ANNALES DE LA FACULTÉ DES SCIENCES DE TOULOUSE

PASCAL MASSART

Some applications of concentration inequalities to statistics

Annales de la faculté des sciences de Toulouse 6^e série, tome 9, n° 2 (2000), p. 245-303

<http://www.numdam.org/item?id=AFST_2000_6_9_2_245_0>

Annales de la Faculté des Sciences de Toulouse

Vol. IX, n° 2, 2000
pp. 245–303

Some Applications of Concentration Inequalities to Statistics

PASCAL MASSART [1]

RÉSUMÉ. — Nous présentons quelques applications d'inégalités de concentration à la résolution de problèmes de sélection de modèles en statistique. Nous étudions en détail deux exemples pour lesquels cette approche s'avère particulirement fructueuse. Nous considérons tout d'abord le classique mais délicat problème du choix d'un bon histogramme. Nous présentons un extrait de travail de Castellan sur la question, mettant en évidence que la structure même des inégalités de concentration de Talagrand pour des processus empiriques influence directement la construction d'un critère de sélection de type Akaike modifié. Nous présentons également un nouveau théorème de sélection de modèles bien adapté à la résolution de problèmes d'apprentissage. Ce résultat permet de réinterpréter et d'améliorer la méthode dite de *minimisation structurelle du risque* due à Vapnik.

ABSTRACT. — The purpose of this paper is to illustrate the power of concentration inequalities by presenting some striking applications to various model selection problems in statistics. We shall study in details two main examples. We shall consider the old-standing problem of optimal selection of an histogram (following the lines of Castellan's work on this topic) for which the structure of Talagrand's concentration inequalities for empirical processes directly influences the construction of a modified Akaike criterion. We shall also present a new model selection theorem which can be applied to improve on Vapnik's *structural minimization of the risk method* for the statistical learning problem of pattern recognition.

[1] Equipe de "Probabilités, Statistique et Modélisation", Laboratoire de Mathématique UMR 8628, Bât. 425, Centre d'Orsay, Université de Paris-Sud 91405 Orsay Cedex. E-mail: Pascal.Massart@math.u-psud.fr

1. Introduction

Since the last ten years, the phenomenon of the concentration of measure has received much attention mainly due to the remarkable series of works by Michel Talagrand which led to a variety of new powerfull inequalities (see in particular [37] and [39]). Our purpose in this paper is to explain why concentration inequalities for functionals of independent variables are important in statistics. We shall also present some applications to random combinatorics which in turn can lead to new results in statistics. Our choice here is to present some selected applications in details rather than provide a more or less exhaustive review of applications. Some of these results are borrowed from very recent papers (mainly from Castellan [23], Birgé and Massart [12] and Boucheron, Lugosi and Massart [16]) and some others are new (see Section 4).

Since the seminal works of Dudley in the seventies, the theory of probability in Banach spaces has deeply influenced the development of asymptotic statistics, the main tools involved in these applications being limit theorems for empirical processes. This led to decisive advances for the theory of asymptotic efficiency in semi-parametric models for instance and the interested reader will find numerous results in this direction in the books by van der Vaart and Wellner [43] or van der Vaart [42]. The main interesting feature of concentration inequalities is that, unlike central limit theorems or large deviations inequalities, they are *nonasymptotic*. We believe that the introduction of these new tools is an important step towards the construction of a non asymptotic theory in statistics. By non asymptotic, we do not mean that large samples of observations are not welcome but that, for instance, it is of great importance to allow the number of parameters of a parametric model to depend on the sample size in order to be able to warrant that the statistical model is not far from the truth. Let us now introduce a framework where this idea can be developed in great generality.

1.1. Introduction to model selection

Suppose that one observes independent variables $\xi_1, ..., \xi_n$ taking their values in some measurable space Ξ. Let us furthermore assume, for the sake of simplicity, that these variables are identically distributed with common distribution P depending on some unknown "parameter" $s \in \mathcal{S}$ (note that the results we are presenting here are non asymptotic and remain valid even in the non stationary case, where P is the arithmetic mean of the distributions of the variables $\xi_1, ..., \xi_n$ and therefore may depend on n). The aim is to estimate s by using as few prior information as possible. One can typically think of s as a function belonging to some infinite dimensional

space $\mathcal{S}$. The two main examples that we have in mind are respectively the density and the regression frameworks. More precisely:

- In the density framework, $P = s\mu$, s is some unknown non negative function and $\mathcal{S}$ can be taken as the set of probability densities with respect to μ.
- In the regression framework, the variables $\xi_i = (X_i, Y_i)$ are independent copies of a pair of random variables (X, Y), where X takes its values in some measurable space $\mathcal{X}$. Assuming the variable Y to be square integrable, one defines the regression function s as $s(x) = \mathbb{E}[Y \mid X = x]$ for every $x \in \mathcal{X}$. Denoting by μ the distribution of X, one can set $\mathcal{S} = \mathbb{L}^2(\mu)$.

One of the most common used method to estimate s is minimum contrast estimation.

1.1.1. *Minimum contrast estimation*

Let us consider some *contrast* function γ, defined on $\mathcal{S} \times \Xi$, which means that

$$l(s,t) = \mathbb{E}[\gamma(t,\xi_1) - \gamma(s,\xi_1)] \geqslant 0, \text{ forall } t \in \mathcal{S}. \tag{1}$$

In other words, the functional $t \to \mathbb{E}[\gamma(t,\xi_1)]$ achieves a minimum at point s. The heuristics of minimum contrast estimation is that, if one substitutes the empirical criterion

$$\gamma_n(t) = P_n[\gamma(t,.)] = \frac{1}{n}\sum_{i=1}^{n}\gamma(t,\xi_i),$$

to its expectation $P[\gamma(t,.)] = \mathbb{E}[\gamma(t,\xi_1)]$ and minimizes the empirical criterion γ_n on some subset S of $\mathcal{S}$ (that we call a *model*), there is some hope to get a sensible estimator of s, at least if s belongs (or is close enough) to model S. This estimation method is widely used and has been extensively studied in the asymptotic parametric setting for which one assumes that S is a given parametric model, s belongs to S and n is large. Probably, the most popular examples are maximum likelihood and least squares estimation. For the density and the regression frameworks, the corresponding contrast functions can be defined as follows.

- Assume that one observes independent and identically distributed random variables $(\xi_1, ...\xi_n)$ with common distribution $s\mu$. Then the contrast function leading to maximum likelihood estimation is simply

$$\gamma(t,x) = -\log(t(x))$$

for every density t and the corresponding loss function l is given by

$$l(s,t) = K(s,t),$$

where $K(s,t)$ denotes the Kullback-Leibler information number between the probabilities $s\mu$ and $t\mu$, i.e.

$$K(s,t) = \int s \log\left(\frac{s}{t}\right)$$

if $s\mu$ is absolutely continuous with respect to $t\mu$ and $K(s,t) = +\infty$ otherwise.

- Assume that one observes independent and identically distributed copies of a pair (X,Y), the distribution of X being denoted by μ. Then the contrast function leading to least squares estimation is defined for every $t \in \mathbb{L}_2(\mu)$ by

$$\gamma(t,(x,y)) = (y - t(x))^2,$$

and the loss function l is given by

$$l(s,t) = \|s - t\|^2,$$

where $\|.\|$ denotes the norm in $\mathbb{L}_2(\mu)$.

The main problem which arises from minimum contrast estimation in a parametric setting is the choice of a proper model S on which the minimum contrast estimator is to be defined. In other words, it may be difficult to guess what is the right parametric model to consider in order to reflect the nature of data from the real life and one can get into problems whenever the model S is false in the sense that the true s is too far from S. One could then be tempted to choose S as big as possible. Taking S as $\mathcal{S}$ itself or as a "huge" subset of $\mathcal{S}$ is known to lead to inconsistent (see [5]) or suboptimal estimators (see [8]). We see that choosing some model S in advance leads to some difficulties

- If S is a "small" model (think of some parametric model, defined by 1 or 2 parameters for instance) the behavior of a minimum contrast estimator on S is satisfactory as long as s is close enough to S but the model can easily turn to be false.
- On the contrary, if S is a "huge" model (think of the set of all continuous functions on $[0,1]$ in the regression framework for instance), the minimization of the empirical criterion leads to a very poor estimator of s.

It is therefore interesting to consider a family of models instead of a single one and try to select some appropriate model from the data in order to estimate s by minimizing some empirical criterion on the selected model. The idea of estimating s by *model selection via penalization* has been extensively developed in [10] and [7].

1.1.2. *Model selection via penalization*

Let us describe the method. Let us consider some countable or finite (but possibly depending on n) collection of models $(S_m)_{m\in\mathcal{M}_n}$ and the corresponding collection of minimum contrast estimators $(\widehat{s}_m)_{m\in\mathcal{M}_n}$. Ideally, one would like to consider $m(s)$ minimizing the risk $\mathbb{E}[l(s,\widehat{s}_m)]$ with respect to $m\in\mathcal{M}_n$. The minimum contrast estimator $\widehat{s}_{m(s)}$ on the corresponding model $S_{m(s)}$ is called an *oracle* (according to the terminology introduced by Donoho and Johnstone, see [21] for instance). Unfortunately, since the risk depends on the unknown parameter s, so does $m(s)$ and the oracle is not an estimator of s. However, the risk of an oracle is a benchmark which will be useful in order to evaluate the performance of any data driven selection procedure among the collection of estimators $(\widehat{s}_m)_{m\in\mathcal{M}_n}$. The purpose is therefore to define a data driven selection procedure within this family of estimators, which tends to mimic an oracle, i.e. one would like the risk of the selected estimator $\widetilde{s}$ to be as close as possible to the risk of an oracle. The trouble is that it is not that easy to compute explicitly the risk of a minimum contrast estimator and therefore of an oracle (except for special cases of interest such as least square estimators on linear models for example). This is the reason why we shall rather compare the risk of $\widetilde{s}$ to an upper bound for the risk of an oracle. In a number of cases (see [9]), a good upper bound (up to constant) for the risk of $\widehat{s}_m$ has the following form

$$l(s,S_m)+\frac{D_m}{n} \tag{2}$$

where D_m is a measure of the "size" of the model (something like the number of parameters defining the model S_m) and $l(s,S_m)=\inf_{t\in S_m} l(s,t)$. This means that, at a more intuitive level, an oracle is making a good compromise between the "size" of the model (that one would like to keep as small as possible) and the fidelity to the data. The model selection via penalization method can be described as follows. One considers some function $\mathrm{pen}_n : \mathcal{M}_n \to \mathbb{R}_+$ which is called the *penalty function*. Note that pen_n can possibly depend on the observations $\xi_1,...,\xi_n$ but not of course on s. Then, for every $m\in\mathcal{M}_n$, one considers the minimum contrast estimator within model S_m

$$\widehat{s}_m=\operatorname*{argmin}_{t\in S_m}\gamma_n(t).$$

Selecting $\widehat{m}$ as a minimizer of

$$\gamma_n\left(\widehat{s}_m\right) + \mathrm{pen}_n\left(m\right) \tag{3}$$

over $\mathcal{M}_n$, one finally estimates s by the *minimum penalized contrast estimator*

$$\widetilde{s} = \widehat{s}_{\widehat{m}}.$$

Since some problems can occur with the existence of a solution to the preceding minimization problems, it is useful to consider approximate solutions (note that even if $\widehat{s}_m$ does exist, it is relevant from a practical point of view to consider approximate solutions since $\widehat{s}_m$ will typically be approximated by some numerical algorithm). Therefore, given $\rho_n \geqslant 0$ (in practice, taking $\rho_n = n^{-2}$ makes the introduction of an approximate solution painless), we shall consider for every $m \in \mathcal{M}_n$ some approximate minimum contrast estimator $\widehat{s}_m$ satisfying

$$\gamma_n\left(\widehat{s}_m\right) \leqslant \gamma_n\left(t\right) + \rho_n/2$$

and say that $\widetilde{s}$ is a ρ_n-*minimum penalized contrast estimator* of s if

$$\gamma_n\left(\widetilde{s}\right) + \mathrm{pen}_n\left(\widehat{m}\right) \leqslant \gamma_n\left(t\right) + \mathrm{pen}_n\left(m\right) + \rho_n, \forall m \in \mathcal{M}_n \text{ and } \forall t \in S_m. \tag{4}$$

The reader who is not familiar with model selection via penalization can legitimately ask the question: where does the idea of penalization come from? It is possible to answer this question at two different levels:

- at some intuitive level by presenting the heuristics of one of the first criterion of this kind which has been introduced by Akaike (1973);
- at some technical level by explaining why such a strategy of model selection has some chances to succeed.

We shall now develop these two points. We first present the heuristics of Akaike's criterion on the case example of histogram selection which will be studied in details in Section 3, following the lines of [23].

1.2. A case example and Akaike's criterion

We consider here the density framework where one observes n independent and identically distributed random variables with common density s with respect to the Lebesgue measure on $[0,1]$. Let $\mathcal{M}_n$ be some finite (but possibly depending on n) collection of partitions of $[0,1]$ into intervals. For any partition m, we consider the corresponding histogram estimator $\widehat{s}_m$ defined by

$$\widehat{s}_m = \sum_{I \in m} \left(n\mu\left(I\right)\right)^{-1} \left[\sum_{i=1}^{n} \mathbb{1}_I\left(\xi_i\right)\right] \mathbb{1}_I,$$

where μ denotes the Lebesgue measure on $[0,1]$, and the purpose is to select "the best one". Recall that The histogram estimator on some partition m is known to be the maximum likelihood estimator on the model S_m of densities which are piecewise constants on the corresponding partition m and therefore falls into our analysis. Then the natural loss function to be considered is the Kullback-Leibler loss and in order to understand the construction of Akaike's criterion, it is essential to describe the behavior of an oracle and therefore to analyze the Kullback-Leibler risk. First it easy to see that the Kullback-Leibler projection s_m of s on the histogram model S_m (i.e. the minimizer of $t \to K(s,t)$ on S_m) is simply given by the orthogonal projection of s on the linear space of piecewise constant functions on the partition m and that the following Pythagore's type decomposition holds

$$K(s,\widehat{s}_m) = K(s,s_m) + K(s_m,\widehat{s}_m). \tag{5}$$

Hence the oracle should minimize $K(s,s_m)+\mathbb{E}[K(s_m,\widehat{s}_m)]$ or equivalently, since $s - s_m$ is orthogonal to $\log(s_m)$,

$$K(s,s_m) + \mathbb{E}[K(s_m,\widehat{s}_m)] - \int s\log(s) \quad = \quad -\int s_m \log(s_m) \\ +\mathbb{E}[K(s_m,\widehat{s}_m)].$$

Since $\int s_m \log s_m$ depends on s, it has to be estimated. One could think of $\int \widehat{s}_m \log \widehat{s}_m$ as being a good candidate for this purpose but since $\mathbb{E}[\widehat{s}_m] = s_m$, the following identity holds

$$\mathbb{E}\left[\int \widehat{s}_m \log(\widehat{s}_m)\right] = \mathbb{E}[K(\widehat{s}_m,s_m)] + \int s_m \log(s_m)\ ,$$

which shows that it is necessary to remove the bias of $\int \widehat{s}_m \log \widehat{s}_m$ if one wants to use it as an estimator of $\int s_m \log s_m$. In order to summarize the preceding analysis of the oracle procedure (with respect to the Kullback-Leibler loss), let us set

$$R_m = \int \widehat{s}_m \log(\widehat{s}_m) - \mathbb{E}\left[\int \widehat{s}_m \log(\widehat{s}_m)\right].$$

Then the oracle minimizes

$$-\int \widehat{s}_m \log(\widehat{s}_m) + \mathbb{E}[K(s_m,\widehat{s}_m)] + \mathbb{E}[K(\widehat{s}_m,s_m)] + R_m. \tag{6}$$

The idea underlying Akaike's criterion relies on two heuristics:

- neglecting the remainder term R_m (which is centered at its expectation),

- replacing $\mathbb{E}\left[K\left(s_m, \widehat{s}_m\right)\right] + \mathbb{E}\left[K\left(\widehat{s}_m, s_m\right)\right]$ by its asymptotic equivalent when n goes to infinity which is equal to D_m/n, where $1+D_m$ denotes the number of pieces of the partition m (see [23] for a proof of this result).

Making these two approximations leads to Akaike's method which amounts to replace (6) by

$$-\int \widehat{s}_m \log\left(\widehat{s}_m\right) + \frac{D_m}{n} \tag{7}$$

and proposes to select a partition $\hat{m}$ minimizing Akaike's criterion (7). An elementary computation shows that

$$P_n\left[-\log\left(\widehat{s}_m\right)\right] = -\int \widehat{s}_m \log\left(\widehat{s}_m\right).$$

If we denote by γ_n the empirical criterion corresponding to maximum likelihood estimation, i.e. $\gamma_n(t) = P_n\left[-\log(t)\right]$, we derive that Akaike's criterion can be written as

$$\gamma_n\left(\hat{s}_m\right) + \frac{D_m}{n},$$

and is indeed a penalized model selection criterion of type (3) with $\mathrm{pen}_n(m) = D_m/n$. It will be one of the main issues of Section 3 to discuss whether this heuristic approach can be validated or not but we can right now try to guess why concentration inequalities will be useful and in what circumstances Akaike's criterion should be corrected. Indeed, we have seen that Akaike's heuristics rely on the fact that some quantities R_m stay close to their expectations (they are actually all centered at 0). Moreover, this should hold with a certain uniformity over the list of partitions $\mathcal{M}_n$. This means that if the collection of partitions is not too rich, we can hope that the R_m's will be concentrated enough around their expectations to warrant that Akaike's heuristics works, while if the collection is too rich, concentration inequalities will turn to be an essential tool to understand how one should correct (substantially) Akaike's criterion.

1.3. The role of concentration inequalities

The role of the penalty function is absolutely fundamental in the definition of the penalized criterion (3) or (4), both from a theoretical and a practical point of view. Ideally, one would like to understand how to choose this penalty function in an optimal way, that is in order that the performance of the resulting penalized estimator be as close as possible as that of an oracle. Moreover one would like to provide explicit formulas for these

penalty functions to allow the implementation of the corresponding model selection procedures. We shall see that concentration inequalities are not only helpful to analyze the mathematical performance of the minimum penalized contrast estimator in terms of risk bounds but also to specify what kind of penalty functions are sensible. In fact these questions are two different aspects of the same problem since in our analysis, sensible penalty functions will be those for which we shall be able to provide efficient risk bounds for the corresponding penalized estimators. The key of this analysis is to take $l(s,t)$ as a loss function and notice that the definition of the penalized procedure leads to a very simple but fundamental control for $l(s,\widetilde{s})$. Indeed, by the definition of $\widetilde{s}$ we have, whatever $m \in \mathcal{M}_n$ and $s_m \in S_m$,

$$\gamma_n(\widetilde{s}) + \mathrm{pen}_n(\widehat{m}) \leqslant \gamma_n(s_m) + \mathrm{pen}_n(m) + \rho_n,$$

and therefore

$$\gamma_n(\widetilde{s}) \leqslant \gamma_n(s_m) + \mathrm{pen}_n(m) - \mathrm{pen}_n(\widehat{m}) + \rho_n. \tag{8}$$

If we introduce the centered empirical process

$$\overline{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}[\gamma(t,\xi_1)], t \in \mathcal{S}$$

and notice that $\mathbb{E}[\gamma(t,\xi_1)] - \mathbb{E}[\gamma(u,\xi_1)] = l(s,t) - l(s,u)$ for all $t,u \in \mathcal{S}$, we readily get from (8)

$$l(s,\widetilde{s}) \leqslant l(s,s_m) + \overline{\gamma}_n(s_m) - \overline{\gamma}_n(\widetilde{s}) - \mathrm{pen}_n(\widehat{m}) + \mathrm{pen}_n(m) + \rho_n. \tag{9}$$

Roughly speaking, starting from inequality (9), one would like to choose a penalty function in such a way that it dominates the fluctuations of the variable $\overline{\gamma}_n(s_m) - \overline{\gamma}_n(\widetilde{s})$. The trouble is that this variable is not that easy to control since we do not know where $\widehat{m}$ is located. This is the reason why we shall rather control $\overline{\gamma}_n(s_m) - \overline{\gamma}_n(\widehat{s}_{m'})$, uniformly over $m' \in \mathcal{M}_n$. At this stage we shall use empirical processes techniques and more precisely concentration inequalities that will help us to understand how the "complexity" of the collection of models has to be taken into account in the definition of the penalty function.

