ANNALES DE L'I. H. P., SECTION B

GILLES BLANCHARD

The "progressive mixture" estimator for regression trees

Annales de l'I. H. P., section B, tome 35, n° 6 (1999), p. 793-820
<http://www.numdam.org/item?id=AIHPB_1999__35_6_793_0>

Ann. Inst. Henri Poincaré,

Vol. 35, n° 6, 1999, p. 793-820. *Probabilités et Statistiques*

The "progressive mixture" estimator for regression trees

by

Gilles BLANCHARD [1]

DMI, École Normale Supérieure, 45 rue d'Ulm, 75230 Paris Cedex, France

Article received on 31 August 1998, revised 10 May 1999

ABSTRACT. – We present a version of O. Catoni's "progressive mixture estimator" (1999) suited for a general regression framework. Following basically Catoni's steps, we derive strong non-asymptotic upper bounds for the Kullback–Leibler risk in this framework. We give a more explicit form for this bound when the models considered are regression trees, present a modified version of the estimator in an extended framework and propose an approximate computation using a Metropolis algorithm. © Elsevier, Paris

RÉSUMÉ. – Nous donnons une version, adaptée à un cadre de régression, de l'estimateur dit de "mélange progressif" introduit par O. Catoni (1999). De façon analogue à Catoni, nous donnons une borne à horizon fini pour la perte de Kullback–Leibler de l'estimateur dans ce cadre. Nous explicitons la forme de cette borne dans le cas d'arbres de régression, présentons une variante dans un cadre étendu, et proposons une méthode de calcul approché par algorithme de Metropolis. © Elsevier, Paris

[1] E-mail: gblancha@dmi.ens.fr

1. INTRODUCTION

The so-called "progressive estimator" was introduced by Catoni [6] (1997) using ideas inspired by the recent work of Willems et al. [11, 12] on universal coding. The principle of a progressive estimator has also been proposed independently by Barron and Yang [4,3]. One of the main attracting features of this estimator is that we can easily derive strong non-asymptotic bounds on its mean Kullback–Leibler risk in an extremely general framework.

Our goal in this paper is to present a version of this estimator first in a general regression estimation framework, then more precisely for classification and regression trees (CARTs). As the computation of this estimator involves performing sums of a Bayesian type on a very large (and possibly non-finite) set of models, we are naturally led in practice to compute these sums approximately using a Monte Carlo algorithm. A very similar method (a Bayesian search among CARTs using a stochastic procedure) has been proposed by Chipman et al. [7] (1997). It is interesting to note that if we start from a different theoretical point of view, which allows us to derive non asymptotic bounds on the Kullback–Leibler risk, we are led to a very similar algorithm (see discussion at the end of the paper). On the other hand, the work of Chipman et al. suggested to us the idea of a data-dependent Bayesian prior on the set of models which we develop in Section 5.

Finally we also want to point out the relation of this algorithm of random walk in the model tree space with the method of tree selection by local entropy minimization developed by Amit, Geman and Wilder [1,2]. This work has been a large source of inspiration and in some regards, the local entropy minimization can be seen as the deterministic (or "zero-temperature") analogous of the Monte Carlo method presented here. This will also be discussed in some more detail in Section 7.

The structure of the paper is as follows. In Section 2, we present a version of the progressive mixture estimator for regression estimation in a very general framework. Section 3 focuses on classification and regression trees and gives a precise and explicit form of the estimator in this case. Section 4 briefly deals with the case of a "fixed-design" tree where an exact computation of the estimator can be performed thanks to the tree weighting recursive algorithm developed by Tjalkens, Willems and Shtarkov in [11] (other authors have also used this technique for classification trees in a different framework, see [10]). Section 5 presents an extension of the framework to a possibly uncountable family of model

trees, using a data-dependent prior, and we show for a precise and non trivial example that the estimator thus obtained still achieves the minimax rate of convergence. In Section 6, we explain how to construct a Monte Carlo chain to obtain an approximate computation of the estimator for general tree models. Section 7 concludes the paper with a discussion of the results.

2. CATONI'S PROGRESSIVE ESTIMATOR IN A REGRESSION FRAMEWORK

In this section we present a version of Catoni's estimator adapted for regression estimation purposes.

Let us specify the framework and notations. Let

$$(X_i, Y_i)_{i=1}^N = (Z_i)_{i=1}^N = Z_1^N$$

be a sequence of random variables, taking their values in some measurable space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with probability distribution P_N. The only hypothesis we will make on P_N is that it is exchangeable, i.e., that this distribution is invariant under any permutation of the variables Z_i. Our goal in a regression framework is, given a sample $z_2, \ldots, z_N$, to get an estimation of the conditional probability of a new observation Y_1 given X_1 (the fact that this "new" observation is actually given the index 1 is purely for later notational convenience and should not be misleading). We thus want to estimate $P_N(Y_1 \mid X_1, Z_2^N)$. This can include any case where the order of the observations (including the test example Z_1) is not of importance, because we can then re-draw them in a random order. This case is of course of but minor practical relevance since in such a predictive framework, the test example is usually given separately and cannot be drawn at random (in which case the hypothesis of exchangeability has to hold fully). However, it covers for example the customary procedure used for classifier testing, when one has a fixed database of examples, that is split at random between a set used to train the classifier and a set used to validate its accuracy.

We will assume that there exists a regular version of this conditional probability, and, more generally throughout this paper, we will assume that all of the conditional probabilities that we are dealing with have a regular version, allowing us to make use of all standard integration properties. For that purpose we will assume that $\mathcal{Z}$ is a Borel space.

To begin with, we will split the sample data set z_2^N into two subsamples $E = z_{L+1}^N$ called "estimation set" and $T = z_1^L$ called "test set" for some fixed integer $L > 0$ (note that we consider the new observation z_1 to be part of the test set). The reason for this will become clearer in the sequel and the issue of the choice of L will be discussed later on. Note that the ordering of the variables is purely arbitrary since the probability distribution P according to which they have been drawn is exchangeable.

We then assume that we have at our disposal a countable family of estimators $(Q_m)_{m\in\mathcal{M}}$, that is, a family of conditional probabilities depending on the observations of the estimation set $Q_m(Y \in \cdot \mid X, E)$, also denoted by $Q_m^E(Y \in \cdot \mid X)$ for a more compact writing (in a more general way in this paper we will often use a superscript E for probability distributions depending on the estimation set, thus writing equally $P^E(\cdot \mid \cdot)$ for $P(\cdot \mid \cdot, E)$). We will speak of $m \in \mathcal{M}$ as a "model".

Our purpose is to find an estimator in this family which will minimize (approximately) the Kullback–Leibler (K-L) risk with regard to $P_N(Y_1 \mid X_1, Z_2^N)$, i.e., find a $m \in \mathcal{M}$ achieving

$$\inf_{m\in\mathcal{M}} E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N),\ Q_m^E(Y_1 \in \cdot \mid X_1)\big),$$

where H is the Kullback–Leibler divergence: for ρ, ν two probability distributions on the space $\mathcal{Y}$:

$$H(\rho, \nu) = \begin{cases} \displaystyle\int \log \frac{d\rho}{d\nu}\, d\rho & \text{if } \rho \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

To get a solution to our estimation problem we will actually not perform a model selection as first announced, but rather build a composite estimator for which we will be able to explicitly upper bound the Kullback–Leibler risk.

