ECOLE NORMALE SUPERIEURE

The Mixture Approach to Universal Model Selection

Olivier CATONI

LIENS - 97 - 22

Département de Mathématiques et Informatique

CNRS URA 1327

The Mixture Approach to Universal Model Selection

Olivier CATONI

LIENS - 97 - 22

September 1997

Laboratoire d'Informatique de l'Ecole Normale Supérieure
45 rue d'Ulm 75230 PARIS Cedex 05

Tel : (33)(1) 44 32 00 00

Adresse électronique : catoni@dmi.ens.fr

THE MIXTURE APPROACH TO UNIVERSAL MODEL SELECTION

OLIVIER CATONI

ABSTRACT. We build a model selection algorithm which has a mean Kullback risk upper bounded by the mean risk for the best model applied to half the sample augmented by an explicit penalty term of order one over the sample size. This estimator is not a "true" selection rule, but instead an adaptive convex combination of the models: this mixture approach is inspired by the context tree weighting method of information theory by Willems, Shtarkov and Tjalkens [6, 7], which is a "universal" data compression algorithm for stationary binary sources. Our algorithm is "universal" with respect to the statistical mean Kullback risk in the sense that it is almost optimal for any exchangeable sample distribution.

We give an application to the estimation of a probability measure by adaptive histograms. We detail a factorised computation of the mixture which weights M models performing a number of operations of order $\log(M)$, when the subdivisions underlying the histograms have nested cells.

The end of the paper shows that in the case of a dense family of models, a true selection rule can be built in a second step. We give an upper bound for the mean square Hellinger distance of the estimator from the sample distribution, which tends to zero at the optimal rate, up to an explicit multiplicative constant, in the case when the square Hellinder distance and the Kullback divergence are comparable.

1. INTRODUCTION

The purpose of this paper is to adapt some methods and results from information theory to the statistical estimation of a probability distribution. We will measure the quality of estimators with respect to the Kullback divergence, and sometimes also with respect to the Hellinger distance. More precisely we will be interested in "universal" model selection algorithms: the true sample distribution will only be supposed to be exchangeable, and we will seek an adaptive way to select the most efficient among a family of estimators taking their values in a family of models. The mixture approach consists in mixing together the estimators using appropriate adaptive weights, instead of actually choosing one. It comes from coding theory, in which some "double mixture codes" have been proved to have a nearly optimal redundancy for any stationary source. These codes are called "universal", because they do not use any special knowledge about the source. Their design, study and

Date: September 1997.

Key words and phrases. Adaptive Statistical Model Selection, Adaptive Mixtures of Models, Adaptive Histograms, Context-Tree Weighting Method, Mean Kullback Risk.

I am glad to thank L. Birgé and P. Massart for the joint organisation of a seminar on Adaptive Statistics at the Ecole Normale Supérieure, and for the beautiful course of lectures they gave on the subject during the first semester 1997 and the numerous fruitful discussions we had on this occasion.

implementation has known a remarkable breakthrough with the works of Willems, Shtarkov and Tjalkens [6, 7] on the "context-tree weighting method".

Although any coding algorithm performs an implicit estimation of the probability distribution of the source, applying methods of information theory to statistics requires some further thinking, since the redundancy criterion of coding theory does not coincide with the Kullback contrast function (it is a Cesaro mean of Kullback risks for subsamples of growing sizes). The main result of this paper is an explicit universal model selection method. We call it the progressive mixture method, because it combines Bayesian estimators based on subsamples of increasing lengths (at least in its basic implementation, some alternatives of lower algorithmic complexity will also be mentioned).

If a true selection rule is needed, it is possible to approach the mixture estimator by a distribution drawn from one of the models in a second step. This gives an efficient upper bound for the mean square of the Hellinger distance, when the true sample distribution is in the closure of the family of models.

To illustrate this general approach and show that it can lead to efficient algorithms, we develop the case of the estimation of a probability measure by adaptive histograms.

To point out that a true selection rule (such as a penalised maximum likelihood estimator) could not always reach the same performance as our progressive mixture estimator with the weak assumptions we make, we give a toy counter example in which the increase in the average risk due to model selection is of order at least $1/\sqrt{N}$ for any true selection rule, and is of order at most $1/N$ for the progressive mixture estimator, where N is the size of the sample. This bears some resemblance with the well known result from game theory, saying that in general the optimal strategies are mixed. The reason for this failure of a true model selection is that we consider situations in which the true sample distribution does not necessarily lies in the closure of the family of models. This situation is very common in practice in signal or image analysis, where the models are far to (and need not either) capture all the complexity of the data.

It is also interesting to note that it is possible to build adaptive estimators for the mean risk for exchangeable sample distributions. As the analysis of the now well understood penalised minimum of contrast adaptive estimators is based on concentration theorems for product measures (see Birgé and Massart [1, 2, 3]), the mixture approach opens a different line of proofs.

2. A universal mixture estimator for model selection according to the Kullback divergence

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Let $\mathcal{B}_N$ be the product sigma algebra on $\mathcal{X}^N$. For each N, let P_N be a probability distribution on $(\mathcal{X}^N, \mathcal{B}_N)$. Let $(X_i)_{i=1}^N = X_1^N$ be the canonical process on $\mathcal{X}^N$. For any permutation $\sigma \in \mathfrak{S}_N$ of $\{1, \dots, N\}$, let σX be the exchanged process

$$(\sigma X)_i = X_{\sigma(i)}, \quad i = 1, \dots, N.$$

Let us assume that for each $N \in \mathbb{N}$, P_N is exchangeable. This means that for any $\sigma \in \mathfrak{S}_N$, any $A \in \mathcal{B}_N$,

$$P_N(X_1^N \in A) = P_N((\sigma X)_1^N \in A).$$

Let us consider a countable family $(Q_m^N)_{m\in\mathbb{N},N\in\mathbb{N}}$ of estimators. More precisely, we assume that for each $x_1^N \in \mathcal{X}^N$, $Q_m^N(\cdot|x_2^N)$ is a probability measure on $(\mathcal{X},\mathcal{B})$ and that for each $A\in\mathcal{B}$ the map $x_2^N \mapsto Q_m^N(A|x_2^N)$ is measurable.