We can indeed derive from this approach some quite general way of describing these penalty functions. Let $(x_m)_{m\in\mathcal{M}_n}$ be a family of nonnegative weights such that, for some absolute constant Σ

$$\sum_{m\in\mathcal{M}_n} e^{-x_m} \leqslant \Sigma.$$

Most of the time Σ can be taken as 1, but it is useful to keep some flexibility with the choice of this normalizing constant. We should think of

$(e^{-x_m})_{m\in\mathcal{M}_n}$ as some prior finite measure on the collection of models which in some sense measures its "complexity". For every $m \in \mathcal{M}_n$, let σ_m^2 be some quantity measuring the difficulty for estimating within model S_m. As suggested by (2), one should expect that $\sigma_m^2 = D_m/n$, if D_m is a properly defined "dimension" of the model S_m (typically, in our case example of histograms above $1 + D_m$ is the number of pieces of partition m). Then

$$\mathrm{pen}_n(m) = K_1\sigma_m^2 + K_2\frac{x_m}{n} \tag{10}$$

where K_1 and K_2 are proper constants. Obviously one would like to know what is the right definition for σ_m^2 and what values for K_1 and K_2 are allowed. The advances for a better understanding of the calibration of penalty functions are of three different types.

- When the centered empirical process $\overline{\gamma}_n$ is acting linearly on S_m which is itself some part of a D_m dimensional linear space, it is possible to get a very precise idea of what the minimal penalty functions to be used are. This idea, first introduced in [10], was based on a preliminary version of Talagrand's deviation inequalities for empirical processes established in [38]. It really became successful with the major improvement obtained by Talagrand in [39]. This new version is truly a concentration inequality around its expectation for the supremum of an empirical process and not a deviation inequality from some unknown (or unrealistic) numerical constant as was the previous one. As a result, in the linear situations described above, it is possible to compute explicitly σ_m^2 together with a minimal value for the constant K_1. This, of course, is of great importance in the situation where x_m behaves like a corrective term and not like a leading term in formula (10), which concretely means that the list of models is not too rich (one model per dimension is a typical example of that kind like in the problem of selecting the best regular histogram for instance). This program has been successfully applied in various frameworks (see [2], [23], [24],[12] where random penalty functions are also considered and [3] where the context of weakly dependent data is even more involved).
- When the collection of models is too rich, like in the problem of selecting irregular histograms for instance, x_m no longer behaves like a corrective term but like a leading term and it becomes essential to evaluate the constant K_2. This program can be achieved whenever one has at one's disposal some precise exponential concentration bounds, that is involving explicit numerical constants. This is exactly the case for the Gaussian frameworks considered in [11], where one can use the Gaussian concentration inequality due to Cirelson, Ibragimov and

Sudakov (see [17] and Inequality 11 below). More than that, thanks to Ledoux's way of proving Talagrand's inequality (as initiated in [28]) it is also possible to evaluate the constants involved in Talagrand's concentration inequalities for empirical processes (see [33]), which leads to some evaluations of K_2. This idea is exploited in [12] and will be illustrated in Section 3 following the lines of Castellan's work [23].

- For the linear situations mentioned above, there is some obvious definition for the dimension D_m of S_m. In the nonlinear case, things are not that clear and one can find in the literature various ways of defining analogues of the linear dimension. In [7], various notions of metric dimensions are introduced (defined from covering by different types of brackets) while an alternative notion of dimension is the Vapnik-Chervonenkis dimension (see [40]). Since the early eighties Vapnik has developed a *statistical learning theory* (see the pioneering book [40] and more recent issues in [41]). One of the main methods introduced by Vapnik is what he called the *structural minimization of the risk* which is a model selection via penalization method in the above sense but with a calibration for the penalty function which differs substantially from (10). A minor difference with what can be found in [7] is that in Vapnik's theory, the Vapnik-Chervonenkis dimension is used to measure the size of a given model instead of "bracketing" dimensions. But there exists some much more significant difference concerning the order of magnitude of the penalty functions since the calibration of the penalty in Vapnik's structural minimization of the risk method is rather of the order of the square root of (10). Our purpose in Section 4 will be to show that the reason for this is that this calibration for the penalty function is related to "global" (that is of Hoeffding type) concentration inequalities for empirical processes. We shall also propose some general theorem (for bounded contrast functions) based on Talagrand's inequality in [39] (which is "local" , that is of Bernstein's type) which allows to deal with any kind of dimension (bracketing dimension or VC-dimension). From this new result we can recover some of the results in [7] and remove the square root in Vapnik's calibration of the penalty function which leads to improvements on the risk bounds for the corresponding penalized estimators. Moreover, following the idea introduced in [16], we show that it is possible to use a random combinatorial entropy number rather than the VC dimension in the definition of the penalty function. This is made possible by using again a concentration argument for the random combinatorial entropy number around its expectation. The concentration inequality which is used for this purpose is

established in [16]. It is an extension to non negative functionals of independent variables of a Poissonian bound given in [33] for the supremum of non negative empirical processes. There exist applications of this bound to random combinatorics which are developed in [16] and that we shall not reproduce here. However, this inequality is so simple to state and to prove that we could not resist to the temptation of presenting its proof in Section 2 of the present paper dedicated to Michel Talagrand.

2. Concentration inequalities and some direct applications

The oldest striking result illustrating the concentration of product probability measures phenomenon is the concentration of the standard Gaussian measure on $\mathbb{R}^D$. Let P denote the canonical Gaussian measure on the Euclidean space $\mathbb{R}^D$ and let ζ be some Lipschitz function on $\mathbb{R}^D$ with Lipschitz constant L. Then, for every $x \geqslant 0$,

$$P[\zeta - M \geqslant x] \leqslant \exp\left(-\frac{x^2}{2L^2}\right), \tag{11}$$

where M denotes either the mean or the median of ζ with respect to P (of course the same inequality holds when replacing ζ by $-\zeta$ which implies that ζ concentrates around M). This inequality has been established independently by Cirelson and Sudakov in [18] and Borell in [15] when M is a median and by Cirelson, Ibragimov and Sudakov in [17] when M is the mean. The striking feature of these inequalities is the fact that they do not depend on the dimension D (or more precisely only through M and L) which allows to use them for studying infinite dimensional Gaussian measures and Gaussian processes for instance (see [27]). Extending such results to more general product measures is not easy. Talagrand's approach to this problem relies on isoperimetric ideas in the sense that concentration inequalities for functionals around their median are derived from probability inequalities for enlargements of sets with respect to various distances. A typical result which can be obtained by his methods is as follows (see [36] and [37] for the best constant). Let $[a,b]^n$ be equipped with the canonical Euclidean distance and let ζ be some convex and Lipschitz function on $[a,b]^n$ with Lipschitz constant L. Let P be some product probability measure on $[a,b]^n$ and M be some median of ζ (with respect to the probability P), then, for every $x \geqslant 0$

$$P\left[\zeta - M \geqslant x + L(b-a)\sqrt{\frac{\log(2)}{2}}\right] \leqslant \exp\left(-\frac{x^2}{2L^2(b-a)^2}\right). \tag{12}$$

Moreover, the same inequality holds for $-\zeta$ instead of ζ. For this problem, the isoperimetric approach developed by Talagrand consists in proving that for any convex set A of $[a,b]^n$,

$$P\left[d(.,A) \geqslant x + (b-a)\sqrt{\frac{\log(1/P(A))}{2}}\right] \leqslant \exp\left(-\frac{x^2}{2(b-a)^2}\right),$$

where $d(.,A)$ denotes the Euclidean distance function to A . The latter inequality can be proved by at least two methods: the original proof by Talagrand in [37] relies on a control of the Laplace transform of $d(.,A)$ which is proved by induction on the number of coordinates while Marton's or Dembo's proofs (see [31] and [20] respectively) are based on some information inequalities. An alternative approach to this question has been proposed by Ledoux in [28]. It consists in focusing on the functional ζ, rather than starting from sets, and proving a logarithmic Sobolev type inequality which leads to some differential inequality on the Laplace transform of ζ. Integrating this differential inequality yields, for every $\lambda \geqslant 0$,

$$\log E_P\left[\exp\left(\lambda\left(\zeta - E\left[\zeta\right]\right)\right)\right] \leqslant \frac{\lambda^2 L^2 (b-a)^2}{2},$$

which in turn implies via Chernoff's inequality

$$P\left[\zeta - E\left[\zeta\right] \geqslant x\right] \leqslant \exp\left(-\frac{x^2}{2L^2(b-a)^2}\right). \tag{13}$$

It is instructive to compare (12) and (13). Since (12) holds when replacing ζ by $-\zeta$, some straightforward integration shows that for some appropriate positive numerical constant C

$$P\left[\zeta - E\left[\zeta\right] \geqslant x + CL(b-a)\right] \leqslant \exp\left(-\frac{x^2}{2L^2(b-a)^2}\right)$$

which is of the same nature as (13) but is obviously weaker from the point of view of getting as good constants as possible. For the applications that we have in view, deviation inequalities of a functional from its mean (rather than from its median) are more suitable. For the sake of presenting inequalities with the best available constants, we shall record below a number of inequalities of that type obtained by direct methods such as martingale differences (see [35]) or logarithmic Sobolev inequalities (see [28]). We begin with Hoeffding type inequalities for empirical processes.

2.0.1. *Hoeffding type inequalities for empirical processes*

The connection between convex Lipschitz functionals and Hoeffding type inequalities comes from the following elementary observation. Let $\xi_1, ..., \xi_n$ be independent $[a,b]$-valued random variables and let $(\alpha_{i,t})_{i \leqslant n, t \in T}$ be some finite family of real numbers, then by Cauchy-Schwarz inequality, the function ζ defined on $[a,b]^n$ by

$$\zeta : x \to \sup_{t \in T} \sum_{i=1}^{n} \alpha_{i,t} x_i$$

is convex and Lipschitz on $[a,b]^n$ with Lipschitz constant σ, where $\sigma^2 = \sup_{t \in T} \sum_{i=1}^{n} \alpha_{i,t}^2$. Therefore (13) implies that the random variable Z defined by

$$Z = \sup_{t \in T} \sum_{i=1}^{n} \alpha_{i,t} \xi_i$$

satisfies for every $x \geqslant 0$

$$\mathbb{P}\left[Z - \mathbb{E}\left[Z\right] \geqslant x\right] \leqslant \exp\left(-\frac{x^2}{2\sigma^2 (b-a)^2}\right). \tag{14}$$

This inequality is due to Ledoux (see inequality (1.9) in [28]) and can be easily extended to the following more general framework. Let $Z = \sup_{t \in T} \sum_{i=1}^{n} X_{i,t}$, where $a_{i,t} \leqslant X_{i,t} \leqslant b_{i,t}$ for some real numbers $a_{i,t}$ and $b_{i,t}$, for all $i \leqslant n$ and all $t \in T$. Then, setting $L^2 = \sup_{t \in T} \sum_{i=1}^{n} (b_{i,t} - a_{i,t})^2$, one has for every $x \geqslant 0$ (see [33]),

$$\mathbb{P}\left[Z \geqslant \mathbb{E}\left[Z\right] + x\right] \leqslant \exp\left(-\frac{x^2}{2L^2}\right). \tag{15}$$

The classical Hoeffding inequality (see [25]) ensures that when $T = \{1\}$, the variable $Z = \sum_{i=1}^{n} X_{i,1}$ satisfies

$$\mathbb{P}\left[Z \geqslant \mathbb{E}\left[Z\right] + x\right] \leqslant \exp\left(-\frac{2x^2}{L^2}\right), \tag{16}$$

for every positive x. When we compare (15) to (16), we notice that some factor 4 has been lost in the exponent of the upper bound. As a matter of fact, at the price of changing

$$\sup_{t \in T} \sum_{i=1}^{n} (b_{i,t} - a_{i,t})^2 \quad \text{into} \quad \sum_{i=1}^{n} \sup_{t \in T} (b_{i,t} - a_{i,t})^2 ,$$

Inequality (16) can be shown to hold for $Z = \sup_{t \in T} \sum_{i=1}^{n} X_{i,t}$. This result derives from the martingale difference method as shown by McDiarmid in [35]. A useful consequence of this result concerns empirical processes. Indeed, if $\xi_1, ..., \xi_n$ are independent random variables and $\mathcal{F}$ is a finite or countable class of functions such that, for some real numbers a and b, one has $a \leqslant f \leqslant b$ for every $f \in \mathcal{F}$, then setting $Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(\xi_i) - \mathbb{E}[f(\xi_i)]$, we get by monotone convergence

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geqslant x) \leqslant 2\exp\left(-\frac{2x^2}{n(b-a)^2}\right). \tag{17}$$

It should be noticed that (17) does not generally provide a subgaussian inequality. The reason is that the maximal variance

$$\sigma^2 = \sup_{f \in \mathcal{F}} \mathrm{Var}\left[\sum_{i=1}^{n} f(\xi_i)\right]$$

can be substantially smaller than $n(b-a)^2/4$ and therefore (17) can be much worse than its "sub-Gaussian" version which should make σ^2 appear instead of $n(b-a)^2/4$. It is precisely our purpose now to provide sharper bounds than Hoeffding type inequalities. We begin by considering nonnegative functionals.

2.1. A Poissonian inequality for nonnegative functionals

We intend to present here some extension to nonnegative functionals of independent random variables due to Boucheron, Lugosi and Massart (see [16]) of a Poissonian bound for the supremum of non negative empirical processes established in [33] by using Ledoux's approach to concentration inequalities. The motivations for considering general nonnegative functionals of independent random variables came from random combinatorics. Several illustrations are given in [16] but we shall focus here on the case example of random combinatorial entropy since the corresponding concentration result will turn out to be very useful for designing random penalties to solve the model selection problem for pattern recognition (see Section 4). Roughly speaking, under some condition (**C**) to be given below, we shall show that a nonnegative functional Z of independent variables concentrates around its expectation like a Poisson random variable with expectation $\mathbb{E}[Z]$ (this comparison being expressed in terms of Laplace transform). This Poissonian inequality can be deduced from the integration of a differential inequality for the Laplace transform of Z which derives from a key information bound.

A very remarkable fact is that Han's inequality for Kullback-Leibler information is at the heart of the proof of this bound and is also deeply involved in the verification of condition (**C**) for combinatorial entropies.

2.1.1. *Han's inequality for entropy*

Let us first recall some well known fact about Shannon's entropy and Kullback-Leibler information. Given some random variable Y taking its values in some finite set $\mathcal{Y}$, Shannon entropy is defined by

$$h_S(Y) = -\sum_y \mathbb{P}(Y = y) \log [\mathbb{P}(Y = y)] .$$

Setting, $q_y = \mathbb{P}(Y = y)$ for any point y in the support of Y, Shannon entropy can also be written as $h_S(Y) = \mathbb{E}[-\log q_Y]$, from which one readily sees that it is a nonnegative quantity. The relationship between Shannon entropy and Kullback-Leibler information is given by the following identity. Let Q be the distribution of Y, P be the uniform distribution on $\mathcal{Y}$ and N be the cardinality of $\mathcal{Y}$, then

$$K(Q, P) = -h_S(Y) + \log(N) . \tag{18}$$

We derive from this equation and the nonnegativity of the Kullback-Leibler information that

$$h_S(Y) \leqslant \log(N) , \tag{19}$$

with equality if and only if Y is uniformly distributed on its support. Han's inequality for Shannon entropy can be stated as follows (see [19], p. 491 for a proof).

PROPOSITION 2.1 (Han's inequality). — *Let us consider some random variable Y with values in some finite product space $\mathcal{Y}^n$ and write $Y^{(i)} = (Y_1, ..., Y_{i-1}, Y_{i+1}, ...Y_n)$ for every $i \in \{1, ..., n\}$. Then*

$$h_S(Y) \leqslant \frac{1}{n-1} \sum_{i=1}^{n} h_S\left(Y^{(i)}\right) .$$

In view of (18) Han's inequality for discrete variables can be naturally extended to arbitrary distributions in the following way. Let $(\Omega^n, \mathcal{A}^n, P^n) = (\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{A}_i, \bigotimes_{i=1}^n \mu_i)$ be some product probability space and Q be some probability distribution on Ω^n which is absolutely continuous with respect to P^n. Let Y be the identity map on Ω^n, $Y^{(i)} = (Y_1, ..., Y_{i-1}, Y_{i+1}, ...Y_n)$ for every $i \in \{1, ..., n\}$ and denote by $P^{(i)}$ (resp. $Q^{(i)}$) the distribution of

$Y^{(i)}$ under P^n (resp. Q). Then, Han's inequality can be written as

$$K(Q,P^n) \geqslant \frac{1}{n-1}\sum_{i=1}^{n} K\left(Q^{(i)},P^{(i)}\right)$$

or equivalently as

$$K(Q,P^n) \leqslant \sum_{i=1}^{n}\left[K(Q,P^n) - K\left(Q^{(i)},P^{(i)}\right)\right]. \tag{20}$$

Now, let X be some random variable taking its values in Ω^n with distribution P^n and $G = g(X)$ be some nonnegative and integrable random variable. If we define Q to be the probability distribution with density $g/\mathbb{E}[G]$ with respect to P^n, denoting by $\mathbb{E}^{(i)}$ the expectation operator conditionally to $X^{(i)} = (X_1,...,X_{i-1},X_{i+1},...X_n)$ and by Φ the function $t \to t\log t$ another equivalent way of formulating (20) is

$$\mathbb{E}[\Phi(G)] - \Phi(\mathbb{E}[G]) \leqslant \mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}^{(i)}[\Phi(G)] - \Phi\left(\mathbb{E}^{(i)}[G]\right)\right]. \tag{21}$$

Inequality (21) is exactly what is called the tensorisation inequality for entropy in [28] (see also [14] for more general tensorisation inequalities). Then, for every positive measurable function $G^{(i)}$ of $X^{(i)}$, using $\log(x) \leqslant x-1$ with $x = G^{(i)}/\mathbb{E}^{(i)}[G]$, one has

$$\mathbb{E}^{(i)}[\Phi(G)] - \Phi\left(\mathbb{E}^{(i)}[G]\right) \leqslant \mathbb{E}^{(i)}\left[G\left(\log G - \log G^{(i)}\right) - \left(G - G^{(i)}\right)\right].$$

Hence, if Z is some measurable function of X and for every $i \in \{1,...,n\}$, $Z^{(i)}$ is some measurable function of $X^{(i)}$, applying the above inequality to the variables $G = e^{\lambda Z}$ and $G^{(i)} = e^{\lambda Z^{(i)}}$, one gets

$$\mathbb{E}^{(i)}[\Phi(G)] - \Phi\left(\mathbb{E}^{(i)}[G]\right) \leqslant \mathbb{E}^{(i)}\left[e^{\lambda Z}\phi(-\lambda(Z-Z^{(i)}))\right],$$

where ϕ denotes the function $z \to \exp(z) - z - 1$. Therefore, we derive from (21), that

$$\lambda\mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right]\log\mathbb{E}\left[e^{\lambda Z}\right] \leqslant \sum_{i=1}^{n}\mathbb{E}\left[e^{\lambda Z}\phi\left(-\lambda(Z-Z^{(i)})\right)\right], \tag{22}$$

for any λ such that $\mathbb{E}\left[e^{\lambda Z}\right] < \infty$. This inequality and its symmetrized version are the main tools used in [33] to evaluate the constants in Talagrand's inequalities for empirical processes (see Lemma 2.3 therein).

2.1.2. *The Poissonian bound*

It can be fruitfully applied to nonnegative functionals following [16].

