Submitted to the Bernoulli
arXiv: math.PR/0909.1863v1

A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression

YANNICK BARAUD[1,*]

[1] *Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice cedex 02*
E-mail: [*] baraud@unice.fr

Let $(X_t)_{t\in T}$ be a family of real-valued centered random variables indexed by a countable set T. In the first part of this paper, we establish exponential bounds for the deviation probabilities of the supremum $Z = \sup_{t\in T} X_t$ by using the generic chaining device introduced in Talagrand (1995). Compared to concentration-type inequalities, these bounds offer the advantage to hold under weaker conditions on the family $(X_t)_{t\in T}$. The second part of the paper is oriented towards statistics. We consider the regression setting $Y = f + \xi$ where f is an unknown vector of $\mathbb{R}^n$ and ξ is a random vector the components of which are independent, centered and admit finite Laplace transforms in a neighborhood of 0. Our aim is to estimate f from the observation of Y by mean of a model selection approach among a collection of linear subspaces of $\mathbb{R}^n$. The selection procedure we propose is based on the minimization of a penalized criterion the penalty of which is calibrated by using the deviation bounds established in the first part of this paper. More precisely, we study suprema of random variables of the form $X_t = \sum_{i=1}^n t_i\xi_i$ when t varies in the unit ball of a linear subspace of $\mathbb{R}^n$. We finally show that our estimator satisfies some oracle-type inequality under suitable assumptions on the metric structures of the linear spaces of the collection.

Keywords: Bernstein's inequality, Model selection, Regression, Supremum of a random process.

1. introduction

1.1. What is this paper about?

The present paper contains two parts. The first one is oriented towards probability. We consider a family $(X_t)_{t\in T}$ of real-valued centered random variables indexed by a countable set T and give an exponential bound for the probability of deviation of the supremum $Z = \sup_{t\in T} X_t$. The result is established under the assumption that the Laplace transforms of the increments $X_t - X_s$ for $s, t \in T$ satisfy some Bernstein-type bounds. This assumption is convenient to handle simultaneously the cases of subgaussian increments (which is the typical case in the literature) as well as more "heavy tailed" ones for which the Laplace transform of $(X_s - X_t)^2$ may be infinite in a neighborhood

of 0. Under additional assumptions on the X_t, our result allows to recover (with worse constants) some deviation bounds based on concentration-type inequalities of Z around its expectation. However our general result cannot be deduced from those inequalities. As we shall see, concentration-type inequalities could be false under the kind of assumptions we consider on the family $(X_t)_{t\in T}$.

The second part is oriented towards statistics. We consider the regression framework

$$Y_i = f_i + \xi_i,\ i = 1, \ldots, n \tag{1.1}$$

where $f = (f_1, \ldots, f_n)$ is an unknown vector of $\mathbb{R}^n$ and $\xi = (\xi_1, \ldots, \xi_n)$ is a random vector the components of which are independent, centered and admit suitable exponential moments. Our aim is to estimate f from the observation of $Y = (Y_1, \ldots, Y_n)$ by mean of a model selection approach. More precisely, we start with a collection $\mathcal{S} = \{S_m,\ m \in \mathcal{M}\}$ of finite dimensional linear spaces S_m to each of which we associate the least-squares estimator $\hat{f}_m \in S_m$ of f. From the same data Y, our aim is to select some suitable estimator $\tilde{f} = \hat{f}_{\hat{m}}$ among the collection $\mathcal{F} = \{\hat{f}_m,\ m \in \mathcal{M}\}$ in such a way that the (squared) Euclidean risk of $\tilde{f}$ is as close as possible to the infimum of the risks over $\mathcal{F}$. The selection procedure we propose is based on the minimization of a penalized criterion the penalty of which is calibrated by using the deviation bounds established in the first part of this paper. More precisely, the penalty is obtained by studying the deviations of χ^2-type random variables, that is, random variables of the form $|\Pi_S \xi|_2^2$ where $|\ |_2$ denotes the Euclidean norm and Π_S the orthogonal projector onto a linear subspace S of $\mathbb{R}^n$. To our knowledge, these deviation bounds in probability are new. We finally show that $\tilde{f}$ satisfies some oracle-type inequality under suitable assumptions on the metric structures of the S_m.

In the following sections, we situate the results of the present paper within the literature.

1.2. Controlling suprema of random processes

Among the most common deviation inequalities, let us recall

Theorem 1.1 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables and set $X = \sum_{i=1}^n (X_i - \mathbb{E}(X_i))$. Assume that there exist nonnegative numbers v, c such that for all $k \geq 3$*

$$\sum_{i=1}^n \mathbb{E}\left[|X_i|^k\right] \leq \frac{k!}{2} v^2 c^{k-2}, \tag{1.2}$$

then for all $u \geq 0$

$$\mathbb{P}\left(X \geq \sqrt{2v^2 u} + cu\right) \leq e^{-u}. \tag{1.3}$$

Besides, for all $x \geq 0$,

$$\mathbb{P}\left(X \geq x\right) \leq \exp\left(-\frac{x^2}{2(v^2 + cx)}\right). \tag{1.4}$$

In the literature, (1.2) together with the fact that the X_i are independent is sometime replaced by the weaker condition on $X = \sum_{i=1}^n (X_i - \mathbb{E}(X_i))$

$$\mathbb{E}\left(e^{\lambda X}\right) \leq \exp\left[\frac{\lambda^2 v^2}{2(1-\lambda c)}\right], \quad \forall \lambda \in (0, 1/c) \tag{1.5}$$

with the convention $1/0 = +\infty$. Bernstein's inequality allows to derive deviation inequalities for a large class of distributions among which the Poisson, Laplace, Gamma or the Gaussian distributions (once suitably centered). In this latter case, (1.5) holds with $c = 0$. Another situation of interest is the case where the X_i are i.i.d. with values in $[-c, c]$. Then (1.2) and (1.5) hold with $v^2 = n\mathrm{var}(X_1)$.

Given a countable family $(X_t)_{t\in T}$ of such random variables X, many efforts have been done in view of extending Bernstein's inequality to the supremum $Z = \sup_{t\in T} X_t$. When T is a bounded subset of a metric space $(\mathcal{X}, d)$, a common technique is to use a *chaining device*. This approach seems to go back to Kolmogorov and was very popular in statistics in the 90s to control suprema of empirical processes with regard to the entropy of T, see van de Geer (1990) for example. However, this approach leads to pessimistic numerical constants that are in general too large to be used in statistical procedures. An alternative to chaining is the use of concentration inequalities. For example, when the X_t are Gaussian, for all $u \geq 0$ we have

$$\mathbb{P}\left(Z \geq \mathbb{E}\left(Z\right) + \sqrt{2v^2 u}\right) \leq e^{-u} \quad \text{where} \quad v^2 = \sup_{t\in T} \mathrm{var}(X_t). \tag{1.6}$$

This inequality, due to Sudakov & Cirel'son (1974), allows to recover (1.5) with $c = 0$ whenever T reduces to a single element. Compared to chaining, (1.6) provides a powerful tool for controlling suprema of Gaussian processes as soon as one is able to evaluate $\mathbb{E}(Z)$ sharply enough.