We assume also that for each m, there is a dominating measure μ_m such that $Q_m^N(\cdot\,|\,x_2^N) \ll \mu_m$ for any $x_2^N \in \mathcal{X}^{N-1}$. Replacing if necessary μ_m by $\sum_{m\in\mathbb{N}} 2^{-(m+1)}\mu_m$ we will assume in the following that $\mu_m = \mu$ is independent of m. We will also assume that there is a measurable version $q_m^N(x_1\,|\,x_2^N)$ of $\dfrac{dQ_m^N(\cdot\,|\,x_2^N)}{d\mu}(x_1)$.

The model selection problem for the Kullback risk is to solve approximately, knowing the sample X_2^N, the minimisation problem

$$\inf_{m\in\mathbb{N}} E_{P_N} H(P_N(X_1\in\cdot\,|\,X_2^N), Q_m^N(X_1\in\cdot\,|\,X_2^N)),$$

where, for any probability measures ρ and $\nu\in\mathcal{M}_+^1(\mathcal{X})$, $H(\rho,\nu)$ is the Kullback divergence function (also called relative entropy)

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

We will build an approximate mixture solution to this problem. Let π be a probability distribution on $\mathbb{N}$, let $K<N$ be a positive integer. We consider the following estimator:

$$q_\pi^N(x_1\,|\,x_2^N) = \frac{1}{N-K+1}\sum_{M=K}^{N} \frac{\displaystyle\sum_{m\in\mathbb{N}} \pi(m)\left(\prod_{n=K+1}^{M} q_m^K(x_n\,|\,x_2^K)\right) q_m^K(x_1\,|\,x_2^K)}{\displaystyle\sum_{m\in\mathbb{N}} \pi(m)\prod_{n=K+1}^{M} q_m^K(x_n\,|\,x_2^K)}$$

$$\frac{dQ_\pi^N(\cdot\,|\,x_2^N)}{d\mu} = q_\pi^N(\cdot\,|\,x_2^N).$$

Let us remark immediately that this definition is independent of the choice of the dominating measure μ.

Theorem 2.1. *Under the previous assumptions*

$$(1)\quad E_{P_N} H(P_N(X_1\in\cdot\,|\,X_2^N), Q_\pi^N(X_1\in\cdot\,|\,X_2^N))$$
$$\le \inf_{m\in\mathbb{N}}\left\{ E_{P_N} H(P_N(X_1\in\cdot\,|\,X_2^N), Q_m^K(X_1\in\cdot\,|\,X_2^K)) + \frac{1}{N-K+1}\log\frac{1}{\pi(m)}\right\}.$$

Remark 2.1. We use the supplementary observations X_{K+1}^N to choose between the estimators Q_m^K, which are computed from X_2^N. Therefore X_{K+1}^N plays the role of a "test set".

Remark 2.2. We will see on examples that it is sometimes possible to balance the two terms of the sum on the right hand side of the previous equation by an appropriate choice of K and π.

Remark 2.3. Of course, the most interesting case is when P_N is a product measure. Anyhow, the proof requires only that P_N should be exchangeable. In practice also, the support of π will be finite, and we will thus use a finite number of models.

Proof. It is easy to see that

$$H(P_N(X_1 \in \cdot \mid X_2^N), Q_\pi^N(X_1 \in \cdot \mid X_2^N)) \\ = f(X_2^N) - \int_{\mathcal{X}} \log q_\pi^N(x_1 \mid X_2^N) P_N(dx_1 \mid X_2^N),$$

where $f(X_2^N)$ is independent of Q_π^N. In the same way

$$H(P_N(X_1 \in \cdot \mid X_2^N), Q_m^K(X_1 \in \cdot \mid X_2^K)) = f(X_2^N) \\ - \int_{\mathcal{X}} \log q_m^K(x_1 \mid X_2^K) P_N(dx_1 \mid X_2^N).$$

Therefore equation (1) is equivalent to

$$(2) \quad - E_{P_N} \log q_\pi^N(X_1 \mid X_2^N) \\ \leq \inf_{m \in \mathbb{N}} -E_{P_N} \log q_m^K(X_1 \mid X_2^K) + \frac{1}{N-K+1} \log \frac{1}{\pi(m)}.$$

Now, using the fact that $-\log$ is a convex function, we see that

$$- E_{P_N} \log q_\pi^N(X_1 \mid X_2^N) \\ \leq -\frac{1}{N-K+1} \sum_{M=K}^{N} E_{P_N} \log \frac{\sum_m \pi(m) \prod_{n=K+1}^{M} q_m^K(X_n \mid X_2^K) q_m^K(X_1 \mid X_2^K)}{\sum_m \pi(m) \prod_{n=K+1}^{M} q_m^K(X_n \mid X_2^K)}.$$

We can then use the fact that P_N is exchangeable to swap X_1 and X_{M+1} in the Mth term of the sum. To simplify the notations we will introduce $X_{N+1} = X_1$. We get that

$$\begin{aligned} -E_{P_N} \log q_\pi^N(X_1 \mid X_2^N) &\leq -\frac{1}{N-K+1} \sum_{M=K}^{N} E_{P_N} \log \frac{\sum_m \pi(m) \prod_{n=K+1}^{M+1} q_m^K(X_n \mid X_2^K)}{\sum_m \pi(m) \prod_{n=K+1}^{M} q_m^K(X_n \mid X_2^K)} \\ &= -\frac{1}{N-K+1} E_{P_N} \log \sum_m \pi(m) \prod_{n=K+1}^{N+1} q_m^K(X_n \mid X_2^K) \\ &\leq \frac{1}{N-K+1} \left(\log \frac{1}{\pi(m)} - E_{P_N} \sum_{n=K+1}^{N+1} \log q_m^K(X_n \mid X_2^K) \right). \end{aligned}$$

Eventually, we exchange X_n and X_1 in the right hand side and get that for any $m \in \mathbb{N}$

$$-E_{P_N} \log q_\pi^N(X_1 \mid X_2^N) \leq -E_{P_N} \log q_m^K(X_1 \mid X_2^K) + \frac{1}{N-K+1} \log \frac{1}{\pi(m)}.$$

