\beta,\xi}(d\theta) = \frac{\displaystyle\prod_{i=1}^{N-1} p_\theta\left(y_i|x_i\right)^\beta p_\theta\left(y_N|x_N\right)^\xi \pi(d\theta)}{\displaystyle\int_\Theta \prod_{i=1}^{N-1} p_\theta\left(y_i|x_i\right)^\beta p_\theta\left(y_N|x_N\right)^\xi \pi(d\theta)}.$$

This estimator can be considered as a "thermalized" version of both the Bayesian ($\beta = 1$) and the maximum likelihood ($\beta = +\infty$) estimators. Catoni studied in [9] this estimator in the high temperature region $\beta < 1$ which is equivalent to a deliberate underestimation of the sample size : to compute the Gibbs estimator, the empirical distribution of $N-1$ observations is plugged into the Bayes estimator for a sample of size $\beta(N-1)$. The reason to consider high temperatures is that the estimator gains stability with respect to the empirical process when β decreases (at the limit, it is constant when $\beta = 0$).

In order to control the risk of the Gibbs estimator let us introduce the following notations:

$$\chi = -\left(0 \wedge \inf_{\xi\in[0,1]} \frac{\mathbf{E}_{P_N} {\mathbf{M}^3}_{\rho_{\beta,\xi}^{Z_1^N}(d\theta)} \log p_\theta\left(y_N|x_N\right)}{\mathbf{E}_{P_N} \mathbf{Var}_{\rho_{\beta,\xi}^{Z_1^N}(d\theta)} \log p_\theta\left(y_N|x_N\right)}\right),$$

and

$$\gamma_\beta\left(\theta\right) = \mathbf{E}_{P_N} \mathbf{E}_{\rho_{\beta,\beta}^{Z_1^N}(\theta')} \left(\log \frac{\displaystyle\prod_{i=1}^{N} p_\theta(y_i|x_i)^\beta}{\displaystyle\prod_{i=1}^{N} p_{\theta'}(y_i|x_i)^\beta}\right).$$

The following result, which we will be used to estimate the performance of our double mixture estimator, is a particular form of the main theorem of [9]:

THEOREM 3 CATONI, [9]. *If the inverse temperature β is such that*

$$\beta < \frac{1}{\chi - 1}\left(\sqrt{1 + (\chi - 1)\left(2 - \frac{\log\chi}{\chi}\right)\frac{\log\chi}{\chi}} - 1\right),$$

then the risk of the Gibbs estimator at inverse temperature β is upper bounded by

$$\begin{aligned}&\mathbf{E}_{P_N\left(dZ_1^{N-1}\right)}\mathcal{K}\left(P_N\left(dY_N|X_N;Z_1^{N-1}\right), G_\beta^N\left(dY_N|X_N;Z_1^{N-1}\right)\right)\\ &\qquad\le \inf_{\theta\in\Theta}\left\{\mathbf{E}_{P_N\left(dZ_1^{N-1}\right)}\mathcal{K}\left(P_N\left(dY_N|X_N;Z_1^{N-1}\right), P_\theta\left(dY_N|X_N\right)\right) + \frac{\gamma_\beta(\theta)}{\beta N}\right\}.\end{aligned}$$

In case Θ is discrete one can show that $\gamma_\beta(\theta)$ is upper bounded by $\log \pi\left(\{\theta\}\right)^{-1}$. However the interesting point of this estimator is that it can be computed for any set of parameters Θ, not necessarily discrete. In our case, we propose to apply this estimator for the set of parameters

$$\Theta = \left\{(m, \theta_m), m \in \mathcal{M}, \theta_m \in \Theta_m\right\},$$

endowed with a probability density expressed as a product:

$$\pi\left(d\theta\right) = \bar{\pi}(m) \times \prod_{s \in \mathcal{S}_m} \mu(d\theta_s), \tag{6.11}$$

where $\bar{\pi}$ is a prior probability on $\mathcal{M}$ (we will show that it has to be taken somehow different from π). Theorem 1 will be a direct consequence of theorem 3 after estimating an upper bound on the inverse temperature β to be used and on the risk bound γ_β.

6.2. *Choice of the inverse temperature β* In this section we prove the following lemma:

LEMMA 1.

$$\chi \leq \bar{\chi} = 24 + 8 \log\left(N + \frac{\alpha}{2} + 1\right).$$

A direct consequence of this lemma is the possibility to chose the inverse temperature β as

$$\begin{aligned} \beta &< \frac{1}{\bar{\chi} - 1}\left(\sqrt{1 + (\bar{\chi} - 1)\left(2 - \frac{\log \bar{\chi}}{\bar{\chi}}\right)\frac{\log \bar{\chi}}{\bar{\chi}}} - 1\right) \\ &\underset{N \to \infty}{\sim} \frac{\sqrt{2 \log \log(N)}}{8 \log(N)}, \end{aligned}$$

in order to fulfill the conditions of theorem 3.