Suppose that we have chosen some probability distribution π on the model set $\mathcal{M}$. If we keep the estimation set E "frozen", we can consider the estimator probabilities $Q_m^E(Y \mid X)$ as simple product (conditional) probabilities on the test set T. We then define the mixture conditional probability on the test set $Q_{\text{mix.}}^E$ obtained by summing these product probabilities Q_m^E weighted by the prior π:

$$Q_{\text{mix.}}^E(dY_1^L \mid x_1^L) \triangleq \sum_{m\in\mathcal{M}} \pi(m) \prod_{i=1}^{L} Q_m^E(dY_i \mid x_i).$$

We eventually build Catoni's progressive mixture estimator $Q_\pi(Y_1 \in \cdot \mid x_1, z_2^N)$ also denoted by $Q_\pi^E(Y_1 \in \cdot \mid x_1, z_2^L)$ in the following way:

$$Q_\pi(dY_1 \mid x_1, z_2^N) = Q_\pi^E(dY_1 \mid x_1, z_2^L) \triangleq \frac{1}{L} \sum_{M=1}^{L} Q_{\text{mix.}}^E(dY_1 \mid x_1, z_2^M). \tag{1}$$

In order to make the last formula somewhat clearer, let us consider for a moment the simple case where the space $\mathcal{Y}$ is discrete (this is actually the case that will be considered in the following sections, for classification problems). In that case the definition (1) yields the following expansion (everywhere the denominator does not vanish):

$$Q_\pi(Y_1 \mid x_1, z_2^N) = \frac{1}{L} \sum_{M=1}^{L} \frac{\sum_{m \in \mathcal{M}} \pi(m) \prod_{i=1}^{M} Q_m^E(Y_i \mid x_i)}{\sum_{m' \in \mathcal{M}} \pi(m') \prod_{j=2}^{M} Q_m^E(Y_j \mid x_j)}. \tag{2}$$

Let us point out that each individual term of the sum over M is nothing but a (predictive) Bayesian estimator based on the probabilities $Q_m^E(dY_1 \mid X_1)$ (when the estimation sample E is kept frozen), using the prior π. Thus Q_π^E is a Cesàro mean of Bayesian estimators with a sample of growing size z_2^M. The subsample E is used for estimation within each model $m \in \mathcal{M}$ ("estimation set") whereas the other subsample z_2^L is then used to "choose" between the models in a Bayesian way ("test set").

We now state a theorem upper-bounding the risk for this estimator.

THEOREM 1. – *With the previous assumptions we get the inequality*:

$$\begin{aligned} &E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N),\ Q_\pi(Y_1 \in \cdot \mid X_1, Z_2^N)\big) \\ &\leqslant \inf_{m \in \mathcal{M}} \left\{ E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N),\ Q_m(Y_1 \in \cdot \mid X_1, Z_{L+1}^N)\big) \right. \\ &\qquad \left. + \frac{1}{L} \log \frac{1}{\pi(m)} \right\}. \end{aligned} \tag{3}$$

Proof. – There is actually little to be changed in comparison to [6] so we will follow quite closely Catoni's steps.

Assume that the estimators $Q_m^E(Y_1 \in \cdot \mid x_1)$ have densities $q_m^E(y_1 \mid x_1)$ with respect to some dominating conditional measure μ. We similarly will denote by lower-case letters the densities with respect to μ of the other probability distributions. In case there is no "obvious" choice for μ we can build the following dominating measure depending on the

estimation set E:

$$\mu^E(Y_1 \in \cdot \mid x_1) = \sum_{m \in \mathcal{M}} 2^{-\sigma(m)} Q_m^E(Y_1 \in \cdot \mid x_1),$$

where σ is some injective mapping of $\mathcal{M}$ into $\mathbb{Z}^+$.

Take an $m \in \mathcal{M}$; if

$$E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N),\ Q_m^E(Y_1 \in \cdot \mid X_1)\big) = +\infty,$$

the inequality is true for this m; else we consider the difference (whose second term is now finite)

$$\begin{aligned} &E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N),\ Q_\pi^E(Y_1 \in \cdot \mid X_1, Z_2^L)\big) \\ &\quad - E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N),\ Q_m^E(Y_1 \in \cdot \mid X_1)\big) \\ &\quad = -E_{P_N} \log \frac{q_\pi^E(Y_1 \mid X_1, Z_2^L)}{q_m^E(Y_1 \mid X_1)}. \end{aligned}$$

Because $-\log$ is a convex function we have

$$-E_{P_N} \log \frac{q_\pi^E(Y_1 \mid X_1, Z_2^L)}{q_m^E(Y_1 \mid X_1)} \leqslant -\frac{1}{L} \sum_{M=1}^{L} E_{P_N} \log \frac{q_{\text{mix.}}^E(Y_1 \mid X_1, Z_2^M)}{q_m^E(Y_1 \mid X_1)}. \quad (4)$$

Now using the hypothesis that P_N is an exchangeable probability distribution, we can swap the role played by X_1 and X_{M+1} in the Mth term of the sum (formally replacing X_{L+1}, Y_{L+1} by X_1, Y_1 whenever necessary).

We thus get to:

$$-E_{P_N} \log \frac{q_\pi^E(Y_1 \mid X_1, Z_2^L)}{q_m^E(Y_1 \mid X_1)} \leqslant -\frac{1}{L} E_{P_N} \sum_{M=1}^{L} \log \frac{q_{\text{mix.}}^E(Y_{M+1} \mid X_{M+1}; Z_2^M)}{q_m^E(Y_{M+1} \mid X_{M+1})},$$

where by definition

$$q_{\text{mix.}}^E(y_{M+1} \mid x_{M+1}, z_2^M) = \frac{q_{\text{mix.}}^E(y_2^{M+1} \mid x_2^{M+1})}{q_{\text{mix.}}^E(y_2^M \mid x_2^M)}.$$

Let us point out that we have no more conditioning with respect to x_{M+1} in the last denominator because of the relation coming straightforward from the definition of $Q_{\text{mix.}}^E$:

$$\int_{\mathcal{Y}} q^E_{\text{mix.}}(y_2^{M+1} \mid x_2^{M+1})\, d\mu(y_{M+1}) = q^E_{\text{mix.}}(y_2^M \mid x_2^M). \tag{5}$$

Thus the sum in (4) reduces to the first component of its last term since the other terms cancel each other successively, leading to:

$$-E_{P_N} \log \frac{q^E_\pi(Y_1 \mid X_1, Z_2^L)}{q^E_m(Y_1 \mid X_1)} \leqslant -\frac{1}{L} E_{P_N} \log \frac{q^E_{\text{mix.}}(Y_1^L \mid X_1^L)}{q^E_m(Y_1^L \mid X_1^L)}$$
$$\leqslant -\frac{1}{L} E_{P_N} \log \frac{\sum_{m' \in \mathcal{M}} \pi(m') q^E_{m'}(Y_1^L \mid X_1^L)}{q^E_m(Y_1^L \mid X_1^L)} \leqslant -\frac{\log \pi(m)}{L}. \qquad \square$$

It has been pointed out that an "aesthetic" flaw of the progressive mixture estimator is that it is not a symmetric function of the data (different orderings give different estimators), because of the arbitrary division of the sample in two sets and because of the "progressive" sum in the definition of the estimator, which is a Cesàro average of Bayesian estimators with a growing sample. The reason for this apparently surprising progressive sum is to be able to bound the Kullback–Leibler distance between probability distributions on the only variable Z_1 by the K-L distance between full multivariate distributions on the whole test set (for which the idea of a mixture distibution comes from the compression and information theory litterature), as appears in the proof of the theorem. However, it is possible to define theoretically a fully symmetric estimator by averaging $Q_\pi(dY_1 \mid x_1, z_2^N)$ over all the $(N-1)!$ possible orderings of the sample z_2^N, that will satisfy the same majorations as Q_π (as can be seen using once more the convexity of the K-L distance $H(\rho, \mu)$ with respect to μ).

3. APPLICATION TO A CART MODEL

In this section we deal with a classification problem, that is, the variable Y will take its values in $\mathcal{Y} = \{0, 1, \ldots, c\}$ (thus there are $(c+1)$ distinct classes and c is the number of *free parameters* for a distribution on $\mathcal{Y}$), and we want to "recognize" the observation X taking its values in a certain space $\mathcal{X}$, deciding to which class it belongs, having already observed a "learning set" z_2^N. In order to do that we want to make use of a classification tree using the "20-questions game" principle: suppose we have at our disposal a countable set of "questions" $\mathcal{Q}$ we can ask on the observation X, to which we can only answer by 0 or 1 (for example, we can think of X as a finite or infinite binary sequence, the questions being

of the form "what is the nth bit of X?"). Having asked a first question, we are then allowed to choose a second one depending on the first answer, and so on. At each step we can decide whether to ask a new question or to stop and guess X's class (or, alternatively, to estimate the probability of X belonging to a certain class, if we prefer a regression framework). If the rule is that we can only ask at most d questions, all the possible sequences of questions and answers can be represented as a complete binary tree of depth at most d, whose internal nodes are labeled with questions (CART tree model).