Taking the infimum in m in the right hand side ends the proof of the theorem. □

3. Dichotomic progressive mixture estimators

The progressive mixture estimator of the previous section has the drawback of requiring lengthy computations, since $N-K+1$ mixtures have to be computed. This number can be reduced to $\log_2(N-K+1)$, if we proceed in a dichotomic way:

Let us assume that $N-K+1=2^r$, and let us define by backward induction the following sequence of probability measures on $\mathbb{N}$:

$$\alpha^r(m)=\pi(m)$$

$$\vdots$$

$$\alpha^k(m)=\frac{1}{2}\left(\frac{\alpha^{k+1}(m)\displaystyle\prod_{n=K+2^k}^{K+2^{k+1}-1}q_m^K(x_n\,|\,x_2^K)}{\displaystyle\sum_{m'\in\mathbb{N}}\alpha^{k+1}(m')\prod_{n=K+2^k}^{K+2^{k+1}-1}q_{m'}^K(x_n\,|\,x_2^K)}+\alpha^{k+1}(m)\right)$$

$$\vdots$$

$$\alpha^0(m)=\frac{1}{2}\left(\frac{\alpha^1(m)q_m^K(x_{K+1}\,|\,x_2^K)}{\displaystyle\sum_{m'\in\mathbb{N}}\alpha^1(m')q_{m'}^K(x_{K+1}\,|\,x_2^K)}+\alpha^1(m)\right)$$

Let us consider the estimator:

$$\begin{cases} q_d^N(x_1\,|\,x_2^N) & =\sum_{m\in\mathbb{N}}\alpha^0(m)q_m(x_1\,|\,x_2^K),\\ \frac{dQ_d^N(\cdot\,|\,x_2^N)}{d\mu} & =q_d^N(\cdot\,|\,x_2^N).\end{cases}$$

It is easy to show by induction that

$$\begin{aligned}(3)\quad &-E_{P_N}\log q_d^N(X_1\,|\,X_2^N)\\ &\qquad\le-\frac{1}{2^r}E_P\log\left(\sum_{m\in\mathbb{N}}\pi(m)\prod_{n=K+1}^{N+1}q_m^K(X_n\,|\,X_2^K)\right)\\ &\qquad\qquad\le\inf_{m\in\mathbb{N}}\left(-E_P\log q_m^K(X_1\,|\,X_2^K)+\frac{1}{2^r}\log\frac{1}{\pi(m)}\right),\end{aligned}$$

and therefore that

Theorem 3.1. *Under the previous assumptions*

$$\begin{aligned}(4)\quad &E_{P_N}H(P_N(X_1\in\cdot\,|\,X_2^N),Q_d^N(X_1\in\cdot\,|\,X_2^N))\\ &\le\inf_{m\in\mathbb{N}}\left\{E_{P_N}H(P_N(X_1\in\cdot\,|\,X_2^N),Q_m^K(X_1\in\cdot\,|\,X_2^K))+\frac{1}{N-K+1}\log\frac{1}{\pi(m)}\right\}.\end{aligned}$$

4. Randomised progressive mixture estimators

Another way to modify the progressive mixture estimator is to randomise the choice of the number of observations used for model selection. Let us consider

for some integer r an i.i.d. sample of integers $(M_i)_{i=1}^r$ drawn with respect to the uniform distribution on $\{K+1, \dots, N\}$. Let us put

$$q_r^N(x_1 \mid x_2^N) = \frac{1}{r} \sum_{i=1}^r \frac{\sum_{m \in \mathbb{N}} \pi(m) \left(\prod_{n=K+1}^{M_i} q_m^K(x_n \mid x_2^K) \right) q_m^K(x_1 \mid x_2^K)}{\sum_{m \in \mathbb{N}} \pi(m) \left(\prod_{n=K+1}^{M_i} q_m^K(x_n \mid x_2^K) \right)}$$

$$\frac{dQ_r^N(x_1 \mid x_2^N)}{d\mu} = q_r^N(x_1 \mid x_2^N).$$

It is easy to see that

Theorem 4.1.

$$(5) \quad E_{P_N} H(P_N(X_1 \in \cdot \mid X_2^N), Q_r^N(X_1 \in \cdot \mid X_2^N))$$
$$\leq \inf_{m \in \mathbb{N}} \left\{ E_{P_N} H(P_N(X_1 \in \cdot \mid X_2^N), Q_m^K(X_1 \in \cdot \mid X_2^K)) + \frac{1}{N-K+1} \log \frac{1}{\pi(m)} \right\}.$$

Various combinations of the randomisation and dichotomic improvements to the progressive mixture estimator can also easily be imagined.

5. Mixture and penalisation: a counter example to their similarity

We give this example to show that a true selection rule, which selects a distribution within one of the models of a family of models, cannot be substituted to the progressive mixture estimator in theorem 2.1. Some supplementary assumptions on the structure of the family of models have to be added.

We consider a sample $(X_1^N) \in \{0,1\}^N$ of binary variables, and two simple models containing only one distribution. For any real number $\lambda \in [0,1]$, let B_λ be the Bernoulli distribution with parameter λ: $B_\lambda = \lambda\delta_1 + (1-\lambda)\delta_0$. Let us consider two models, the first one containing only the distribution $B_{1/4}$ and the second one only the distribution $B_{3/4}$. The best estimators for these two models are obviously the constant estimators

$$Q_0^K(x_1 \mid x_2^K) = Q_0(x_1) = B_{1/4},$$
$$Q_1^K(x_1 \mid x_2^K) = Q_1(x_1) = B_{3/4}.$$

Now let us assume that the true distribution is $B_{1/2-1/\sqrt{N}}$. The central limit theorem shows that for N large enough and some positive constant α independent of N,

$$P_N \left(\frac{1}{N-1} \sum_{i=2}^N X_i > 1/2 \right) \geq \alpha.$$

Therefore the maximum likelihood estimator will choose Q_1 with a probability at least equal to α. Moreover $H(B_{1/2-1/\sqrt{N}}, B_{3/4}) - H(B_{1/2-1/\sqrt{N}}, B_{1/4}) = \dfrac{2\log 3}{\sqrt{N}}$,

therefore, with the notations of theorem 2.1, if $Q_\ell(x_1 \mid x_2^N)$ is the maximum likelihood estimator

$$E_{P_N}(H(P(X_1 \in \cdot), Q_\ell(X_1 \in \cdot \mid X_2^N)) \\ \geq \inf_{m \in \{0,1\}} \left\{ E_{P_N} H(P(X_1 \in \cdot), Q_m^N(X_1 \in \cdot \mid X_2^N)) + \frac{2\alpha \log 3}{\sqrt{N}} \right\}.$$