It is the merit of Talagrand (1995) to extend this approach for the purpose of controlling suprema of bounded empirical processes, that is, for X_t of the form $X_t = \sum_{i=1}^n t(\xi_i) - \mathbb{E}\left(t(\xi_i)\right)$ where $\xi_1, \ldots, \xi_n$ are independent random variables and T a set of uniformly bounded functions, say with values in $[-c, c]$. From Talagrand's inequality, one can deduce deviation bounds with respect to $\mathbb{E}(Z)$ of the form

$$\mathbb{P}\left[Z \geq C\left(\mathbb{E}(Z) + \sqrt{v^2 u} + cu\right)\right] \leq \exp\left(-u\right) \quad \text{for all } u \geq 0 \tag{1.7}$$

where $v^2 = \sup_{t\in T} \mathrm{var}\left(X_t\right)$ and C is a positive numerical constant. Apart from the constants, (1.7) and (1.3) have a similar flavor even though the boundness assumption on the elements of T seems too strong compared to conditions (1.2) or (1.5).

As the original result by Talagrand involved suboptimal numerical constants, many efforts were made to recover it with sharper ones. A first step in this direction is due to Ledoux (1996) by mean of nice entropy and tensorisation arguments. Then, further refinements were made on Ledoux's result by Massart (2000), Rio (2002) and Bousquet (2002), the latter author achieving the best possible result in terms of constants. For a nice introduction to these inequalities (and their applications to statistics) we refer

the reader to the book by Massart (2007). Other improvements upon (1.7) have been done in the recent years. In particular Klein & Rio (2005) generalized the result to the case

$$X_t = \sum_{i=1}^{n} \overline{X}_{i,t} \tag{1.8}$$

where for each $t \in T$, $\left(\overline{X}_{i,t}\right)_{i=1,\ldots,n}$ are independent (but not necessarily i.i.d.) centered random with values in $[-c, c]$.

In the present paper, the result we establish holds under different assumptions than the ones leading to inequalities such as (1.7). Actually, an inequality such as (1.7) could be false under our set of assumptions on $(X_t)_{t\in T}$. This fact was communicated to us by Jonas Kahn. The counter-example we give in Section 2, which is is a slight modification of the one Jonas Kahn gave us, shows that Z may deviate from $\mathbb{E}(Z)$ on a set the probability of which may not be exponentially small. Moreover, even in the more common situation where X_t is of the form (1.8), we establish deviation inequalities that are available for possibly unbounded random variables $\overline{X}_{i,t}$ which is beyond the scope of the concentration inequalities proven in Bousquet (2002) and Klein & Rio (2005).

Even though it was originally introduced to bound $\mathbb{E}(Z)$ from above, *generic chaining* as described in Talagrand's book (2005) provides another way of establishing deviation bounds for Z. Talagrand's approach relies on the idea of decomposing T into partitions rather than into nets as it was usually done before with the classical chaining device. Denoting by $e_1, \ldots, e_k$ the canonical basis of $\mathbb{R}^k$ and $\xi^{(1)}, \ldots, \xi^{(k)}$ i.i.d. random vectors of $\mathbb{R}^n$ with common distribution μ, generic chaining was used in Mendelson *et al* (2007) and Mendelson (2008) to study the properties of the random operator $\Gamma : t \mapsto k^{-1/2} \sum_{i=1}^{k} \langle \xi^{(i)}, t\rangle e_i$ defined for t in the unit sphere T of $\mathbb{R}^n$ (which we endow with its usual scalar product $\langle ., .\rangle$). Their results rely on the control of suprema of random variables of the form $X_t = k^{-1} \sum_{i=1}^{k} \langle \xi^{(i)}, t\rangle$ for $t \in T$. When $k = 1$, this form of X_t is analogous to that we consider in our statistical application. However, the deviation bounds obtained in Mendelson *et al* (2007) and Mendelson (2008) require that μ be subgaussian which we do not want to assume here. Closer to our result is Theorem 3.3 in Klartag & Mendelson (2005) which bounds on a set of probability at least $1 - \delta$ (for some $\delta \in (0, 1)$) the supremum $Z = \sup_{t\in T} |X_t|$. Unfortunately, their bound involves non-explicit constants (that depend on δ) which makes it useless for statistical issues.

Our approach also uses generic chaining. With such a technique, the inequalities we get suffer from the usual drawback that the numerical constants are non-optimal but at least allow a suitable control of the χ^2-type random variables we consider in the statistical part of this paper. To our knowledge, these inequalities are new.

1.3. From the control of χ^2-type random variables to model selection in regression

The reason why χ^2-type random variables naturally emerge in the regression setting is the following one. Let S be a linear subspace of $\mathbb{R}^n$. The classical least-squares estimator

of f in S is given by $\hat{f} = \Pi_S Y = \Pi_S f + \Pi_S \xi$ and since the Euclidean (squared) distance beween f and $\hat{f}$ decomposes as

$$\left| f - \hat{f} \right|_2^2 = |f - \Pi_S f|_2^2 + |\Pi_S \xi|_2^2$$

the study of the quadratic loss $|f - \hat{f}|_2^2$ requires that of its random component $|\Pi_S \xi|_2^2$. This quantity is called a χ^2-type random variable by analogy to the Gaussian case. Its study is connected to that of suprema of random variables by the formula

$$|\Pi_S \xi|_2 = \sup_{t \in T} X_t = Z \quad \text{with} \quad X_t = \sum_{i=1}^{n} \xi_i t_i \tag{1.9}$$

where T is the unit ball of S (or a countable and dense subset of it). The control of such random variables is at the heart of the model selection scheme. When ξ is a standard Gaussian vector of $\mathbb{R}^n$, Birgé & Massart (2001) used (1.6) to control the probability of deviation of $|\Pi_S \xi|_2$ with respect to its expectation. The strong integrability properties of the ξ_i allows to handle very general collections of models. By using chaining techniques, these results were extended to the subgaussian case (that is for $X = \pm \xi_i$ satisfying (1.5) with $c = 0$ for all i) in Baraud, Comte & Viennet (2001). Similarly, very few assumptions were required on the collection to perform model selection. Baraud (2000) considered the case where the ξ_i only admit few finite moments. There, the weak integrability properties of the ξ_i induced severe restrictions on the collection of models $\mathcal{S}$. Typically, for all $D \in \{1, \ldots, n\}$ the number of models S_m of a given dimension D had to be at most polynomial with respect to D, the degree of the polynomial depending on the number of finite moments of ξ_1.