PROOF OF LEMMA 1. For any given $z_1^N \in \left(\mathcal{X} \times \mathcal{Y}\right)^N$ and $\beta \in [0, 1]$, let η and f be defined for $\xi \in [0, 1]$ by:

$$\left\{ \begin{array}{ll} \eta(\xi) & = \mathbf{E}_{\pi(d\theta)} \displaystyle\prod_{i=1}^{N-1} p_\theta\left(y_i | x_i\right)^\beta p_\theta\left(y_N | x_N\right)^\xi, \\ f(\xi) & = \log \eta(\xi) \end{array} \right.$$

The function f is related to the Gibbs estimator through the following equality:

$$\log g_\beta^N\left(y_N | x_N; z_1^N\right) = f(1) - f(0).$$

Moreover a simple computation shows that the first three derivatives of f are equal to the moments of $\log p_\theta(y_N|x_N)$ under $\rho_{\beta,\xi}^{z_1^N}(d\theta)$:

$$f'(\xi) = \mathbf{E}_{\rho_{\beta,\xi}^{z_1^N}(d\theta)} \log p_\theta\left(y_N|x_N\right),$$

$$f''(\xi) = \mathbf{Var}_{\rho_{\beta,\xi}^{z_1^N}(d\theta)} \log p_\theta\left(y_N|x_N\right),$$

$$f^{(3)}(\xi) = \mathbf{M}^3_{\rho_{\beta,\xi}^{z_1^N}(d\theta)} \log p_\theta\left(y_N|x_N\right).$$

Using (6.11) and (3.4) we see that for $\xi \in [0,1]$,

$$\begin{aligned}\eta\left(\xi\right) &= \sum_{m\in\mathcal{M}} \bar{\pi}(m) \int_{\Theta_m} \prod_{i=1}^{N-1} p_{m,\theta_m}\left(y_i|x_i\right)^\beta p_{m,\theta_m}\left(y_N|x_N\right)^\xi \mu(d\theta_m)\\ &= \sum_{m\in\mathcal{M}} \bar{\pi}(m) \prod_{s\in\mathcal{S}_m} \int_\Sigma \exp\left[-n_s(\beta,\xi)\left(h(\bar{\theta}_s(\beta,\xi)) + d(\bar{\theta}_s(\beta,\xi)||\theta_s)\right)\right]\mu\left(d\theta_s\right).\end{aligned}$$

However for every model $m \in \mathcal{M}$ the variables $n_s(\beta,\xi)$ and $\bar{\theta}_s(\beta,\xi)$ only depend on ξ for $s = s(x_N)$. Besides, using (3.3), the integral involved in the preceding formula is known to be (for $s = s(x_N)$):

$$\int_\Sigma \exp\left[-n_s(\beta,\xi)\left(h(\bar{\theta}_s(\beta,\xi)) + d(\bar{\theta}_s(\beta,\xi)||\theta_s)\right)\right]\mu\left(d\theta_s\right) = Cte \times \frac{\Gamma\left(\beta a_s^{y_N} + \frac{1}{2} + \xi\right)}{\Gamma\left(n_s + \frac{\alpha}{2} + \xi\right)},$$

where Cte is a term which does not depend on ξ. As a result, if we introduce the functions:

$$\forall (m,\xi) \in \mathcal{M}\times[0,1] \qquad \mu_m\left(\xi\right) = \frac{\Gamma\left(\beta a_{s_m(s_N)}^{y_N} + \frac{1}{2} + \xi\right)}{\Gamma\left(n_s + \frac{\alpha}{2} + \xi\right)},$$

then η can be decomposed in the following way:

$$\eta(\xi) = \sum_{m\in\mathcal{M}} \lambda_m \mu_m(\xi), \tag{6.12}$$

where the $\left(\lambda_m\right)_{m\in\mathcal{M}}$ do not depend on ξ.

In order to express the derivatives of μ and η let us introduce the Polygamma functions:

$$\forall i \in \mathbb{N} \qquad \psi_i(z) = \frac{d^{i+1}}{dz^{i+1}} \log\Gamma(z).$$

Indeed, if we use the notation:

$$\forall\,(i,m) \in \mathbb{N} \times \mathcal{M} \qquad \phi_i^m(\xi) = \psi_i\left(\beta a_{s_m(x_N)}^{y_N} + \frac{1}{2} + \xi\right) - \psi_i\left(n_{s_m(x_N)} + \frac{\alpha}{2} + \xi\right),$$

then we get:

$$\mu'_m = \mu_m \phi_0^m, \tag{6.13}$$

$$\mu''_m = \mu_m \left(\phi_1^m + (\phi_0^m)^2\right), \tag{6.14}$$

$$\mu_m^{(3)} = \mu_m \left((\phi_0^m)^3 + 3\phi_0^m \phi_1^m + \phi_2^m\right). \tag{6.15}$$

In order to control γ_β we will need the following controls on ϕ_i^m :

Lemma 2. *For all (z_1^N, ξ, m) in $\mathcal{Z}_1^N \times [0,1] \times \mathcal{M}$ the following inequalities hold:*

$$0 \geq \phi_0^m(\xi) \geq -\left(\log\left(N + \frac{\alpha}{2} + 1\right) + 3\right),$$

and for any integer i, $i \geq 1$:

$$0 \geq \frac{\phi_{i+1}^m(\xi)}{\phi_i^m(\xi)} \geq -2(i+2).$$

Proof of Lemma 2:. In order to prove the first inequality concerning ϕ_0^m we use the fact (see [18]) that the Psi function ψ_0 is increasing on $\mathbf{R}_*^+$, that $\psi_0(1/2) = -\gamma - 2\log 2$ and that $\psi_0(z) \leq \log z + 1$ Therefore the following inequality holds $\forall (z_1^N, \xi, m) \in \mathcal{Z}_1^N \times [0,1] \times \mathcal{M}$:

$$\begin{aligned} 0 \geq \phi_0^m(\xi) &\geq \psi_0\left(\frac{1}{2}\right) - \psi_0\left(\beta N + \frac{\alpha}{2} + 1\right) \\ &\geq -\gamma - 2\log 2 - \log\left(\beta N + \frac{\alpha}{2} + 1\right) - 1 \\ &\geq -\left(\log\left(N + \frac{\alpha}{2} + 1\right) + 3\right) \end{aligned}$$