This is to be compared with theorem 2.1 applied with $\pi(0) = \pi(1) = 1/2$ and $K = 1$, which gives

$$E_{P_N}(H(P(X_1 \in \cdot), Q_\pi^N(X_1 \in \cdot \mid X_2^N)) \\ \leq \inf_{m \in \{0,1\}} \left\{ E_{P_N} H(P(X_1 \in \cdot), Q_m^N(X_1 \in \cdot \mid X_2^N)) + \frac{\log 2}{N} \right\}.$$

The symmetry of the problem shows that any other selection rule between $B_{1/4}$ and $B_{3/4}$ based on the observations X_2^N would not do better than the maximum likelihood estimator.

The reason why this "counter example" works is of course that the map $B_\lambda \mapsto \inf\left\{ H(B_\lambda, B_{1/4}), H(B_\lambda, B_{3/4}) \right\}$ is not differentiable at its extremal point $\lambda = 1/2$, where it has two non zero directional first derivatives. In other words if a true selection rule gave an increase in the risk of order $1/N$, it would be possible to estimate λ with a precision of order $1/N$, which is clearly in contradiction with the central limit theorem.

6. Universal Bayesian estimation of a Bernoulli variable

In this section, we will show that the Bayesian estimate corresponding to an a priori uniform mixture of the parameter is universal. The technique of the proof is borrowed from [6] where a more subtle proof is provided for the Krichevsky-Trofimov estimator. We will use notations analogous to those of section 1, although here the parameter θ will be in the continuous space $[0, 1]$. We consider on $[0, 1]$ the uniform mixture π equal to the Lebesgue measure. We put

$$Q_B^N(x_1^N) = \int_0^1 B_\theta^{\otimes N}(x_1^N) d\theta = \int_0^1 \theta^a (1-\theta)^b d\theta = \frac{a!\, b!}{(a+b+1)!},$$

where $a = \sum_{i=1}^N x_i$ and $b = N - a$. The Bayesian estimator corresponding to π is the well known Laplace estimator

$$Q_B^N(x_1 \mid x_2^N) = \frac{a}{a+b+1}\delta_1(x_1) + \frac{b}{a+b+1}\delta_0(x_1).$$

In order to study its performance, we will use only the fact that the true distribution is assumed to be exchangeable. Let us introduce the random variables

$a = \frac{1}{N}\sum_{k=1}^{N} X_k$ and $b = N - a$. We have

$$\begin{aligned}
-E_{P_N} \log Q_B^N(X_1 \mid X_2^N) &= -\frac{1}{N} E_{P_N} \sum_{k=1}^{N} \log Q_B^N(X_k \mid X_i, i \neq k, 1 \leq i \leq N) \\
&= -E_{P_N}\left(\frac{a}{a+b}\log\frac{a}{a+b+1} + \frac{b}{a+b}\log\frac{b}{a+b+1}\right) \\
&= E_{P_N} \inf_{\theta\in[0,1]} -\frac{1}{N}\log B_\theta^{\otimes N}(X_1^N) + \log\left(1+\frac{1}{a+b}\right) \\
&\leq \inf_{\theta\in[0,1]} E_{P_N} \log\frac{1}{B_\theta(X_1)} + \frac{1}{N}.
\end{aligned}$$

We have proved that

Theorem 6.1. *For any exchangeable distribution P_N on $\{0,1\}^N$,*

$$\begin{aligned}
(6)\quad & E_{P_N} H(P_N(X_1 \in \cdot \mid X_2^N), Q_B^N(X_1 \in \cdot \mid X_2^N)) \\
&\qquad \leq \inf_{\theta\in[0,1]} E_{P_N} H(P_N(X_1 \in \cdot \mid X_2^N), B_\theta(X_1 \in \cdot)) + \frac{1}{N}
\end{aligned}$$

Remark 6.1. If we had used a progressive mixture estimator based on the Krichevsky-Trofimov prior distribution $\pi(d\theta) = \dfrac{d\theta}{\pi\sqrt{\theta(1-\theta)}}$, we would have got an upper bound with a penalty equal to $\frac{1}{2}\frac{\log N}{N} + \frac{1}{N}$. This is an easy consequence of the fact that

$$\log \frac{\left(\frac{a}{a+b}\right)^a \left(\frac{b}{a+b}\right)^b}{\displaystyle\int_\theta \frac{\theta^a(1-\theta)^b d\theta}{\sqrt{\theta(1-\theta)}}} \leq \frac{1}{2}\log(a+b) + 1.$$

(See [6] for a proof.) Therefore in this case of a continuous parameter space the universal upper bound we can prove for the Bayesian estimator is better than for the progressive mixture estimator.

Remark 6.2. The Laplace estimator is well known since a long time, so we do not know whether the result we give here is new or not. The only thing we can say is that our source of inspiration for this proof was [6].

7. Generalisation to random variables taking a finite number of values

Let us consider now the case when $\mathcal{X} = \{0, \cdots, d\}$. Let D_θ, $\theta \in [0,1]^{d+1}$, $\sum \theta_i = 1$ be the distribution $D_\theta(i) = \theta_i$, $i = 0, \ldots, d$. Let $U(d\theta)$ be the uniform (Lebesgue) probability measure on the parameter space $\Theta = \{\theta \in [0,1]^{d+1}, \sum_i \theta_i = 1\}$. We have now for this uniform prior

$$Q_B^N(x_1^N) = \int_\Theta D_\theta^{\otimes N}(x_1^N) U(d\theta) = \int_\Theta \prod_{i=0}^{d} \theta_i^{a_i} U(d\theta) = \frac{a_0! \cdots a_d!}{(a_0 + \cdots + a_d + d)!}$$

and

$$Q_B^N(x_1 \mid x_2^N) = \frac{a_i}{a_0 + \cdots + a_d + d} = \frac{a_i}{N+d},$$

where $a_i = \sum_{k=1}^{N} \mathbf{1}(x_k = i)$. For any exchangeable distribution P_N on $\mathcal{X}^N$ we have

$$\begin{aligned}
-E_{P_N} \log Q_B^N(X_1 \mid X_2^N) &= -E_{P_N} \frac{1}{N} \sum_{k=1}^{N} Q_B^N(X_k \mid X_j, 1 \le j \le N, j \neq k) \\
&= -E_{P_N} \sum_{i=0}^{d} \frac{a_i}{N} \log \frac{a_i}{N+d} \\
&= E_{P_N} \inf_{\theta \in \Theta} -\frac{1}{N} \log D_\theta^{\otimes N}(X_1^N) + \log\left(1 + \frac{d}{N}\right) \\
&\le \inf_{\theta \in \Theta} E_{P_N} \log \frac{1}{D_\theta(X_1)} + \frac{d}{N}.
\end{aligned}$$