To our knowledge, the intermediate case where the random variables $\pm \xi_i$ admit exponential moments of the form (1.5) for all i (with $c \neq 0$ to exclude the already known subgaussian case) has remained open for general collections of models. In this context, the concentration-type inequality obtained in Klein & Rio (2005) cannot be used to control $|\Pi_S \xi|_2$ since it would require that the ξ_i be bounded. An attempt at relaxing this boundedness assumption on the ξ_i can be found in Bousquet (2003). There, the author considered the situation where T is a subset of $[-1, 1]^n$ and the ξ_i independent and centered random variables satisfying

$$\mathbb{E}\left[|\xi_i|^k \right] \leq \frac{k!}{2} \sigma^2 c^{k-2}, \quad \forall\, k \geq 2. \tag{1.10}$$

Note that (1.10) implies that the X_t satisfy (1.5) with $v^2 = v^2(t) = |t|_2^2 \, \sigma^2$. The result by Bousquet provides an analogue of (1.7) with v^2 replaced by $n\sigma^2$ although one would expect the smaller (and usual) quantity $v^2 = \sup_{t \in T} v^2(t)$. Because of this, the resulting inequality turns out to be useless at least for the statistical applications we have in mind. This fact has already been pointed out in Sauvé (2008). Sauvé also tackled the problem of model selection when the ξ_i satisfy (1.10). Compared to Baraud (2000), her condition on the collection of models is weaker in the sense that the number of models with a

given dimension D is allowed to be exponentially large with respect to D. However, the collection she considered only consists of linear spaces S_m with a specific form (leading to regressogram estimators). Besides, her selection procedure was relying on a known upper bound on $\max_{i=1,\ldots,n}|f_i|$ which can be unrealistic in practice. A similar assumption was made in Barron *et al* (1999) (Theorem 4) in the related context of regression with i.i.d. design points. Unlike these authors, our procedure does not depend on such an upper bound on f.

1.4. Organisation of the paper and main notations

The paper is organized as follows. We present our deviation bound for Z in Section 2. The statistical application is developed in Sections 3 and 4. In Section 3 we consider particular cases of collections $\mathcal{S}$ of interest, the general case being considered in Section 4. Section 5 is devoted to the proofs.

Along the paper we assume that $n \geq 2$ and use the following notations. We denote by $e_1, \ldots, e_n$ the canonical basis of $\mathbb{R}^n$ which we endow with the Euclidean inner product denoted $\langle ., . \rangle$. For $x \in \mathbb{R}^n$, we set $|x|_2 = \sqrt{\langle x, x \rangle}$, $|x|_1 = \sum_{i=1}^n |x_i|$ and $|x|_\infty = \max_{i=1,\ldots,n} |x_i|$. The linear span of a family $u_1, \ldots, u_k$ of vectors is denoted by $\mathrm{Span}\{u_1, \ldots, u_k\}$. The quantity $|I|$ is the cardinality of a finite set I. Finally, κ denotes the numerical constant 18. It appears first in the control of the deviation of Z when applying Talagrand's chaining argument and then all along the paper. It seemed interesting to emphasize the influence of this constant in the model selection procedure we propose.

2. A Talagrand-type Chaining argument for controlling suprema of random variables

Let $(X_t)_{t\in T}$ be a family of real valued and centered random variables indexed by a countable and nonempty set T. Fix some t_0 in T and set

$$Z = \sup_{t\in T} (X_t - X_{t_0}) \quad \text{and} \quad \overline{Z} = \sup_{t\in T} |X_t - X_{t_0}| .$$

Our aim is to give a probabilistic control of the deviations of Z (and $\overline{Z}$). We make the following assumptions

***Assumption* 2.1.** There exist two distances d and δ on T and a nonnegative constant c such that for all $s, t \in T$ $(s \neq t)$

$$\mathbb{E}\left[e^{\lambda(X_t - X_s)}\right] \leq \exp\left[\frac{\lambda^2 d^2(s,t)}{2(1 - \lambda c\delta(s,t))}\right], \quad \forall \lambda \in \left[0, \frac{1}{c\delta(s,t)}\right) \tag{2.1}$$

with the convention $1/0 = +\infty$.

Note that $c = 0$ corresponds to the particular situation where the increments of the process X_t are *subgaussian*.

Besides Assumption 2.1, we also assume in this section that d and δ derive from norms. This is the only case we need to consider to handle the statistical problem described in Section 3. Nevertheless, a more general result with arbitrary distances can be found in Section 5.

***Assumption* 2.2.** Let S be a linear space with finite dimension D endowed with two *arbitrary* norms denoted $\| \ \|_2$ and $\| \ \|_\infty$ respectively. Define for $s, t \in S$, $d(s,t) = \|t-s\|_2$ and $\delta(s,t) = \|s-t\|_\infty$ and assume that for constants $v > 0$ and $c \geq 0$,

$$T \subset \left\{ t \in S \ \middle| \ \|t - t_0\|_2 \leq v, \ \ c\|t-t_0\|_\infty \leq b \right\}.$$

Then, the following result holds.

Theorem 2.1. *Under Assumptions 2.1 and 2.2,*

$$\mathbb{P}\left[Z \geq \kappa \left(\sqrt{v^2(D+x)} + b(D+x) \right) \right] \leq e^{-x}, \quad \forall x \geq 0 \tag{2.2}$$

with $\kappa = 18$. *Moreover*

$$\mathbb{P}\left[\overline{Z} \geq \kappa \left(\sqrt{v^2(D+x)} + b(D+x) \right) \right] \leq 2e^{-x}, \quad \forall x \geq 0. \tag{2.3}$$

Since S is separable, the result easily extends to the case where $T \subset S$ is not countable provided the paths $t \mapsto X_t$ are continuous with probability 1 (with respect to $\| \ \|_2$ or $\| \ \|_\infty$, both norms being equivalent on S).

2.1. Connections with deviations inequalities with respect to $\mathbb{E}(Z)$

In this section we make some connections between our bound (2.2) and inequalities (1.6) and (1.7). Along this section, T is the unit ball of the linear span S of an orthonormal system $\{u_1, \ldots, u_D\}$. Both norms $| \ |_2$ and $| \ |_\infty$ being equivalent on S, we set

$$\Lambda_2(S) = \sup_{t \in T \setminus \{0\}} \frac{|t|_\infty}{|t|_2} < +\infty.$$

Note that $\Lambda_2(S)$ depends on the metric structure of S. In all cases, $\Lambda_2(S) \leq 1$, this bound being achieved for $S = \mathrm{Span}\{e_1, \ldots, e_D\}$ for example. However, $\Lambda_2(S)$ can be much smaller, equal to $\sqrt{D/n}$ for example, when $n = kD$ for some positive integer k and $u_j = \left(e_{(j-1)k+1}, \ldots, e_{jk}\right)/\sqrt{k}$ for $j = 1, \ldots, D$. The set T fulfills Assumption 2.2 with $t_0 = 0$, $d(s,t) = |s-t|_2$, $\delta(s,t) = |s-t|_\infty$, $v = 1$ and $b = c\Lambda_2(S)$. Let $\xi = (\xi_1, \ldots, \xi_n)$ be a random vector of $\mathbb{R}^n$ with i.i.d. components of common variance 1. We consider the

process defined on T by $X_t = \langle t, \xi \rangle$ and note that in this case $Z = \sup_{t \in T} X_t = |\Pi_S \xi|_2$. Besides, by using Jensen's inequality

$$\mathbb{E}[Z] = \mathbb{E}\left[\sqrt{\sum_{j=1}^{D} \langle u_j, \xi \rangle^2}\right] \leq \sqrt{D}. \tag{2.4}$$