In order to prove the second inequality of lemma 2 we can use the expression of ψ_i in terms of the Hurwitz Zeta function, for $i \geq 1$:

$$\psi_i(u) = (-1)^{i+1} i! \sum_{k=0}^{\infty} \frac{1}{(k+u)^{i+1}}.$$

This shows that for any $u > 1/2$ and $i \geq 1$:

$$0 \leq -\frac{\psi_{i+1}(z)}{\psi_i(z)} \leq 2(i+1).$$

Therefore, $\forall (z_1^N, \xi, m) \in \mathcal{Z}_1^N \times [0,1] \times \mathcal{M}$ and $i \geq 1$:

$$0 \geq \frac{\phi_{i+1}^m(\xi)}{\phi_i^m(\xi)} = \frac{\displaystyle\int_{\beta a^{y_N}_{s_m(x_N)}+\frac{1}{2}+\xi}^{n_{s_m(x_N)}+\frac{\alpha}{2}+\xi} \psi_{i+2}(u)du}{\displaystyle\int_{\beta a^{y_N}_{s_m(x_N)}+\frac{1}{2}+\xi}^{n_{s(x_N)}+\frac{\alpha}{2}+\xi} \psi_{i+1}(u)du}$$

$$\geq -2(i+2) \qquad \square$$

We can now concentrate on the problem of upper bounding χ. Using the fact that $f = \log \eta$, one easily gets, for any given z_1^N:

$$\begin{aligned}
\frac{\mathbf{M}^3_{\rho^{z_1^N}_{\beta,\xi}(d\theta)} \log p_\theta\left(y_N|x_N\right)}{\mathbf{Var}_{\rho^{z_1^N}_{\beta,\xi}(d\theta)} \log p_\theta\left(y_N|x_N\right)} &= \frac{f^{(3)}}{f''}(\xi) \\
&= \frac{\eta^{(3)}\eta - \eta''\eta'}{\eta''\eta - (\eta')^2}(\xi) - \frac{2\eta'}{\eta}(\xi) \qquad (6.16) \\
&\geq \frac{\eta^{(3)}\eta - \eta''\eta'}{\eta''\eta - (\eta')^2}(\xi),
\end{aligned}$$

the last inequality holding because $\eta'/\eta = f' = \mathbf{E}_\rho \log p_\theta(y_N|x_N) \leq 0$.

Let us now consider an ordered list of models : $\mathcal{M} = (m_1, \dots)$. In order to simplify the notations, let us write ϕ_i^j for $\phi_i^{m_j}$, for $i \in \mathbb{N}$, and let us note:

$$\forall (i,j) \in \mathbb{N}^2, \qquad q_{i,j} = \begin{cases} \lambda_{m_i}\lambda_{m_j}\mu_{m_i}(\xi)\mu_{m_j}(\xi) & \text{if } i \neq j; \\ \frac{1}{2}\lambda_{m_i}^2\mu_{m_i}(\xi)^2 & \text{if } i = j. \end{cases}$$

Using (6.16), (6.12) and (6.13) we finally get:

$$\begin{aligned}
\chi &\leq - \inf_{z_1^N \in \mathcal{Z}_1^N} \frac{\displaystyle\sum_{(i,j)\in\mathbb{N}^2} q_{i,j}\left[(\phi_0^i + \phi_0^j)(\phi_0^i - \phi_0^j)^2 + \phi_1^i(3\phi_0^i - \phi_0^j) + \phi_1^j(3\phi_0^j - \phi_0^i) + \phi_2^i + \phi_2^j\right]}{\displaystyle\sum_{(i,j)\in\mathbb{N}^2} q_{i,j}\left[(\phi_0^i - \phi_0^j)^2 + \phi_1^i + \phi_1^j\right]} \\
&\leq - \inf_{z_1^N \in \mathcal{Z}_1^N} \inf_{(i,j)\in\mathbb{N}^2} \frac{(\phi_0^i + \phi_0^j)(\phi_0^i - \phi_0^j)^2 + \phi_1^i(3\phi_0^i - \phi_0^j) + \phi_1^j(3\phi_0^j - \phi_0^i) + \phi_2^i + \phi_2^j}{(\phi_0^i - \phi_0^j)^2 + \phi_1^i + \phi_1^j} \\
&\leq - \inf_{z_1^N \in \mathcal{Z}_1^N} \inf_{(i,j)\in\mathbb{N}^2} \left(\frac{(\phi_0^i + \phi_0^j)(\phi_0^i - \phi_0^j)^2}{(\phi_0^i - \phi_0^j)^2} + \frac{\phi_1^i(3\phi_0^i - \phi_0^j)}{\phi_1^i} + \frac{\phi_1^j(3\phi_0^j - \phi_0^i)}{\phi_1^j} + \frac{\phi_2^i}{\phi_1^i} + \frac{\phi_2^j}{\phi_1^j}\right) \\
&\leq - \inf_{z_1^N \in \mathcal{Z}_1^N} \inf_{(i,j)\in\mathbb{N}^2} \left(4\phi_0^i + 4\phi_0^j + \frac{\phi_2^i}{\phi_1^i} + \frac{\phi_2^j}{\phi_1^j}\right) \\
&\leq 24 + 8\log\left(N + \frac{\alpha}{2} + 1\right),
\end{aligned}$$

which proves lemma 1. □.