Theorem 7.1. *For any exchangeable distribution P_N on $\{0, \dots, d\}^N$,*

$$\begin{aligned}
&E_{P_N} H(P_N(X_1 \in \cdot \mid X_2^N), Q_B^N(X_1 \in \cdot \mid X_2^N)) \\
&\qquad \le \inf_{\theta \in \Theta} E_{P_N} H(P_N(X_1 \in \cdot \mid X_2^N), D_\theta(X_1 \in \cdot)) + \frac{d}{N}.
\end{aligned}$$

8. Estimation by adaptive histograms

We consider as in section 2 a measurable space $(\mathcal{X}, \mathcal{B})$ and a reference probability measure μ on this space. We assume again that the sample distribution P_N on $(\mathcal{X}^N, \mathcal{B}^{\otimes N})$ is exchangeable.

We let $\mathcal{S}$ be a countable family of subdivisions of $\mathcal{X}$. Here we call a subdivision S of $\mathcal{X}$ a partition of $\mathcal{X}$ into a finite number of measurable sets $I \in S$ such that $\mu(I) > 0$. For any $S \in \mathcal{S}$, any parameter $\theta_S \in \mathcal{M}_+^1(S) = \Theta_S$ we consider the measure on $\mathcal{X}$ with density with respect to the reference measure μ

$$\chi(S, \theta_S)(x) = \sum_{I \in S} \frac{\theta_S(I)}{\mu(I)} \mathbf{1}(x \in I).$$

We consider the estimator $Q_S^K(x_1 \mid x_2^K)$ with density

$$\begin{aligned}
\frac{dQ_S^K(x_1 \in \cdot \mid x_2^K)}{d\mu} &= q_S^K(x_1 \mid x_2^K) \\
&= \frac{\displaystyle\int_{\Theta_S} \prod_{n=1}^{K} \chi(S, \theta_S)(x_n) U(d\theta_S)}{\displaystyle\int_{\Theta_S} \prod_{n=2}^{K} \chi(S, \theta_S)(x_n) U(d\theta)}.
\end{aligned}$$

From the previous section, we see that

$$q_S^K(x_1 \mid x_2^K) = \sum_{I \in S} \frac{a(I) + 1}{K + d(S)} \frac{\mathbf{1}(x_1 \in I)}{\mu(I)},$$

where $a(I) = \sum_{n=2}^{K} \mathbf{1}(X_n \in I)$ and $d(S) = |S| - 1$, and that

$$-E_{P_N} \log q_S^K(X_1 \mid X_2^K) \le \inf_{\theta_S \in \Theta_S} E_{P_N} \log \frac{1}{\chi(S, \theta_S)(X_1)} + \frac{d(S)}{K}.$$

Consider now the progressive mixture estimator based on $(q_S^K)_{S\in\mathcal{S}}$ and on some prior probability distribution π on $\mathcal{S}$:

$$q_\pi^N(x_1 \mid x_2^N) = \frac{1}{N-K+1}\sum_{M=K}^{N} \frac{\displaystyle\sum_{S\in\mathcal{S}} \pi(S) \prod_{n=K+1}^{M} q_S^K(x_n \mid x_2^K) q_S^K(x_1 \mid x_2^K)}{\displaystyle\sum_{S\in\mathcal{S}} \pi(S) \prod_{n=K+1}^{M} q_S^K(x_n \mid x_2^K)}.$$

We have proved in section 2 that

$$-E_{P_N} \log q_\pi^N(X_1 \mid X_2^N) \leq \inf_{S\in\mathcal{S}} \left\{ \inf_{\theta_S\in\Theta_S} E_{P_N} \log \frac{1}{\chi(S,\theta_S)(X_1)} + \frac{d(S)}{K} + \frac{1}{N-K+1}\log\frac{1}{\pi(S)} \right\}$$

If we want to balance the two penalty terms, we can take $\pi(S) = \frac{1}{Z(\beta)} e^{-\beta\, d(S)}$. This gives a penalty of

$$\frac{d(S)}{K} + \frac{\beta\, d(S)}{N-K+1} + \frac{\log Z(\beta)}{N-K+1}.$$

Proposition 8.1. *If we take* $\beta = \dfrac{N-K+1}{K}$ *we get*

$$-E_{P_N} \log q_\pi^N(X_1 \mid X_2^N) \leq \inf_{S\in\mathcal{S}} \left\{ \inf_{\theta_S\in\Theta_S} E_{P_N} \log \frac{1}{\chi(S,\theta_S)(X_1)} + (1+\beta)\frac{2d(S) + \dfrac{\log Z(\beta)}{\beta}}{N+1} \right\}.$$

Remark 8.1. The same result is of course also true for the dichotomic and the randomised progressive mixture estimators.

9. Binary trees of histograms

We will restrict ourselves here to families of dichotomic subdivisions of $\mathcal{X}$, leading to fast algorithms. In these cases, some progressive mixture estimators can be computed efficiently by adapting the context tree weighting algorithm of Willems, Shtarkov and Tjalkens [6, 7].

We will not exactly base the estimation on $q_S^K(x_1 \mid x_2^K)$, but on the approximation

$$\begin{aligned}
\tilde q_S^K(x_1 \mid x_2^K) &= \sum_{I\in S} \frac{(a(I)+1)e^{-\frac{|S|}{K-1}}}{(K-1)\mu(I)} \mathbf{1}(x_1 \in I) \\
&= \left(1 + \frac{|S|}{K-1}\right) e^{-\frac{|S|}{K-1}} q_S^K(x_1 \mid x_2^K) \\
&\leq q_S^K(x_1 \mid x_2^K).
\end{aligned}$$

The advantage of this approximation will be seen in the computation of the mixtures.