The Gaussian case: Assume that the ξ_i are standard Gaussian random variables. On the one hand, since $\sup_{t \in T} \mathrm{var}(X_t) = 1$ we deduce from Sudakov & Cirel'son's bound (1.6) together with (2.4)

$$\mathbb{P}\left(Z \geq \sqrt{D} + \sqrt{2x}\right) \leq e^{-x}, \ \forall x \geq 0. \tag{2.5}$$

On the other hand, since (1.5) holds with $c = 0$, for all $s, t \in S$ and $\lambda \geq 0$

$$\begin{aligned} \mathbb{E}\left[e^{\lambda(X_t - X_s)}\right] &= \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda \xi_i (t_i - s_i)}\right] \leq \prod_{i=1}^{n} \exp\left[\frac{\lambda^2 |t_i - s_i|^2}{2}\right] \\ &\leq \exp\left[\frac{\lambda^2 |t - s|_2^2}{2}\right]. \end{aligned}$$

Consequently, (2.1) holds with $c = 0$ and one can apply Theorem 2.1 to get

$$\mathbb{P}\left[Z \geq \kappa\left(\sqrt{D} + \sqrt{x}\right)\right] \leq \mathbb{P}\left(Z \geq \kappa\sqrt{D + x}\right) \leq e^{-x}, \ \forall x \geq 0. \tag{2.6}$$

Apart from the numerical constants, it turns out that (2.5) and (2.6) are similar in this case.

The bounded case: Let us assume that the ξ_i take their values in $[-a, a]$ for some $a \geq 1$. We can apply the bound given by Klein & Rio (2005) with $v = 1$ and $c = a\Lambda_2(S)$ in (1.7) which together with (2.4) gives for a suitable constant $C > 0$,

$$\mathbb{P}\left[Z \geq C\left(\sqrt{D} + \sqrt{x} + a\Lambda_2(S)x\right)\right] \leq \exp(-x) \quad \text{for all } x \geq 0. \tag{2.7}$$

When the ξ_i are bounded, there are actually two ways of applying Theorem 2.1. One relies on the fact that the random variables $\pm\xi_i$ satisfy (1.5) with $v = 1$ and $c = a$ for all i. Hence, whatever $s, t \in S$ and $\lambda \leq (a|s - t|_\infty)^{-1}$,

$$\begin{aligned} \mathbb{E}\left[e^{\lambda(X_t - X_s)}\right] &= \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda \xi_i (t_i - s_i)}\right] \leq \prod_{i=1}^{n} \exp\left[\frac{\lambda^2 |t_i - s_i|^2}{2(1 - \lambda a |t - s|_\infty)}\right] \\ &\leq \exp\left[\frac{\lambda^2 |t - s|_2^2}{2(1 - \lambda a |t - s|_\infty)}\right] \end{aligned}$$

and since Assumption 2.1 holds with $c = a$ and we get from Theorem 2.1

$$\mathbb{P}\left[Z \geq \kappa\left(\sqrt{D} + \sqrt{x} + a\Lambda_2(S)x + a\Lambda_2(S)D\right)\right] \leq e^{-x}, \ \forall x \geq 0. \tag{2.8}$$

Inequalities (2.7) and (2.8) essentially differ by the fact that the latter involves the extra term $a\Lambda_2(S)D$ whenever $x \leq D$. In this case, we recover (2.7) only for those S bearing some specific metric structure for which $\Lambda_2(S) \leq C'(a\sqrt{D})^{-1}$ for some numerical constant $C' > 0$.

The other way of using Theorem 2.1 is to note that the random variables $\pm\xi_i$ are subgaussian (because they are bounded) and therefore satisfy (1.5) with $v = a$ and $c = 0$. By arguing as in the Gaussian case, Assumption 2.1 holds with $d(s,t) = a\,|s-t|_2$ for all $s, t \in S$, $c = 0$ and Assumption 2.2 is fulfilled with $v = a$ and $b = 0$. We deduce from Theorem 2.1

$$\mathbb{P}\left[Z \geq \kappa\left(a\sqrt{D} + a\sqrt{x}\right)\right] \leq e^{-x} \quad \forall x \geq 0. \tag{2.9}$$

Note that whenever a is not too large compared to 1, this bound improves (2.7) by avoiding the linear term $a\Lambda_2(S)x$.

2.2. A counter-example

In this section we show that the supremum Z of a random process $\mathbf{X} = (X_t)_{t\in T}$ satisfying (2.1) may not concentrate around $\mathbb{E}(Z)$. More precisely, let us show that (1.7) could be false under (2.1). A simple counter-example is the following one. For $D \geq 1$, let $S = \mathrm{Span}\,\{e_1, \ldots, e_D\}$, T be the unit ball of S and $\mathbf{X}' = (X'_t)_{t\in T}$ the Gaussian process defined for $t \in T$ by $t \mapsto \langle t, \xi\rangle$ where ξ is a standard Gaussian vector of $\mathbb{R}^n$. For some $p \in (0,1)$ to be chosen later on, define $\mathbf{X}$ as either $\mathbf{X}'$ with probability p or as the process $\mathbf{X}''$ identically equal to 0 with probability $1-p$. On the one hand, note that both processes $\mathbf{X}'$ and $\mathbf{X}''$ satisfy (2.1) with $c = 0$, $d(s,t) = |s-t|_2$ for all $s, t \in S$ and therefore so does $\mathbf{X}$ (whatever p). On the other hand, since

$$\mathbb{E}(Z) = p\mathbb{E}\left[\sup_{t\in T} X'_t\right] = p\mathbb{E}\left[\sqrt{\sum_{i=1}^D \xi_i^2}\right] \leq p\sqrt{D}$$

and $\sup_{t\in T}\mathrm{var}(X_t) \leq 1$, (1.7) would imply that for some positive numerical constant C (that we can take larger than 1 with no loss of generality) and all $u \geq 0$,

$$\begin{aligned}\mathbb{P}\left[Z \geq Cp\sqrt{D} + C\left(\sqrt{u}+u\right)\right] &= p\mathbb{P}\left[\sqrt{\sum_{i=1}^D \xi_i^2} \geq Cp\sqrt{D} + C\left(\sqrt{u}+u\right)\right] \\ &\leq e^{-u}.\end{aligned}$$

By choosing $p = (2C)^{-1} \in (0,1)$ and $u = \log(2/p)$, we would get

$$\mathbb{P}\left[\sqrt{\frac{1}{D}\sum_{i=1}^D \xi_i^2} \geq \frac{1}{2} + \frac{C}{\sqrt{D}}\left(\sqrt{\log(2/p)} + \log(2/p)\right)\right] \leq \frac{1}{2}$$

which is of course false by the law of large numbers for large values of D.