THEOREM 2.2. — *Let $X_1, ..., X_n$ be independent random variables and define for every $i \in \{1, ..., n\}$ $X^{(i)} = (X_1, ..., X_{i-1}, X_{i+1}, ...X_n)$. Let Z be some nonnegative and bounded measurable function of $X = (X_1, ..., X_n)$. Assume that for every $i \in \{1, ..., n\}$, there exists some measurable function $Z^{(i)}$ of $X^{(i)}$ such that*

$$0 \leqslant Z - Z^{(i)} \leqslant 1. \tag{23}$$

Assume furthermore that

$$\sum_{i=1}^{n} \left(Z - Z^{(i)}\right) \leqslant Z. \tag{C}$$

Defining h as $h(u) = (1+u)\log(1+u) - u$, for $u \geqslant -1$, the following inequalities hold:

$$\mathbb{P}[Z \geqslant \mathbb{E}[Z] + x] \leqslant \exp\left[-\mathbb{E}[Z]\, h\left(\frac{x}{\mathbb{E}[Z]}\right)\right], \text{ for all } x > 0 \tag{24}$$

and

$$\mathbb{P}[Z \leqslant \mathbb{E}[Z] - x] \leqslant \exp\left[-\mathbb{E}[Z]\, h\left(-\frac{x}{\mathbb{E}[Z]}\right)\right], \text{ for } 0 < x \leqslant \mathbb{E}[Z]. \tag{25}$$

Proof. — We know that (22) holds for any λ. Since the function ϕ is convex with $\phi(0) = 0$, $\phi(-\lambda u) \leqslant u\phi(-\lambda)$ for any λ and any $u \in [0,1]$. Hence it follows from (23) that for every λ, $\phi(-\lambda(Z - Z^{(i)})) \leqslant (Z - Z^{(i)})\phi(-\lambda)$ and therefore we derive from (22) and (**C**) that

$$\begin{aligned} \lambda\mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right]\log\mathbb{E}\left[e^{\lambda Z}\right] &\leqslant \mathbb{E}\left[\phi(-\lambda)e^{\lambda Z}\sum_{i=1}^{n}\left(Z - Z^{(i)}\right)\right] \\ &\leqslant \phi(-\lambda)\mathbb{E}\left[Ze^{\lambda Z}\right]. \end{aligned}$$

We introduce $\widetilde{Z} = Z - \mathbb{E}[Z]$ and define for any λ, $F(\lambda) = \mathbb{E}\left[e^{\lambda\widetilde{Z}}\right]$. Setting $v = \mathbb{E}[Z]$, the preceding inequality becomes

$$[\lambda - \phi(-\lambda)]\frac{F'(\lambda)}{F(\lambda)} - \log F(\lambda) \leqslant v\phi(-\lambda),$$

which in turn implies

$$\left(1 - e^{-\lambda}\right)\Psi'(\lambda) - \Psi(\lambda) \leqslant v\phi(-\lambda) \quad \text{with} \quad \Psi(\lambda) = \log F(\lambda). \tag{26}$$

Now observe that $v\phi$ is a solution of the ordinary differential equation $\left(1-e^{-\lambda}\right) f'(\lambda) - f(\lambda) = v\phi(-\lambda)$. In order to show that $\Psi \leqslant v\phi$, we set

$$\Psi(\lambda) = v\phi(\lambda) + \left(e^{\lambda}-1\right) g(\lambda), \tag{27}$$

for every $\lambda \neq 0$ and derive from (26) that

$$\left(1-e^{-\lambda}\right)\left[e^{\lambda} g(\lambda) + \left(e^{\lambda}-1\right) g'(\lambda)\right] - \left(e^{\lambda}-1\right) g(\lambda) \leqslant 0,$$

which yields

$$\left(1-e^{-\lambda}\right)\left(e^{\lambda}-1\right) g'(\lambda) \leqslant 0.$$

We derive from this inequality that g' is nonpositive which means that g is nonincreasing. Now, since $\widetilde{Z}$ is centered at expectation $\Psi'(0) = \phi(0) = 0$ and it comes from (27) that $g(\lambda)$ tends to 0 as λ goes to 0. This shows that g is nonnegative on $(-\infty, 0)$ and nonpositive on $(0, \infty)$ which in turn means by (27) that $\Psi \leqslant v\phi$ and we have proved that

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leqslant v\phi(\lambda) \quad \text{for every } \lambda \in \mathbb{R}. \tag{28}$$

Thus, by Markov's inequality,

$$\mathbb{P}[Z - \mathbb{E}[Z] \geqslant x] \leqslant \exp\left[-\sup_{\lambda>0}\left(x\lambda - v\phi(\lambda)\right)\right]$$

and

$$\mathbb{P}[Z - \mathbb{E}[Z] \leqslant -x] \leqslant \exp\left[-\sup_{\lambda<0}\left(-x\lambda - v\phi(\lambda)\right)\right].$$

The proof can be completed by using the easy to check (and well known) relations: $\sup_{\lambda>0}\left[x\lambda - v\phi(\lambda)\right] = vh(x/v)$ for every $x > 0$ and $\sup_{\lambda<0}\left[-x\lambda - v\phi(\lambda)\right] = vh(-x/v)$ for every $0 < x \leqslant v$. □

The above inequalities are exactly the classical Cramér-Chernoff upper bounds for the Poisson distribution with mean $\mathbb{E}[Z]$ and in this sense this Theorem establishes a comparison between the concentration around its mean of a non negative functional Z satisfying the above assumptions and that of the Poisson distribution with the same mean. Let us give some further comments.

- This theorem can typically be applied to the supremum of sums of nonnegative random variables. Indeed let $X_1, \ldots, X_n$ be independent $[0,1]^N$-valued random variables and consider

$$Z = \sup_{1 \leqslant t \leqslant N} \sum_{i=1}^{n} X_{i,t}$$

with $Z^{(i)} = \sup_{1 \leqslant t \leqslant N} \sum_{j \neq i} X_{j,t}$ for all $i \leqslant n$. Then denoting by τ some random number such that $Z = \sum_{i=1}^{n} X_{i,\tau}$, one obviously has

$$0 \leqslant Z - Z^{(i)} \leqslant X_{i,\tau} \leqslant 1,$$

and therefore

$$\sum_{i=1}^{n} \left(Z - Z^{(i)} \right) \leqslant \sum_{i=1}^{n} X_{i,\tau} = Z.$$

It is easy to see on this example that (24) and (25) are in some sense unimprovable. Indeed, if $N = 1$, and $X_1, ..., X_n$ are Bernoulli trials with parameter θ, then Z follows the binomial distribution $\mathcal{B}(n, \theta/n)$ and its asymptotic distribution is actually a Poisson distribution with mean θ.

- Inequality (25) readily implies the sub-Gaussian inequality

$$\mathbb{P}[Z \leqslant \mathbb{E}[Z] - x] \leqslant \exp\left[-\frac{x^2}{2\mathbb{E}[Z]} \right] \tag{29}$$

which holds for every $x > 0$. Indeed, (29) is trivial when $x > \mathbb{E}[Z]$ and follows from (25) otherwise since, for every $\varepsilon \in [0,1]$ one has $h(-\varepsilon) \geqslant \varepsilon^2/2$.

Let us turn now to a somehow more subtle application of Theorem 2.2 to combinatorial entropy. Surprisingly, Han's inequality will be involved again to show that the combinatorial entropy satisfies condition (**C**).

2.1.3. *Application to combinatorial entropies*

Let $\mathcal{F}$ be some class of measurable functions defined on some set $\mathcal{X}$ and taking their values in $\{1, ..., k\}$. We define the combinatorial entropy of $\mathcal{F}$ at point $x \in \mathcal{X}^n$ by

$$\zeta(x) = \log_k |Tr(x)|,$$

where $Tr(x) = \{(f(x_1), ..., f(x_n)), f \in \mathcal{F}\}$ and $|Tr(x)|$ denotes the cardinality of $Tr(x)$. It is quite remarkable that, given some independent variables, $X_1, ..., X_n$, $Z = \zeta(X)$ satisfies to the assumptions of our Theorem 2.2. Indeed, let $Z^{(i)} = \zeta\left(X^{(i)}\right)$ for every i. Obviously $0 \leqslant Z - Z^{(i)} \leqslant 1$ for all i. On the other hand, given $x \in \mathcal{X}^n$, let us consider some random variable Y with uniform distribution on the set $Tr(x)$. It comes from Han's inequality (see Proposition 2.1) that,

$$\log |Tr(x)| = h_S(Y) \leqslant \frac{1}{n-1} \sum_{i=1}^{n} h_S\left(Y^{(i)}\right).$$

Now for every i, $Y^{(i)}$ takes its values in $Tr\left(x^{(i)}\right)$ and therefore by (19) we have $h_S\left(Y^{(i)}\right) \leqslant \log\left|Tr\left(x^{(i)}\right)\right|$. Hence

$$\log|Tr(x)| \leqslant \frac{1}{n-1}\sum_{i=1}^{n}\log\left|Tr\left(x^{(i)}\right)\right|,$$

which means that

$$\zeta(x) \leqslant \frac{1}{n-1}\sum_{i=1}^{n}\zeta\left(x^{(i)}\right).$$

Thus condition (2.2) is satisfied and Theorem 2.2 applies to the combinatorial entropy $\log_k|Tr(X)|$, which, in particular implies that

$$\mathbb{P}\left[Z \leqslant \mathbb{E}[Z] - \sqrt{2\mathbb{E}[Z]x}\right] \leqslant e^{-x}, \text{ for all } x > 0. \tag{30}$$

This inequality will be of great importance for our study of statistical learning in pattern recognition in Section 4.

Of course Theorem 2.2 is designed for nonnegative functionals and does not solve the problem of improving on Hoeffding type bounds for the supremum of a centered empirical process.

2.2. Talagrand's inequalities for empirical processes

In [39] (see Theorem 4.1), Talagrand obtained some striking concentration inequality for the supremum of an empirical process which is an infinite dimensional analogue of Bernstein's inequality. There exists several ways of expressing his result. We choose the one which is the most convenient for our needs.

THEOREM 2.3 (Talagrand's inequality). — *Consider n independent and identically distributed random variables $\xi_1, ..., \xi_n$ with values in some measurable space Ξ. Let $\mathcal{F}$ be some countable family of real valued measurable functions on Ξ, such that $\|f\|_\infty \leqslant b < \infty$ for every $f \in \mathcal{F}$. Let $Z = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} f(\xi_i)\right|$ and $v = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f^2(\xi_i)\right]$. Then*

$$\mathbb{P}\left[Z \geqslant \mathbb{E}[Z] + c_1\sqrt{vx} + c_2 bx\right] \leqslant Ke^{-x} \quad \textit{for all } x > 0, \tag{31}$$

where K, c_1 and c_2 are universal positive constants. Moreover the same inequality holds when replacing Z by $-Z$.

In order to use this inequality, it is desirable to get a more tractable formulation of (31), involving

$$\sup_{f\in\mathcal{F}}\mathbb{E}\left[\sum_{i=1}^{n} f^2(\xi_i)\right] \text{ instead of } \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f^2(\xi_i)\right].$$

This can be done for centered empirical processes at the price of additional technicalities related to classical symmetrization and contraction inequalities as described in [29]. One indeed has (see [33] for more details)

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}f^2(\xi_i)\right] \leqslant \sup_{f\in\mathcal{F}}\mathbb{E}\left[\sum_{i=1}^{n}f^2(\xi_i)\right] + 16b\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}f(\xi_i)-\mathbb{E}\left[f(\xi_i)\right]\right|\right]. \quad (32)$$

In particular if every function $f \in \mathcal{F}$ is centered at expectation, setting $\sigma^2 = \sup_{f\in\mathcal{F}}\mathbb{E}\left[\sum_{i=1}^{n}f^2(\xi_i)\right]$, the preceding inequality becomes

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}f^2(\xi_i)\right] \leqslant \sup_{f\in\mathcal{F}}\mathbb{E}\left[\sum_{i=1}^{n}f^2(\xi_i)\right] + 16b\mathbb{E}\left[Z\right].$$

Plugging this inequality into (31) leads to

$$\mathbb{P}\left[Z \geqslant \mathbb{E}\left[Z\right] + c_1\sigma\sqrt{x} + 4c_1\sqrt{b\mathbb{E}\left[Z\right]x} + c_2bx\right] \leqslant Ke^{-x},$$

and therefore for every positive ε

$$\mathbb{P}\left[Z \geqslant (1+\varepsilon)\,\mathbb{E}\left[Z\right] + c_1\sigma\sqrt{x} + \left(\frac{8c_1}{\varepsilon}+c_2\right)bx\right] \leqslant Ke^{-x}.$$

The same kind of inequality could be obtained for the left tail of Z.

It is indeed possible and useful for some applications (see Section 3 below) to have an idea of the value of the numerical constants involved in the above inequality. This is done in [33] by following Ledoux's approach to concentration (see [28]).

THEOREM 2.4. — *Consider n independent random variables $\xi_1, ..., \xi_n$ with values in some measurable space Ξ. Let $\mathcal{F}$ be some countable family of real valued measurable functions on Ξ, such that $\|f\|_\infty \leqslant b < \infty$ for every $f \in \mathcal{F}$. Let*

$$Z = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}f(\xi_i)-\mathbb{E}\left[f(\xi_i)\right]\right| \text{ and } \sigma^2 = \sup_{f\in\mathcal{F}}\left[\sum_{i=1}^{n}\mathrm{Var}\left(f(\xi_i)\right)\right].$$

Then, for any positive real numbers ε and x,

$$\mathbb{P}\left[Z \geqslant (1+\varepsilon)\,\mathbb{E}\left[Z\right] + \sigma\sqrt{2\kappa x} + \kappa(\varepsilon)\,bx\right] \leqslant \exp\left(-x\right), \quad (33)$$

where κ *and* $\kappa(\varepsilon)$ *can be taken equal to* $\kappa = 4$ *and* $\kappa(\varepsilon) = 2.5 + 32\varepsilon^{-1}$. *Moreover, one also has*

$$\mathbb{P}\left[Z \leqslant (1-\varepsilon)\,\mathbb{E}[Z] - \sigma\sqrt{2\kappa' x} - \kappa'(\varepsilon)\,bx\right] \leqslant \exp(-x), \tag{34}$$

where $\kappa' = 5.4$ *and* $\kappa'(\varepsilon) = 2.5 + 43.2\varepsilon^{-1}$.

A very simple example of application of such concentration inequalities is the study of chi-square statistics which will turn out to be at the heart of Castellan's work presented in Section 3.

2.2.1. *A first application to chi-square statistics*

One very remarkable feature of the concentration inequalities stated above is that, despite of their generality, they turn out to be sharp when applied to the particular and apparently simple problem of getting non asymptotic exponential for chi-square statistics. Following [10], we denote by ν_n the centered empirical measure $P_n - P$, given some finite set of bounded functions $\{\varphi_I\}_{I\in m}$, we can indeed write

$$\sqrt{\sum_{I\in m} \nu_n^2[\varphi_I]} = \sup_{|a|_2=1} \nu_n\left[\sum_{I\in m} a_I\varphi_I\right], \text{ where } |a|_2^2 = \sum_{I\in m} a_I^2.$$

Let $Z^2 = n\sum_{I\in m}\nu_n^2[\varphi_I]$. Applying Theorem 2.4 to the countable class of functions

$$\mathcal{F} = \left\{\sum_{I\in m} a_I\varphi_I : a \in \mathcal{S}'_m\right\},$$

where $\mathcal{S}'_m$ denotes some countable and dense subset of the unit sphere $\mathcal{S}_m$ in $\mathbb{R}^m$, one derives from (33) that, for every positive numbers ε and x,

$$\mathbb{P}\left[Z \geqslant (1+\varepsilon)\sqrt{\mathbb{E}[Z]} + \sqrt{2\kappa x v_m} + \kappa(\varepsilon)\sqrt{\frac{\|\Phi_m\|_\infty}{n}}x\right] \leqslant \exp(-x), \tag{35}$$

where

$$\Phi_m = \sum_{I\in m}\varphi_I^2 \text{ and } v_m = \sup_{a\in\mathcal{S}_m}\left[\mathrm{Var}\left(\sum_{I\in m} a_I\varphi_I(\xi_1)\right)\right].$$

Moreover by Cauchy-Schwarz inequality

$$\mathbb{E}[Z] \leqslant \sqrt{n\sum_{I\in m}\mathbb{E}[\nu_n^2[\varphi_I]]} \leqslant \sqrt{\sum_{I\in m}\mathrm{Var}(\varphi_I(\xi_1))}. \tag{36}$$

Let us now turn to the case example of "classical" chi-square statistics, which is of special interest by itself and also in view of the application to histogram selection given in Section 3. Let us take m to be some finite partition of $[0,1]$ which elements are intervals and define for every interval $I \in m$

$$\varphi_I = P(I)^{-1/2} \mathbb{1}_I,$$

then, the resulting functional Z^2 is the chi-square statistics

$$\chi_n^2(m) = \sum_{I \in m} \frac{n[P_n(I) - P(I)]^2}{P(I)}. \tag{37}$$

In this case, we derive from (36) that

$$\mathbb{E}[\chi_n(m)] \leqslant \sqrt{\sum_{I \in m} (1 - P(I))} \leqslant \sqrt{D_m},$$

where $1 + D_m$ denotes the number of pieces of m. We also notice that $v_m \leqslant 1$ and setting $\delta_m = \sup_{I \in m} P(I)^{-\frac{1}{2}}$, that

$$\|\Phi_m\|_\infty \leqslant \delta_m^2.$$

Therefore (35) becomes

$$\mathbb{P}\left[\chi_n(m) \geqslant (1+\varepsilon)\sqrt{D_m} + \sqrt{2\kappa x} + \kappa(\varepsilon)\frac{\delta_m}{\sqrt{n}}x\right] \leqslant \exp(-x). \tag{38}$$

For the left tail we get in the same way, for $\varepsilon \in (0,1)$,

$$\mathbb{P}\left[\chi_n(m) \leqslant (1-\varepsilon)\mathbb{E}[\chi_n(m)] - \sqrt{2\kappa' x} - \kappa'(\varepsilon)\frac{\delta_m}{\sqrt{n}}x\right] \leqslant \exp(-x),$$

and it remains to bound $\mathbb{E}[Z]$ from below. But this can be done by using again a concentration argument. The easiest one is probably the following Poincaré type inequality due to Ledoux [28].

PROPOSITION 2.5. — *Let $\xi_1, \ldots, \xi_n$ be independent random variables with values in some measurable space $(\Xi, \mathcal{X})$ and $\mathcal{F}$ be some countable class of real valued measurable functions on Ξ. Let*

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n f(\xi_i) - \mathbb{E}[f(\xi_i)] \right|.$$

Let $(\xi_1', \ldots, \xi_n')$ be independent from $(\xi_1, \ldots, \xi_n)$ and with the same distribution. Then

$$v = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n (f(\xi_i) - f(\xi_i'))^2\right] \geqslant \mathrm{Var}(Z).$$

Using the same kind of symmetrization and contraction arguments as before, whenever $\|f\|_\infty \leqslant b$ one shows that

$$v \leqslant 2\sigma^2 + 32b\mathbb{E}(Z).$$

When applied to chi-square statistics this inequality combined with Proposition 2.5 yields

$$D_m - [\mathbb{E}[\chi_n(m)]]^2 \leqslant 2 + 32\frac{\delta_m}{\sqrt{n}}\mathbb{E}[\chi_n(m)]$$

which implies that

$$\mathbb{E}[\chi_n(m)] \geqslant \sqrt{D_m - 2 + \frac{(16)^2\delta_m^2}{n}} - \frac{16\delta_m}{\sqrt{n}}.$$

Hence, provided that

$$D_m \geqslant \frac{2}{\varepsilon} \vee \frac{(32)^2\delta_m^2}{n\varepsilon^2}, \tag{39}$$

one derives that

$$\mathbb{E}[\chi_n(m)] \geqslant (1-\varepsilon)\sqrt{D_m}$$

and therefore the inequality on the left tail becomes

$$\mathbb{P}\left[\chi_n(m) \leqslant (1-\varepsilon)^2\sqrt{D_m} - \sqrt{2\kappa' x} - \kappa'(\varepsilon)\frac{\delta_m}{\sqrt{n}}x\right] \leqslant \exp(-x). \tag{40}$$

We do not know of any other way for deriving this inequality while on the other hand there already exists some deviation inequality on the right tail, obtained by Mason and Van Zwet. As compared to Mason and Van Zwet's inequality in [32], (38) is sharper but however not sharp enough for our needs in Section 3 since for irregular partitions the linear term $\delta_m x/\sqrt{n}$ can become too large. To do better, the above argument needs to be substantially refined and this is precisely what we shall perform in the next section.