6.3. *Upper bound for the risk* Let us first state a lemma in order to be able to control the Psi function. Remember that the Psi function ψ_0 (also called Digamma function) is defined by

$$\psi_0(x) = \frac{\partial}{\partial x} \log \Gamma(x).$$

Lemma 3.

$$\begin{aligned} \forall x > 0 \qquad & \psi_0\left(x + \frac{1}{2}\right) \geq \log x, \\ \forall \alpha \in \mathbb{N}, \alpha \geq 2, \forall x > 0 \qquad & \psi_0\left(x + \frac{\alpha}{2}\right) \leq \log x + \frac{\alpha - 1}{2x}. \end{aligned}$$

Proof of lemma 3:. To prove the first inequality, we write ψ_0 as an integral:

$$\forall x > 0 \qquad \psi_0(x) = \int_0^\infty \left(\frac{e^{-t}}{t} - \frac{e^{-tx}}{1 - e^{-t}}\right) dt,$$

and do the same for $\log x$:

$$\forall x > 0 \qquad \log x = \int_0^\infty \frac{e^{-t} - e^{-tx}}{t} dt.$$

Therefore, for all $x > 0$,

$$\begin{aligned} \log x - \psi_0\left(x + \frac{1}{2}\right) &= \int_0^\infty \frac{e^{-xt-t/2}}{1 - e^{-t}} - \frac{e^{-tx}}{t} dt \\ &= \int_0^\infty e^{-tx} \phi(t) dt, \end{aligned}$$

with, for all $t > 0$,

$$\begin{aligned} \phi(t) &= \frac{e^{-t/2}}{1 - e^{-t}} - \frac{1}{t} \\ &= \frac{1}{2 \sinh\left(\frac{t}{2}\right)} - \frac{1}{t}. \end{aligned}$$

Now it suffices to notice that $\sinh(t) \geq t$, which implies that $2\sinh(t/2) \geq t \geq 0$, and therefore $\phi(t) \leq 0$ for all $t > 0$. This proves the first inequality of the lemma.

For the second inequality we can use the following, proved for instance in [2]:

$$\forall x > 0 \qquad \psi_0(x) < \log x - \frac{1}{2x}.$$

Therefore we can write, for all $x > 0$,

$$\psi_0\left(x+\frac{\alpha}{2}\right)-\log x-\frac{\alpha-1}{2x}\leq\phi_\alpha(x),$$

with

$$\phi_\alpha(x)=\log\left(1+\frac{\alpha}{2x}\right)-\frac{\alpha-1}{2x}-\frac{1}{2x+\alpha}.$$

Let us introduce $y=\alpha/(2x)$. Then we can write :

$$\phi_\alpha(x)=\phi_\alpha^y(y)=\log(1+y)-\frac{\alpha-1}{\alpha}y-\frac{y}{\alpha(1+y)}.$$

whose derivative is

$$(\phi_\alpha^y)'(y)=\frac{y}{\alpha(1+y)^2}\left[y(1-\alpha)+2-\alpha\right].$$

But α is supposed to be an integer larger or equal to 2, so $(\phi_\alpha^y)'(y)\leq 0$ for $y>0$. In other words ϕ_α^y as a function of y is decreasing on $\mathbb{R}^+$, and $\phi_\alpha^y(0)=0$. As a result, $\phi_\alpha^y(y)\leq 0$ for all $y>0$. This is sufficient to prove the second inequality of lemma 3. □

Let us now evaluate the risk defined by

$$\forall m\in\mathcal{M},\theta_m\in\Theta_m\qquad\gamma_\beta(m,\theta_m)=\mathbf{E}_{P_N}\mathbf{E}_{\rho_{\beta,\beta}^{Z_1^N}(m',d\theta')}\left(\log\frac{\prod_{i=1}^N p_{m,\theta_m}(y_i|x_i)^\beta}{\prod_{i=1}^N p_{m',\theta_{m'}}(y_i|x_i)^\beta}\right).$$

Using equation (3.4) it is possible to express the posterior Gibbs distribution as:

$$\rho_{\beta,\xi}\left(d\theta_m|m\right)=\frac{\prod_{s\in\mathcal{S}_m}\exp\left[-n_s(\beta,\xi)d\left(\bar{\theta}_s(\beta,\xi)||\theta_s\right)\right]\mu\left(\theta_s\right)d\theta_s}{\prod_{s\in\mathcal{S}}\int_\Sigma\exp\left[-n_s(\beta,\xi)d\left(\bar{\theta}_s(\beta,\xi)||\theta_s\right)\right]\mu\left(\theta_s\right)d\theta_s},\tag{6.17}$$

and

$$\rho_{\beta,\xi}\left(\mathcal{S}\right)\sim\bar{\pi}\left(\mathcal{S}\right)\prod_{s\in\mathcal{S}}\int_\Sigma\exp\left[-n_s(\beta,\xi)d\left(\bar{\theta}_s(\beta,\xi)||\theta_s\right)\right]\mu\left(\theta_s\right)d\theta_s$$
$$\times\exp\left[-\sum_{s\in\mathcal{S}}n_s(\beta,\xi)h\left(\bar{\theta}_s(\beta,\xi)\right)\right].\tag{6.18}$$

In order to simplify the notations let us write $n_s = n_s(\beta,\beta)$ and $\bar{\theta}_s = \bar{\theta}_s(\beta,\beta)$ in the rest of this section. As a first step let us prove the following lemma:

LEMMA 4. *For all m in $\mathcal{M}$,*

$$\mathbf{E}_{\rho_{\beta,\beta}^{Z_1^N}(d\theta_m|m)} \log \prod_{i=1}^{N} p_{m,\theta_m}(y_i|x_i)^{-\beta} \leq \sum_{s\in\mathcal{S}_m} n_s h\left(\bar{\theta}_s\right) + \frac{\alpha-1}{2} D_m.$$