3. Applications to model selection in regression

Consider the regression framework given by (1.1) and assume that for some known nonnegative numbers σ and c

$$\log \mathbb{E}\left[e^{\lambda \xi_i}\right] \leq \frac{\lambda^2 \sigma^2}{2(1-|\lambda| c)} \quad \text{for all} \quad \lambda \in (-1/c, 1/c) \text{ and } i = 1, \ldots, n. \tag{3.1}$$

Inequality (3.1) holds for a large class of distributions (once suitably centered) including Gaussian, Poisson, Laplace or Gamma (among others). Besides, (3.1) is fulfilled when the ξ_i satisfy (1.10) and therefore whenever these are bounded.

Our estimation strategy is based on model selection. We start with a (possibly large) collection $\{S_m,\ m \in \mathcal{M}\}$ of linear subspaces (models) of $\mathbb{R}^n$ and associate to each of these the least-squares estimators $\hat{f}_m = \Pi_{S_m} Y$. Given a penalty function pen from $\mathcal{M}$ to $\mathbb{R}_+$, we define the penalized criterion crit(.) on $\mathcal{M}$ by

$$\text{crit}(m) = \left|Y - \hat{f}_m\right|_2^2 + \text{pen}(m). \tag{3.2}$$

In this section, we propose to establish risk bounds for the estimator of f given by $\hat{f}_{\hat{m}}$ where the index $\hat{m}$ is selected from the data among $\mathcal{M}$ as any minimizer of crit(.).

In the sequel, the penalty pen will be based on some *a priori* choice of nonnegative numbers $\{\Delta_m,\ m \in \mathcal{M}\}$ for which we set

$$\Sigma = \sum_{m \in \mathcal{M}} e^{-\Delta_m} < +\infty.$$

When $\Sigma = 1$, the choice of the Δ_m can be viewed as that of a prior distribution on the models S_m. For related conditions and their interpretation, see Barron and Cover (1991) or Barron *et al* (1999).

In the following sections, we present some applications of our main result (to be presented in Subsection 4.2) for some collections of linear spaces $\{S_m,\ m \in \mathcal{M}\}$ of interest.

3.1. Selecting among histogram-type estimators

For a partition m of $\{1, \ldots, n\}$, S_m denotes the linear span of vectors of $\mathbb{R}^n$ the coordinates of which are constants on each element I of m. In the sequel, we shall restrict to partitions m the elements of which consist of consecutive integers.

Consider a partition $\mathfrak{m}$ of $\{1, \ldots, n\}$ and $\mathcal{M}$ a collection of partitions m such that $S_m \subset S_{\mathfrak{m}}$. We obtain the following result.

***Proposition* 3.1.** Let $a, b > 0$. Assume that

$$|I| \geq a^2 \log^2 n, \quad \forall I \in \mathfrak{m}. \tag{3.3}$$

If for some $K > 1$,

$$\mathrm{pen}(m) \geq K\kappa^2 \left(\sigma^2 + 2c\frac{(\sigma+c)(b+2)}{a\kappa}\right)(|m| + \Delta_m), \quad \forall m \in \mathcal{M} \tag{3.4}$$

the estimator $\hat{f}_{\hat{m}}$ satisfies

$$\mathbb{E}\left(\left|f - \hat{f}_{\hat{m}}\right|_2^2\right) \leq C(K)\left[\inf_{m\in\mathcal{M}}\left[\mathbb{E}\left(\left|f - \hat{f}_m\right|_2^2\right) + \mathrm{pen}(m)\right] + R\right] \tag{3.5}$$

where $C(K)$ is given by (4.4) and

$$R = \kappa^2\left(\sigma^2 + 2c\frac{(c+\sigma)(b+2)}{a\kappa}\right)\Sigma + 2\frac{(c+\sigma)^2(b+2)^2}{a^2 n^b}.$$

Note that when $c = 0$, inequality (3.4) holds as soon as

$$\mathrm{pen}(m) = K\kappa^2\sigma^2(|m| + \Delta_m), \quad \forall m \in \mathcal{M}. \tag{3.6}$$

Besides, by taking $a = (\log n)^{-1}$ we see that condition (3.3) becomes automatically satisfied and by letting b tend to $+\infty$, inequality (3.5) holds with pen given by (3.6) and $R = \kappa^2\sigma^2\Sigma$.

The problem of selecting among histogram-type estimators in this regression setting has recently been investigated in Sauvé (2008). Her selection procedure is similar to ours with a different choice of the penalty term. Unlike hers, our penalty does not involve any known upper bound on $|f|_\infty$.

3.2. Families of piecewise polynomials

In this section, we assume that $f = (F(x_1), \ldots, F(x_n))$ where $x_i = i/n$ for $i = 1, \ldots, n$ and F is an unknown function on $(0, 1]$. Our aim is to estimate F by a piecewise polynomial of degree not larger than d based on a data-driven choice of a partition of $(0, 1]$.

In the sequel, we shall consider partitions m of $\{1, \ldots, n\}$ such that each element $I \in m$ consists of at least $d + 1$ consecutive integers. For such a partition, S_m denotes the linear span of vectors of the form $(P(1/n), \ldots, P(n/n))$ where P varies among the space of piecewise polynomials with degree not larger than d based on the partition of $(0, 1]$ given by

$$\left\{\left(\frac{\min I - 1}{n}, \frac{\max I}{n}\right],\ I \in m\right\}.$$

Consider a partition $\mathfrak{m}$ of $\{1, \ldots, n\}$ and $\mathcal{M}$ a collection of partitions m such that $S_m \subset S_{\mathfrak{m}}$. We obtain the following result.

***Proposition* 3.2.** Let $a, b > 0$. Assume that

$$|I| \geq (d+1)a^2 \log^2 n \geq d + 1, \quad \forall I \in \mathfrak{m}. \tag{3.7}$$

If for some $K > 1$,

$$\mathrm{pen}(m) \geq K\kappa^2 \left(\sigma^2 + c\frac{4\sqrt{2}(\sigma + c)(d+1)(b+2)}{a\kappa} \right) (D_m + \Delta_m) \quad \forall m \in \mathcal{M}$$

the estimator $\hat{f}_{\hat{m}}$ satisfies (3.5) with

$$R = \kappa^2 \left(\sigma^2 + c\frac{4\sqrt{2}(\sigma + c)(d+1)(b+2)}{a\kappa} \right) \Sigma + 4\frac{(c+\sigma)^2(b+2)^2}{a^2 n^b}.$$

3.3. Families of trigonometric polynomials

We assume that f has the same form as in Subsection 3.2. Here, our aim is to estimate F by a trigonometric polynomial of degree not larger than some $\overline{D} \geq 0$.

Consider the (discrete) trigonometric system $\{\phi_j\}_{j \geq 0}$ of vectors in $\mathbb{R}^n$ defined by

$$\begin{aligned} \phi_0 &= (1/\sqrt{n}, \ldots, 1/\sqrt{n}) \\ \phi_{2j-1} &= \sqrt{\frac{2}{n}} \left(\cos\left(2\pi j x_1\right), \ldots, \cos\left(2\pi j x_1\right) \right), \ \forall j \geq 1 \\ \phi_{2j} &= \sqrt{\frac{2}{n}} \left(\sin\left(2\pi j x_1\right), \ldots, \sin\left(2\pi j x_1\right) \right), \ \forall j \geq 1. \end{aligned}$$

Let $\mathcal{M}$ be a family of subsets of $\left\{0, \ldots, 2\overline{D}\right\}$. For $m \in \mathcal{M}$, we define S_m as the linear span of the ϕ_j with $j \in m$ (with the convention $S_m = \{0\}$ when $m = \varnothing$).