3. Selecting the best histogram

We shall present in this section some of the results concerning the old standing problem of selecting "the best partition" when constructing some histogram, obtained by Castellan in [23] . Here $\mathcal{M}_n$ will denote some finite collection of partitions of $[0,1]$ into intervals (we restrict ourselves to $[0,1]$ for simplicity but Castellan's results hold for more general situations) and

$\xi_1, ..., \xi_n$ are independent and identically distributed random variables with distribution $s\mu$, where μ denotes the uniform distribution on $[0,1]$. We consider the collection of histogram estimators $\{\hat{s}_m,\ m \in \mathcal{M}_n\}$ and we intend to study penalized maximum likelihood selection criteria. More precisely let for every density t

$$\gamma_n(t) = -P_n[\log(t)]$$

be the empirical criterion defining the maximum likelihood estimation procedure. Let, for every partition m

$$\hat{s}_m = \sum_{I \in m} \frac{P_n(I)}{\mu(I)} \mathbb{1}_I.$$

Then, $\hat{s}_m$ minimizes γ_n over the model S_m of densities which are piecewise constants on the partition m and

$$\gamma_n(\hat{s}_m) = -\int \hat{s}_m \log(\hat{s}_m)\, d\mu.$$

Our purpose is to study the penalized selection procedure which consists in retaining a partition $\widehat{m}$ minimizing

$$\gamma_n(\hat{s}_m) + \mathrm{pen}_n(m),$$

over $m \in \mathcal{M}_n$. We recall that Akaike's criterion corresponds to the choice $\mathrm{pen}_n(m) = D_m/n$, where D_m+1 denotes the number of pieces of m and that Castellan's results presented below will allow to correct this criterion. Let us first explain the connection between the study of the penalized estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$ and the problem considered just above of controlling some chi-square statistics. The key is that in this case (9) allows to control the Kullback-Leibler information between s and $\tilde{s}$ in the following way

$$K(s, \tilde{s}) \leqslant K(s, s_m) + \nu_n(\log \tilde{s} - \log s_m) + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}), \quad (41)$$

for every $m \in \mathcal{M}_n$, where

$$s_m = \sum_{I \in m} \frac{P(I)}{\mu(I)} \mathbb{1}_I.$$

Now the main task is to bound $\nu_n(\log \tilde{s} - \log s_m)$ as sharply as possible in order to determine what is the minimal value of $\mathrm{pen}_n(\widehat{m})$ which is allowed for deriving a risk bound from (41). This will result from a uniform control of $\nu_n(\log \widehat{s}_{m'} - \log s_m)$ with respect to $m' \in \mathcal{M}_n$. We write

$$\nu_n\left(\log \frac{\widehat{s}_{m'}}{s_m}\right) = \nu_n\left(\log \frac{\hat{s}_{m'}}{s_{m'}}\right) + \nu_n\left(\log \frac{s_{m'}}{s}\right) + \nu_n\left(\log \frac{s}{s_m}\right)$$

and notice that the first term is the most delicate to handle since it involves the action of the empirical process on the estimator $\hat{s}_{m'}$ which is of course a random variable. This is precisely the control of this term which leads to the introduction of chi-square statistics. Indeed, setting

$$V^2(f,g) = \int s \left(\log \left(\frac{f}{g} \right) \right)^2 d\mu,$$

for every densities f and g such that $\log(f/g) \in \mathbb{L}_2(P)$, one derives that

$$\nu_n \left(\log \frac{\hat{s}_{m'}}{s_{m'}} \right) \leqslant \sup_{t \in S_{m'}} \left| \nu_n \left(\frac{\log t - \log s_{m'}}{V(t, s_{m'})} \right) \right| V(\hat{s}_{m'}, s_{m'}).$$

and if we set $\varphi_I = P(I)^{-\frac{1}{2}} \mathbb{1}_I$ for all $I \in m'$ then,

$$\begin{aligned} \sup_{t \in S_{m'}} \left| \nu_n \left(\frac{\log t - \log s_{m'}}{V(t, s_{m'})} \right) \right| &= \sup_{a \in \mathbb{R}^{m'}, |a|_2 = 1} \left| \sum_{I \in m'} a_I \nu_n(\varphi_I) \right| \\ &= \left[\sum_{I \in m'} \nu_n^2(\varphi_I) \right]^{1/2} = \frac{1}{\sqrt{n}} \chi_n(m'). \end{aligned}$$

Hence (41) becomes

$$\begin{aligned} K(s, \tilde{s}) \leqslant\ & K(s, s_m) + \mathrm{pen}_n(m) + \nu_n \left(\log \frac{s}{s_m} \right) \\ & + n^{-1/2} V \left(\hat{s}_{\widehat{m}}, s_{\widehat{m}} \right) \chi_n(\widehat{m}) + \nu_n \left(\log \frac{s_{\widehat{m}}}{s} \right) - \mathrm{pen}_n(\hat{m}). \quad (42) \end{aligned}$$

At this stage it becomes clear that what we need is a uniform control of $\chi_n(m')$ over $m' \in \mathcal{M}_n$. The key idea for improving on (38) is to majorize $\chi_n(m')$ only on some part of the probability space where $P_n(\varphi_I)$ remains close to $P(\varphi_I)$ for every $I \in m'$.

3.1. Some deepest analysis of chi-square statistics

This idea introduced in [12] in the context of subset selection within a conveniently localized basis can be fruitfully applied here. More precisely, Castellan proves in [23] the following inequality.

PROPOSITION 3.1.— *Let m be some partition of $[0,1]$ with $D_m + 1$ pieces and $\chi_n^2(m)$ be the chi-square statistics given by (37). Then for any positive real numbers ε and x,*

$$\mathbb{P} \left[\chi_n(m) \mathbb{1}_{\Omega_m(\varepsilon)} \geqslant (1+\varepsilon) \left(\sqrt{D_m} + \sqrt{2\kappa x} \right) \right] \leqslant \exp(-x) \qquad (43)$$

where $\kappa = 4$ *and* $\kappa(\varepsilon)$ *are the constants of Theorem 2.4 and* $\Omega_m(\varepsilon) = \{|P_n(I) - P(I)| \leqslant 2\kappa\varepsilon P(I)/\kappa(\varepsilon)$, *for every* $I \in m\}$.

Proof. — Let $\xi = 2\kappa\varepsilon/\kappa(\varepsilon)$ and z be some positive number to be chosen later. Setting $\varphi_I = P(I)^{-\frac{1}{2}} \mathbb{1}_I$ for every $I \in m$ and denoting by $\mathcal{S}_m$ the unit sphere in $\mathbb{R}^m$ as before, we have on the one hand

$$\Omega_m(\varepsilon) = \left\{|\nu_n[\varphi_I]| \leqslant \xi\sqrt{P(I)}, \text{ for every } I \in m\right\}$$

and on the other hand, , by Cauchy-Schwarz inequality

$$n^{-1/2}\chi_n(m) = \left[\sum_{I\in m} \nu_n^2[\varphi_I]\right]^{1/2} \geqslant \left|\sum_{I\in m} a_I \nu_n[\varphi_I]\right| \text{ for all } a \in \mathcal{S}_m,$$

with equality when $a_I = \nu_n[\varphi_I]\left(n^{-1/2}\chi_n(m)\right)^{-1}$ for all $I \in m$. Hence, defining $\mathcal{A}_m$ to be the set of those elements $a \in \mathcal{S}_m$ satisfying $\sup_{I\in m}|a_I| \leqslant \xi/z$, we have that on the event $\Omega_m(\varepsilon) \cap \{\chi_n(m) \geqslant z\}$

$$n^{-1/2}\chi_n(m) = \sup_{a\in\mathcal{A}_m}\left|\sum_{I\in m} a_I\nu_n[\varphi_I]\right| = \sup_{a\in\mathcal{A}_m}\left|\nu_n\left[\sum_{I\in m} a_I\varphi_I\right]\right|. \tag{44}$$

Moreover the same identity holds when replacing $\mathcal{A}_m$ by some countable and dense subset $\mathcal{A}'_m$ of $\mathcal{A}_m$, so that applying (33) to the countable set of functions

$$\left\{\sum_{I\in m} a_I\varphi_I, a \in \mathcal{A}'_m\right\},$$

we derive that for every positive x

$$\mathbb{P}\left[\sup_{a\in\mathcal{A}_m}\left|\nu_n\left[\sum_{I\in m} a_I\varphi_I\right]\right| \geqslant (1+\varepsilon)E_m + \sigma_m\sqrt{\frac{2\kappa x}{n}} + \kappa(\varepsilon)\frac{b_m}{n}x\right] \leqslant e^{-x}, \tag{45}$$

where

$$E_m = \mathbb{E}\left[\sup_{a\in\mathcal{A}_m}\left|\nu_n\left[\sum_{I\in m} a_I\varphi_I\right]\right|\right] \leqslant \frac{\mathbb{E}(\chi_n(m))}{\sqrt{n}} \leqslant \sqrt{\frac{D_m}{n}},$$

$$\sigma_m^2 = \sup_{a\in\mathcal{A}_m}\operatorname{Var}\left[\sum_{I\in m} a_I\varphi_I(\xi_1)\right] \leqslant \sup_{a\in\mathcal{S}_m}\mathbb{E}\left[\sum_{I\in m} a_I\varphi_I(\xi_1)\right]^2 \leqslant 1,$$

and

$$b_m = \sup_{a\in\mathcal{A}_m}\left\|\sum_{I\in m} a_I\varphi_I\right\|_\infty = \sup_{a\in\mathcal{A}_m}\sup_{m\in I}\frac{|a_I|}{\sqrt{P(I)}} \leqslant \frac{\xi}{z}.$$

Hence we get by (44) and (45),

$$\mathbb{P}\left[\chi_n(m)\,\mathbb{1}_{\Omega_m(\varepsilon)\cap\{\chi_n(m)\geqslant z\}} \geqslant (1+\varepsilon)\sqrt{D_m} + \sqrt{2\kappa x} + \kappa(\varepsilon)\frac{\xi}{\sqrt{nz}}x\right] \leqslant e^{-x}.$$

If we now choose $z = \sqrt{2\kappa x/n}$ and take into account the definition of ξ, we get

$$\mathbb{P}\left[\chi_n(m)\,\mathbb{1}_{\Omega_m(\varepsilon)} \geqslant (1+\varepsilon)\left(\sqrt{D_m} + \sqrt{2\kappa x}\right)\right] \leqslant e^{-x}. \qquad \square$$

REMARK 1. — *It is well known that given some partition m of $[0,1]$, when $n \to +\infty$, $\chi_n(m)$ converges in distribution to $\|Y\|$, where Y is a standard Gaussian vector in $\mathbb{R}^{D_m}$. An easy exercise consists in deriving from (11) a tail bound for the chi-square distribution with D_m degrees of freedom. Indeed by Cauchy-Schwarz inequality $\mathbb{E}[\|Y\|] \leqslant \sqrt{D_m}$ and therefore*

$$\mathbb{P}\left[\|Y\| \geqslant \sqrt{D_m} + \sqrt{2x}\right] \leqslant e^{-x}. \tag{46}$$

One can see that (43) is very close to (46), especially if one has in mind that a reasonable conjecture about the constant κ which comes from Theorem 2.4 is that κ should be equal to 1 instead of 4.

We are now in a position to control uniformly a collection of square roots of chi-square statistics $\{\chi_n(m),\, m \in \mathcal{M}_n\}$, under the following mild restriction on the collection of partitions $\mathcal{M}_n$.

(H_0) : *Let N be some integer such that $N \leqslant n(\log(n))^{-2}$ and m_N be a partition of $[0,1]$ the elements of which are intervals with equal length $(N+1)^{-1}$. We assume that every element of any partition m belonging to $\mathcal{M}_n$ is the union of pieces of m_N.*

Assume that **(H_0)** holds. Given $\eta \in (0,1)$, setting

$$\Omega(\eta) = \{|P_n(I) - P(I)| \leqslant \eta P(I), \text{ for every } I \in m_N\} \tag{47}$$

one has

$$\Omega(\eta) \subset \bigcap_{m\in\mathcal{M}_n} \{|P_n(I) - P(I)| \leqslant \eta P(I), \text{ for every } I \in m\}.$$

Therefore, given some arbitrary family of positive numbers $(y_m)_{m\in\mathcal{M}_n}$, provided that $\eta \leqslant 2\kappa\varepsilon/\kappa(\varepsilon)$, we derive from (43) that

$$\mathbb{P}\left[\bigcup_{m\in\mathcal{M}_n}\left\{\chi_n(m)\mathbb{1}_{\Omega(\eta)} \geqslant (1+\varepsilon)\left(\sqrt{D_m} + \sqrt{2\kappa y_m}\right)\right\}\right] \leqslant \sum_{m\in\mathcal{M}_n} e^{-y_m}. \tag{48}$$

This inequality is the required tool to evaluate the penalty function and establish a risk bound for the corresponding penalized maximum likelihood estimator.

3.2. A model selection result

Another advantage brought by the restriction to $\Omega(\eta)$ when assuming that $(\mathbf{H_0})$ holds, is that for every $m \in \mathcal{M}_n$, the ratios $\widehat{s}_m/s_m$ remain bounded on this set, which implies that $V^2(\widehat{s}_m, s_m)$ is of the order of $K(s_m, \hat{s}_m)$. More precisely, on the set $\Omega(\eta)$, one has $P_n(I) \geqslant (1-\eta) P(I)$ for every $I \in \mathcal{M}_n$ and therefore

$$\widehat{s}_m \geqslant (1-\eta) s_m,$$

which implies by Lemma 5.3 (see the Appendix) that

$$K(s_m, \hat{s}_m) \geqslant \frac{1-\eta}{2} \int s_m \log^2\left(\frac{\hat{s}_m}{s_m}\right) d\mu.$$

Since $\log^2(\hat{s}_m/s_m)$ is piecewise constant on the partition m

$$\int s_m \log^2\left(\frac{\hat{s}_m}{s_m}\right) d\mu = \int s \log^2\left(\frac{\hat{s}_m}{s_m}\right) d\mu = V^2(\widehat{s}_m, s_m),$$

and therefore, for every $m \in \mathcal{M}_n$

$$K(s_m, \hat{s}_m) \mathbb{1}_{\Omega(\eta)} \geqslant \frac{1-\eta}{2} V^2(\widehat{s}_m, s_m) \mathbb{1}_{\Omega(\eta)}. \tag{49}$$

This allows to better understand the structure of the proof of Theorem 3.2 below. Indeed, provided that

$$\eta \leqslant \frac{\varepsilon}{1+\varepsilon}, \tag{50}$$

one derives from (42) and (49) that on the set $\Omega(\eta)$,

$$\begin{aligned} K(s, \tilde{s}) \leqslant\ & K(s, s_m) + \mathrm{pen}_n(m) + \nu_n\left(\log \frac{s}{s_m}\right) \\ & + n^{-1/2}\sqrt{2(1+\varepsilon) K(s_{\hat{m}}, \hat{s}_{\hat{m}})}\chi_n(\widehat{m}) + \nu_n\left(\log \frac{s_{\hat{m}}}{s}\right) \\ & - \mathrm{pen}_n(\widehat{m}). \end{aligned}$$

Now, by (5) $K(s, \tilde{s}) = K(s, s_{\hat{m}}) + K(s_{\hat{m}}, \tilde{s})$, hence, taking into account that

$$n^{-1/2}\sqrt{2(1+\varepsilon) K(s_{\hat{m}}, \hat{s}_{\hat{m}})}\chi_n(\widehat{m}) \leqslant (1+\varepsilon)^{-1} K(s_{\hat{m}}, \hat{s}_{\hat{m}}) + \frac{\chi_n^2(\widehat{m})(1+\varepsilon)^2}{2n},$$

one derives that on the set $\Omega(\eta)$,

$$K(s, s_{\hat{m}}) + \frac{\varepsilon}{1+\varepsilon} K(s_{\hat{m}}, \tilde{s}) \leqslant K(s, s_m) + \mathrm{pen}_n(m) + \nu_n\left(\log \frac{s}{s_m}\right) + \frac{\chi_n^2(\widehat{m})(1+\varepsilon)^2}{2n} + \nu_n\left(\log \frac{s_{\hat{m}}}{s}\right) \quad (51)$$

$$-\mathrm{pen}_n(\hat{m}). \quad (52)$$

Neglecting the terms $\nu_n(\log s/s_m)$ and $\nu_n(\log s_{\hat{m}}/s)$, we see that the penalty $\mathrm{pen}_n(\hat{m})$ should be large enough to compensate $\chi_n^2(\widehat{m})(1+\varepsilon)^2/2n$ with high probability. Since we have at our disposal the appropriate exponential bound to control chi-square statistics uniformly over the family of partitions $\mathcal{M}_n$, it remains to control

$$\nu_n[\log(s/s_m)] + \nu_n[\log(s_{\hat{m}}/s)].$$

The trouble is that there is no way to warrant that the ratios $s_{\hat{m}}/s$ remain bounded except by making some extra unpleasant preliminary assumption on s. This makes delicate the control of $\nu_n[\log(s_{\hat{m}}/s)]$ as a function of $K(s, s_{\hat{m}})$ as one should expect. This is the reason why we shall rather pass to the control of Hellinger loss rather than Kullback-Leibler loss.

Let us recall that the Hellinger distance $h(f,g)$ between two densities f and g on $[0,1]$ is defined by

$$h^2(f,g) = \frac{1}{2}\int \left(\sqrt{f} - \sqrt{g}\right)^2.$$

It is known that

$$2h^2(f,g) \leqslant K(f,g), \quad (53)$$

and that a converse inequality exists whenever $\|\log(f/g)\|_\infty < \infty$. This in some sense confirms that it is slightly easier (although very close by essence) to control Hellinger risk as compared to Kullback-Leibler risk. The following result is due to Castellan (it is in fact a particular case of Theorem 3.2. in [23]).