PROOF OF LEMMA 4:. Using (6.17) we get:

$$\mathbf{E}_{\rho_{\beta,\beta}^{Z_1^N}(d\theta_m|m)} \log \prod_{i=1}^{N} p_{m,\theta_m}(y_i|x_i)^{-\beta}$$

$$= \sum_{s\in\mathcal{S}_m} n_s h\left(\bar{\theta}_s\right) + \sum_{s\in\mathcal{S}_m} \frac{\displaystyle\int_{\Sigma} n_s d\left(\bar{\theta}_s||\theta_s\right) e^{-n_s d\left(\bar{\theta}_s||\theta_s\right)} \mu(d\theta_s)}{\displaystyle\int_{\Sigma} e^{-n_s d\left(\bar{\theta}_s||\theta_s\right)} \mu(d\theta_s)}.$$

Consider now the function defined for $x \in \mathbb{R}^+$ by

$$f(x) = \prod_{s\in\mathcal{S}} \int_{\Sigma} e^{-x n_s d\left(\bar{\theta}_s||\theta_s\right)} \mu(d\theta_s).$$

All integrals being absolutely convergent the derivation under the integral is possible around $x = 1$, and one gets:

$$\mathbf{E}_{\rho_{\beta,\beta}^{Z_1^N}(d\theta_m|m)} \log \prod_{i=1}^{N} p_{m,\theta_m}(y_i|x_i)^{-\beta} = \sum_{s\in\mathcal{S}_m} n_s h\left(\bar{\theta}_s\right) - \frac{f'(1)}{f(1)}. \tag{6.19}$$

But $f'/f = (\log f)'$, so let us compute $\log f(x)$ for $x > 0$:

$$\log f(x) = \sum_{s\in\mathcal{S}} \log \int_{\Sigma} e^{-x n_s d\left(\bar{\theta}_s||\theta_s\right)} \mu(d\theta_s)$$

$$= \sum_{s\in\mathcal{S}} \left\{ x n_s h\left(\bar{\theta}_s\right) + \log \int_{\Sigma} e^{-x n_s \left[h\left(\bar{\theta}_s\right) + d\left(\bar{\theta}_s||\theta_s\right)\right]} \mu(d\theta_s) \right\}.$$

The exact value of the integral is known in terms of the Gamma function (thanks to ((3.3)):

$$\log \int_{\Sigma} e^{-x n_s \left[h\left(\bar{\theta}_s\right) + d\left(\bar{\theta}_s||\theta_s\right)\right]} \mu(d\theta_s) = C + \sum_{i=1}^{\alpha} \log\Gamma\left(x a_s^i + \frac{1}{2}\right) - \log\Gamma\left(x n_s + \frac{\alpha}{2}\right),$$

where C does not depend on x. Taking the derivatives in $x = 1$ of these expressions and using lemma 3 we finally get:

$$\begin{aligned}
-\frac{f'(1)}{f(1)} &= -\sum_{s\in\mathcal{S}}\left[\sum_{i=1}^{\alpha} a_s^i\left(\psi_0\left(a_s^i+\frac{1}{2}\right)-\log a_s^i\right)-n_s\left(\psi_0\left(n_s+\frac{\alpha}{2}\right)-\log n_s\right)\right]\\
&\leq \sum_{s\in\mathcal{S}}\frac{\alpha-1}{2}\\
&\leq \frac{\alpha-1}{2}D_m,
\end{aligned}$$

and coming back to (6.19) we obtain:

$$\mathbf{E}_{\rho_{\beta,\beta}^{z_1^N}(d\theta_m|m)}\log\prod_{i=1}^{N}p_{m,\theta_m}(y_i|x_i)^{-\beta} \leq \sum_{s\in\mathcal{S}_m} n_s h\left(\bar{\theta_s}\right)+\frac{\alpha-1}{2}D_m. \qquad \square$$

Let us use the notation:

$$\forall m\in\mathcal{M}, \forall z_1^N\in(\mathcal{X}\times\mathcal{Y})^N \qquad \lambda(m,z_1^N)=\prod_{s\in\mathcal{S}_m}\int_{\Sigma}e^{-n_s d\left(\bar{\theta}_s||\theta_s\right)}\mu\left(\theta_s\right)d\theta_s. \quad (6.20)$$

Using (6.18) and (6.19) we get the following equality, for any m in $\mathcal{M}$:

$$\begin{aligned}
&\mathbf{E}_{\rho_{\beta,\beta}(m,d\theta_m)}\log\prod_{i=1}^{N}p_{m,\theta_m}(y_i|x_i)^{-\beta}\\
&\quad\leq \frac{\sum_{m\in\mathcal{M}}\bar{\pi}\left(m\right)\lambda\left(m,z_1^N\right)e^{-\sum_{s\in m}n_s h\left(\bar{\theta}_s\right)}\left(\sum_{s\in m}n_s h\left(\bar{\theta}_s\right)+\frac{\alpha-1}{2}D_m\right)}{\sum_{m\in\mathcal{M}}\bar{\pi}\left(m\right)\lambda\left(m,z_1^N\right)e^{-\sum_{s\in m}n_s h\left(\bar{\theta}_s\right)}}\\
&\quad\leq \frac{\sum_{m\in\mathcal{M}}\tilde{\pi}\left(m\right)\tilde{g}\left(m\right)e^{-\tilde{g}(m)}}{\sum_{m\in\mathcal{M}}\tilde{\pi}\left(m\right)e^{-\tilde{g}(m)}},
\end{aligned}$$

with, for all m in $\mathcal{M}$:

$$\begin{aligned}
\tilde{\pi}\left(m\right) &= \bar{\pi}\left(m\right)e^{\frac{\alpha-1}{2}D_m}\lambda\left(m,z_1^N\right),\\
\tilde{g}\left(m\right) &= \sum_{m\in\mathcal{M}}n_s(\beta,\beta)h\left(\bar{\theta}_s(\beta,\beta)\right)+\frac{\alpha-1}{2}D_m\\
&= \sup_{\theta_m\in\Sigma^{D_m}}\log\prod_{i=1}^{N}p_{m,\theta_m}(y_i|x_i)^{-\beta}+\frac{\alpha-1}{2}D_m.
\end{aligned}$$