In a first step we have to explain how to use the $\tilde q_S^K$ instead of the q_S^K and to prove that we do not loose much by doing this. Let us first define the randomised

progressive mixture based on $\tilde{q}_S^K$:

$$\tilde{q}_\pi^N(x_1 \mid x_2^N) = \frac{1}{r}\sum_{i=1}^{r} \frac{\sum_{S\in\mathcal{S}} \pi(S) \prod_{n=K+1}^{M_i} \tilde{q}_S^K(x_n \mid x_2^K)\tilde{q}_S^K(x_1 \mid x_2^K)}{\sum_{S\in\mathcal{S}} \pi(S) \prod_{n=K+1}^{M_i} \tilde{q}_S^K(x_n \mid x_2^K)},$$

where $(M_i)_{1\leq i\leq r}$ is an i.i.d sample from the uniform distribution on $\{K+1,\ldots,N\}$, independent of everything else. To get a statistical estimator, we have to normalise $\tilde{q}_\pi^N$. Therefore our estimator will be

$$\bar{q}_\pi^N(x_1 \mid x_2^N) = \tilde{q}_\pi^N(x_1 \mid x_2^N)\left(\int_{x\in\mathcal{X}} \tilde{q}_\pi^N(x \mid x_2^N)\mu(dx)\right)^{-1}.$$

As for any $S \in \mathcal{S}$

$$\tilde{q}_S^K(x_1 \mid x_2^K) \leq q_S^K(x_1 \mid x_2^K),$$

we have that

$$\int_{x\in\mathcal{X}} \tilde{q}_\pi^N(x \mid x_2^N)\mu(dx) \leq 1,$$

therefore

$$E - \log \bar{q}_\pi^N(X_1 \mid X_2^N) \leq E - \log \tilde{q}_\pi^N(X_1 \mid X_2^N).$$

Replacing q_S^K by $\tilde{q}_S^K$ in the proof of section 2, we see moreover that

$$\begin{aligned}
-E\log \tilde{q}_\pi^N(X_1 \mid X_2^N) &\leq \inf_{S\in\mathcal{S}} -E\log \tilde{q}_S^K(X_1 \mid X_2^K) + \frac{1}{N-K+1}\log\frac{1}{\pi(S)} \\
&= \inf_{S\in\mathcal{S}} -E\log q_S^K(X_1 \mid X_2^K) \\
&\quad + \frac{d(S)+1}{K-1} - \log\left(1+\frac{d(S)+1}{K-1}\right) + \frac{1}{N-K+1}\log\frac{1}{\pi(S)} \\
&\leq \inf_{S\in\mathcal{S}} \inf_{\theta_S\in\Theta_S} E\log\frac{1}{\chi(S,\theta_S)(X_1)} + \frac{d(S)}{K} \\
&\quad + \frac{1}{2}\left(\frac{d(S)+1}{K-1}\right)^2 + \frac{1}{N-K+1}\log\frac{1}{\pi(S)}.
\end{aligned}$$

Let us now describe the family $\mathcal{S}$ and the prior distribution π. Let us assume for this purpose that we have divided $\mathcal{X}$ into a binary tree of nested cells $\{I_s : s \in \{0,1\}^*\}$, where we use the notation $\{0,1\}^* = \bigcup_{i=0}^{+\infty}\{0,1\}^i$. The cells are thus indexed by the set of finite binary strings, which we will call words in the future. We assume that

- $I = \mathcal{X}$, where $\emptyset$ is the empty string,
- for any word s, $I_{s0} \cup I_{s1} = I_s$ and $I_{s0} \cap I_{s1} = \emptyset$.

For example, in the case when $\mathcal{X} = [0,1[$ is the unit interval, we can choose

$$I_s = \left[\sum_{k=1}^{l(s)} s(k)2^{-k},\ 2^{-l(s)} + \sum_{k=1}^{l(s)} s(k)2^{-k}\right[,$$

where $l(s)$ is the length of the string s.

We recall that a complete prefix code $S \in \{0,1\}^*$ is a finite set of finite binary strings such that:

- if s,t are in S, then for some $k \leq (l(s) \wedge l(t))$ we have $s(k) \neq t(k)$ (no word in S is the prefix of another word).
- S is maximal for inclusion.

A complete prefix code can also be seen as a complete binary tree. To see this, it suffices to arrange $\{0,1\}^*$ into a binary tree, by considering that the two sons of s are $s0$ and $s1$. Then a prefix code is made of the leaves of a finite subtree of the infinite tree $\{0,1\}^*$ and a complete prefix code is made of the leaves of a finite complete subtree of $\{0,1\}^*$ (one in which any interior node has two sons).

It is easy to see that for each complete prefix code S, $\{I_s,\ s \in S\}$ is a subdivision of $\mathcal{X}$. We will write that $S_1 \leq S_2$ if each word in S_2 has a prefix in S_1, or equivalently if the subdivision corresponding to S_2 is a refinement of the subdivision corresponding to S_1.

Now let $\mathcal{C}$ be the set of all complete prefix codes and consider for some $\bar{S} \subset \mathcal{C}$ the family of subdivisions

$$\mathcal{S} = \{S \in \mathcal{C} \mid S \leq \bar{S}\}.$$

For some survival probability $\rho \in]0,1[$, we consider the a priori distribution on $\mathcal{S}$

$$\pi(S) = \rho^{|S|-1}(1-\rho)^{|S|-|S\cap\bar{S}|}, \quad S \in \mathcal{S}.$$

This is the distribution of the genealogy tree of a branching process where each particle either gives birth to two sons with probability ρ on the next generation, or dies without heirs with probability $1-\rho$, except when it is located on $\bar{S}$ in the genealogy, in which case it dies without posterity with probability one.