Our set of models $\mathcal{M}$ will therefore be a set of couples (T, F_T), where T is a tree and $F_T = (f_s)_{s \in \mathrm{int}(T)}$ the set of questions indexed by the internal nodes of T. A question $f \in \mathcal{Q}$ is nothing but the indicator of a certain measurable set $A_f \subset \mathcal{X}$. We will think of a binary tree T as a subset of $\Gamma = \{\emptyset\} \cup \{0,1\}^*$ (the set of all finite, possibly empty strings formed of zeroes and ones) such that if $s \in T$, every ancestor (that is, every prefix string) t of s also belongs to T. By definition, $\emptyset$ is the root of the tree, and if s is a node, its two sons are $s0$ and $s1$. In our case we will only consider complete binary trees (such that every $s \in T$ either is a terminal node or has two sons). If m is such a model and x is an observation, we denote by $m(x)$ the terminal node associated to x by m (obtained by asking questions on x and letting him follow the tree branches until we reach a terminal node of m). Also, we will denote by ∂m the set of terminal nodes (or leaves) of the model tree m.

Given $m = (T, F_T)$ we now have to define our set of estimators $Q_m(Y_1 \mid X_1, E)$. For the remainder of the paper, when dealing with tree models we will take for Q_m a conditional Laplace estimator at every terminal node: given a terminal node $s \in T$ we define the following counters on the estimation set $E = z_{L+1}^N$:

$$n_s^i \triangleq \sum_{k=L+1}^{N} \mathbf{1}\{m(x_k) = s, y_k = i\} \quad \text{for } i = 0, \ldots, c,$$

where $\mathbf{1}_A$ denotes the indicator function of the set A. Then, if $m(X_1) = s$, we take for $Q_m^E(Y_1 \mid X_1)$ a multivariate distribution defined by

$$Q_m^E(Y = i \mid m(X) = s) \triangleq \frac{n_s^i + 1}{\sum_i n_s^i + c + 1}.$$

Let us then denote by $M_{m,\theta^m}(Y \mid X)$ the general conditional multivariate distribution based on the labeled tree m, where $\theta^m = (\theta_s)_{s \in \partial m}$ is a set

of real vectors of parameters all belonging to the c-dimensional simplex $\mathcal{S}_c = \{\theta \in [0,1]^{(c+1)} \mid \sum_i \theta_i = 1\}$, and indexed by the terminal nodes of m, defined by

$$M_{m,\theta^m}(Y = i \mid m(X) = s) = (\theta_s)_i$$

(where $(\theta_s)_i$ denotes the ith coordinate of the vector parameter θ_s).

Once again following very closely [6], we give the following upper bound for the risk of the Laplace estimator:

THEOREM 2. – *For any exchangeable distribution P_N on the variables $Z_1^N \in (\{0,\dots,c\} \times \mathcal{X})^N$, for a given model m, the conditional Laplace estimator Q_m satisfies*

$$E_{P_N} H\big(P_N(Y_1 \mid X_1, Z_2^N), Q_m(Y_1 \mid X_1, Z_{L+1}^N)\big)$$
$$\leqslant \inf_{\theta^m} E_{P_N} H\big(P_N(Y_1 \mid X_1, Z_2^N), M_{m,\theta^m}(Y_1 \mid X_1)\big) + \frac{c\,|\partial m|}{N-L+1},$$

where $|\partial m|$ is the number of terminal nodes of the tree model m.

Proof. – To shorten the equations below we will write in the sequel equally X_{N+1}, Y_{N+1} for X_1, Y_1.

For a given terminal node $s \in \partial m$, let us first introduce the modified counters a_s^i, $i = 0,\dots,c$:

$$a_s^i = \sum_{k=L+1}^{N+1} \mathbf{1}\{m(X_k) = s, Y_k = i\} = n_s^i + \mathbf{1}\{m(X_1) = s, Y_1 = i\},$$

and denote by $\Sigma_s = \sum_i a_s^i$ the total number of examples at node s.

With these new counters we get

$$Q_m^E(Y_1 = i \mid m(X_1) = s; Z_{L+1}^N) = \frac{a_s^i}{\Sigma_s + c},$$

thus

$$E_{P_N} - \log Q_m(Y_1 \mid X_1, E)$$
$$= -\frac{1}{N-L+1}$$
$$\times E_{P_N} \sum_{j=L+1}^{N+1} \log Q_m(Y_j \mid X_j; Z_k, k \neq j, L+1 \leqslant k \leqslant N+1)$$
$$= -\frac{1}{N-L+1} E_{P_N} \sum_{s \in \partial m} \sum_{i=0}^{c} a_s^i \log \frac{a_s^i}{\Sigma_s + c},$$

where the first inequality follows from the exchangeability of the distribution P_N. In the second equality we put $0 \log 0 = 0$ if necessary. Now for a fixed $s \in \partial m$, if at least one example reached this node, that is, if $\Sigma_s > 0$, we have

$$\begin{aligned}\sum_{i=0}^{c} a_s^i \log \frac{a_s^i}{\Sigma_s + c} &= \sum_{i=0}^{c} a_s^i \log \frac{a_s^i}{\Sigma_s} - \Sigma_s \log\left(1 + \frac{c}{\Sigma_s}\right) \\ &= -\inf_{\theta \in \mathcal{S}_c} \left[-\sum_{i=0}^{c} a_s^i \log \theta_i \right] - \Sigma_s \log\left(1 + \frac{c}{\Sigma_s}\right) \\ &\geqslant -\inf_{\theta \in \mathcal{S}_c} \left[-\sum_{i=0}^{c} a_s^i \log \theta_i \right] - c.\end{aligned}$$

Now this last inequality obviously remains true if $a_s^i = 0$ for all i and thus

$$\begin{aligned}&E_{P_N} - \log Q_m(Y_1 \mid X_1, E) \\ &\leqslant \frac{1}{N-L+1} E_{P_N}\left(\inf_{\theta^m} - \log M_{m,\theta^m}^{\otimes(N-L+1)}\left(Y_{L+1}^{N+1} \mid X_{L+1}^{N+1}\right) + c\,|\partial m|\right) \\ &\leqslant \inf_{\theta^m} E_{P_N} - \log M_{m,\theta^m}(Y_1 \mid X_1) + \frac{c\,|\partial m|}{N-L+1}. \qquad \square\end{aligned}$$

Now suppose we have chosen some *a priori* measure π on the tree models $\mathcal{M}$ ($\mathcal{M}$ is countable since $\mathcal{Q}$ is countable); we thus get as a straightforward consequence of Theorems 1 and 2:

COROLLARY 1. – *With the previous choice of the estimators Q_m for labeled tree models, and a prior π on $\mathcal{M}$ we get for the estimator Q_π defined in Section 2 the inequality*

$$\begin{aligned}&E_{P_N} H\left(P_N(Y_1 \mid X_1, Z_2^N), Q_\pi(Y_1 \mid X_1, Z_2^N)\right) \qquad (6) \\ &\leqslant \inf_{m,\theta^m} \left\{ E_{P_N} H\left(P_N(Y_1 \mid X_1, Z_2^N), M_{m,\theta^m}(Y_1 \mid X_1)\right) \right. \\ &\qquad \left. + \frac{\log \pi(m)^{-1}}{L} + \frac{c|\partial m|}{N-L+1} \right\}.\end{aligned}$$

For example, let us take for prior π the distribution of the genealogy tree of a branching process of parameter $\rho \leqslant 1/2$ (an individual gets two sons with probability ρ, or dies without descendance with probability $1-\rho$)—this prior will be used several times in the next sections. This prior is explictly given by

$$\pi(m) = \frac{(\rho(1-\rho))^{|\partial m|}}{\rho},$$

so that the first penalization term in Eq. (6) is

$$\frac{\log \pi(m)^{-1}}{L} = -\frac{|\partial m|(\log \rho(1-\rho))}{L} + \frac{\log \rho}{L},$$

thus the natural choice for L (the size of the test sample) in this case is of the form kN, with $k \in [0, 1]$ such that the two penalization terms of Eq. (6) are balanced.

Remark. – Although the K-L risk is interesting in itself and has been customarily used for classification tree construction (through "local entropy minimization", see, e.g., [1,2]), one may be legitimately interested by the true classification error of the estimator. A thorough discussion on a precise control of this error is beyond the scope of this paper; we simply recall some basic inequalities allowing to control the classification error through the K-L risk.