Introducing a threshold ϵ to be optimized afterwards and using the fact that xe^{-x} is upper bounded by e^{-1} on $\mathbf{R}^{+}$, this expression can be upper bounded for any particular $\bar{m} \in \mathcal{M}$ by:

$$\begin{aligned}
&\frac{\sum\limits_{m\in\mathcal{M}} \tilde{\pi}\left(m\right) \tilde{g}\left(m\right) e^{-\tilde{g}(m)}}{\sum\limits_{m\in\mathcal{M}} \tilde{\pi}\left(m\right) e^{-\tilde{g}(m)}} \\
&\qquad \le \epsilon + \frac{\sum\limits_{m\in\mathcal{M}} \tilde{\pi}\left(m\right) \left[\left(\tilde{g}\left(m\right)-\epsilon\right)_{+}\right] e^{-\tilde{g}(m)}}{\sum\limits_{m\in\mathcal{M}} \tilde{\pi}\left(m\right) e^{-\tilde{g}(m)}} \\
&\qquad \le \epsilon + \exp\left(-\epsilon\right) \frac{\sum\limits_{m\in\mathcal{M}} \tilde{\pi}\left(m\right) \left[\left(\tilde{g}\left(m\right)-\epsilon\right)_{+}\right] e^{-(\tilde{g}(m)-\epsilon)}}{\sum\limits_{m\in\mathcal{M}} \tilde{\pi}\left(m\right) e^{-\tilde{g}(m)}} \\
&\qquad \le \epsilon + \exp\left(-\epsilon\right) \frac{e^{-1} \sum\limits_{m\in\mathcal{M}} \tilde{\pi}\left(m\right)}{\tilde{\pi}\left(\bar{m}\right) e^{-\tilde{g}(\bar{m})}}.
\end{aligned}$$

Taking $\epsilon = -1 + \log \sum_{m\in\mathcal{M}} \tilde{\pi}\left(m\right) - \log \tilde{\pi}\left(\bar{m}\right) + \tilde{g}\left(\bar{m}\right)$, we finally get:

$$\mathbf{E}_{\rho_{\beta,\beta}(m,d\theta_m)} \log \prod_{i=1}^{N} p_{m,\theta_m}\left(y_i|x_i\right)^{-\beta} \le \tilde{g}\left(\bar{m}\right) + \log \frac{1}{\tilde{\pi}\left(\bar{m}\right)} + \log \sum_{m\in\mathcal{M}} \tilde{\pi}\left(m\right).$$

This proves the following lemma:

LEMMA 5. *For any m' in $\mathcal{M}$,*

$$\sup_{\theta_{m'}\in\Theta_{m'}} \gamma_{\beta}\left(m', \theta_{m'}\right) \le \frac{\alpha-1}{2} D_m + \mathbf{E}_{P^N}\left(\log \frac{1}{\tilde{\pi}\left(m'\right)} + \log \sum_{m\in\mathcal{M}} \tilde{\pi}\left(m\right)\right),$$

with

$$\tilde{\pi}\left(m\right) = \bar{\pi}\left(m\right) \exp\left(\frac{\alpha-1}{2} D_m\right) \prod_{s\in\mathcal{S}_m} \int_{\Sigma} \exp\left[-n_s(\beta,\beta) d\left(\bar{\theta}_s(\beta,\beta)||\theta_s\right)\right] \mu\left(\theta_s\right) d\theta_s.$$

Lemma 5 shows that the bound on $\gamma_{\beta}\left(m,\theta_m\right)$ is the sum of a parameter risk $(\alpha-1)D_m/2$ and a model risk $-\log \tilde{\pi}(m)/\left(\sum_{\mathcal{M}} \tilde{\pi}(m)\right)$, but with a functional $\tilde{\pi}$ different from the prior distribution $\bar{\pi}$. Besides, the ratio between $\tilde{\pi}\left(m\right)$ and $\bar{\pi}\left(m\right)$ is the product of two terms:

- a term that only depends on the size of m : $\exp\left((\alpha-1)D_m/2\right)$;

- a term that depends on the *unobserved* $\bar{\theta}_m(\beta, \beta)$:

$$\prod_{s \in \mathcal{S}_m} \int_\Sigma \exp\left[-n_s(\beta,\beta) d\left(\bar{\theta}_s(\beta,\beta) || \theta_s\right)\right] \mu\left(\theta_s\right) d\theta_s$$

The reason why we decided to take μ equal to Jeffrey's prior is that it makes the second term asymptotically independent of $\bar{\theta}$ (at least inside of the simplex), thanks to the following lemma:

LEMMA 6. *For any* $(n, \bar{\theta}) \in \mathbb{R}_*^+ \times \Sigma$ *let* f *be the function,*

$$f(n,\bar{\theta}) = \log \frac{\Gamma\left(\frac{\alpha}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^\alpha} + \frac{\alpha-1}{2}\log\left(\frac{2\pi}{n}\right) - \log\left(\int_\Sigma e^{-nd(\bar{\theta}||\theta)} \mu\left(d\theta\right)\right).$$

This function satisfies:

$$\forall (n,\bar{\theta}) \in \mathbb{R}_*^+ \times \Sigma \qquad 0 \leq f(n,\bar{\theta}) \leq \frac{\alpha^2}{4n} + \frac{\alpha}{4n \min_i(\bar{\theta}^i) + 2}.$$

PROOF OF LEMMA 6:. The upper bound is proven in [26, lemma 1] for $n \in \mathbb{N}$ and $\bar{\theta}$ being of the form $(a_1/n, \ldots, a_\alpha/n)$ with $(a_1, \ldots, a_\alpha) \in \mathbb{N}^\alpha$. The proof, based on Stirling's formula to approximate the Γ function, still works in the general case $(n, \bar{\theta}) \in \mathbb{R}_*^+ \times \Sigma$.