***Proposition* 3.3.** Let $a, b > 0$. Assume that $2\overline{D} + 1 \leq \sqrt{n}/(a \log n)$. If for some $K > 1$,

$$\mathrm{pen}(m) \geq K\kappa^2 \left(\sigma^2 + \frac{4c(c+\sigma)(b+2)}{a} \right) (D_m + \Delta_m), \quad \forall m \in \mathcal{M}$$

then $\hat{f}_{\hat{m}}$ satisfies (3.5) with

$$R = \kappa^2 \left(\sigma^2 + \frac{4c(c+\sigma)(b+2)}{a} \right) \Sigma + \frac{4(b+2)^2(c+\sigma)^2}{a^2(2\overline{D}+1)n^b}.$$

4. Towards a more general result

We consider the statistical framework presented in Section 3 and give a general result that allows to handle Propositions 3.1, 3.2 and 3.3 simultaneously. It will rely on some geometric properties of the linear spaces S_m that we describe below.

4.1. Some metric quantities

Let S be a linear subspace of $\mathbb{R}^n$. We associate to S the following quantities

$$\Lambda_2(S) = \max_{i=1,\dots,n} |\Pi_S e_i|_2 \quad \text{and} \quad \Lambda_\infty(S) = \max_{i=1,\dots,n} |\Pi_S e_i|_1. \tag{4.1}$$

It is not difficult to see that these quantities can be interpreted in terms of norm connexions, more precisely

$$\Lambda_2(S) = \sup_{t\in S\setminus\{0\}} \frac{|t|_\infty}{|t|_2} \quad \text{and} \quad \Lambda_\infty(S) = \sup_{t\in \mathbb{R}^n\setminus\{0\}} \frac{|\Pi_S t|_\infty}{|t|_\infty}.$$

Clearly, $\Lambda_2(S) \leq 1$. Besides, since $|x|_1 \leq \sqrt{n}\,|x|_2$ for all $x \in \mathbb{R}^n$, $\Lambda_\infty(S) \leq \sqrt{n}\Lambda_2(S)$. Nevertheless, these bounds can be rather rough and turn out to be much smaller for the linear spaces S_m presented in Subsections 3.1, 3.2 and 3.3.

4.2. The main result

Let $\{S_m,\ m \in \mathcal{M}\}$ be family of linear spaces and $\{\Delta_m,\ m \in \mathcal{M}\}$ a family of nonnegative weights. We define $\mathcal{S}_n = \sum_{m\in\mathcal{M}} S_m$ and

$$\overline{\Lambda}_\infty = \left(\sup_{m,m'\in\mathcal{M}} \Lambda_\infty(S_m + S_{m'})\right) \vee 1.$$

Theorem 4.1. *Let $K > 1$ and $z \geq 0$. Assume that for all $i = 1,\dots,n$, inequality (3.1) holds. Let* pen *be some penalty function satisfying*

$$\mathrm{pen}(m) \geq K\kappa^2 \left(\sigma^2 + \frac{2cu}{\kappa}\right)(D_m + \Delta_m), \quad \forall m \in \mathcal{M} \tag{4.2}$$

where

$$u = (c+\sigma)\overline{\Lambda}_\infty \Lambda_2(\mathcal{S}_n)\log(n^2 e^z). \tag{4.3}$$

If one selects $\hat{m}$ among $\mathcal{M}$ as any minimizer of crit(.) *defined by (3.2) then*

$$\mathbb{E}\left[\left|f - \hat{f}_{\hat{m}}\right|_2^2\right] \leq C(K)\left[\inf_{m\in\mathcal{M}}\left(\mathbb{E}\left[\left|f - \hat{f}_m\right|_2^2\right] + \mathrm{pen}(m)\right) + R\right]$$

where

$$C(K) \quad = \quad \frac{K(K^2+K-1)}{(K-1)^3} \tag{4.4}$$

and $R = \kappa^2\left(\sigma^2 + 2\kappa^{-1}cu\right)\Sigma + 2u^2\overline{\Lambda}_\infty^{-2}e^{-z}$.

When $c = 0$ we derive the following corollary by letting z grow towards infinity.

***Corollary* 4.1.** Let $K > 1$. Assume that the ξ_i for $i = 1, \ldots, n$ satisfy inequality (3.1) with $c = 0$. If one selects $\hat{m}$ among $\mathcal{M}$ as a minimizer of crit defined by (3.2) with pen satisfying

$$\mathrm{pen}(m) \geq K\kappa^2\sigma^2 \left(D_m + \Delta_m\right), \quad \forall m \in \mathcal{M}$$

then

$$\mathbb{E}\left[\left|f - \hat{f}_{\hat{m}}\right|_2^2\right] \leq \frac{K(K^2+K-1)}{(K-1)^3} \inf_{m\in\mathcal{M}} \left(\mathbb{E}\left[\left|f - \hat{f}_m\right|_2^2\right] + \mathrm{pen}(m)\right) + R$$

where $R = K^3(K-1)^{-2}\kappa^2\sigma^2\Sigma$.

5. Proofs

We start with the following result generalizing Theorem 2.1 when d and δ are not induced by norms. We assume that T is finite and take numbers v and b such that

$$\sup_{s\in T} d(s, t_0) \leq v, \quad \sup_{s\in T} c\delta(s, t_0) \leq b. \tag{5.1}$$

We consider now a family of finite partitions $(\mathcal{A}_k)_{k\geq 0}$ of T, such that $\mathcal{A}_0 = \{T\}$ and for $k \geq 1$ and $A \in \mathcal{A}_k$

$$d(s,t) \leq 2^{-k}v \ \text{ and } \ c\delta(s,t) \leq 2^{-k}b, \quad \forall s,t \in A.$$

Besides, we assume $\mathcal{A}_k \subset \mathcal{A}_{k-1}$ for all $k \geq 1$, which means that all elements $A \in \mathcal{A}_k$ are subsets of an element of $\mathcal{A}_{k-1}$. Finally, we define for $k \geq 0$

$$N_k = |\mathcal{A}_{k+1}||\mathcal{A}_k|.$$

Theorem 5.1. *Let T be some finite set. Under Assumption 2.1,*

$$\mathbb{P}\left(Z \geq H + 2\sqrt{2v^2x} + 2bx\right) \leq e^{-x}, \quad \forall x > 0 \tag{5.2}$$

where

$$H = \sum_{k\geq 0} 2^{-k}\left(v\sqrt{2\log(2^{k+1}N_k)} + b\log(2^{k+1}N_k)\right).$$

Moreover,

$$\mathbb{P}\left(\overline{Z} \geq H + 2\sqrt{2v^2x} + 2bx\right) \leq 2e^{-x}, \quad \forall x > 0. \tag{5.3}$$