Let us forget for a moment the conditioning with respect to X and let P be a probability distribution on $\mathcal{Y} = \{0, \ldots, c\}$. If P is the "true" distribution, the best classification rule is to predict the class of highest probability. Let $P_* = \max_i P(i)$, then the lowest attainable average classification error is $L_* = 1 - P_*$. Now suppose we have some estimate Q of P and predict the class $\alpha = \arg\max_i Q(i)$; the average classification error for this rule is $L(Q) = 1 - P(\alpha)$. Then the following elementary inequalities hold:

$$L(Q) - L_* \leqslant \sum_{i=0}^{c} |P(i) - Q(i)| \leqslant \sqrt{2H(P, Q)}.$$

The first inequality can be found in [9] and the second one, for example, in [8], p. 300.

Now if me make this reasoning conditional to some X, taking the expectation over X of the above quantities we get (somewhat informally)

$$\begin{aligned} EL\big(Q(Y \mid X)\big) - EL_*(X) &\leqslant E\left(\sqrt{2H(P, Q \mid X)}\right) \\ &\leqslant \sqrt{2}\sqrt{E(H(P, Q \mid X))}. \end{aligned}$$

Using the bounds derived for the K-L distance, we can thus obtain a coarse upper bound for the difference between the classification error obtained with the estimator $Q(Y \mid X)$ and the optimal average error $EL_*(X)$.

4. EXACT COMPUTATION FOR A FIXED-DESIGN TREE

In this section we will assume that there is actually no choice in the question to be asked at each internal node of the model tree. In this case we will speak of a "fixed-design" tree model; the set of models is then the set of all complete subtrees of the maximal complete binary tree of depth d denoted by $\mathcal{T}$. The most classical example for that is a "context tree" model where X is a binary string of length d and all nodes at depth r are labeled with the question "what is the rth bit of X?". Recall that since we are dealing with a discrete space $\mathcal{Y}$, the developed formula for the progressive mixture estimator is given by Eq. (2).

We will take for prior π the distribution of the genealogy tree of a branching process of parameter ρ stopped at depth d, meaning that every node of the tree will have two sons with probability ρ or will be a terminal node with probability $1-\rho$, except for the nodes at depth d which are always terminal. In this framework we can use an efficient algorithm found by Willems, Shtarkov and Tjalkens [11] for universal coding using weighted context trees to compute a sum of the form

$$\sum_{m\in\mathcal{M}} \pi(m) \prod_{i=1}^{M} Q_m^E(y_i \mid x_i). \tag{7}$$

Namely, let us define for every node (internal or terminal) s of $\mathcal{T}$ the counters n_s^i, b_s^i for $i = 0, \ldots, c$ containing the number of examples x_k of class i whose "path" passes through node s (i.e., for which s is an ancestor of or is equal to $\mathcal{T}(x_k)$), respectively, for the estimation set $E = z_{L+1}^N$ and the truncated test set z_1^M:

$$\begin{cases} n_s^i = \displaystyle\sum_{j=L+1}^{N} \mathbf{1}\{s \preceq \mathcal{T}(x_j), y_j = i\}, \\ b_s^i = \displaystyle\sum_{j=1}^{M} \mathbf{1}\{s \preceq \mathcal{T}(x_j), y_j = i\}, \end{cases}$$

where $\preceq$ means "to be an ancestor of or to be equal to".

Let us now define the local (Laplace) estimator at node s (estimated using E, and applied to the set z_1^M):

$$L(s) = \prod_{i=0}^{c} \left(\frac{n_s^i + 1}{\sum_i n_s^i + c + 1} \right)^{b_s^i}$$

(note that any other local estimator could actually be taken here instead of the Laplace estimator; this is only for the sake of simplicity and coherence with the previous section).

By backward induction on the depth of the nodes $s \in \mathcal{T}$ we build the quantities

$$\sigma(s) = \begin{cases} L(s) & \text{if } s \text{ is a terminal node of } \mathcal{T}, \\ (1-\rho)L(s) + \rho\,\sigma(s0)\,\sigma(s1) & \text{if } s \text{ is an internal node of } \mathcal{T} \end{cases} \tag{8}$$

(we recall that $s0$ and $s1$ are the two sons of the node s).

It has been proved in [11] that with this construction

$$\sigma(\emptyset) = \sum_{m \in \mathcal{M}} \pi(m) \prod_{i=1}^{M} Q_m^E(y_i \mid x_i).$$

To compute all the terms of the Cesàro sum defining the progressive estimator, we have to perform $2L$ such sums, but having performed one sum the next one is obtained only by adding an observation (x_i, y_i); so it is sufficient to update the calculations for the counters and $\sigma(s)$ along the branch it follows, giving rise to d computations. However, to determine completely the distribution Q_π we have to consider every 2^d possibilities for $\mathcal{T}(x_1)$ (that is, it can be any terminal node of $\mathcal{T}$). In conclusion the complexity of the algorithm is of order $d2^dN$; this is to be compared to the number of models considered which is greater than $2^{2^{d-1}}$.

It has to be noted that the form of the prior π is crucial in order this algorithm to be used (namely, the principle is that it allows a nice factorization of Eq. (7)). However, we can make a straightforward extension to a more general set of priors by allowing the branching parameter ρ to depend on the node s considered in the tree, in which case one has just to replace the constant ρ by the the local branching parameter ρ_s in Eq. (8). This set of priors is large enough to allow a wide range of flexibility, for example, one can choose to give less *a priori* weight to "shallow" trees which we expect not to be so relevant, by taking ρ_s close to 1 for those nodes s that are close to the root.

5. DATA-DEPENDENT SCALING FOR CONTINUOUS-VALUED OBSERVATIONS

In this section we will consider the case when X takes its values in $\mathbb{R}^p$, where each coordinate corresponds to a "feature" of the observation. Denoting by x^k the kth coordinate of x, we will restrict ourselves to questions of the form "$x^k < a$?" depending on k and a. We can thus represent all possible questions by couples $(k, a) \in \{1, \ldots, p\} \times \mathbb{R}$. In principle it would be possible to discretize the values taken by a over $\mathbb{R}$, for instance by restricting $a \in \mathbb{Q}$, in order to get a countable set $\mathcal{Q}$ of questions. If we know nothing about the range of values taken by x however, it seems clear that choosing some *a priori* distribution on $\mathbb{Q}$ would give quite bad results in pratice unless we have a huge number of examples.

Hence we propose to modify slightly the construction of Catoni's estimator by letting the prior depend on the data. A model is now a triplet $m = (T, K_T, A_T)$ where T is a tree structure, and K_T, A_T are families of integers and real numbers, respectively, both indexed by the internal nodes of T. A quite natural way to define a distribution on the set of models $\mathcal{M}$ is to choose a distribution on the tree structure T, then on K_T given T, then on A_T given (T, K_T). In the sequel we will denote by $\tau = (T, K_T)$ the tree with its internal nodes labeled with the "type" of question to ask (i.e., on which coordinate of X the question should be asked).

In our case we will choose the two first distributions in a deterministic (data-independent) way which we will denote by $\pi(\tau)$. Now for the last one we notice that in fact the value of the estimator $Q_m^E(y_1 \mid x_1, z_2^L)$ depends only on the statistic $m(x_1), \ldots, m(x_n)$, that is, on the way the sample x_1^N is "split" across the tree. In other words, we can group the possible values of A_T into statistically (with regard to the sample) equivalent sets of questions, corresponding to the different possible splittings of the sample x_1^N (a "splitting" is here a mapping from the set $\{x_i;\ i = 1, \ldots, N\}$ to the set $\partial\tau$ of the leaves of the tree τ). Let $\Delta^\tau(x_1^N)$ be the set of all such possible splittings when A_T varies in $\mathbb{R}^{|\partial\tau|}$. What we propose is simply to give equal weight to each one of these splittings. We thus define the data-dependent mixture probability

$$\widetilde{Q}^E_{\text{mix.}}(y_1^L \mid x_1^L) = \sum_{\tau=(T,K_T)} \pi(\tau) \sum_{\sigma \in \Delta^\tau(x_1^N)} \frac{1}{|\Delta^\tau(x_1^N)|} \prod_{i=1}^{L} Q^E_{\tau,\sigma}(y_i \mid x_i),$$

where $Q^E_{\tau,\sigma}$ denotes the common value of the estimators Q^E_m for the tree τ and the splitting σ.