We have proved the following theorem:

Theorem 9.1. *With the previous assumptions and notations, the modified randomised progressive mixture estimator $\bar{q}_\pi^N$ satisfies*

$$-E\log\bar{q}_\pi^N(X_1 \mid X_2^N)$$
$$\leq \inf_{S\in\mathcal{S}}\inf_{\theta_S\in\Theta_S} E\log\frac{1}{\chi(S,\theta_S)(X_1)} + \frac{d(S)}{K} + \frac{1}{2}\left(\frac{d(S)+1}{K-1}\right)^2 + \frac{d(S)+1}{N-K+1}\log\frac{1}{\rho(1-\rho)}.$$

We are now going to see how to compute $\bar{q}_\pi^N(x_1 \mid x_2^N)$. We have to compute $2r$ mixtures of the type

$$\begin{aligned} w_M(\mathcal{S}) &= \sum_{S\in\mathcal{S}} \pi(S) \prod_{n=K+1}^{M} \tilde{q}_S^K(x_n \mid x_2^K) \\ &= \sum_{S\in\mathcal{S}} \pi(S) e^{\frac{M-K}{K-1}|S|} \prod_{I\in S}\left(\frac{a(I)+1}{(K-1)\mu(I)}\right)^{b(I)}, \end{aligned}$$

where

$$a(I) = \sum_{k=2}^{K} \mathbf{1}(x_k \in I),$$
$$b(I) = \sum_{k=K+1}^{M} \mathbf{1}(x_k \in I).$$

We see now the advantage of using $\tilde{q}_S^K$: in the product $\prod_{I\in S}$, the factors are functions of I only (and not of I and S). Thus it will be possible to factorise the computation of $w(\mathcal{S})$.

Let $T(\bar{S})$ be the tree corresponding to $\bar{S}$, i.e the set of all the prefixes of the words of $\bar{S}$, i.e. the finite sigma algebra generated by the subdivision $\bar{S}$. On $T(\bar{S})$ we define the backward recursion

$$w(s) = \begin{cases} \left(\frac{a(I_s)+1}{(K-1)\mu(I_s)}\right)^{b(I_s)} e^{-(M-K)/(K-1)}, & \text{if } s \in \bar{S}, \\ (1-\rho)\left(\frac{a(I_s)+1}{(K-1)\mu(I_s)}\right)^{b(I_s)} e^{-(M-K)/(K-1)} + \rho w(s0)w(s1), & \text{if } s \notin \bar{S}. \end{cases}$$

Then it is easy to check that

Proposition 9.1.

$$w(\emptyset) = w(\mathcal{S}),$$

where $\emptyset$ stands for the empty string.

Hence to compute $w(\mathcal{S})$, we need to perform a number of operations of order $|T(\bar{S})| = 2|\bar{S}| - 1$. This is to be compared with the number of models $|\mathcal{S}|$. It is given by the following backward recursion on $T(\bar{S})$:

$$\begin{cases} n(s) = \begin{cases} 1, & \text{if } s \in \bar{S}, \\ 1 + n(s0)n(s1), & \text{if } s \notin \bar{S}, \end{cases} \\ n(\emptyset) = |\mathcal{S}|. \end{cases}$$

In the case when $\bar{S}$ is the complete tree of depth δ (for example when $\mathcal{X} = [0,1[$ and the cells are the dyadic intervals, the complete tree of depth δ defines the uniform subdivision of $[0,1[$ of step size $2^{-\delta}$), we see easily that $n(s) = f(l(s))$, where $f(\delta) = 1$ and $f(k-1) = f(k)^2 + 1$. Therefore we have $f(k-1) \geq f(k)^2$ and $f(k-1) + \alpha \leq (f(k) + \alpha)^2$ with $\alpha = \frac{1}{2}(1 + \sqrt{5})$. Thus in this case

$$2^{2^{\delta-1}} \leq |\mathcal{S}| \leq (2+\alpha)^{2^{\delta-1}} - \alpha.$$

In conclusion, using the modified randomised progressive mixture estimator $\bar{q}_\pi^N$, we perform almost as well as the best of more than $2^{2^{\delta-1}}$ models in a time of order $\delta(r2^\delta + N)$ where r is the number of trials for the randomised length of the second subsample. The justification of this complexity estimation is the following: when you add or you modify either an observation x_i, $K < i \leq N$ or the point x_1 where you want to compute $\bar{q}(x_1 \mid x_2^N)$, you have to update the computations for $b(I_s)$ and $w(s)$ in δ nested cells. Now you will use at most $N-K+1 \sim N$ observations, which you will introduce progressively in the computations, and you have to consider the case when x_1 falls in each of the 2^δ finer cells, and this for the r random lengths of the second subsample, this accounts for the $r2^\delta$ factor. If the cells are "well balanced" a reasonable choice is to take 2^δ of order N, and if moreover r is kept bounded, this results in a complexity in N of order $N \log N$.

Remark 9.1. All this section could be generalised to more general trees. More flexibility in the choice of π is also possible. We refer for comparison to the generalisations of the context tree weighting method in [7].

10. Model selection from a dense family of models

Using the same notations as in section 2, let us assume now that $Q_m^N(\cdot \mid X_2^N) \in \mathcal{M}_m$ (a set of probability distributions forming a "model" indexed by m). Let us assume that P_N is a product measure, $P_N = P^{\otimes N}$. This we do for simplicity, straightforward generalisations to the case when P is only exchangeable are left to the reader. We assume that

- $\inf_{m,\mu_m\in\mathcal{M}_m} H(P,\mu_m)=0$. Thus we assume that the sample distribution P is in the closure of the family of models $(\mathcal{M}_m)_{m\in\mathbb{N}}$ with respect to the Kullback distance. (In other words the family of models is assumed to be dense in the set of all possible sample distributions.)
- $E(H(P,Q_m^N(.\mid X_2^N)))\leq \inf_{m\in\mathbb{N}}\inf_{\mu_m\in\mathcal{M}_m} H(P,\mu_m)+\dfrac{c(m)}{N}$, for some family $c(m)$ of positive weights. We make this assumption to give a more suggestive result. It is easily seen that $\dfrac{c(m)}{N}$ could be replaced by an arbitrary function $c(m,N)$ throughout this section.