However the proof used for the lower bound in that case ([26, lemma 2]) does not work in the general case. Therefore let us just prove the lower bound. Using (3.3) f can be rewritten as:

$$f(n,\bar{\theta}) = -nh\left(\bar{\theta}\right) + \frac{\alpha-1}{2}\log\frac{2\pi}{n} - \sum_{i=1}^{\alpha} \log\Gamma\left(n\bar{\theta}^i + \frac{1}{2}\right) + \log\Gamma\left(n + \frac{\alpha}{2}\right),$$

whose derivative w.r.t. n is equal to :

$$\begin{aligned}
\frac{\partial f}{\partial n}(n,\bar{\theta}) &= -h\left(\bar{\theta}\right) - \frac{\alpha-1}{2n} - \sum_{i=1}^{\alpha} \bar{\theta}^i \psi_0\left(n\bar{\theta}^i + \frac{1}{2}\right) + \psi_0\left(n + \frac{\alpha}{2}\right) \\
&= -\frac{\alpha-1}{2n} - \sum_{i=1}^{\alpha} \bar{\theta}^i \left[\psi_0\left(n\bar{\theta}^i + \frac{1}{2}\right) - \log\left(n\bar{\theta}^i\right)\right] + \psi_0\left(n + \frac{\alpha}{2}\right) - \log n \\
&\leq -\frac{\alpha-1}{2n} + 0 + \frac{\alpha-1}{2n} \\
&\leq 0
\end{aligned}$$

where we used lemma 3 in order to obtain the inequality. As a result, for any given $\bar{\theta} \in \Sigma$, the function $n \mapsto f(n, \bar{\theta})$ is decreasing on $\mathbb{R}_*^+$. Besides, Laplace method of integration shows that for any $\bar{\theta}$ in the interior of the simplex,

$$\lim_{n \to \infty} f(n, \bar{\theta}) = 0.$$

As a result, $f(n, \bar{\theta}) \geq 0$ for any $n > 0$ and $\bar{\theta}$ in the interior of the simplex. Now if $\bar{\theta}$ is on the boundary of the simplex, one can consider a sequence $(\theta_k)_{k \geq 0}$ of points in the interior of the simplex which converges to $\bar{\theta}$. By the theorem of dominated convergence for a fixed $n > 0$ the integrals $\int_\Sigma \exp\left(-nd(\theta_k||\theta)\right)\mu(d\theta)$ converge to $\int_\Sigma \exp\left(-nd(\bar{\theta}||\theta)\right)\mu(d\theta)$ as $k \to \infty$. As a result the lower bound that we proved in the interior of the simplex remains true on its border, for any $n > 0$. This proves lemma 6. □

If π is a prior on $\mathcal{M}$, consider now the Gibbs estimator formed with the prior $\bar{\pi}$ such that:

$$\forall m \in \mathcal{M} \qquad \bar{\pi}(m) = \frac{1}{Z} \frac{\pi(m) \exp\left(-\frac{\alpha - 1}{2} D_m\right)}{\prod_{s \in \mathcal{S}_m} \left(\frac{2\pi}{n_s}\right)^{\frac{\alpha-1}{2}} C_\alpha^{-1}} = \frac{1}{Z} \pi(m) \prod_{s \in \mathcal{S}_m} C_\alpha \times \left(\frac{n_s}{2\pi e}\right)^{\frac{\alpha-1}{2}},$$

where Z is a normalizing constant and $C_\alpha = \Gamma(1/2)^\alpha / \Gamma(\alpha/2) = \pi^{\alpha/2}/\Gamma(\alpha/2)$.

Note that X_N is observed so n_s is an observed variable which is invariant under permutation of z_1^N, and therefore $\bar{\pi}$ can be taken as a prior to form the Gibbs estimator. For such a choice, lemma 5 is valid with the following $\tilde{\pi}$:

$$\tilde{\pi}(m) = \bar{\pi}(m) \exp\left(\frac{\alpha - 1}{2} D_m\right) \lambda\left(m, z_1^N\right) = \frac{1}{Z} \pi(m) \prod_{s \in \mathcal{S}_m} \frac{\int_\Sigma \exp\left(n_s d\left(\bar{\theta}_s || \theta_s\right)\right) \mu(\theta_s) d\theta_s}{C_\alpha^{-1} \left(\frac{2\pi}{n_s}\right)^{\frac{\alpha-1}{2}}}.$$

Using lemma 6 we obtain the following bound:

$$\log \frac{\sum_{m' \in \mathcal{M}} \tilde{\pi}(m')}{\tilde{\pi}(m)} \leq \log \frac{1}{\pi(m)} + \sum_{s \in \mathcal{S}_m} \left(\frac{\alpha^2}{4n_s} + \frac{\alpha}{4n_s \min_i(\bar{\theta}_i) + 2}\right).$$

Finally, using lemmas 5 we get:

$\forall\,(m, \theta_m) \in \Theta,$

$$\gamma_\beta\left(m,\theta_m\right) \leq \frac{\alpha-1}{2} D_m + \log\frac{1}{\pi\left(m\right)} + \sum_{s\in\mathcal{S}_m}\left(\frac{\alpha^2}{4n_s} + \frac{\alpha}{4n_s \min_i(\bar{\theta}_i)+2}\right).$$

Applying theorem 3 finishes the proof of theorem 1. □

7. Conclusion Our goal in this paper was to adapt the idea of twice-universal codes studied in universal compression to the problem of statistical density estimation. The similarity between the redundancy criterion in compression and the cumulated statistical risk justifies this goal, but some technical works has to be done in order to get a bound on the statistical risk of the estimator, and not on the cumulated statistical risk for samples of increasing sizes. We could get a result for a mixture estimator by using a Gibbs estimator as studied by Catoni in [9] and translating double-mixture codes ([14], [24]) into double-mixture statistical estimators.