It is important to note here that with this definition, the value of the mixture probability on a single point $\widetilde{Q}^E_{\text{mix.}}(y_i \mid x_1^L)$ still depends on the *whole* X-sample x_1^N, since the data-dependent prior still depends on the number of different ways we can split the *entire* X-sample, and that is why the conditioning will always involve x_1^L. (Recall that on the other hand, the dependency of $\widetilde{Q}^E_{\text{mix.}}$ with respect to the "estimation" examples ranging from $L+1$ to N, is denoted by the superscript E.)

We can then define the data-dependent progressive mixture estimator $\widetilde{Q}_\pi$ by Eq. (1) just replacing $Q^E_{\text{mix.}}$ with $\widetilde{Q}^E_{\text{mix.}}$ and introducing the suitable modification as for the conditioning, that is:

$$\widetilde{Q}^E_\pi(dY_1 \mid x_1, z_2^L) = \frac{1}{L}\sum_{M=1}^{L} \widetilde{Q}^E_{\text{mix.}}(dY_1 \mid x_1^L, y_2^M). \tag{9}$$

We now give a result similar to Theorem 1:

THEOREM 3. – *With the above definitions, the progressive mixture estimator with data-dependent prior $\widetilde{Q}_\pi$ satisfies, for any exchangeable probability distribution P_N,*

$$E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N), \widetilde{Q}_\pi(Y_1 \in \cdot \mid X_1, Z_2^N)\big) \tag{10}$$
$$\leqslant \inf_{m=(\tau,A_T)\in\mathcal{M}} \left\{ E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N), Q_m(Y_1 \in \cdot \mid X_1, Z_{L+1}^N)\big) + \frac{1}{L}\log\frac{1}{\pi(\tau)} + \frac{|\partial m|\log(N+1)}{L} \right\},$$

where $|\partial m|$ denotes the number of terminal nodes of the tree model m.

If we choose for the estimators Q_m the same conditional Laplace estimators as in the previous section, then using Theorem 2 we immediately deduce

COROLLARY 2. – *Under the same hypotheses the estimator $\widetilde{Q}_\pi$ satisfies*

$$E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N), \widetilde{Q}_\pi(Y_1 \in \cdot \mid X_1, Z_2^N)\big) \tag{11}$$
$$\leqslant \inf_{m,\theta^m} \left\{ E_{P_N} H\big(P_N(Y_1 \in \cdot \mid X_1, Z_2^N), M_{m,\theta^m}(Y_1 \mid X_1)\big) + \frac{\log\pi(m)^{-1}}{L} + \frac{c\,|\partial m|}{N-L+1} + \frac{|\partial m|\log(N+1)}{L} \right\},$$

where M_{m,θ^m} is the conditional multivariate distribution associated with the tree model m and the set of parameters θ^m, defined in Section 3.

Proof of Theorem 3. – We can apply the same first steps as in the proof of Theorem 1 (however, in this case, since we are dealing with discrete distributions, we can skip the considerations about relative densities). Namely, let m be a given model, we want to upper-bound the quantity

$$
\begin{aligned}
-E_{P_N} \log \frac{\widetilde{Q}^E_\pi(Y_1 \mid X_1, Z_2^L)}{Q^E_m(Y_1 \mid X_1)} &\leqslant -\frac{1}{L}\sum_{M=1}^{L} E_{P_N} \log \frac{\widetilde{Q}^E_{\text{mix.}}(Y_1 \mid X_1^L, Y_2^M)}{Q^E_m(Y_1 \mid X_1)} \\
&= -\frac{1}{L}\sum_{M=1}^{L} E_{P_N} \log \frac{\widetilde{Q}^E_{\text{mix.}}(Y_{M+1} \mid X_1^L, Y_2^M)}{Q^E_m(Y_{M+1} \mid X_{M+1})} \\
&= -\frac{1}{L} E_{P_N} \log \frac{\widetilde{Q}^E_{\text{mix.}}(Y_1^L \mid X_1^L)}{Q^E_m(Y_1^L \mid X_1^L)},
\end{aligned}
$$

where the first inequality follows from the concavity of the logarithm, the following equality is obtained by swapping the role of (X_1, Y_1) and (X_{M+1}, Y_{M+1}) using the exchangeability of P_N (replacing (X_{L+1}, Y_{L+1}) by (X_1, Y_1) if necessary), and the last one by the chain rule.

Finally the following inequality holds:

$$
-\frac{1}{L} E_{P_N} \log \frac{\widetilde{Q}^E_{\text{mix.}}(Y_1^L \mid X_1^L)}{Q^E_m(Y_1^L \mid X_1^L)} \leqslant -\frac{\log \pi(\tau)}{L} + \frac{E_{P_N} \log |\Delta^\tau(x_1^N)|}{L}.
$$

It can be derived in the following way: for our fixed model $m = (\tau, A_T)$ there exists a splitting $\sigma_0 \in \Delta^\tau(x_1^N)$ which separates the data exactly in the same way as does the tree model m. Thus

$$
Q^E_m(X_1^L \mid Y_1^L) = Q^E_{\tau,\sigma_0}(X_1^L \mid Y_1^L),
$$

and keeping only the corresponding term in the sum defining $\widetilde{Q}^E_{\text{mix.}}$ as in the proof of Theorem 1, we get the desired inequality.

Now the sample reaching an internal node s is at most of size N and thus can be split at most in $N+1$ ways at this node since we only make use of order statistics for a certain coordinate. Thus

$$
\left|\Delta^\tau\left(x_1^N\right)\right| \leqslant (N+1)^{|\operatorname{int}(m)|}
$$

where $|\operatorname{int}(m)|$ denotes the number of internal nodes of the model tree m; but $|\operatorname{int}(m)| \leqslant |\partial m|$ since for a complete binary tree $|\partial m| = |\operatorname{int}(m)| + 1$; thus we get to the desired conclusion. □

This result can be enlarged in several ways. First, one can notice that we can easily extend this result in a more general framework. Namely, assume we have a possibly uncountable family of models $\mathcal{M}$, but that every model can be written in the from $m = (\tau, \lambda_\tau)$ where τ takes its values in some countable set. In the same way as above, we can choose a fixed prior distribution for τ and a data-dependent distribution for λ_τ, grouping the values of λ_τ into statistically equivalent sets to which we assign an equal weight. We can then apply the same reasoning and the main remaining point is to get a bound for $E_{P_N} \log |\Delta^\tau(x_1^N)|$. For instance, if instead of questions of the type "$x^k < a$?" we choose at each node a set of questions of Vapnik–Cervonenkis dimension bounded by D, then we easily come to

$$\log|\Delta^\tau(x_1^N)| \leqslant C|\partial m| D \log N,$$

where C is a real constant.

Secondly, in practice it seems uneasy to compute $|\Delta^\tau(x_1^N)|$. Instead of a uniform distribution on all the global possible splittings, we would like to take the distribution given by the product of the uniform distributions on all local possible splits at each node. To state it in a more formal (and probably clearer) way, let us denote by σ a given partitioning of the sample over the tree τ and by I_s^σ the number of examples reaching an internal node s using the partitioning σ. If we choose uniformly at each internal node between the $I_s^\sigma + 1$ different possible splits, then the new distribution ω on the possible partitionings is given by

$$\omega(\sigma) = \prod_{s \in \text{int(T)}} \frac{1}{I_s^\sigma + 1}.$$

The advantage of this distribution is that it can be computed easily in a recursive way for every σ. Furthermore, the bound of Theorem 3 still holds true since

$$-\log \omega(\sigma) = \sum_{s \in \text{int}(T)} \log(I_s^\sigma + 1) \leqslant |\partial m| \log(N+1).$$

Finally, notice that the bound in Theorem 3 is not as good as in Theorem 1, namely, there is a term of order $\log(N)/L$ in the new bound whereas the penalty term was only of order $1/L$ in Theorem 1. If we take $L = kN$, $0 < k < 1$, this means that if the real conditional distribution $P_N(Y_1^N \mid X_1^N)$ is in the model (that is, is of the form

$M_{m,\theta^m}^{\otimes N}$), the modified estimator converges towards it at a rate of order $\log N/N$, to be compared with $1/N$ in Theorem 1. This is because we dropped the hypothesis of a countable family of models. However, in the particular case considered in this section, namely, conditional multivariate distributions M_{m,θ^m} depending on the tree model m and the set of parameters θ^m, the following theorem states that this rate of convergence is actually the minimax rate within the set of models, and therefore cannot be significantly improved.