The quantity H can be related to the entropies of T with respect to the distances d and $c\delta$ (when $c \neq 0$) in the following way. We first recall that for a distance $e(.,.)$ on T and $\varepsilon > 0$, the entropy $H(T, e, \varepsilon)$ is defined as logarithm of the minimum number of balls of radius ε with respect to e which are necessary to cover T. For $\varepsilon > 0$, let us set $H(T, \varepsilon) = \max\{H(T, d, \varepsilon v), H(T, c\delta, \varepsilon b)\}$. Note that $H(T, \varepsilon) = 0$ for $\varepsilon > 1$ because

of (5.1). For $\varepsilon < 1$, one can bound $H(T,\varepsilon)$ from above as follows. For $k \geq 0$, each element A of the partition $\mathcal{A}_{k+1}$ is both a subset of a ball of radius $2^{-(k+1)}v$ with respect to d and of a ball of radius $2^{-(k+1)}b$ with respect $c\delta$. Since $|\mathcal{A}_{k+1}| \leq N_k$, we obtain for all $\varepsilon \in [2^{-(k+1)}, 2^{-k})$, $H(T,\varepsilon) \leq \log N_k$ and by integrating with respect to ε and summing over $k \geq 0$, we get

$$\int_0^1 \left(\sqrt{2v^2 H(T,\varepsilon)} + bH(T,\varepsilon) \right) d\varepsilon \leq H.$$

5.1. Proof of Theorem 5.1

Note that we obtain (5.3) by using (5.2) twice (once with X_t and then with $-X_t$). Let us now prove (5.2). For each $k \geq 1$ and $A \in \mathcal{A}_k$, we choose some arbitrary element $t_k(A)$ in A. For each $t \in T$ and $k \geq 1$, there exists a unique $A \in \mathcal{A}_k$ such that $t \in A$ and we set $\pi_k(t) = t_k(A)$. When $k = 0$, we set $\pi_0(t) = t_0$.

We consider the (finite) decomposition

$$X_t - X_{t_0} = \sum_{k \geq 0} X_{\pi_{k+1}(t)} - X_{\pi_k(t)}$$

and set for $k \geq 0$

$$z_k = 2^{-k} \left(v\sqrt{2\left(\log(2^{k+1}N_k) + x\right)} \; + \; b\left(\log(2^{k+1}N_k) + x\right) \right)$$

Since $\sum_{k\geq 0} z_k \leq z = H + 2v\sqrt{2x} + 2bx$,

$$\begin{aligned} \mathbb{P}\left(Z \geq z\right) &\leq \mathbb{P}\left(\exists t,\ \exists k \geq 0,\ \ X_{\pi_{k+1}(t)} - X_{\pi_k(t)} \geq z_k\right) \\ &\leq \sum_{k\geq 0} \sum_{(s,u)\in E_k} \mathbb{P}\left(X_u - X_s \geq z_k\right) \end{aligned}$$

where

$$E_k = \left\{ (\pi_k(t), \pi_{k+1}(t)) \mid t \in T \right\}.$$

Since $\mathcal{A}_{k+1} \subset \mathcal{A}_k$, $\pi_k(t)$ and $\pi_{k+1}(t)$ belong to a same element of $\mathcal{A}_k$ and therefore $d(s,u) \leq 2^{-k}v$ and $c\delta(s,u) \leq 2^{-k}b$ for all pairs $(s,u) \in E_k$. Besides, under Assumption 2.1, the random variable $X = X_u - X_s$ with $(s,u) \in E_k$ is centered and satisfies (1.5) with $2^{-k}v$ and $2^{-k}b$ in place of v and c. Hence, by using Bernstein's inequality (1.3), we get for all $(s,u) \in E_k$ and $k \geq 0$

$$\mathbb{P}\left(X_u - X_s \geq z_k\right) \leq 2^{-(k+1)} N_k^{-1} e^{-x} \leq 2^{-(k+1)} |E_k|^{-1} e^{-x}.$$

Finally, we obtain inequality (5.2) summing up this inequalities over $(s,u) \in E_k$ and $k \geq 0$.

5.2. Proof of Theorem 2.1

We only prove (2.2), the argument for proving (2.3) being the same as that for proving (5.3). For $t \in S$ and $r > 0$, we denote by $B_2(t,r)$ and $B_\infty(t,r)$ the balls centered at t of radius r associated to $\|\ \|_2$ and $\|\ \|_\infty$ respectively. In the sequel, we shall use the following result on the entropy of those balls.

***Proposition* 5.1.** Let $\|\ \|$ be an arbitrary norm on S and $B(0,1)$ the corresponding unit ball. For each $\delta \in (0,1]$, the minimal number $\mathcal{N}(\delta)$ of balls of radius δ (with respect to $\|\ \|$) which are necessary to cover $B(0,1)$ satisfies

$$\mathcal{N}(\delta) \leq \left(1 + 2\delta^{-1}\right)^D.$$

The proof of this classical lemma can be found in Birgé (1983) (Lemma 4.5, p. 209). Let us now turn to the proof of (2.2). Note that it is enough to prove that for some $u < H + 2\sqrt{2v^2x} + 2bx$ and all finite sets T satisfying inequalities (2.1) and (5.1)

$$\mathbb{P}\left(\sup_{t\in T}(X_t - X_{t_0}) > u\right) \leq e^{-x}.$$

Indeed, for any sequence $(T_n)_{n\geq 0}$ of finite subsets of T increasing towards T, that is, satisfying $T_n \subset T_{n+1}$ for all $n \geq 0$ and $\bigcup_{n\geq 0} T_n = T$, the sets

$$\left\{\sup_{t\in T_n}(X_t - X_{t_0}) > u\right\}$$

increases (for the inclusion) towards $\{Z > u\}$. Therefore,

$$\mathbb{P}(Z > u) = \lim_{n\to+\infty} \mathbb{P}\left(\sup_{t\in T_n}(X_t - X_{t_0}) > u\right).$$

Consequently, we shall assume hereafter that T is finite.

For $k \geq 0$ and $j \in \{2,\infty\}$ define the sets $\mathcal{A}_{j,k}$ as follows. We first consider the case $j = 2$. For $k = 0$, $\mathcal{A}_{2,0} = \{T\}$. By applying Proposition 5.1 with $\|\ \| = \|\ \|_2/v$ and $\delta = 1/4$, we can cover $T \subset B_2(t_0, v)$ with at most 9^D balls with radius $v/4$. From such a finite covering $\{B_1, \ldots, B_N\}$ with $N \leq 9^D$, it is easy to derive a partition $\mathcal{A}_{2,1}$ of T by at most 9^D sets of diameter not larger than $v/2$. Indeed, $\mathcal{A}_{2,1}$ can merely consist of the non-empty sets among the family

$$\left\{\left(B_k \setminus \bigcup_{1\leq\ell<k} B_\ell\right) \cap T, \quad k = 1, \ldots, N\right\}$$