THEOREM 3.2.— *Let $\xi_1, ..., \xi_n$ be some independent $[0,1]$-valued random variables with common distribution $P = s\mu$, where μ denotes the Lebesgue measure. Consider a finite family $\mathcal{M}_n$ of partitions of $[0,1]$ satisfying to assumption $(\boldsymbol{H}_0)$. Let, for every partition m*

$$\hat{s}_m = \sum_{I \in m} \frac{P_n(I)}{\mu(I)} 1\!\!1_I \quad \textit{and} \quad s_m = \sum_{I \in m} \frac{P(I)}{\mu(I)} 1\!\!1_I$$

be respectively the histogram estimator and the histogram projection of s, based on m. Consider some absolute constant Σ and some family of non-negative weights $(x_m)_{m\in\mathcal{M}_n}$ such that

$$\sum_{m\in\mathcal{M}_n} e^{-x_m} \leqslant \Sigma. \tag{54}$$

Let $c_1 > 1/2$ and $c_2 = 2\left(\kappa + c_1^{-1}\right)$ ($\kappa = 4$ works) and consider some penalty function $\mathrm{pen}_n(.) : \mathcal{M}_n \to \mathbb{R}_+$ such that

$$\mathrm{pen}_n(m) \geqslant \frac{c_1}{n}\left(\sqrt{D_m} + \sqrt{c_2 x_m}\right)^2, \quad \text{for all } m \in \mathcal{M}_n,$$

where $D_m + 1$ denotes the number of elements of partition m. Let $\widehat{m}$ minimizing the penalized likelihood criterion

$$-\int \hat{s}_m \log(\hat{s}_m)\, d\mu + \mathrm{pen}_n(m)$$

over $m \in \mathcal{M}_n$ and define the penalized maximum likelihood estimator by $\widetilde{s} = s_{\widehat{m}}$. If $\operatorname{essinf}\{s(x),\ x \in [0,1]\} \geqslant \rho > 0$ is positive and $\int s(\log s)^2\, d\mu \leqslant L < \infty$, then for some constant $C(c_1, \rho, L, \Sigma)$,

$$\mathbb{E}\left[h^2(s,\tilde{s})\right] \leqslant \frac{(2c_1)^{1/5}}{(2c_1)^{1/5} - 1} \inf_{m\in\mathcal{M}_n}\left\{K(s, s_m) + \mathrm{pen}_n(m)\right\} + \frac{C(c_1,\rho,L,\Sigma)}{n}.$$

Proof. — Let $\xi > 0$ be given, $\varepsilon > 0$ to be chosen later and

$$\eta = \frac{\varepsilon}{1+\varepsilon} \wedge \frac{2\kappa\varepsilon}{\kappa(\varepsilon)}.$$

Hellinger distance will appear naturally in the above analysis of the Kullback-Leibler risk through the control of $\nu_n\left[\log\left(s_{\widehat{m}}/s\right)\right]$ which can be performed via Proposition 5.4 of the Appendix. Indeed, one has

$$\mathbb{P}\left[\bigcup_{m\in\mathcal{M}_n}\left\{\nu_n\left[\log\left(\frac{s_m}{s}\right)\right] \geqslant K(s,s_m) - 2h^2(s,s_m) + 2\frac{y_m}{n}\right\}\right] \tag{55}$$
$$\leqslant \sum_{m\in\mathcal{M}_n} e^{-y_m},$$

which means that, except on a set of probability less than $\sum_{m\in\mathcal{M}_n} e^{-y_m}$, the following inequality is valid:

$$\nu_n\left[\log(s_{\hat{m}}) - \log(s)\right] \leqslant K(s, s_{\hat{m}}) - 2h^2(s, s_{\hat{m}}) + 2\frac{y_{\hat{m}}}{n}.$$

Let $\Omega(\eta)$ be defined by (47). Setting for every $m \in \mathcal{M}_n$, $y_m = x_m + \xi$, since (50) holds because of our choice of η, it comes from (51) and (55) that on the set $\Omega(\eta)$ and except on a set with probability less than $\Sigma e^{-\xi}$, one has for every $m \in \mathcal{M}_n$

$$\begin{aligned} K(s, s_{\hat{m}}) + \frac{\varepsilon}{1+\varepsilon} K(s_{\hat{m}}, \tilde{s}) \leqslant\ & K(s, s_m) + \mathrm{pen}_n(m) + \nu_n\left(\log\left(\frac{s}{s_m}\right)\right) \\ & + \frac{\chi_n^2(\widehat{m})(1+\varepsilon)^2}{2n} + K(s, s_{\hat{m}}) \\ & - 2h^2(s, s_{\hat{m}}) + 2\frac{x_{\hat{m}} + \xi}{n} - \mathrm{pen}_n(\hat{m}). \end{aligned}$$

Equivalently

$$\begin{aligned} 2h^2(s, s_{\hat{m}}) + \frac{\varepsilon}{1+\varepsilon} K(s_{\hat{m}}, \tilde{s}) \leqslant\ & K(s, s_m) + \mathrm{pen}_n(m) \\ & + \nu_n\left(\log\left(\frac{s}{s_m}\right)\right) + \frac{(1+\varepsilon)^2 \chi_n^2(\widehat{m})}{2n} \\ & + 2\frac{x_{\hat{m}} + \xi}{n} - \mathrm{pen}_n(\hat{m}). \end{aligned}$$

Now, by the triangle inequality,

$$h^2(s, \tilde{s}) \leqslant 2h^2(s, s_{\hat{m}}) + 2h^2(s_{\hat{m}}, \tilde{s}).$$

Hence, using (53), we derive that on $\Omega(\eta)$ and except on a set with probability less than $\Sigma e^{-\xi}$, the following inequality holds:

$$\begin{aligned} \frac{\varepsilon}{1+\varepsilon} h^2(s, \tilde{s}) \leqslant\ & K(s, s_m) + \mathrm{pen}_n(m) + \nu_n\left(\log\left(\frac{s}{s_m}\right)\right) \\ & + \frac{(1+\varepsilon)^2 \chi_n^2(\widehat{m})}{2n} + 2\frac{x_{\hat{m}} + \xi}{n} - \mathrm{pen}_n(\hat{m}). \end{aligned} \tag{56}$$

Now we can use the above uniform control of chi-square statistics and derive from (48) that on the set $\Omega(\eta)$ and except on a set with probability less than $\Sigma e^{-\xi}$

$$\begin{aligned} \chi_n^2(\widehat{m}) &\leqslant (1+\varepsilon)^2 \left(\sqrt{D_{\hat{m}}} + \sqrt{2\kappa(x_{\hat{m}} + \xi)}\right)^2 \\ &\leqslant (1+\varepsilon)^2 \left[(1+\varepsilon)\left(\sqrt{D_{\hat{m}}} + \sqrt{2\kappa x_{\hat{m}}}\right)^2 + 2\kappa\xi\left(1 + \varepsilon^{-1}\right)\right]. \end{aligned}$$

Plugging this inequality in (56) implies that on the set $\Omega(\eta)$ and except on a set with probability less than $2\Sigma e^{-\xi}$,

$$\frac{\varepsilon}{1+\varepsilon} h^2(s, \tilde{s}) \leqslant K(s, s_m) + \mathrm{pen}_n(m) + \nu_n\left(\log\left(\frac{s}{s_m}\right)\right)$$

$$+\frac{(1+\varepsilon)^5}{2n}\left(\sqrt{D_{\hat{m}}}+\sqrt{2\kappa x_{\hat{m}}}\right)^2+2\frac{x_{\hat{m}}}{n}-\mathrm{pen}_n(\hat{m})$$
$$+\frac{\xi}{n}\left(\kappa\varepsilon^{-1}(1+\varepsilon)^5+2\right).$$

Now we can notice that choosing ε adequately, i.e. such that $c_1=(1+\varepsilon)^5/2$ ensures that

$$\frac{(1+\varepsilon)^5}{2n}\left(\sqrt{D_{\hat{m}}}+\sqrt{2\kappa x_{\hat{m}}}\right)^2+2\frac{x_{\hat{m}}}{n}-\mathrm{pen}_n(\hat{m})\leqslant 0.$$

Hence, except on a set of probability less than $2\Sigma e^{-\xi}$, the following inequality is available: □

$$\frac{\varepsilon}{1+\varepsilon}h^2(s,\tilde{s})\,\mathbb{1}_{\Omega(\eta)} \leqslant K(s,s_m)+\mathrm{pen}_n(m)+\nu_n\left(\log\left(\frac{s}{s_m}\right)\right)\mathbb{1}_{\Omega(\eta)}$$
$$+\frac{\xi}{n}\left(\kappa\varepsilon^{-1}(1+\varepsilon)^5+2\right).$$

Integrating this inequality with respect to ξ leads to

$$\frac{\varepsilon}{1+\varepsilon}\mathbb{E}\left[h^2(s,\tilde{s})\,\mathbb{1}_{\Omega(\eta)}\right] \leqslant K(s,s_m)+\mathrm{pen}_n(m)$$
$$+\mathbb{E}\left[\nu_n\left(\log\left(\frac{s}{s_m}\right)\right)\mathbb{1}_{\Omega(\eta)}\right]$$
$$+\frac{2\Sigma}{n}\left(\kappa\varepsilon^{-1}(1+\varepsilon)^5+2\right).$$

Since $\nu_n(\log(s/s_m))$is centered at expectation and the Hellinger distance is bounded by 1, it follows from the above inequality that

$$\frac{\varepsilon}{1+\varepsilon}\mathbb{E}\left[h^2(s,\tilde{s})\right] \leqslant K(s,s_m)+\mathrm{pen}_n(m)+\frac{2\Sigma}{n}\left(\kappa\varepsilon^{-1}(1+\varepsilon)^5+2\right)$$
$$+\mathbb{E}\left[\left(-\nu_n\left(\log\left(\frac{s}{s_m}\right)\right)+1\right)\mathbb{1}_{\Omega^c(\eta)}\right]. \qquad (57)$$

It remains to bound the last term of the righthand side of the above inequality. By Cauchy-Schwarz inequality

$$\mathbb{E}\left[-\nu_n\left(\log\left(\frac{s}{s_m}\right)\right)\mathbb{1}_{\Omega^c(\eta)}\right]\leqslant\left[\frac{1}{n}\int s\left(\log\left(\frac{s}{s_m}\right)\right)^2 d\mu\right]^{\frac{1}{2}}\left(\mathbb{P}\left[\Omega^c(\eta)\right]\right)^{\frac{1}{2}},$$

and

$$\int s\left(\log\frac{s}{s_m}\right)^2 d\mu \leqslant 2\left[\int s(\log s)^2\,d\mu+\int s(\log s_m)^2\,d\mu\right]$$
$$\leqslant 2\left[\int s(\log s)^2\,d\mu+\left(\log\left(\frac{1}{\rho}\vee n\right)\right)^2\right],$$

since $\rho \leqslant s_m \leqslant n$. Moreover, setting $\delta = \inf_{I \in m_N} P(I)$ it follows from Bernstein's inequality that

$$\mathbb{P}\left[\Omega^c(\eta)\right] \leqslant 2(N+1)\exp\left(-\frac{n\eta^2\delta}{2(1+\eta/3)}\right)$$

yielding, because of the restriction $N+1 \leqslant n(\log(n))^{-2}$ (see $(\mathbf{H}_0)$),

$$\mathbb{P}\left[\Omega^c(\eta)\right] \leqslant 2n\exp\left(-\frac{\eta^2\rho(\log(n))^2}{2(1+\eta/3)}\right).$$

This shows that, as a function of n, $\mathbb{P}\left[\Omega^c(\eta)\right]$ tends to 0 faster than any power of n. Collecting the above inequalities and plugging them into (57) finishes the proof of the theorem. □

Theorem 3.2 suggests to take a penalty function of the form:

$$\mathrm{pen}_n(m) = \frac{c_1}{n}\left(\sqrt{D_m} + \sqrt{c_2 x_m}\right)^2,$$

where the weights x_m satisfy (54) and, of course, the constant c_1 and c_2 are independent of the density s. The choice $c_1 > 1/2$ provides an upper bound for the Hellinger risk of the penalized maximum likelihood estimator:

$$\mathbb{E}\left[h^2(s,\tilde{s})\right] \leqslant C_1 \inf_{m \in \mathcal{M}_n}\left\{K(s,s_m) + \mathrm{pen}_n(m)\right\} + \frac{C_2}{n} \tag{58}$$

where the constant C_1 does not depend on s whereas the constant C_2 depends on s (via ρ and L) and on the family of models (via Σ). Furthermore, the constant C_1, which depends only on c_1, converges to infinity when c_1 tends to $1/2$. This suggests that on the one hand c_1 should be chosen substantially larger than $1/2$ and on the other hand that one could get into trouble when choosing $c_1 < 1/2$. Using further refinements of the above method, it is proved in [23] that the special choice $c_1 = 1$ optimizes the risk bound (58). Moreover, following Castellan in [23], we shall show below, as a consequence of the exponential inequalities for chi-square statistics of Section 2 (both on the left and the right tails), that one cannot dispense from the condition $c_1 > 1/2$, at least for the case example of regular histograms.

3.3. Choice of the weights $\{x_m, m \in \mathcal{M}_n\}$

The penalty function depends on the family $\mathcal{M}_n$ through the choice of the weights x_m satisfying (54). A reasonable way of choosing those weights is to make them depend on m only through the dimension D_m. More precisely, we are interested in weights of the form

$$x_m = L(D_m)\, D_m.$$

With such a definition the number of histogram models S_m having the same dimension plays a fundamental role for bounding the series (54) and therefore to decide what value of $L(D)$ should be taken in order to get a reasonable value for Σ. Let us consider two extreme examples.

- **Case of regular histograms.** Let J be the largest integer such that 2^J is not larger than $n(\log(n))^{-2}$. Let $\mathcal{M}_J^r$ be the collection of regular partitions with 2^j pieces with $j \leqslant J$. Then assumption $(\mathbf{H}_0)$ is satisfied and since there is only one model per dimension, $L(D)$ can be taken as some arbitray positive constant η and

$$\sum_{m \in \mathcal{M}_J^r} e^{-\eta D_m} \leqslant \sum_{j=0}^{\infty} e^{-\eta 2^j} \leqslant \eta^{-1}.$$

  Consequently, all penalties of the form

$$\mathrm{pen}_n(m) = c\frac{D_m}{n},$$

  with $c > 1/2$ are allowed, including that of Akaike, namely $c = 1$. Since $K(s, s_m) + D_m/2$ represents actually the order of the Kullback-Leibler risk of the histogram estimator $\hat{s}_m$ (see ([23])), the meaning of (58) is that, up to constant, $\widetilde{s}$ behaves like an oracle. This is not exactly true in terms of the Kullback loss since we have bounded the Hellinger risk instead of the Kullback-Leibler risk. However when the log-ratios $\log(s/s_m)$ remain uniformly bounded, then the Kullback bias $K(s, s_m)$ is of the order of $h^2(s, s_m)$ and (58) can be interpreted as an oracle inequality for the Hellinger loss. It should be noticed that the statement of the Theorem provides some flexibility concerning the choice of the penalty function so that we could take as well

$$\mathrm{pen}_n(m) = c\frac{D_m}{n} + c'\frac{D_m^\alpha}{n}$$

  for some $\alpha \in (0,1)$. As already mentionned, the choice $c = 1$ can be shown to optimize the risk bound for the corresponding penalized estimator and the structure of the proof made in ([23]) tends to indicate that it would be desirable to choose a penalty function which is slightly heavier than what is proposed in Akaike's criterion. This is indeed confirmed by simulations in [13], the gain being especially spectacular for small or moderate values of the sample size n (we mean less than 200).

- **Case of irregular histograms.** We consider here the family $\mathcal{M}_N^{ir}$ of all partitions built from a single regular partition m_N with $N+1$

pieces where N is less than $n\left(\log\left(n\right)\right)^{-2}$. Then the cardinality of the family of partitions belonging to $\mathcal{M}_N^{ir}$ with a number of pieces equal to $D+1$ is bounded by $\binom{N}{D}$. Hence

$$\sum_{m\in\mathcal{M}_N^{ir}} e^{-x_m} \leqslant \sum_{D=1}^{N} \binom{N}{D} e^{-L(D)D} \leqslant \sum_{D\geqslant 1} \left(\frac{eN}{D}\right)^{D} e^{-L(D)D},$$

and the choice $L\left(D\right) = L + \log\left(eN/D\right)$ implies that condition (54) holds with $\Sigma = \left(e^{L}-1\right)^{-1}$. This leads to a penalty function of the form

$$\mathrm{pen}_n\left(m\right) = c\frac{D_m}{n}\log\left(\frac{N}{D_m}\right) + c'\frac{D_m}{n},$$

for large enough constants c and c'. The corresponding risk bound can be written as:

$$\mathbb{E}\left[h^2\left(s,\tilde{s}\right)\right] \leqslant C\left[\inf_{D\leqslant N}\inf_{\mathcal{M}_N^{ir}(D)}\left\{K\left(s,s_m\right) + \frac{D}{n}\left(1+\log\left(\frac{N}{D}\right)\right)\right\}\right]$$

where $\mathcal{M}_N^{ir}\left(D\right)$ denotes the set of partitions m with dimension $D_m = D$. This means that, given some integer D, whenever s belongs to $\mathcal{S}_D = \cup_{m\in\mathcal{M}_N^{ir}(D)} S_m$, the Hellinger risk of s is bounded by $CD/n\left(1+\log\left(N/D\right)\right)$. This shows that, because of the extra logarithmic factor, the penalized estimator fails to mimic the oracle in terms of Hellinger loss. One can wonder whether this is due to a weakness of the method or not. The necessity of this extra logarithmic factor is proved in [9] (see Proposition 2 therein) where the minimax risk over the set $\mathcal{S}_D$ is shown to be bounded from below by $D/n\left(1+\log\left(N/D\right)\right)$, up to some constant. In this sense the above risk bound is optimal.

3.4. Lower bound for the penalty function

One can also wonder whether the condition $c_1 > 1/2$ in Theorem 3.2 is necessary or not. We cannot answer this question in full generality. The following result shows that, when there are only a few models per dimension, taking $\mathrm{pen}_n\left(m\right) = cD_m/n$ for some arbitrary constant $c < 1/2$ leads to a disaster in the sense that, if the true s is uniform, the penalized maximum likelihood selection criterion will choose models of large dimension with high probability and the Hellinger risk will be bounded away from 0 when n goes to infinity. The proof of this result heavily relies on the inequalities for the right and also for the left tails of chi-square statistics established in Section 2 (namely (38) and (40)). The proof being quite similar to that of Theorem

3.2, we skip it and refer the interested reader to [23] (and also to [11]) where a similar result is proved in the Gaussian framework).

THEOREM 3.3. — *Let $\xi_1, ..., \xi_n$ be some independent $[0,1]$-valued random variables with common distribution $P = s\mu$ with $s = 1\!\!1_{[0,1]}$. Consider some finite family of partitions $\mathcal{M}_n$ such that for each integer D, there exists only one partition m such that $D_m = D$. Moreover, let us assume that $\mu(I) \geqslant (\log(n))^2 / n$ for every $I \in m$ and $m \in \mathcal{M}_n$.*

Assume that for some partition $m_N \in \mathcal{M}_n$ with $N+1$ pieces one has

$$\mathrm{pen}_n(m_N) = c\frac{N}{n}$$

with $c < 1/2$. Let $\widehat{m}$ be the minimizer over $\mathcal{M}_n$ of the penalized criterion

$$-\int \widehat{s}_m \log(\widehat{s}_m) + \mathrm{pen}_n(m).$$

Then, whatever the values of $\mathrm{pen}_n(m)$ for $m \neq m_N$ there exist positive numbers N_0 and L, depending only on c, such that, for all $N \geqslant N_0$,

$$\mathbb{P}\left(D_{\widehat{m}} \geqslant \frac{1-4c^2}{4}N\right) \geqslant 1 - \beta(c),$$

where

$$\beta(c) = \Sigma(L)\exp\left[-\frac{L}{2}\left(\left(1-4c^2\right)N\right)^{\frac{1}{2}}\right] + \frac{C(c)}{n} \text{ with } \Sigma(L) = \sum_{D\geqslant 1} e^{-L\sqrt{D}}.$$

Moreover, if $\tilde{s} = \widehat{s}_{\widehat{m}}$,

$$\mathbb{E}\left[K(s,\tilde{s})\right] \geqslant \delta(c)\left[1-\beta(c)\right]\frac{N}{n}$$

where $\delta(c) = (1-2c)(1+2c)^2/16$.

4. Model selection and statistical learning

The purpose of this section is to provide general model selection theorems for bounded contrast functions. The proofs will heavily rely on the concentration inequalities for empirical processes recalled in Section 2. First we shall use the Hoeffding type inequalities to propose an other look at the celebrated *Vapnik structural minimization of the risk method* (initiated in [40]). Then, we shall present a new result, based on the Bernstein type inequalities of Section 2. We shall apply this result to improve on the risk bounds derived from Vapnik's method for the pattern recognition problem. We shall also recover some of the results given in [7].

4.1. A first model selection theorem for bounded contrast functions

Let us first see what can be derived from (9) by using only the following boundedness assumption on the contrast function γ

A1 There exists some absolute constant $b > 0$ such that, for every t belonging to some set $\mathcal{S}$, one has for some function $a(t)$, $a(t) \leqslant \gamma(t,.) \leqslant a(t) + b$.