THEOREM 4. – *Assume X is a real random variable drawn according to the uniform distribution U in $[0,1]$, then with the set of tree models $\mathcal{M}$ defined in this section, there exists a real constant C such that for any estimator $Q(y \mid x, z_1^N)$ depending of a sample of size N, the following lower bound holds true for sufficiently big N:*

$$\sup_{m,\theta^m} E H\left(M_{m,\theta^m}(Y \mid X), Q\left(Y \mid X, Z_1^N\right)\right) \geqslant C \frac{\log N}{N},$$

where the expectation on Z_1^N is taken with respect to the product distribution $(U(dX).M_{m,\theta^m}(dY \mid X))^{\otimes N}$ (in other words, when Z_1^N is drawn i.i.d. according to $U(dX).M_{m,\theta^m}(dY \mid X)$).

To prove this theorem we will make use of the following version of Fano's lemma (see, e.g. [5], Corollary 2.9):

LEMMA 1. – *Let $\mathcal{P}$ be a finite set of probability distributions on a measurable space $\mathcal{X}$ such that $|\mathcal{P}| = J \geqslant 4$, and $\psi : \mathcal{P} \to F$ a function taking its values in some metric space (F, d). Assume there exist positive constants K, γ, t such that*

(i) $$\sup_{P,Q \in \mathcal{P}} H(P,Q) \leqslant K,$$

(ii) $$\inf_{\substack{P,Q \in \mathcal{P} \\ P \neq Q}} d\left(\psi(P), \psi(Q)\right) > t > 0,$$

(iii) $$K + \log 2 \leqslant (1-\gamma)\log(J-1)$$

then for any function $\Phi : \mathcal{X} \to F$ and any $p > 0$ the following lower bound holds

$$\sup_{P \in \mathcal{P}} E_P\left[d^p\left(\Phi(X), \psi(P)\right)\right] \geqslant \gamma \left(\frac{t}{2}\right)^p.$$

Proof of Theorem 4. – We can restrict ourselves to the case $c = 1$ since the minimax rate of convergence can only increase with c. Thus the distributions we are dealing with are conditional Bernoulli, which we will denote by $B_{m,\theta^m}(dY \mid X)$.

We will apply Fano's lemma in the following framework: as the metric space F we will take the set of all distributions on $[0, 1] \times \{0, 1\}$ and for d the Hellinger distance on E. Below we will construct the set $\mathcal{P}$ as a finite subset of the the set of product probability distributions

$$\mathcal{D} = \left\{ \left(U(dX).B_{m,\theta^m}(dY \mid X) \right)^{\otimes N} \right\}$$

on $([0, 1] \times \{0, 1\})^N$. A natural function $\psi : \mathcal{D} \to F$ is $\psi : P^{\otimes N} \mapsto P$ and any estimator depending on a sample of size N is indeed a function $\Phi : ([0, 1] \times \{0, 1\})^N \to F$.

Let us choose $\alpha, \beta \in [0, 1]$ which will remain fixed for the rest of the proof. Let us then consider the conditional probability of Y given by a Bernoulli distribution of parameter β if X belongs to some subinterval $I \subset [0, 1]$ of length η, and of parameter α if X does not belong to I. This distribution clearly belongs to our model since it can be represented as a tree with two questions corresponding to the endpoints of I. If $P_1^{\otimes N}$, $P_2^{\otimes N}$ are Nth direct products of two such distributions on (X, Y) (drawing X uniformly on $[0, 1]$), obtained with two disjoint intervals I_1 and I_2 of the same length η, then we get easily

$$H\left(P_1^{\otimes N}, P_2^{\otimes N}\right) \leqslant N\eta\left(H(\beta, \alpha) + H(\alpha, \beta)\right),$$

$$d^2(P_1, P_2) = \eta\left(\sqrt{\alpha} - \sqrt{\beta}\right)^2.$$

Now we can find as many as $J = \lfloor \frac{1}{\eta} \rfloor$ such disjoint intervals in $[0, 1]$ and take for $\mathcal{P}$ the set of distributions thus obtained. Taking

$$\eta = \frac{1}{H(\alpha, \beta) + H(\beta, \alpha)} \frac{\log N}{2N},$$

the hypotheses of Fano's lemma are now satisfied for $K = \frac{1}{2} \log N$, $\gamma = 1/4$, $t = C_1 \sqrt{\log N / N}$ for a constant

$$C_1 = \left|\sqrt{\alpha} - \sqrt{\beta}\right| / \sqrt{(2H(\alpha, \beta) + H(\beta, \alpha))},$$

and N big enough so that (iii) is satisfied.

This yields the desired result if we choose $p = 2$ in the inequality obtained, and finally apply the well-known inequality $H(P, Q) \geqslant 2d^2(P, Q)$ as a final step. □

Remark. – Note that the only probability distributions for Y used in this proof are Bernoulli with parameter α or β which can be arbitrarily fixed. Therefore the extra $\log N$ factor in the rate of convergence does not come from the "pathologic" behavior of the Kullback–Leibler distance for parameter values near 0 and 1, as one might have thought at first glance.

6. APPROXIMATE COMPUTATION USING A MONTE-CARLO METHOD

Let us have a new look on Eq. (1) and focus our attention on one particular term $Q^E_{\text{mix.}}(dY_1 \mid x_1, z_2^M)$ in the Cesàro sum defining Q^N_π. Since we will still be working in this section in the framework of classification trees, the space $\mathcal{Y}$ is discrete and we can identify probability distributions with their densities. Let us write

$$\begin{aligned} Q^E_{\text{mix.}}(y_1 \mid x_1, z_2^M) &= \frac{\sum_{m\in\mathcal{M}} \pi(m) Q^E_m(y_1^M \mid x_1^M)}{\sum_{m'\in\mathcal{M}} \pi(m') Q^E_{m'}(y_2^M \mid x_2^M)} \\ &= \sum_{m\in\mathcal{M}} Q^E_m(y_1 \mid x_1) \frac{\pi(m) Q^E_m(y_2^M \mid x_2^M)}{\sum_{m'\in\mathcal{M}} \pi(m') Q^E_{m'}(y_2^M \mid x_2^M)} \\ &= \sum_{m\in\mathcal{M}} Q^E_m(y_1 \mid x_1) \omega^E(m \mid z_2^M), \end{aligned}$$

where ω^E is, from a Bayesian point of view, the *a posteriori* distribution of the models given the data subset z_2^M of the test set, when the estimation set is fixed. We have shown in Section 4 that in the case of a fixed-design tree, and when the prior π is the distribution of the genealogy tree of a branching process with constant parameter, this sum can be recursively computed in an efficient way.

In the general case however, the tree design is not fixed because we have to choose between several questions at each node, and we did not find any similar factorization for the computations. We therefore propose to simulate the *a posteriori* distribution of the models using a Monte-Carlo algorithm.

In order to do this we will need to define a reversible transition kernel $\Gamma(m, m')$ on the set of models $\mathcal{M}$, such that the associated stationary distribution is precisely ω^E. Γ should therefore be irreducible and acyclic and satisfy for any $m, m' \in \mathcal{M}$: either

$$\Gamma(m, m') = \Gamma(m', m) = 0$$

or

$$\frac{\Gamma(m', m)}{\Gamma(m, m')} = \frac{\omega^E(m \mid z_2^M)}{\omega^E(m' \mid z_2^M)} = \frac{\pi(m)}{\pi(m')} \cdot \frac{Q_m^E(y_2^M \mid x_2^M)}{Q_{m'}^E(y_2^M \mid x_2^M)}. \tag{12}$$

Our goal is that at each transition of the MCMC chain there should be as few computations as possible to perform. For that purpose we will allow $\Gamma(m, m') \neq 0$ only if m' is a "neighbor" of m. We will call $m = (T, F_T)$ a neighbor of $m' = (T', F'_{T'})$ if the trees T and T' only differ by two added or removed terminal nodes (leafs), i.e., if T' is obtained from T by turning one of its leafs into an internal node and adjoining to it two leafs, or vice-versa, and if $F_{T \cap T'} = F'_{T \cap T'}$, or in other words, if the questions attached to the common internal nodes of T and T' are the same.