Then we have seen that

$$E(H(P,Q_\pi^N(X_1\in\cdot\mid X_2^N))\leq \inf_m\inf_{\mu_m\in\mathcal{M}_m} H(P,\mu_m)+\frac{c(m)}{K}+\frac{1}{N-K+1}\log\frac{1}{\pi(m)}.$$

Let h be the Hellinger distance between probability measures,

$$h(\mu,\nu)=\sqrt{\frac{1}{2}\int(\sqrt{d\mu}-\sqrt{d\nu})^2}.$$

It is well known that $2h^2(\mu,\nu)\leq H(\mu,\nu)$ and that, h being a "true" distance, $h^2(\mu,\nu)\leq 2(h^2(\mu,\xi)+h^2(\xi,\nu))$.

Let us select $\hat m(X_2^N)$ and $\hat\mu_{\hat m}(X_1\in\cdot\mid X_2^N)\in\mathcal{M}_m$ by minimising

$$\inf_{m,\mu_m\in\mathcal{M}_m} h^2(Q_\pi^N(\cdot\mid X_2^N),\mu_m)+\frac{\gamma(m)}{N},$$

where $\gamma(m)$ is a family of positive weights (here again we could use an arbitrary positive function $\gamma(m,N)$). We let also the reader figure out by himself the obvious modifications required in the case when the infimum is not reached in the previous equation.

We have that for any $\mu_m\in\mathcal{M}_m$

$$\begin{aligned}
&E_{P_N}\left(h^2(P,\hat\mu_{\hat m}(X_1\in\cdot\mid X_2^N))+2\frac{\gamma(\hat m(X_2^N))}{N}\right)\\
&\quad\leq 2Eh^2(P,Q_\pi^N(X_1\in\cdot\mid X_2^N))\\
&\qquad+2Eh^2(Q_\pi^N(X_1\in\cdot\mid X_2^N),\hat\mu_{\hat m}(X_1\in\cdot\mid X_2^N))+2\frac{\gamma(\hat m(X_2^N))}{N}\\
&\quad\leq 2Eh^2(P,Q_\pi^N(X_1\in\cdot\mid X_2^N))+2Eh^2(Q_\pi^N(X_1\in\cdot\mid X_2^N),\mu_m)+2\frac{\gamma(m)}{N}\\
&\quad\leq 6Eh^2(P,Q_\pi^N(X_1\in\cdot\mid X_2^N))+4h^2(P,\mu_m)+2\frac{\gamma(m)}{N}\\
&\quad\leq 3EH(P,Q_\pi^N(X_1\in\cdot\mid X_2^N))+4h^2(P,\mu_m)+2\frac{\gamma(m)}{N}\\
&\quad\leq 5H(P,\mu_m)+3\frac{c(m)}{K}+\frac{3}{N-K+1}\log\frac{1}{\pi(m)}+2\frac{\gamma(m)}{N}
\end{aligned}$$

We have proved that

Theorem 10.1. *The rate of approximation of P by the adaptive estimator $\hat\mu_{\hat m}$ according to the mean square of the Hellinger distance is upper bounded by*

$$E\left(h^2(P,\hat\mu_{\hat m})\right)\leq \inf_{m\in\mathbb{N},\mu_m\in\mathcal{M}_m} 5H(P,\mu_m)+3\frac{c(m)}{K}+\frac{3}{N-K+1}\log\frac{1}{\pi(m)}+2\frac{\gamma(m)}{N}.$$

Remark 10.1. This theorem shows that in many cases the rate of decrease of $E(h^2(P,\hat{\mu}_{\tilde{m}}))$ will be optimal. This will be the case when $H(P,\mu_m)$ is of the same order as $h^2(P,\mu_m)$ where the infimum in m is reached, when $K/(N-K)$ is chosen of order 1, when $\gamma(m)$ is chosen of order at most $c(m)$ and when $\sum_m e^{-c(m)} < +\infty$, so that it is possible to choose $\log\frac{1}{\pi(m)}$ of order $c(m)$.

Remark 10.2. A choice of $\gamma(m) = 0$ is possible, but in practice, we will like to select a model as simple as possible, and therefore we will prefer a choice of $\gamma(m)$ increasing with the dimension of the model as fast as possible. Usually this dimension is related to $c(m)$, and we will choose $\gamma(m)$ of order $c(m)$.

11. Conclusion

The progressive mixture estimator is interesting from the theoretical point of view, since it proves that it is always possible to perform almost as well as the best of a countable family of estimators with a loss of performance only due to the fact of splitting the sample in two and due to the presence of an extra term $\frac{1}{N-K+1}\log\frac{1}{\pi(m)}$, where π is an arbitrary probability distribution. It is remarkable that this could be achieved without any special assumption on the estimators or on the true sample distribution.

It also can be a fast practical algorithm in the case when the computation of the mixtures factorises well. We have treated here the case of adaptive histograms. This may suggest already a variety of applications, since we worked with a broad definition of histograms. We are planning to give further applications and generalisations of the mixture approach to regression and classification problems in the near future.

References

[1] A. Barron, L. Birgé and P. Massart, Risk bounds for model selection via penalisation, *preprint*, Equipe de Modélisation Stochastique et Statistique, Université Paris-Sud Orsay, 1995.

[2] L. Birgé and P. Massart, From model selection to adaptive estimation, *preprint*, Equipe de Modélisation Stochastique et Statistique, Université Paris-Sud Orsay, 1995.

[3] L. Birgé and P. Massart, Minimum contrast estimators on sieves, *preprint*, Equipe de Modélisation Stochastique et Statistique, Université Paris-Sud Orsay, 1995.

[4] M. Feder and N. Merhav, Hierarchical Universal Coding, *IEEE Trans. Inform. Theory*, vol 42, no 5, Sept, 1996.

[5] B. Y. Ryabko, Twice-universal coding, *Probl. Inform. Transm.*, vol 20, no 3, pp. 24-28, July-Sept 1984.

[6] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, The Context-Tree Weighting Method: Basic Properties, *IEEE Trans. Inform. Theory*, vol 41, no 3, May, 1995.

[7] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens, Context Weighting for General Finite-Context Sources, *IEEE Trans. Inform. Theory*, vol 42, no 5, Sept, 1996.

D.I.A.M. - Intelligence Artificielle et Mathématiques, Laboratoire de Mathématiques de l'Ecole Normale Supérieure, U.A. 762 du C.N.R.S., 45 rue d'Ulm, 75 005 Paris, France