In order to avoid any measurability problem, let us first assume that each of the models S_m is countable. Given some constant Σ, let us consider some preliminary collection of nonnegative weights $(x_m)_{m \in \mathcal{M}_n}$ such that

$$\sum_{m \in \mathcal{M}_n} e^{-x_m} \leqslant \Sigma$$

and let $\xi > 0$ be given. It follows from Mac Diarmid's Inequality (see (17) above) that for every $m' \in \mathcal{M}_n$,

$$\mathbb{P}\left[\sup_{t \in S_{m'}} (-\overline{\gamma}_n(t)) \geqslant \mathbb{E}\left[\sup_{t \in S_{m'}} (-\overline{\gamma}_n(t))\right] + b\sqrt{\frac{x_{m'} + \xi}{2n}}\right] \leqslant e^{-x_{m'} - \xi},$$

and therefore, setting $\mathbb{E}\left[\sup_{t \in S_{m'}} (-\overline{\gamma}_n(t))\right] = E_{m'}$, except on a set of probability not larger than $\Sigma e^{-\xi}$, one has for every $m' \in \mathcal{M}_n$,

$$\sup_{t \in S_{m'}} (-\overline{\gamma}_n(t)) \leqslant E_{m'} + b\sqrt{\frac{x_{m'} + \xi}{2n}}.$$

Hence, (9) implies that the following inequality holds, except on a set of probability not larger than $\Sigma e^{-\xi}$:

$$\begin{aligned} l(s, \widetilde{s}) \leqslant\ & l(s, s_m) + \overline{\gamma}_n(s_m) + E_{\widehat{m}} + b\sqrt{\frac{x_{\widehat{m}}}{2n}} - \mathrm{pen}_n(\widehat{m}) + \mathrm{pen}_n(m) \\ & + b\sqrt{\frac{\xi}{2n}} + \rho_n. \end{aligned} \tag{59}$$

It is tempting to choose $\mathrm{pen}_n(m') = E_{m'} + b\sqrt{x_{m'}/2n}$ for every $m' \in \mathcal{M}_n$ but we should not forget that $E_{m'}$ typically depends on the unknown s. Thus, we are forced to consider some upper bound $\widetilde{E}_{m'}$ of $E_{m'}$ which does not depend on s. This upper bound can be either deterministic (we shall discuss below the drawbacks of this strategy) or random and in such a case we shall take benefit of the fact that it is enough to assume that $\widetilde{E}_{m'} \geqslant E_{m'}$

holds on a set with sufficiently high probability. More precisely, assuming that for some constant K and for every $m' \in \mathcal{M}_n$

$$\mathrm{pen}_n(m') \geqslant E_{m'} + b\sqrt{\frac{x_{m'}}{2n}} - K\sqrt{\frac{\xi}{2n}} \tag{60}$$

holds, except on set of probability not larger than $\exp(-x_{m'} - \xi)$, we derive from (59) and (60) that

$$l(s,\widetilde{s}) \leqslant l(s,s_m) + \overline{\gamma}_n(s_m) + \mathrm{pen}_n(m) + (b+K)\sqrt{\frac{\xi}{2n}} + \rho_n$$

holds except on a set of probability not larger than $2\Sigma e^{-\xi}$. Thus, integrating with respect to ξ leads to

$$\mathbb{E}\left[\left(l(s,\widetilde{s}) - l(s,s_m) - \overline{\gamma}_n(s_m) - \mathrm{pen}_n(m) - \rho_n\right)^+\right] \leqslant \Sigma(b+K)\sqrt{\frac{\pi}{2n}}$$

and therefore, since $\overline{\gamma}_n(s_m)$ is centered at expectation

$$\mathbb{E}[l(s,\widetilde{s})] \leqslant l(s,s_m) + \mathbb{E}[\mathrm{pen}_n(m)] + \Sigma(b+K)\sqrt{\frac{\pi}{2n}}.$$

Hence, we have proven the following result.

THEOREM 4.1. — *Let $\xi_1, \ldots, \xi_n$ be independent observations taking their values in some measurable space Ξ and with common distribution P depending on some unknown parameter $s \in \mathcal{S}$. Let $\gamma : \mathcal{S} \times \Xi \to \mathbb{R}$ be some contrast function satisfying assumption **A1**. Let $(S_m)_{m \in \mathcal{M}_n}$ be some at most countable collection of countable subsets of $\mathcal{S}$ and $\rho_n \geqslant 0$ be given. Consider some absolute constant Σ, some family of nonnegative weights $(x_m)_{m \in \mathcal{M}_n}$ such that*

$$\sum_{m \in \mathcal{M}_n} e^{-x_m} \leqslant \Sigma$$

and some (possibly data-dependent) penalty function $\mathrm{pen}_n : \mathcal{M}_n \to \mathbb{R}_+$. Let $\widetilde{s}$ be a ρ_n-minimum penalized contrast estimator of s as defined by (4). Then, if for some nonnegative constant K, for every $m \in \mathcal{M}_n$ and every positive ξ

$$\mathrm{pen}_n(m) \geqslant \mathbb{E}\left[\sup_{t \in S_m}\left(-\overline{\gamma}_n(t)\right)\right] + b\sqrt{\frac{x_m}{2n}} - K\sqrt{\frac{\xi}{2n}}$$

holds with probability larger than $1 - \exp(-x_m - \xi)$, the following risk bound holds for all $s \in \mathcal{S}$

$$\mathbb{E}[l(s,\widetilde{s})] \leqslant \left[\inf_{m \in \mathcal{M}_n}\left(l(s,S_m) + \mathbb{E}[\mathrm{pen}_n(m)]\right) + \Sigma(b+K)\sqrt{\frac{\pi}{2n}} + \rho_n\right], \tag{61}$$

where l is defined by (1) and $l(s,S_m) = \inf_{t \in S_m} l(s,t)$.

It is not that easy to discuss whether this result is sharp or not in the generality where it is stated here. Nevertheless we shall see that, at the price of making an extra assumption on the contrast function γ, it is possible to improve on (61) by weakening the constraint on the penalty function. This will be the purpose of our next section.

4.1.1. *Vapnik's learning theory revisited*

We would like here to explain how Vapnik's *structural minimization of the risk method* (as described in [40] and further developed in [41]) fits in the above framework of penalized minimum contrast model selection. More precisely, we shall consider some *pattern classification* problem and show how to recover (or refine in the spirit of Boucheron, Lugosi and Massart in [16]) some of Vapnik's results from Theorem 4.1. The data $\xi_1 = (X_1, Y_1), ..., \xi_n = (X_n, Y_n)$ consist of independent, identically distributed copies of the random variable pair (X, Y) taking values in $\mathbb{R}^d \times \{0,1\}$. Let the models $(S_m)_{m \in \mathcal{M}_n}$ being defined for every $m \in \mathcal{M}_n$ as

$$S_m = \{\mathbb{1}_C : C \in \mathcal{C}_m\},$$

where $\mathcal{C}_m$ is some countable class of subsets of $\mathbb{R}^d$. Let $\mathcal{S}$ be the set of measurable functions taking their values in $[0,1]$. In this case, the least squares contrast function fulfills condition **A1**. Indeed, since $\gamma(t,(x,y)) = (y - t(x))^2$, **A1** is fulfilled with $b = 1$ whenever $t \in \mathcal{S}$ and $y \in [0,1]$. For every $m \in \mathcal{M}_n$, a ρ_n-least squares estimator $\widehat{s}_m$ minimizes over $t \in S_m$, the quantity

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2 + \rho_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i \neq t(X_i)} + \rho_n.$$

Each estimator $\widehat{s}_m$ represents some possible classification rule and the purpose of model selection is here to select what classification rule is the best according to some risk minimization criterion. At this stage it should be noticed that we have the choice here between two different definitions of the statistical object of interest s. Indeed, we can take s to be the minimizer of $t \to \mathbb{E}[Y - t(X)]^2$ subject or not to the constraint that t takes its values in $\{0,1\}$. On the one hand the function $s^{(1)}$ defined for $\delta \in \{0,1\}$, as $s^{(1)}(x) = \delta$ if and only if $\mathbb{P}[Y = \delta \mid X = x] > 1/2$ is a minimizer of $\mathbb{E}[Y - t(X)]^2$ under the constraint that t takes its values in $\{0,1\}$. Then the loss function can be written as

$$l\left(s^{(1)}, t\right) = \mathbb{E}\left[s^{(1)}(X) - t(X)\right]^2 = \mathbb{P}[Y \neq t(X)] - \mathbb{P}[Y \neq s(X)].$$

On the other hand, if $s^{(2)}$ denotes the minimizer of $\mathbb{E}[Y - t(X)]^2$ without the constraint that t takes its values in $\{0,1\}$, then $s^{(2)}(x) = \mathbb{E}(Y \mid X = x)$

and $l\left(s^{(2)},t\right)=\mathbb{E}\left[s^{(2)}\left(X\right)-t\left(X\right)\right]^2$. It turns out that the results presented below are valid for both definitions of s simultaneously. In order to apply Theorem 4.1, it remains to majorize $\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]$. Let us introduce the (random) Vapnik-Chervonenkis entropy (VC-entropy) of $\mathcal{C}_m$

$$H_m=\log\left|\left\{C\cap\left\{X_1,...,X_n\right\},\,C\in\mathcal{C}_m\right\}\right|.$$

If we take some independent copy $(\xi_1',...,\xi_n')$ of $(\xi_1,...,\xi_n)$ and consider the corresponding copy γ_n' of γ_n, we can use the following standard symmetrization argument. By Jensen's inequality

$$\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]\leqslant\mathbb{E}\left[\sup_{t\in S_m}\left(\gamma_n'\left(t\right)-\gamma_n\left(t\right)\right)\right],$$

so that, given independent random signs $(\varepsilon_1,...,\varepsilon_n)$, independent of $(\xi_1,...,\xi_n)$, one has,

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]&\leqslant\frac{1}{n}\mathbb{E}\sup_{t\in S_m}\left[\sum_{i=1}^n\varepsilon_i\left(\mathbb{1}_{Y_i'\neq t(X_i')}-\mathbb{1}_{Y_i\neq t(X_i)}\right)\right]\\&\leqslant\frac{2}{n}\mathbb{E}\sup_{t\in S_m}\left[\sum_{i=1}^n\varepsilon_i\mathbb{1}_{Y_i\neq t(X_i)}\right].\end{aligned}$$

Hence, using Lemma 5.2 (presented in the Appendix), we get

$$\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]\leqslant\frac{2\sqrt{2}}{n}\mathbb{E}\left[H_m\sup_{t\in S_m}\left(\sum_{i=1}^n\mathbb{1}_{Y_i\neq t(X_i)}\right)\right]^{1/2},$$

and by Jensen's inequality

$$\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]\leqslant2\sqrt{\frac{2\mathbb{E}\left[H_m\right]}{n}}.\tag{62}$$

The trouble now is that $\mathbb{E}\left[H_m\right]$ is unknown. Two different strategies can be followed to overcome this difficulty. First, one can assume each $\mathcal{C}_m$ to be a VC-class with VC-dimension V_m, which provides a universal upper bound for H_m of the form ([40]):

$$\mathbb{E}\left[H_m\right]\leqslant V_m\left(1+\log\left(\frac{n}{V_m}\right)\right).\tag{63}$$

If $\mathcal{M}_n$ has cardinality not larger than n, one can take $x_m=\log\left(n\right)$ for each $m\in\mathcal{M}_n$ which leads to a penalty function of the form

$$\mathrm{pen}_n\left(m\right)=2\sqrt{\frac{2V_m\left(1+\log\left(n/V_m\right)\right)}{n}}+\sqrt{\frac{\log\left(n\right)}{2n}},$$

and to the following risk bound for the corresponding penalized estimator $\widetilde{s}$, since then one can take $\Sigma = 1$:

$$\mathbb{E}\left[l\left(s,\widetilde{s}\right)\right] \leqslant \inf_{m\in\mathcal{M}_n}\left(l\left(s,S_m\right)+\mathrm{pen}_n\left(m\right)\right)+\sqrt{\frac{\pi}{2n}}+\rho_n. \tag{64}$$

This approach has two main drawbacks:

- the VC-dimension of a given collection of sets is generally very difficult to compute or even to evaluate (this is especially true for cases of interest such as half algebraic subspaces of a given degree for instance);
- even if the VC-dimension is computable (in the case of affine half spaces of $\mathbb{R}^d$ for instance), inequality (63) is too pessimistic and it would be desirable to define a penalty function from a quantity which is much closer to $\mathbb{E}\left[H_m\right]$ than the right hand side of (63).

The second strategy consists (following ([16])) in substituting H_m to $\mathbb{E}\left[H_m\right]$ by using again a concentration argument. Indeed, by (29), for any positive ξ, one has $H_m \geqslant \mathbb{E}\left[H_m\right] - \sqrt{2\log\left(2\right)\mathbb{E}\left[H_m\right]\left(x_m+\xi\right)}$, on a set of probability not less than $1-\exp\left(-x_m-\xi\right)$. Hence, since

$$\sqrt{2\log\left(2\right)\mathbb{E}\left[H_m\right]\left(x_m+\xi\right)} \leqslant \frac{\mathbb{E}\left[H_m\right]}{2}+\log\left(2\right)\left(x_m+\xi\right),$$

we have on the same set,

$$\mathbb{E}\left[H_m\right] \leqslant 2H_m + 2\log\left(2\right)\left(x_m+\xi\right),$$

which, by (62), yields

$$\mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right] \leqslant 4\left(\sqrt{\frac{H_m}{n}}+\sqrt{\frac{\log\left(2\right)x_m}{n}}+\sqrt{\frac{\log\left(2\right)\xi}{n}}\right).$$

Taking $x_m = \log\left(n\right)$ as before leads to the following choice for the penalty function

$$\mathrm{pen}_n\left(m\right) = 4\sqrt{\frac{H_m}{n}}+4.1\sqrt{\frac{\log\left(n\right)}{n}},$$

which satisfies

$$\mathrm{pen}_n\left(m\right) \geqslant \mathbb{E}\left[\sup_{t\in S_m}\left(-\overline{\gamma}_n\left(t\right)\right)\right]+\sqrt{\frac{\log\left(n\right)}{2n}}-4\sqrt{\frac{\log\left(2\right)\xi}{n}}.$$

The corresponding risk bound can be written as

$$\mathbb{E}\left[l\left(s,\widetilde{s}\right)\right] \leqslant \left[\inf_{m\in\mathcal{M}_n}\left(l\left(s,S_m\right)+\mathbb{E}\left[\mathrm{pen}_n\left(m\right)\right]\right)+4\sqrt{\frac{\pi\log\left(2\right)}{n}}+\rho_n\right],$$

and therefore,

$$\mathbb{E}\left[l\left(s,\widehat{s}\right)\right] \leqslant \left[\inf_{m\in\mathcal{M}_n}\left(l\left(s,S_m\right)+4\sqrt{\frac{\mathbb{E}\left[H_m\right]}{n}}\right)+4.1\sqrt{\frac{\log\left(n\right)}{n}}+\frac{6}{\sqrt{n}}+\rho_n\right]. \tag{65}$$

Note that if we take $s=s^{(1)}$, denoting by L_t the probability of missclassification of the rule t, i.e. $L_t=\mathbb{P}\left[Y\neq t\left(X\right)\right]$, the risk bound (65) can also be written as

$$\mathbb{E}\left[L_{\widehat{s}}\right] \leqslant \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m} L_t+4\sqrt{\frac{\mathbb{E}\left[H_m\right]}{n}}\right)+4.1\sqrt{\frac{\log\left(n\right)}{n}}+\frac{6}{\sqrt{n}}+\rho_n,$$

which is may be a more standard way of expressing the performance of a learning pattern recognition method. Of course, if we follow the first strategy of penalization a similar bound can be derived from (64), namely

$$\begin{aligned}\mathbb{E}\left[L_{\widehat{s}}\right] \quad &\leqslant \quad \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m} L_t+2\sqrt{\frac{2V_m\left(1+\log\left(n/V_m\right)\right)}{n}}\right)\\ &\quad +\sqrt{\frac{\log\left(n\right)}{2n}}+\sqrt{\frac{\pi}{2n}}+\rho_n.\end{aligned}$$

4.2. A more advanced selection theorem for bounded contrast functions

In addition to the loss function l we shall need another way of measuring the closeness between the elements of $\mathcal{S}$ which is directly connected to the variance of the increments of $\overline{\gamma}_n$ and therefore will play an important role in the analysis of the fluctuations of $\overline{\gamma}_n$.

We shall use two assumptions on the contrast function γ. The first one is a boundedness assumption while the second one asserts that the functional $t\longrightarrow\mathbb{E}\left[\gamma\left(t,\xi_1\right)\right]$ behaves quadratically around its minimum s with respect to the "pseudo-distance" d, closely related to the variance of the contrast function.

A2 There exists some pseudo-distance d and some absolute constant c such that for every $t\in\mathcal{S}$ and $u\in\mathcal{S}$ $\mathrm{Var}\left[\gamma\left(t,\xi_1\right)-\gamma\left(u,\xi_1\right)\right]\leqslant d^2\left(u,t\right)$ and $d^2(s,t)\leqslant cl\left(s,t\right)$, where we recall that

$$l\left(s,t\right)=\left(\mathbb{E}\left[\gamma\left(t,\xi_1\right)\right]-\mathbb{E}\left[\gamma\left(s,\xi_1\right)\right]\right).$$

In order to prove our main result, the following lemma will be useful.

We are now in a position to state the main new result of this paper.

THEOREM 4.2. — *Let $\xi_1, ..., \xi_n$ be independent observations taking their values in some measurable space Ξ and with common distribution P depending on some unknown parameter $s \in \mathcal{S}$. Let $\gamma : \mathcal{S} \times \Xi \to \mathbb{R}$ satisfying to assumptions **A1** and **A2**. Let $(S_m)_{m \in \mathcal{M}_n}$ be some at most countable collection of countable subsets of $\mathcal{S}$ and $\rho_n \geqslant 0$ be given. For any positive number σ and any $u \in S_m$, let us define*

$$\mathcal{B}_m(u, \sigma) = \{t \in S_m : d(t, u) \leqslant \sigma\},$$

*where d is the pseudo-distance given by **A2**. We assume that for any $m \in \mathcal{M}_n$, there exists some continuous function ϕ_m mapping $\mathbb{R}_+$ onto $\mathbb{R}_+$ such that $\phi_m(0) = 0$, $\phi_m(x)/x$ is nonincreasing and*

$$\mathbb{E}\left[\sup_{t \in \mathcal{B}_m(u,\sigma)} |\overline{\gamma}_n(t) - \overline{\gamma}_n(u)|\right] \leqslant \phi_m(\sigma), \quad \text{for all } \sigma \geqslant \sigma_m, \tag{66}$$

where σ_m is the solution of the equation

$$\phi_m(x) = x^2, \quad x > 0.$$

Consider some constant Σ, some family of nonnegative weights $(x_m)_{m \in \mathcal{M}}$ such that

$$\sum_{m \in \mathcal{M}_n} e^{-x_m} \leqslant \Sigma$$

*and some (possibly depending on the data) penalty function $\mathrm{pen}_n : \mathcal{M}_n \to \mathbb{R}_+$. Let $\widetilde{s}$ be a ρ_n-minimum penalized contrast estimator of s as defined by (4). Then, given $C > 1$, there exists some positive constants K_1 and K_2 (depending on C and on the constants b and c of assumptions **A1** and **A2**) such that if for some nonnegative constant K_3, for every $m \in \mathcal{M}_n$ and every positive ξ,*

$$pen_n(m) \geqslant K_1 \sigma_m^2 + K_2 \frac{x_m}{n} - K_3 \frac{\xi}{n} \tag{67}$$

holds with probability larger than $1 - \exp(-x_m - \xi)$, the following risk bound holds for all $s \in \mathcal{S}$

$$\mathbb{E}[l(s, \widetilde{s})] \leqslant C\left[\inf_{m \in \mathcal{M}_n}\left(\inf_{t \in S_m} l(s,t) + \mathbb{E}[pen_n(m)]\right) + C' \frac{1 + \Sigma}{n} + \rho_n\right], \tag{68}$$

where $l(s, S_m) = \inf_{t \in S_m} l(s,t)$ and C' is a constant (depending on C, b, c and K_3).