The implementation procedure we suggest in section 5.2 takes the form of a penalized maximum likelihood model selection, justified by the selection of the model with highest posterior Gibbs distribution. However the Gibbs estimator is of a mixture of models and one could also imagine a approximation of this mixture using Monte Carlo simulations, instead of selecting one particular model (see [10], [6]).

As far as applications of such estimators are concerned, we refer to [23] for an example in natural language processing. It is shown how to use adaptive models in order to represent non-stochastic objects, e.g. texts, from which a statistical experiment is carried out. Such a representation can then be used to characterize the original object; as an application the similarity between two objects can be estimated by computing the similarity between the two corresponding models.

8. Acknowledgment The work presented in this paper in part of the PhD thesis that I prepare under the direction of Olivier Catoni. I thank him a lot for his numerous advises and suggestions.

REFERENCES

[1] Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proceedings 2nd International Symposium on Information Theory* (1974), P. Petrov and F. C. (Eds.), Eds., pp. 267–281.

[2] Alzer, H. On some inequalities for the gamma and psi functions. *Mathematics of Computation 66*, 217 (Jan. 1997), 373–389.

[3] Barron, A. Are bayes rules consistent in information ? In *Open Problems in Communication and Computation, T.M. Cover and B. Gopinath Ed.* Springer Verlag, 1987.

[4] Barron, A., Birgé, L., and Massart, P. Risk bounds for model selection via penalization. *Probability Theory and Related Fields 113*, 3 (1999), 301–413.

[5] Barron, A., and Yang, Y. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics 27*, 5 (1999), 1564–1599.

[6] Blanchard, G. The "progressive mixture" estimator for regression trees. *Ann. Inst. Henri Poincaré, Probabilités et Statistiques 35*, 6 (1999), 793–820.

[7] Breiman, L. Bagging predictors. Tech. Rep. 421, Statistics Department, University of California at Berkeley, Sept. 1994.

[8] Catoni, O. A mixture approach to universal model selection. preprint LMENS - 97 - 30, http://www.dmi.ens.fr/preprints, 1997.

[9] Catoni, O. Gibbs estimators. preprint LMENS-98-21 at http://www.dmi.ens.fr/preprints, pages 1–23, May 1998.

[10] Catoni, O. "Universal" aggregation rules with exact bias bounds. *to appear in the Annals of Statistics* (1999).

[11] Clarke, B. S., and Barron, A. R. Jeffrey's prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference 41* (1994), 37–60.

[12] Cover, T. M., and Thomas, J. A. *Elements of information theory.* Wiley, 1991.

[13] Davisson, L. D. Universal noiseless coding. *IEEE Trans. Inform. Theory 19*, 6 (Nov. 1973), 783–795.

[14] Feder, M., and Merhav, N. Hierarchical universal coding. *IEEE Trans. Inform. Theory 42*, 5 (Sept. 1996), 1354–1364.

[15] Krichevsky, R. E., and Trofimov, V. K. The performance of universal encoding. *IEEE Trans. Inform. Theory 27*, 2 (Mar. 1981), 199–207.

[16] Mallows, C. Some comments on c_p. *Technometrics*, 15 (1973), 661–675.

[17] Merhav, N., and Feder, M. Universal prediction. *IEEE Trans. Inform. Theory 44*, 6 (Oct. 1998), 2124–2147.

[18] Nikiforov, A., and Ouvarov, V. *Fonctions spéciales de la physique mathématique.* Mir, 1983.

[19] Rissanen, J. Minimax codes for finite alphabets. *IEEE Trans. Inform. Theory 24*, 3 (May 1978), 389–392.

[20] Ryabko, B. Y. Twice-universal coding. *Problems of Information Transmission 20*, 3 (July 1984), 24–28.

[21] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6 (1978), 461–464.

[22] Vapnik, V. N. *Statistical learning theory.* Wiley, 1998.

[23] Vert, J.-P. Adaptive context trees and text clustering. Tech. rep., Departement de mathématiques et applications, Ecole normale superieure de Paris, 2000. preprint LMENS - 00 - 05, http://www.dmi.ens.fr/preprints.

[24] Willems, F. M., Shtarkov, Y. M., and Tjalkens, T. J. The context tree weighting method : Basic properties. *IEEE Trans. Inform. Theory 41*, 3 (May 1995), 653–664.

[25] Willems, F. M., Shtarkov, Y. M., and Tjalkens, T. J. Context weighting for general finite-context sources. *IEEE Trans. Inform. Theory 42* (Sept. 1996), 1514–1520.

[26] Xie, Q., and Barron, A. R. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inform. Theory 46*, 2 (Mar. 2000), 431–445.

Département de mathématiques et applications
Ecole normale supérieure de Paris
45, rue d'Ulm
75230 Paris cedex 05
France
E-mail : jean-philippe.vert@ens.fr