As in the previous section we will construct our prior π in the form

$$\pi(m) = \pi(T, F_T) = \pi(T)\pi(F_T \mid T).$$

Once again we will choose for the first marginal $\pi(T)$ the probability distribution of the genealogy tree of a branching process with a constant parameter $\rho \leqslant \frac{1}{2}$, restricted (and normalized) to the set of trees of depth less than d (this a is slightly different prior than the one considered in Section 4, which was the distribution of a *stopped* branching process, though the difference is quite negligible in practice). The advantage of this choice is that an MCMC chain simulating this distribution can easily be built using the following algorithm for the transitions:

- Start at the root node, then follow the branches choosing for each internal node you reach one of its sons with a probability of $\frac{1}{2}$, until you reach a terminal node f.
- If f's brother is also a leaf, then destroy them both with probability $\frac{1}{2}$, or give two sons to f with probability $\rho(1 - \rho)/2$ (unless you are at the maximum depth d, in which case you do nothing), or do nothing with the remaining probability.
- If f's brother is an internal node, then give two sons to f with probability $\rho(1 - \rho)/2$, or do nothing with the remaining probability.

It can readily be seen that the transition kernel thus defined on the trees is irreducible and aperiodic and that it is reversible with respect to its stationary distribution, which is the desired one. Again, like in Section 4, this can be extended easily to a more general prior where the parameter ρ is not a constant but depends on the node s considered (for example

one can actually simulate the exact same distribution as in Section 4 if wanted).

The next step is to see how to define $\pi(F_T \mid T)$. Recall that we have in mind to construct a Markov Chain Monte Carlo, and therefore we are interested at the first place in transition probabilities from one tree to one of his neighbors. We thus want to compute in a simple way the ratio

$$\frac{\pi(F'_{T'} \mid T')}{\pi(F_T \mid T)}, \tag{13}$$

when (T, F_T) and $(T', F'_{T'})$ are neighbor labeled trees. (To be sure that this ratio is well-defined we will assume here and for the rest of this section that all trees have strictly positive prior probability.)

To make things clearer, let us assume that we want to perform a transition from T to T', where T is a subtree of T' and that the leaf s of T has been turned into an internal node in T', so a new question must be attached to s in T'. We would like to define a *local rule* $R(f_s \mid F_T)$ giving us the probability distribution of this new question given the rest of the tree (which we will use as transition probability for the MCMC chain). Unfortunately, the ratio (13) is *not* necessarily a probability distribution on the new question, because of the different dependencies on T' and T, respectively, for the numerator and the denominator.

This motivates the introduction of the following definitions and hypotheses: let us write $F_T \prec F_{T'}$ if $T \subset T'$ and if $F_{T'}$ and F_T coincide on the internal nodes of T. Let us then define the marginal

$$\pi(F_{T'} \mid T) = \sum_{F_T:\ F_{T'} \prec F_T} \pi(F_T \mid T).$$

In order the ratio (13) to be a probability distribution, we will make the assumption that

$$\forall T' \subset T \quad \pi(F_{T'} \mid T) = \pi(F_{T'} \mid T'). \tag{14}$$

As a natural consequence, if we define $\widetilde{\pi}$ as a distribution on the questions attached to the maximal complete tree T^* of depth d by $\widetilde{\pi}(F_{T^*}) = \pi(F_{T^*} \mid T^*)$ and $\widetilde{\pi}$'s marginal

$$\widetilde{\pi}(F_T) \triangleq \sum_{F_{T^*}:\ F_T \prec F_{T^*}} \pi(F_{T^*} \mid T^*) = \pi(F_T \mid T^*),$$

then by assumption (14) we have

$$\widetilde{\pi}(F_T) = \pi(F_T \mid T).$$

Therefore any non-vanishing distribution on the questions attached to the maximal tree T^* gives rise to an associated local rule defined by

$$R(f_s \mid F_T) = \frac{\widetilde{\pi}(F'_{T'})}{\widetilde{\pi}(F_T)} \tag{15}$$

(we recall that s is the leaf of T which becomes an internal node of T', and f_s is the question attached to node s). R is now a true conditional distribution on the question f_s.

The natural reciprocal problem to be asked now is, what are the minimal conditions to be satisfied by a set of local rules R (that is to say, a set of conditional distributions on the question f_s to be asked at node s given the questions asked on the internal nodes of a tree T such that s is a leaf of T), so that there exists a distribution $\widetilde{\pi}(F_T)$ such that (15) is satisfied? Obviously, a "local coherence" condition must be satisfied, namely, take a tree T, let s_1 and s_2 be two different leafs of T; T_1, T_2, T_{12} be the trees obtained by adjoining two sons to s_1, s_2, and both s_1 and s_2, respectively; F_T, $F^1_{T_1}$, $F^2_{T_2}$ and $F^{12}_{T_{12}}$ be sets of questions attached to these trees such that they match on the common internal nodes of two of them, then R must satisfy

$$\frac{\widetilde{\pi}(F_{T_{12}})}{\widetilde{\pi}(F_T)} = R(f_{s_2} \mid F^1_{T_1}).R(f_{s_1} \mid F_T) = R(f_{s_1} \mid F^2_{T_2}).R(f_{s_2} \mid F_T) \qquad (*)$$

(this means that the probability distribution of the two new questions given the initial tree T must not depend on the order they were constructed).

The following theorem states that this necessary condition is also sufficient:

THEOREM 5. – *Let a set local rules R be given such such that condition $(*)$ is satisfied for any choice of (T, s_1, s_2), then there exists a distribution $\widetilde{\pi}(F_T)$ on the set of questions attached to the maximal tree T^* satisfying* (15).

Proof. – This is actually quite straightforward. Let (T, F_T) be a model tree. Choose some way of building T node by node, namely a sequence $\{\emptyset\} = T_0 \subsetneq T_1 \subsetneq \cdots \subsetneq T_k = T$ and a sequence $\emptyset = F^0_{T_0}, F^1_{T_1}, \ldots, F^k_{T_k} =$

F_T such that for every $i < k$, $(T_i, F^i_{T_i})$ and $(T_{i+1}, F^{i+1}_{T_{i+1}})$ are neighbors (thus $F^i_{T_i}$ is necessarily the set of questions F_T restricted to the internal nodes of T_i). We will call such a sequence an *construction path* for T. Necessarily $\widetilde{\pi}(F_T)$ should be defined the following way:

$$\widetilde{\pi}(F_T) = \prod_{i=1}^{k} R(f_{s_i} \mid F^{i-1}_{T_{i-1}}),$$

where s_i is of course the leaf of T_{i-1} turned into an internal node of T_i, and f_{s_i} the question attached to s_i.

It is clear now that we should only make sure that this definition does not depend on the path we chose. Obviously the construction paths are in natural bijection with the set of permutations $\mathfrak{S} = (s_0 = \emptyset, s_1, \ldots, s_n)$ of all the internal nodes of T, satisfying the property that for any $i < j$, s_j is not an ancestor of s_i (i.e., $\mathfrak{S}$ is compatible with the order "be ancestor of"). Such a permutation will be called *admissible* and gives us the order of construction of T node by node. Now relation $(*)$ tells us that if s_i is not the father of s_{i+1}, swapping them in the permutation does not change the value of $\widetilde{\pi}(F_T)$ above defined (and the new permutation is still admissible). We will say that the new permutation is equivalent to the first one.

Now it is easy to see that actually all admissible permutations are equivalent using a simple recursion on the number of internal nodes of T. Namely, let $\mathfrak{S}_1$ and $\mathfrak{S}_2$ be two admissible permutations. Let s^* be an internal node of T such that his two sons are both leafs. Thus no internal node is a descendant of s^* and therefore we can swap s^* in $\mathfrak{S}_1$ and $\mathfrak{S}_2$ successively with all the nodes built after him. We thus obtain two admissible permutations $\mathfrak{S}'_1$ and $\mathfrak{S}'_2$, respectively, equivalent to $\mathfrak{S}_1$ and $\mathfrak{S}_2$ and such that s^* is the last node to be built in both cases. We can thus proceed recursively, since the case of a single internal node is obvious. □

In particular, it is clear that we can choose any local rule R such that the probability distribution of the question at a given node only depends on his ancestors. In this case $(*)$ is obviously satisfied.