Proof. — We first assume for sake of simplicity that $\rho_n = 0$ and take $m \in \mathcal{M}_n$. For any $m' \in \mathcal{M}_n$, we consider some point $u_{m'}$ in $S_{m'}$ such that

$$d(s, u_{m'}) \leqslant 2 \inf_{t \in S_{m'}} d(s,t). \tag{69}$$

Let $y_{m'} \geqslant \sigma_{m'}$ to be chosen later, define for any $t \in S_{m'}$

$$w_{m'}(t) = [d(s, s_m) + d(s,t)]^2 + y_{m'}^2,$$

and finally set

$$V_{m'} = \sup_{t \in S_{m'}} \left[\frac{|\overline{\gamma}_n(t) - \overline{\gamma}_n(s_m)|}{w_{m'}(t)} \right].$$

Taking these notations into account, we get from (9)

$$l(s, \widehat{s}) \leqslant l(s, s_m) + w_{\widehat{m}}(\widehat{s}) V_{\widehat{m}} - \mathrm{pen}_n(\widehat{m}) + \mathrm{pen}_n(m). \tag{70}$$

It remains to control the variables $V_{m'}$ for all possible values of m' in $\mathcal{M}_n$. To do this, we use Talagrand's inequality for empirical processes as stated in the preceding section, noticing that

$$V_{m'} = \sup_{t \in S_{m'}} \left| \nu_n \left[\frac{\gamma(t,.) - \gamma(s_m,.)}{w_{m'}(t)} \right] \right|.$$

Hence, since by Assumption **A2**

$$\mathrm{Var}\left[\frac{\gamma(t,.) - \gamma(s_m,.)}{w_{m'}(t)} \right] \leqslant \frac{d^2(t, s_m)}{(d^2(t, s_m) + y_{m'}^2)^2} \leqslant \frac{1}{4y_{m'}^2},$$

and by Assumption **A1**

$$\left| \frac{\gamma(t,.) - \gamma(s_m,.)}{w_{m'}(t)} \right| \leqslant \frac{b}{y_{m'}^2},$$

(33) implies that, for any $x > 0$ and appropriate constants κ_1 and κ_2,

$$\mathbb{P}\left[V_{m'} \geqslant \kappa_1 \mathbb{E}[V_{m'}] + \kappa_2 \left(\sqrt{\frac{x}{4n}} y_{m'}^{-1} + b \frac{x}{n} y_{m'}^{-2} \right) \right] \leqslant e^{-x}. \tag{71}$$

Given $\xi > 0$, we apply (71) with $x = x_{m'} + \xi$ and sum up the resulting inequalities over $m' \in \mathcal{M}_n$. It follows that, on some event Ω_ξ with probability larger than $1 - \Sigma e^{-\xi}$, one has for all $m' \in \mathcal{M}_n$,

$$V_{m'} \leqslant \kappa_1 \mathbb{E}[V_{m'}] + \kappa_2 \left(\sqrt{\frac{x_{m'} + \xi}{4n}} y_{m'}^{-1} + b \frac{x_{m'} + \xi}{n} y_{m'}^{-2} \right). \tag{72}$$

We now use assumption (66) to bound $\mathbb{E}[V_{m'}]$. Indeed

$$\mathbb{E}[V_{m'}] \leqslant \mathbb{E}\left[\sup_{t \in S_{m'}} \left[\frac{|\overline{\gamma}_n(t) - \overline{\gamma}_n(u_{m'})|}{w_{m'}(t)} \right] \right] + \mathbb{E}\left[\frac{|\overline{\gamma}_n(u_{m'}) - \overline{\gamma}_n(s_m)|}{\inf_{t \in S_{m'}} [w_{m'}(t)]} \right] \tag{73}$$

and, since by (69) $d(t,u_{m'}) \leqslant d(s,t)+d(s,u_{m'}) \leqslant 3d(s,t)$ for every $t \in S_{m'}$, we get $w_{m'}(t) \geqslant 9^{-1}d^2(t,u_{m'})+y_{m'}^2$. Thus we derive from (66) and Lemma 5.1 that

$$\mathbb{E}\left[\sup_{t\in S_{m'}}\left[\frac{|\overline{\gamma}_n(t)-\overline{\gamma}_n(u_{m'})|}{w_{m'}(t)}\right]\right] \leqslant 4y_{m'}^{-2}\phi_{m'}(3y_{m'}).$$

Hence, using the monotonicity assumption on $\phi_{m'}(x)/x$, since $y_{m'} \geqslant \sigma_{m'}$ we get by definition of $\sigma_{m'}$

$$\mathbb{E}\left[\sup_{t\in S_{m'}}\left[\frac{|\overline{\gamma}_n(t)-\overline{\gamma}_n(u_{m'})|}{w_{m'}(t)}\right]\right] \leqslant 12y_{m'}^{-1}\sigma_{m'}^{-1}\phi_{m'}(\sigma_{m'}) \leqslant 12y_{m'}^{-1}\sigma_{m'}$$

which achieves the control of the first term in the right hand side of (73). For the second term, we note that from (69)

$$\inf_{t\in S_{m'}}[w_{m'}(t)] \geqslant 2y_{m'}\inf_{t\in S_{m'}}(d(s,s_m)+d(s,t)) \geqslant y_{m'}d(u_{m'},s_m),$$

hence

$$\mathbb{E}\left[\frac{|\overline{\gamma}_n(u_{m'})-\overline{\gamma}_n(s_m)|}{\inf_{t\in S_{m'}}[w_{m'}(t)]}\right] \leqslant y_{m'}^{-1}\mathbb{E}\left[\frac{|\overline{\gamma}_n(u_{m'})-\overline{\gamma}_n(s_m)|}{d(u_{m'},s_m)}\right]$$

and by Jensen's inequality

$$\mathbb{E}\left[\frac{|\overline{\gamma}_n(u_{m'})-\overline{\gamma}_n(s_m)|}{\inf_{t\in S_{m'}}[w_{m'}(t)]}\right] \leqslant y_{m'}^{-1}\frac{\sqrt{\mathrm{Var}\left[\overline{\gamma}_n(u_{m'})-\overline{\gamma}_n(s_m)\right]}}{d(u_{m'},s_m)} \leqslant y_{m'}^{-1}n^{-1/2}.$$

Collecting these inequalities we get from (73):

$$\mathbb{E}[V_{m'}] \leqslant y_{m'}^{-1}\left[12\sigma_{m'}+n^{-1/2}\right], \text{ for all } m' \in \mathcal{M}_n.$$

Hence, (72) implies that on the event Ω_ξ,

$$V_{m'} \leqslant \kappa_1 y_{m'}^{-1}\left[12\sigma_{m'}+n^{-1/2}\right]+\kappa_2\left(\sqrt{\frac{x_{m'}+\xi}{4n}}y_{m'}^{-1}+b\frac{x_{m'}+\xi}{n}y_{m'}^{-2}\right),$$

for all $m' \in \mathcal{M}$. So, if we define

$$y_{m'} = 2K\left[\kappa_1\left[12\sigma_{m'}+n^{-1/2}\right]+\kappa_2\sqrt{\frac{x_{m'}+\xi}{4n}}+\sqrt{\kappa_2 b\frac{x_{m'}+\xi}{n}}\right]$$

so that on Ω_ξ, one has $V_{m'} \leqslant K^{-1}$ for all $m' \in \mathcal{M}$, we derive from (70) that

$$l(s,\widetilde{s}) \leqslant l(s,s_m)+K^{-1}w_{\widehat{m}}(\widetilde{s})-\mathrm{pen}(\widehat{m})+\mathrm{pen}(m),$$

and therefore

$$l(s,\widetilde{s}) \leqslant l(s,s_m) + K^{-1}\left[\left[d(s,s_m)+d(s,\widetilde{s})\right]^2 + y_{\widehat{m}}^2\right] - \mathrm{pen}(\widehat{m}) + \mathrm{pen}(m).$$

Using repeatedly the elementary inequality $(\alpha+\beta)^2 \leqslant 2\alpha^2+2\beta^2$, we derive that, on the one hand,

$$y_{\widehat{m}}^2 \leqslant 8K^2\left[144\kappa_1^2\sigma_{\widehat{m}}^2 + 2\kappa_2\left(\sqrt{\frac{\kappa_2}{4}}+\sqrt{b}\right)^2\left(\frac{x_{\widehat{m}}+\xi}{n}\right)+\frac{2\kappa_1^2}{n}\right],$$

and, on the other hand, by Assumption **A2,**

$$\left[d(s,s_m)+d(s,\widetilde{s})\right]^2 \leqslant 2\left(d^2(s,s_m)+d^2(s,\widetilde{s})\right) \leqslant 2c\left(l(s,\widetilde{s})+l(s,s_m)\right).$$

Hence, on Ω_ξ, the following inequality is valid

$$\begin{aligned} l(s,\widetilde{s}) \leqslant\ & l(s,s_m)\left(1+\frac{2c}{K}\right)+\frac{2c}{K}l(s,\widetilde{s}) \\ & +8K^2\left(144\kappa_1^2\sigma_{\widehat{m}}^2+2\kappa_2\left(\sqrt{\frac{\kappa_2}{4}}+\sqrt{b}\right)^2\left(\frac{x_{\widehat{m}}}{n}\right)\right)-\mathrm{pen}(\widehat{m}) \\ & +\mathrm{pen}(m)+\frac{16K^2}{n}\left(\kappa_1^2+\kappa_2\left(\sqrt{\frac{\kappa_2}{4}}+\sqrt{b}\right)^2\xi\right), \end{aligned}$$

which in turn implies because of the condition (67) on the penalty function $\mathrm{pen}_n(.)$, that if we choose

$$K=\frac{2c(1+C)}{C-1},\ K_1=9\times 2^7K^2 \text{ and } K_2=16K^2\kappa_2\left(\sqrt{\kappa_2/4}+\sqrt{b}\right)^2,$$

one has on a set of probability larger than $1-2\Sigma\exp(-\xi)$

$$\begin{aligned} \left(\frac{2}{1+C}\right)l(s,\widetilde{s}) \leqslant\ & l(s,s_m)\left(\frac{2C}{1+C}\right)+\mathrm{pen}_n(m) \\ & +\frac{16K^2}{n}\left(\kappa_1^2+\kappa_2\left(\sqrt{\frac{\kappa_2}{4}}+\sqrt{b}\right)^2\xi\right)+\frac{K_3}{n}\xi. \end{aligned}$$

Integrating this inequality with respect to ξ straightforwardly leads to the required risk bound (68). □

Remarks. — The countability assumption on each model of the collection is meant to overcome measurability difficulties in order to focus on the essentials. Obviously it could be relaxed and the following assumption could be substituted to countability

A3 For every $m \in \mathcal{M}_n$, there exists some countable subset S'_m of S_m such that, for every $u \in \mathcal{S}$ and any positive number η

$$\sup_{t \in S'_m} \left[\frac{|\overline{\gamma}_n(t) - \overline{\gamma}_n(u)|}{d^2(t,u) + \eta} \right] = \sup_{t \in S_m} \left[\frac{|\overline{\gamma}_n(t) - \overline{\gamma}_n(u)|}{d^2(t,u) + \eta} \right] \quad \text{a.s.}$$

4.3. Application to bounded regression

We consider some regression frameworks, where the boundedness assumption on the response variables Y_i appears as a natural preliminary information. More precisely, if the response variables Y_i are known to belong to some bounded interval $\left[a, a+\sqrt{b}\right]$, then, to estimate the regression function, it will be natural to consider only models which are included in the set $\mathcal{S}$ of measurable functions taking their values in $\left[a, a+\sqrt{b}\right]$. In this case, the least squares contrast function fulfills conditions **A1** and **A2**. Indeed, since $\gamma(t,(x,y)) = (y - t(x))^2$, **A1** is fulfilled whenever $t \in \mathcal{S}$ and $y \in \left[a, a+\sqrt{b}\right]$. Moreover

$$[\gamma(t,(x,y)) - \gamma(s,(x,y))]^2 = [t(x) - s(x)]^2 [2(y - s(x)) - t(x) + s(x)]^2.$$

Hence, since

$$\mathbb{E}[Y - s(X) \mid X] = 0 \quad \text{and} \quad \mathbb{E}\left[(Y - s(X))^2 \mid X\right] \leqslant \frac{b}{4},$$

one has

$$\begin{aligned} \mathbb{E}\left[[2(y - s(x)) - t(x) + s(x)]^2 \mid X\right] &= 4\mathbb{E}\left[(Y - s(X))^2 \mid X\right] \\ &\quad + (-t(X) + s(X))^2 \\ &\leqslant 2b, \end{aligned}$$

and therefore

$$\mathbb{E}[\gamma(t,(X,Y)) - \gamma(s,(X,Y))]^2 \leqslant 2b\mathbb{E}(t(X) - s(X))^2. \tag{74}$$

This implies that **A2** holds with $c = 2b$ and $d^2(t,s) = 2b\|t - s\|^2$. We first go back to the pattern recognition framework for which the response variables Y_i's belong to $\{0;1\}$.

4.3.1. *Pattern recognition: alternative penalty functions and risk bounds*

Let us recall that one observes $(X_1, Y_1), ..., (X_n, Y_n)$ which are independent copies of some pair (X,Y) taking values in $\mathbb{R}^d \times \{0,1\}$ and that the models $(S_m)_{m \in \mathcal{M}_n}$ are defined for every $m \in \mathcal{M}_n$ as

$$S_m = \{\mathbb{1}_C : C \in \mathcal{C}_m\},$$

where $\mathcal{C}_m$ is some countable class of subsets of $\mathbb{R}^d$. For every $m \in \mathcal{M}_n$, a ρ_n-least squares estimator $\widehat{s}_m$ minimizes over $t \in S_m$, the quantity

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - t(X_i))^2 + \rho_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq t(X_i)} + \rho_n.$$

We consider the pattern recognition problem as a regression problem which means that we set $s(x) = \mathbb{E}(Y \mid X = x)$ and therefore the loss function that we choose to consider is

$$l(s,t) = \|t - s\|^2.$$

The results that we intend to derive here will hold for this particular loss function and this very definition of s (unlike the results of the preceding section which were valid for two possible definitions of s). In order to apply Theorem 4.2, our main task is to compute some function ϕ_m fulfilling (66). We have to refine on the symmetrization arguments developed in the preceding section. Given independent random signs $(\varepsilon_1, ..., \varepsilon_n)$, independent from $(\xi_1, ..., \xi_n)$, one has by the same symmetrization argument as before

$$\mathbb{E}\left[\sup_{t \in \mathcal{B}_m(u,\sigma)} |\overline{\gamma}_n(t) - \overline{\gamma}_n(u)|\right] \leqslant \frac{2}{n}\mathbb{E}\left[\sup_{t \in \mathcal{B}_m(u,\sigma)} \left|\sum_{i=1}^{n} \varepsilon_i \left(\mathbb{1}_{Y_i \neq t(X_i)} - \mathbb{1}_{Y_i \neq u(X_i)}\right)\right|\right].$$

Hence, recalling that

$$H_m = \log |\{C \cap \{X_1, ..., X_n\},\ C \in \mathcal{C}_m\}|,$$

using Lemma 5.2 again and symmetry, and setting

$$W_m(\sigma) = \sup_{t \in \mathcal{B}_m(u,\sigma)} |\overline{\gamma}_n(t) - \overline{\gamma}_n(u)|,$$

we get

$$\mathbb{E}[W_m(\sigma)] \leqslant \frac{4\sqrt{2}}{n}\mathbb{E}\sqrt{H_m \sup_{t \in \mathcal{B}_m(u,\sigma)} \sum_{i=1}^{n} \left(\mathbb{1}_{Y_i \neq t(X_i)} - \mathbb{1}_{Y_i \neq u(X_i)}\right)^2}.$$

Then by Cauchy-Schwarz inequality,

$$\mathbb{E}[W_m(\sigma)] \leqslant \frac{4\sqrt{2}}{n}\sqrt{\mathbb{E}[H_m]\,\mathbb{E}\left[\sup_{t \in \mathcal{B}_m(u,\sigma)} \sum_{i=1}^{n} \left(\mathbb{1}_{Y_i \neq t(X_i)} - \mathbb{1}_{Y_i \neq u(X_i)}\right)^2\right]}. \tag{75}$$

Now, noticing by (74) that,

$$\sup_{t\in\mathcal{B}_m(u,\sigma)} \mathbb{E}\left[\sum_{i=1}^{n}\left(\mathbb{1}_{Y_i\neq t(X_i)}-\mathbb{1}_{Y_i\neq u(X_i)}\right)^2\right]\leqslant n\sigma^2,$$

we can use (32) and derive from (75) that the following inequality holds:

$$\mathbb{E}\left[W_m\left(\sigma\right)\right]\leqslant\frac{4\sqrt{2}}{n}\sqrt{\mathbb{E}\left[H_m\right]\left(n\sigma^2+16n\mathbb{E}\left[W_m\left(\sigma\right)\right]\right)}.$$

But the latter inequality is equivalent to

$$\mathbb{E}\left[W_m\left(\sigma\right)\right]\leqslant 4\sqrt{\frac{\mathbb{E}\left[H_m\right]}{n}}\left[64\sqrt{\frac{\mathbb{E}\left[H_m\right]}{n}}+\sqrt{2^{12}\frac{\mathbb{E}\left[H_m\right]}{n}+2\sigma^2}\right],$$

which whenever $\sigma\geqslant 24\sqrt{\mathbb{E}\left[H_m\right]/n}$, implies that

$$\mathbb{E}\left[W_m\left(\sigma\right)\right]\leqslant 4\sigma\sqrt{\frac{\mathbb{E}\left[H_m\right]}{n}}\left[\frac{8}{3}+\sqrt{9.12}\right]\leqslant 24\sigma\sqrt{\frac{\mathbb{E}\left[H_m\right]}{n}}.$$

Hence, setting $\phi_m\left(x\right)=24x\sqrt{\mathbb{E}\left[H_m\right]/n}$, inequality (66) is satisfied with $\sigma_m=24\sqrt{\mathbb{E}\left[H_m\right]/n}$. We can now follow the same strategy as in Section 4.1 and define a proper random penalty function by using the concentration property of H_m around its expectation. More precisely, we know that except on a set of probability not larger than $1-\exp\left(-x_m-\xi\right)$, the following inequality is valid

$$\mathbb{E}\left[H_m\right]\leqslant 2H_m+2\log\left(2\right)\left(x_m+\xi\right).$$

Hence, defining

$$\mathrm{pen}_n\left(m\right)=K_1^{'}\frac{H_m}{n}+K_2^{'}\frac{x_m}{n},$$

for proper constants K_1' and K_2' (the choice $K_1'=9\times 2^7K_1$and $K_2'=K_2+K_1'\log\left(2\right)$ works) and setting $K_3=\log\left(2\right)K_1'$, we can apply Theorem 4.2. Typically, if $\mathcal{M}_n$ has cardinality less than n, we can take $x_m=\log\left(n\right)$ and the risk bound for the corresponding selected classification rule $\widetilde{s}$ can be written as

$$\mathbb{E}\left[\left\|s,\widetilde{s}\right\|^2\right]\leqslant C\left[\inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\left\|s-t\right\|^2+K_1^{'}\frac{\mathbb{E}\left[H_m\right]}{n}\right)+C''\frac{1+\log\left(n\right)}{n}+\rho_n\right],$$

for some appropriate constant C''. We readily see that this bound is (up to constant!) better than the one obtained in Section 4.1 since we have replaced in the right hand side $\sqrt{\mathbb{E}\left[H_m\right]/n}$ by $\mathbb{E}\left[H_m\right]/n$ which can be much smaller.

4.3.2. *Binary images*

Following [26], our purpose is to study the particular regression framework for which the variables X_i's are uniformly distributed on $[0,1]^2$ and $s(x) = \mathbb{E}[Y \mid X = x]$ is of the form

$$s(x_1, x_2) = 1 \text{ if } x_2 \leqslant \partial s(x_1) \text{ and } 0 \text{ otherwise,}$$

where ∂s is some measurable map from $[0,1]$ to $[0,1]$. The function ∂s should be understood as the parametrization of a boundary fragment corresponding to some portion s of a binary image in the plane and restoring this portion of the image from the noisy data $(X_1, Y_1), ..., (X_n, Y_n)$ means estimating s or equivalently ∂s. We assume the regression errors to be bounded, more precisely we suppose that the variable's take their values in some known interval $\left[a, a+\sqrt{b}\right]$ which contains $[0,1]$. Let us introduce some notation which will turn out to be convenient to describe the list of models that we wish to consider. Let $\mathcal{G}$ be the set of measurable maps from $[0,1]$ to $[0,1]$. For any $f \in \mathcal{G}$, let us denote by χ_f the function defined on $[0,1]^2$ by

$$\chi_f(x_1, x_2) = 1 \text{ if } x_2 \leqslant f(x_1) \text{ and } 0 \text{ otherwise.}$$

From this definition we see that $\chi_{\partial s} = s$ and more generally if we define $\mathcal{S} = \{\chi_f : f \in \mathcal{G}\}$, for every $t \in \mathcal{S}$, we denote by ∂t the element of $\mathcal{G}$ such that $\chi_{\partial t} = t$. It is natural to consider here models S_m of the form $S_m = \{\chi_f : f \in \partial S_m\}$, where $(\partial S_m)_{m \in \mathcal{M}_n}$ denotes some collection of subsets of $\mathcal{G}$. Denoting by $\|.\|_1$ the Lebesgue $\mathbb{L}_1$-norm on $\mathcal{G}$, one has for every $f, g \in \mathcal{G}$ $\|\chi_f - \chi_g\|^2 = \|f - g\|_1$ or equivalently for every $s, t \in \mathcal{S}$ $\|s - t\|^2 = \|\partial s - \partial t\|_1$. The application of Theorem 4.2 requires the computation of some function ϕ_m fulfilling (66) and therefore to majorize $\mathbb{E}[W_m(\sigma)]$, where

$$W_m(\sigma) = \sup_{t \in \mathcal{B}_m(u,\sigma)} |\overline{\gamma}_n(t) - \overline{\gamma}_n(u)|.$$

This can be done using entropy with bracketing arguments. Indeed, let us notice that if g belongs to some ball with radius δ in $\mathbb{L}_\infty[0,1]$, then for some function $f \in \mathbb{L}_\infty[0,1]$, one has $f - \delta \leqslant g \leqslant f + \delta$ and therefore, defining $f_L = \sup(f - \delta, 0)$ and $f_U = \inf(f + \delta, 1)$

$$\chi_{f_L} \leqslant \chi_g \leqslant \chi_{f_U}$$

with $\|\chi_{f_L} - \chi_{f_U}\|^2 \leqslant \delta$. This means that, taking (74) into account, the $\mathbb{L}_2$ metric entropy with bracketing of the class of functions

$$\{\gamma(t, .) - \gamma(u, .),\ t \in \mathcal{B}_m(u, \sigma)\}$$

for radius ε (denoted by $H_{[2]}(\varepsilon, \mathcal{B}_m(u,\sigma))$) is bounded by the $\mathbb{L}_\infty$ metric entropy for radius $\varepsilon^2/2b$ of the $\mathbb{L}_1$ ball centered at ∂u with radius $\sigma^2/2b$ in ∂S_m. Hence, defining $H_m(\delta,\rho)$ as the supremum over $g \in \partial S_m$ of the $\mathbb{L}_\infty$ metric entropy for radius δ of the $\mathbb{L}_1$ ball centered at g with radius ρ in ∂S_m, we derive from some by now classical inequality (see [43]) that, for some absolute constant κ_1

$$\begin{aligned}
\kappa_1^{-1}\mathbb{E}[W_m(\sigma)] &\leqslant \frac{1}{\sqrt{n}}\int_0^\sigma \sqrt{H_{[2]}(\varepsilon, \mathcal{B}_m(u,\sigma))}d\varepsilon \\
&\quad + \frac{1}{n}H_{[2]}\left(\frac{\sigma}{\sqrt{2}}, \mathcal{B}_m(u,\sigma)\right) \\
&\leqslant \frac{1}{\sqrt{n}}\int_0^\sigma \sqrt{H_m(\varepsilon^2/2b, \sigma^2/2b)}d\varepsilon \\
&\quad + \frac{1}{n}H_m(\sigma^2/4b, \sigma^2/2b) \\
&\leqslant \frac{\sigma}{2\sqrt{n}}\int_0^1 \sqrt{\frac{H_m(x\sigma^2/2b, \sigma^2/2b)}{x}}dx \\
&\quad + \frac{1}{n}H_m(\sigma^2/4b, \sigma^2/2b).
\end{aligned}$$

The point now is that, whenever ∂S_m is part of a linear finite dimensional subspace of $\mathbb{L}_\infty[0,1]$, $H_m(\delta,\rho)$ is typically bounded by $D_m[B_m + \log(\rho/\delta)]$ for some appropriate constants D_m and B_m. If it is so then

$$\int_0^1 \sqrt{\frac{H_m(x\sigma^2/2b, \sigma^2/2b)}{x}}dx \leqslant \sqrt{D_m}\left[2\sqrt{B_m} + \int_0^1 \left[\frac{|\log(\delta)|}{\delta}\right]^{1/2} d\delta\right],$$

which implies that for some absolute constant κ_2

$$\mathbb{E}\left[\sup_{t\in\mathcal{B}_m(u,\sigma)} |\overline{\gamma}_n(t) - \overline{\gamma}_n(u)|\right] \leqslant \kappa_2\left[\sigma\sqrt{\frac{(1+B_m)D_m}{n}} + \frac{D_m(1+B_m)}{n}\right].$$

Hence, whenever $\sigma \geqslant (1+\kappa_2)\sqrt{(1+B_m)D_m/n}$ we have

$$\mathbb{E}[W_m(\sigma)] \leqslant (1+\kappa_2)\sigma\sqrt{\frac{(1+B_m)D_m}{n}}$$

which means that Theorem 4.2 can be applied with

$$\sigma_m = (1+\kappa_2)\sqrt{(1+B_m)D_m/n}.$$

To be more concrete, let us see what this gives when

$$\mathcal{M}_n = \{(r,J) : r \in \mathbb{N},\ J \in \mathbb{N}^*\},$$

and for each $(r, J) \in \mathcal{M}_n$, ∂S_m is taken to be the set of piecewise polynomials on the regular partition with J pieces on $[0,1]$ with degree not larger than r and which belong to $\mathcal{G}$. Then, it is shown in [7] that

$$H_m(\delta, \rho) \leqslant D_m \left[\log(\rho/\delta) + \log\left(6er\sqrt{r+1}\right)\right],$$

where $D_m = (r+1)J$ is the dimension of the underlying linear space of piecewise polynomials. As a conclusion, if we choose $x_m = (r+1)J$, we can take $\Sigma = 1$ and, given $C > 1$, there exists some constant K depending only on C and b such that, if we set

$$\mathrm{pen}_n(m) = K\frac{D_m}{n}\left[1 + \log(1+r)\right]$$

the following risk bound holds

$$\mathbb{E}\left[\|\partial s - \partial\widetilde{s}\|_1\right] \leqslant C\left[\inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|\partial s - \partial t\|_1 + \mathrm{pen}_n(m)\right) + \frac{2C'}{n} + \rho_n\right],$$

where C' depends on C and b. We recover here the result established in [7]. Starting from this risk bound it is also possible to show that the above penalized estimator $\widetilde{s}$ is adaptive in the minimax sense on some collection of indicators of sets with smooth boundaries. We do not want to go further into details here and the interested reader will find in [7] some precise statements about this phenomenon as well as other examples that could be treated via Theorem 4.2 such as the estimation of the support of a density.

5. Appendix

The following inequalities are more or less classical and well known. We present some (short) proofs for the sake of completeness.

5.1. From increments to weighted processes

LEMMA 5.1. — *Let (S, d) be some at most countable pseudo-metric space and $u \in S$. Assume that for some process Z indexed by S, the nonnegative random variable $\sup_{t\in\mathcal{B}(u,\sigma)}[Z(t) - Z(u)]$ has finite expectation for any positive number σ, where $\mathcal{B}(u,\sigma) = \{t \in S,\ d(t,u) \leqslant \sigma\}$. Then, for any function ϕ on $\mathbb{R}_+$ such that $\phi(x)/x$ is nonincreasing on $\mathbb{R}_+$ and satisfies to*

$$\mathbb{E}\left[\sup_{t\in\mathcal{B}(u,\sigma)}|Z(t) - Z(u)|\right] \leqslant \phi(\sigma), \quad \text{for any } \sigma \geqslant \sigma_* \geqslant 0$$

one has for any positive number $x \geqslant \sigma_*$

$$\mathbb{E}\left[\sup_{t\in S}\left[\frac{|Z(t) - Z(u)|}{d^2(t,u) + x^2}\right]\right] \leqslant 4x^{-2}\phi(x).$$

Proof. — Let us introduce for any integer j

$$\mathcal{C}_j = \left\{t \in S,\, r^j x < d(t,u) \leqslant r^{j+1} x\right\},$$

with $r > 1$ to be chosen later. Then $\left\{\mathcal{B}(u,x), \{\mathcal{C}_j\}_{j\geqslant 0}\right\}$ is a partition of S and therefore,

$$\begin{aligned}\sup_{t\in S}\left[\frac{|Z(t)-Z(u)|}{d^2(t,u)+x^2}\right] &\leqslant \sup_{t\in\mathcal{B}(u,x)}\left[\frac{|Z(t)-Z(u)|}{x^2}\right] \\ &\quad + \sum_{j\geqslant 0}\sup_{t\in\mathcal{C}_j}\left[\frac{|Z(t)-Z(u)|}{d^2(t,u)+x^2}\right],\end{aligned}$$

which in turn implies that

$$\begin{aligned}x^2\sup_{t\in S}\left[\frac{|Z(t)-Z(u)|}{d^2(t,u)+x^2}\right] &\leqslant \sup_{t\in\mathcal{B}(u,x)}|Z(t)-Z(u)| \\ &\quad + \sum_{j\geqslant 0}\left(1+r^{2j}\right)^{-1}\sup_{t\in\mathcal{B}(u,r^{j+1}x)}|Z(t)-Z(u)|.\end{aligned}$$

Taking expectation in the above inequality yields

$$x^2\mathbb{E}\left[\sup_{t\in S}\left[\frac{|Z(t)-Z(u)|}{d^2(t,u)+x^2}\right]\right] \leqslant \phi(x) + \sum_{j\geqslant 0}\left(1+r^{2j}\right)^{-1}\phi\left(r^{j+1}x\right).$$

Now by our monotonicity assumption, $\phi\left(r^{j+1}x\right) \leqslant r^{j+1}\phi(x)$, hence

$$\begin{aligned}x^2\mathbb{E}\left[\sup_{t\in S}\left[\frac{|Z(t)-Z(u)|}{d^2(t,u)+x^2}\right]\right] &\leqslant \phi(x)\left[1+r\sum_{j\geqslant 0}r^j\left(1+r^{2j}\right)^{-1}\right] \\ &\leqslant \phi(x)\left[1+r\left(\frac{1}{2}+\sum_{j\geqslant 1}r^{-j}\right)\right] \\ &\leqslant \phi(x)\left[1+r\left(\frac{1}{2}+\frac{1}{r-1}\right)\right]\end{aligned}$$

and the result follows by choosing $r = 1+\sqrt{2}$. □

5.2. A sub-Gaussian inequality

We now turn to a bound which is useful for taking advantage of symmetrization arguments.

LEMMA 5.2. — *Let $\mathcal{A}$ be some finite subset of $\mathbb{R}^n$ and $(\varepsilon_1, ..., \varepsilon_n)$ be independent Rademacher variables. Let N denote the cardinality of $\mathcal{A}$ and let $R = \sup_{a\in\mathcal{A}} \left[\sum_{i=1}^n a_i^2\right]^{1/2}$, then*

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}} \sum_{i=1}^n \varepsilon_i a_i\right] \leqslant R\sqrt{2\log(N)}. \tag{76}$$

Proof. — Setting $Z_a = \sum_{i=1}^n \varepsilon_i a_i$ and $x = \mathbb{E}\left[\sup_{a\in\mathcal{A}} Z_a\right]$, we have by Jensen's inequality,

$$\exp(\lambda x) \leqslant \mathbb{E}\left[\exp\left(\lambda \sup_{a\in\mathcal{A}} Z_a\right)\right] \leqslant \mathbb{E}\left[\sup_{a\in\mathcal{A}} \exp(\lambda Z_a)\right] \leqslant \sum_{a\in\mathcal{A}} \mathbb{E}\left[\exp(\lambda Z_a)\right],$$

for any $\lambda \in \mathbb{R}_+$. Hence, since
$\mathbb{E}\left[\exp(\lambda Z_a)\right] = \prod_{i=1}^n \cosh(a_i\lambda) \leqslant \prod_{i=1}^n \exp\left(a_i^2\lambda^2/2\right)$

$$\exp(\lambda x) \leqslant \sum_{a\in\mathcal{A}} \exp\left(\sum_{i=1}^n a_i^2\lambda^2/2\right) \leqslant N\exp\left(R^2\lambda^2/2\right).$$

Therefore for all positive λ we have

$$\lambda x - \frac{R^2\lambda^2}{2} \leqslant \log(N),$$

and maximizing the left hand side of this inequality with respect to λ leads to

$$\frac{x^2}{2R^2} \leqslant \log(N)$$

which implies (76). □

5.3. Connecting moments of order 1 and 2 of log-likelihood ratios

The following Lemma is adapted from Lemma 1 of Barron and Sheu [6] and can be found in [23].

LEMMA 5.3. — *For all positive densities p and q with respect to some measure μ*

$$\frac{1}{2}\int p\wedge q\left(\log\frac{p}{q}\right)^2 d\mu \leqslant K(p,q) \leqslant \frac{1}{2}\int p\vee q\left(\log\frac{p}{q}\right)^2 d\mu.$$

Proof. — Let $f = \log(q/p)$ then

$$K(p,q) = \int p\,\left(e^f - 1 - f\right) d\mu = \int p\,\phi(f) d\mu,$$

where ϕ is the function defined by $\phi(x) = e^x - 1 - x$, for all $x \in \mathbf{R}$. Then, since

$$\frac{1}{2}x^2\,(1 \wedge e^x) \leqslant \phi(x) \leqslant \frac{1}{2}x^2\,(1 \vee e^x)$$

for all real number x, one derives that

$$\frac{1}{2}\int f^2\left(1 \wedge e^f\right) p\,d\mu \leqslant K(p,q) \leqslant \frac{1}{2}\int f^2\left(1 \vee e^f\right) p\,d\mu$$

which leads to the result. □

5.4. Large deviations of log-likelihood ratios

Let h denote the Hellinger distance.

PROPOSITION 5.4. — *Let $\xi_1, ..., \xi_n$ be independent random variables with common distribution $P = s\mu$. Then, for every positive density f and any positive number x*

$$\mathbb{P}\left[P_n\left[\log\left(\frac{f}{s}\right)\right] \geqslant -2h^2(s,f) + 2\frac{x}{n}\right] \leqslant e^{-x}.$$

Proof. — The result derives from the control of the Laplace transform at point $\alpha = 1/2$. Indeed, from Markov's inequality

$$\mathbb{P}\left[P_n\left[\log\left(\frac{f}{s}\right)\right] \geqslant a\right] \leqslant e^{-n\alpha a}\left[\mathbb{E}\left[\exp\left(\alpha \log\frac{f(\xi_1)}{s(\xi_1)}\right)\right]\right]^n$$

which, since for $\alpha = 1/2$,

$$\mathbb{E}\left[\exp\left(\alpha \log\frac{f(\xi_1)}{s(\xi_1)}\right)\right] = 1 - h^2(s,f) \leqslant \exp\left(-h^2(s,f)\right)$$

leads to

$$\mathbb{P}\left[P_n\left[\log\left(\frac{f}{s}\right)\right] \geqslant a\right] \leqslant \exp\left(-n\left(\frac{a}{2} + h^2(s,f)\right)\right).$$

The result immediately follows. □

Bibliography

[1] AKAIKE (H.). — Information theory and an extension of the maximum likelihood principle. In P.N. Petrov and F. Csaki, editors, *Proceedings 2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973.

[2] BARAUD (Y.). — Model selection for regression on a fixed design. Technical report #97.49, (1997) Université Paris-Sud (to appear in *Probability Theory and Related Fields).*

[3] BARAUD (Y.), COMTE (F.) and VIENNET (G.) (1999). — Model selection for (auto-)regression with dependent data. Technical Report LMENS 99-12. Ecole Normale Supérieure, Paris.

[4] BENNETT (G.). — Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57** 33-45 (1962).

[5] BAHADUR (R.R.). — Examples of inconsistency of maximum likelihood estimates. *Sankhya Ser.A* **20**, 207-210 (1958).

[6] BARRON (A.R.) and SHEU (C.H.). — Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19:1054–1347, 1991.

[7] BARRON (A.R.), BIRGÉ (L.), MASSART (P.). — Risk bounds for model selection via penalization. *Probab. Th. Rel. Fields.* **113**, 301-415 (1999).

[8] BIRGÉ (L.) and MASSART (P.). — Rates of convergence for minimum contrast estimators. *Probab. Th. Relat. Fields* **97**, 113-150 (1993).

[9] BIRGÉ (L.) and MASSART (P.). — Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, **4**(3), 329-375 (1998).

[10] BIRGÉ (L.) and MASSART (P.). — From model selection to adaptive estimation. In *Festschrift for Lucien Lecam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.), 55-87 (1997) Springer-Verlag, New-York.

[11] BIRGÉ (L.) and MASSART (P.). — Gaussian model selection. Technical report (1999).

[12] BIRGÉ (L.) and MASSART (P.). — A generalized cross-validation criterion for density estimation. Unpublished manuscript.

[13] BIRGÉ (L.) and ROZENHOLC (Y.). — How many bins should be put in a regular histogram. Unpublished manuscript.

[14] BOBKOV (S.). — On Gross' and Talagrand's inequalities on the discrete cube. *Vestnik of Syktyvkar University* **Ser. 1,1** 12-19 (1995) (in Russian).

[15] BORELL (C.). — The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30**, 207-216 (1975).

[16] BOUCHERON (S.), LUGOSI (G.) and MASSART (P.). — A sharp concentration inequality with applications. Technical report # 99.25 (1999), Université de Paris-Sud.

[17] CIREL'SON (B.S.), IBRAGIMOV (I.A.) and SUDAKOV (V.N.). — Norm of Gaussian sample function. In *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*, Lecture Notes in Mathematics **550** 20-41 (1976) Springer-Verlag, Berlin.

[18] CIREL'SON (B.S.) and SUDAKOV (V.N.). — Extremal properties of half spaces for spherically invariant measures. *J. Soviet. Math.* **9**, 9-18 (1978); translated from *Zap. Nauch. Sem. L.O.M.I.* **41**, 14-24 (1974).

[19] COVER (T.M.) and THOMAS (J.A.). — *Elements of Information Theory.* Wiley series in telecommunications. Wiley (1991).

[20] DEMBO (A.). — Information inequalities and concentration of measure. *Ann. Prob.* **25** 927-939 (1997).

[21] DONOHO (D.L.) and JOHNSTONE (I.M.). — Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455 (1994)

[22] GROSS (L.) Logarithmic Sobolev inequalities. *Amer. J. Math.* **97** 1061-1083 (1975).

[23] CASTELLAN (G.). — Modified Akaike's criterion for histogram density estimation. Technical report #99.61, (1999) Université de Paris-Sud.

[24] CASTELLAN (G.). — Density estimation via exponential model selection. Technical report (1999).

[25] HOEFFDING (W.). — Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58** 13-30 (1963).

[26] KOROSTELEV (A.P.) and TSYBAKOV (A.B.). — Minimax theory of of image reconstruction. *Lectures notes in Statistics* **82**, Soringer Verlag, NewYork (1993).

[27] LEDOUX (M.). — Isoperimetry and Gaussian Analysis. In *Probabilités de St-Flour XXIV-1994* (P. Bernard, ed.), 165-294 (1996) Springer, Berlin.

[28] LEDOUX (M.). — On Talagrand deviation inequalities for product measures. *ESAIM: Probability and Statistics* **1**, 63-87 (1996) http://www.emath.fr/ps/.

[29] LEDOUX (M.) and TALAGRAND (M.). — *Probability in Banach spaces (Isoperimetry and processes).* Ergebnisse der Mathematik und ihrer Grenzgebiete (1991) Springer-Verlag.

[30] MARTON (K.). — A simple proof of the blowing up lemma. IEEE Trans. Inform. Theory **IT-32** 445-446 (1986).

[31] MARTON (K.). — Bounding $\overline{d}$-distance by information divergence: a method to prove measure concentration. *Ann. Prob.* **24** 927-939 (1996).

[32] MASON (D.M.) and van ZWET (W.R.) A refinement of the KMT inequality for the uniform empirical process. *Ann. Prob.* **15**, 871-884 (1987).

[33] MASSART (P.). — About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Prob.* (To appear)(2000).

[34] MASSART (P.). — Optimal constants for Hoeffding type inequalities. Technical report (1998).

[35] MCDIARMID (C.). — On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148-188. Cambridge University Press, Cambridge, 1989.

[36] TALAGRAND (M.). — An isoperimetric theorem on the cube and the Khintchine-Kahane inequalities in product spaces. *Proc. Amer. Math. Soc.* **104** 905-909 (1988).

[37] TALAGRAND (M.). — Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.* **81** 73-205 (1995).

[38] TALAGRAND (M.). — Sharper bounds for empirical processes. *Annals of Probability* **22**, 28-76 (1994).

[39] TALAGRAND (M.). — New concentration inequalities in product spaces. *Invent. Math.* **126**, 505-563 (1996).

[40] VAPNIK (V.N.). — *Estimation of dependencies based on empirical data.* Springer, New York.

[41] VAPNIK (V.N.). — *Statistical learning theory.* J. Wiley, New York.

[42] VAN DER VAART (A.). — *Asymptotic statistics.* Cambridge University Press (1998).

[43] VAN DER VAART (A.) and WELLNER (J.) *Weak Convergence and Empirical Processes.* Springer, New York (1996).