We now study the last factor in the right side of (12). Let $m_1 = (T_1, F^1_{T_1})$ and $m_2 = (T_2, F^2_{T_2})$ be two neighbor models such that $T_1 \subsetneq T_2$, u is the leaf of T_1 being turned into an internal node of T_2, and f_u is the new question attached to u. As in Section 4 we define for any internal or terminal node s the counters b^i_s, $i = 0, \ldots, c$, as the number of examples

of the subsample z_2^M that cross or reach node s and are of class i. Since the estimators $Q_m^E(\cdot \mid \cdot)$ are conditional multivariate distributions whose vector parameters θ_s at each leaf s are estimated using the estimation set z_{L+1}^N (here we do not make any hypothesis on the way these parameters are estimated, it can be a Laplace estimator or any other estimator), we get

$$K(m_1, u, f_u) \triangleq \frac{Q_{m_1}^E(y_2^M \mid x_2^M)}{Q_{m_2}^E(y_2^M \mid x_2^M)} = \frac{\prod_{s \in \partial T_1} \prod_{i=0}^{c} (\theta_{s,i})^{b_s^i}}{\prod_{s \in \partial T_2} \prod_{i=0}^{c} (\theta_{s,i})^{b_s^i}}$$
$$= \prod_{i=0}^{c} \frac{(\theta_{u,i})^{b_u^i}}{(\theta_{u0,i})^{b_{u0}^i} (\theta_{u1,i})^{b_{u1}^i}}.$$

Here $u0$ and $u1$ denote of course the two sons of u, and $\theta_{s,i}$ is the ith coordinate of θ_s. Thus this factor only depends on the counters at nodes $u, u0, u1$ and is therefore easy to compute.

To put it together, here is an algorithmic description of an iteration of our MCMC algorithm for the step concerning the transition towards a new model:

(1) Choose a leaf of the current model m and choose whether you adjoin to it two new sons, or destroy it and his brother, or do nothing, according to the algorithm already described sooner.

(2) If you chose to delete two leafs, denote by u their father, then accept the new truncated model m' with probability $K(m', u, f_u) \wedge 1$.

(3) If you chose to add two leafs, denote by u the happy father of the twins, choose a question f_u to be attached to it according to the local distribution R, then accept the new model m' with probability $1/K(m, u, f_u) \wedge 1$.

It is now clear that the transition kernel $\Gamma(m, m')$ thus defined satisfies (12). Furthermore, the number of counters to be updated at each step is of most $c + 1$ (in the case you create two new leafs).

What is left is now to perform the Cesàro sum of the Bayesian estimators while the test set is growing. To do this, we suggest that the examples could actually be introduced progressively while the Monte-Carlo chain is running, so that there is the same number of steps performed in the chain for each new example introduced. However, one could think of other possible methods to achieve a tradeoff between accuracy and computation complexity; we are planning to perform several tests using various methods in the next future.

7. DISCUSSION

The point of view we adopted in this paper is to give some basic theoretical results about the progressive mixture estimator used in regression estimation. Actually this estimator is a modification of the classical Bayesian analysis, for which we divide the sample into two subsets and perform a Cesàro sum over Bayesian mixtures of available estimators as we let the size of the "test set" grow. Restricting our attention on the case of regression trees, we then proposed that the computation of the mixture could be approximated by a MCMC chain, which is quite standard in Bayesian analysis.

We want to discuss here the relation of this work with that of Geman and Amit [1,2] who put forward, in a written character recognition problem, the construction of a classification tree using a local entropy minimization rule. More precisely, the basic algorithm they propose is to grow a tree by selecting at each node the question which reduces the most the empirical entropy at this node (i.e., a local maximum likelihood selection), stopping the construction when reaching a fixed maximum depth or a minimum size of the sample at a given node. (In practice, the optimization is in fact performed on a randomly selected small subset of the available set of questions at each node.) We want to point out the similarity of this procedure with ours: in an informal sense the local entropy reduction procedure can be compared with a MCMC chain (or, to fit the comparison better, a Gibbs sampler) at "zero temperature". In our algorithm the number of examples in the test set plays the role of inverse temperature (as the bigger the test set is, the less likely we are to select questions which are locally "irrelevant"). In this regard, introducing the test examples progressively can be seen as a kind of simulated annealing algorithm.

We hope that allowing more liberty in the search in the tree space can lead to visit trees achieving a better performance. Moreover, the *a priori* distribution acts here as a penalty over the complexity of the models, and thus as a "natural" stopping rule included in the algorithm. However as what we want here is to compute a sum over the models by simulating their *a posteriori* distribution, and not merely to select one of them, the complexity of the computation will of course be much higher than for the local optimization algorithm. We are currently trying to find some answers to this problem; one could think for example of performing a MCMC trajectory once and for all, and then just re-use it to estimate every new example.

We would also like to make some reference to Chipman et al.'s work [7], of which we became aware only recently, which also extensively presents the use of a MCMC algorithm to explore the space of tree models with a Bayesian prior. There is a strong similarity between their procedure and ours; we therefore want to highlight a few points on which we hope our point of view might be relevant. First we started from a theoretical point of view, deriving strong upper bounds for the risk of the progressive mixture estimator in a general framework and for the particular case of regression trees. This has two main consequences on the design of our algorithm: first, our goal is really to compute a sum, and not to select a single tree as that was the case in [7]. In this regard we follow the Bayesian principle more strictly. It is relevant to note that Catoni [6] proved on a simple example that model selection may behave worse than model mixture. Besides, Geman and Amit also show that performance turns out to be much better when they perform a weighted sum over several different trees obtained with the "randomized" version of the local entropy reduction algorithm.

Secondly, because of our construction we divide our sample into two sets and suggest to introduce the test examples progressively. We believe that this could give more liberty in the movements of the MCMC chain at the beginning like in a simulated annealing algorithm (see above). On the other hand, the MCMC implementation we exposed here is still rough and quite sketchy; practical evidence in [7] strongly suggests that allowing more liberty in the chain transitions (allowing to change the question asked at an internal node; swapping the questions of two neighbor nodes) significantly improves the performance of the algorithm. We plan to make more practical experiments in this direction in the next future.

ACKNOWLEDGEMENTS

We would like to thank the reviewer for his close reading of this paper and for his useful remarks and suggestions that led to some additions to the original text.

REFERENCES

[1] Y. AMIT and D. GEMAN, Shape quantization and recognition with randomized trees, *Neural Computation* 9 (1997) 1545–1588.

[2] Y. AMIT, D. GEMAN and K. WILDER, Joint induction of shape features and tree classifiers, *IEEE Trans. PAMI* 19 (11) (1997) 1300–1306.

[3] A. BARRON and Y. YANG, Information theoretic determination of minimax rates of convergence, Department of Statistics, Yale University, 1997.

[4] A.A. BARRON, Are Bayes rules consistent in information? in: T.M. Cover and B. Gopinath (Eds.), *Open Problems in Communication and Computation*, Springer, Berlin, 1987, pp. 85–91.

[5] L. BIRGÉ, Approximation dans les espaces métriques et théorie de l'approximation, *Z. Wahrscheinlichkeitstheor. Verw. Geb.* 65 (1983) 181–237.

[6] O. CATONI, "Universal" aggregation rules with exact bias bounds, Preprint of the Laboratoire de Probabilités et Modèles Aléatoires, Université Pierre et Marie Curie, available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999 (to appear in *Annals of Statistics*), 1999.

[7] H. CHIPMAN, E.I. GEORGE and E. MCCULLOCH, Bayesian CART model search, *JASA* 93 (1998) 935–947.

[8] T.M. COVER and J.A. THOMAS, *Elements of Information Theory*, Wiley Series in Telecommunications, Wiley, New York, 1991.

[9] L. DEVROYE and L. GYÖRFI, *Nonparametric Density Estimation: The L^1 View*, Wiley, New York, 1985.

[10] D. HELMBOLD and R. SHAPIRE, Predicting nearly as well as the best pruning of a decision tree, *Machine Learning* 27 (1997) 51–68.

[11] F.M.J. WILLEMS, Y.M. SHTARKOV and T.J. TJALKENS, The context-tree weighting method: basic properties, *IEEE Trans. Inform. Theory* 41 (3) (1995) 653–664.

[12] F.M.J. WILLEMS, Y.M. SHTARKOV and T.J. TJALKENS, Context weighting for general finite-context sources, *IEEE Trans. Inform. Theory* 42 (5) (1996) 1514–1520.