(with the convention $\bigcup_\varnothing = \varnothing$). Then, for $k \geq 2$, proceed by induction using Proposition 5.1 repeatedly. Each element $A \in \mathcal{A}_{2,k-1}$ is a subset of a ball of radius $2^{-k}v$ and can

be partitioned similarly as before into 5^D subsets of balls of radii $2^{-(k+1)}v$. By doing so, the partitions $\mathcal{A}_{2,k}$ with $k \geq 1$ satisfy $\mathcal{A}_{2,k} \subset \mathcal{A}_{2,k-1}$, $|\mathcal{A}_{2,k}| \leq (1.8)^D \times 5^{kD}$ and for all $A \in \mathcal{A}_{2,k}$,

$$\sup_{s,t \in A} \|s-t\|_2 \leq 2^{-k}v.$$

Let us now turn to the case $j = +\infty$. If $c > 0$, define the partitions $\mathcal{A}_{\infty,k}$ in exactly the same way as we did for the $\mathcal{A}_{2,k}$. Similarly, the partitions $\mathcal{A}_{\infty,k}$ with $k \geq 1$ satisfy $\mathcal{A}_{\infty,k} \subset \mathcal{A}_{\infty,k-1}$, $|\mathcal{A}_{\infty,k}| \leq (1.8)^D \times 5^{kD}$ and for all $A \in \mathcal{A}_{\infty,k}$,

$$\sup_{s,t \in A} c\|s-t\|_\infty \leq 2^{-k}b.$$

When $c = 0$, we simply take $\mathcal{A}_{\infty,k} = \{T\}$ for all $k \geq 0$ and note that the properties above are fulfilled as well.

Finally, define the partition $\mathcal{A}_k$ for $k \geq 0$ as that generated by $\mathcal{A}_{2,k}$ and $\mathcal{A}_{\infty,k}$, that is

$$\mathcal{A}_k = \{A_2 \cap A_\infty |\ A_2 \in \mathcal{A}_{2,k},\ A_\infty \in \mathcal{A}_{\infty,k}\}.$$

Clearly, $\mathcal{A}_{k+1} \subset \mathcal{A}_k$. Besides, $|\mathcal{A}_0| = 1$ and for $k \geq 1$,

$$|\mathcal{A}_k| \leq |\mathcal{A}_{2,k}||\mathcal{A}_{\infty,k}| \leq (1.8)^{2D} \times 5^{2kD}.$$

The set T being finite, we can apply Theorem 5.1. Actually, our construction of the $\mathcal{A}_k$ allows us to slightly gain in the constants. Going back to the proof of Theorem 5.1, we note that

$$|E_k| = |\{(\pi_k(t), \pi_{k+1}(t)) \mid t \in T\}| \leq |\mathcal{A}_{k+1}| \leq 9^{2D} \times 5^{2kD}$$

since the element $\pi_{k+1}(t)$ determines $\pi_k(t)$ in a unique way. This means that one can take $N_k = 9^{2D} \times 5^{2kD}$ in the proof of Theorem 5.1. By taking the notations of Theorem 5.1, we have,

$$\begin{aligned} H &\leq \sum_{k \geq 0} 2^{-k} \left[v\sqrt{2\log(2^{k+1} \times 9^{2D} \times 5^{2kD})} + b\log\left(2^{k+1} \times 9^{2D} \times 5^{2kD}\right) \right] \\ &< 14\sqrt{Dv^2} + 18Db \end{aligned}$$

and using the concavity of $x \mapsto \sqrt{x}$, we get

$$\begin{aligned} H + 2\sqrt{2v^2x} + 2bx &\leq 14\sqrt{Dv^2} + 2\sqrt{2v^2x} + 18b(D+x) \\ &\leq 18\left(\sqrt{v^2(D+x)} + b(D+x)\right). \end{aligned}$$

which leads to the result.

5.3. Control of χ^2-type random variables

We have the following result.

Theorem 5.2. *Let S be some linear subspace of $\mathbb{R}^n$ with dimension D. If the coordinates of ξ are independent and satisfy (3.1), for all $x, u > 0$,*

$$\mathbb{P}\left[|\Pi_S\xi|_2^2 \geq \kappa^2\left(\sigma^2 + \frac{2cu}{\kappa}\right)(D+x)\,,\ |\Pi_S\xi|_\infty \leq u\right] \leq e^{-x} \tag{5.4}$$

with $\kappa = 18$ and

$$\mathbb{P}\left(|\Pi_S\xi|_\infty \geq x\right) \leq 2n\exp\left[-\frac{x^2}{2\Lambda_2^2(S)\left(\sigma^2 + cx\right)}\right] \tag{5.5}$$

where $\Lambda_2(S)$ is defined by (4.1).

Proof. Let us set $\chi = |\Pi_S\xi|_2$. For $t \in S$, let $X_t = \langle \xi, t\rangle$ and $t_0 = 0$. It follows from the independence of the ξ_i and inequality (3.1) that (2.1) holds with $d(t,s) = \sigma|t-s|_2$ and $\delta(t,s) = |t-s|_\infty$, for all $s, t \in S$. The random variable χ equals the supremum of the X_t when t runs among the unit ball of S. Besides, the supremum is achieved for $\hat{t} = \Pi_S\xi/\chi$ and thus, on the event $\{\chi \geq z,\ |\Pi_S\xi|_\infty \leq u\}$

$$\chi = \sup_{t\in T} X_t \quad \text{with } T = \left\{t \in S,\ |t|_2 \leq 1,\ |t|_\infty \leq uz^{-1}\right\}$$

leading to the bound

$$\mathbb{P}\left(\chi \geq z,\ |\Pi_S\xi|_\infty \leq u\right) \leq \mathbb{P}\left(\sup_{t\in T} X_t \geq z\right).$$

We take $z = \kappa\sqrt{(\sigma^2 + 2cu\kappa^{-1})(D+x)}$ and (using the concavity of $x \mapsto \sqrt{x}$) note that

$$z \geq \kappa\left(\sqrt{\sigma^2(D+x)} + cuz^{-1}(D+x)\right).$$

Then, by applying Theorem 2.1 with $v = \sigma$, $b = cu/z$, we obtain (5.4).

Let us now turn to (5.5). Under (3.1), we can apply Bernstein's inequality (1.3) to $X = \langle \xi, t\rangle$ and $X = \langle -\xi, t\rangle$ with $t \in S$, $v^2 = \sigma^2|t|_2^2$ and $c|t|_\infty$ in place of c and get for all $t \in S$ and $x > 0$

$$\mathbb{P}\left(|\langle \xi, t\rangle| \geq x\right) \leq 2\exp\left[-\frac{x^2}{2\left(\sigma^2|t|_2^2 + c|t|_\infty x\right)}\right]. \tag{5.6}$$

Let us take $t = \Pi_S e_i$ with $i \in \{1, \ldots, n\}$. Since $|t|_2 \leq \Lambda_2(S)$ and

$$|t|_\infty = \max_{i,i'=1,\ldots,n} |\langle \Pi_S e_i, e_{i'}\rangle| = \max_{i,i'=1,\ldots,n} |\langle \Pi_S e_i, \Pi_S e_{i'}\rangle| \leq \Lambda_2^2(S),$$

for all $i \in \{1, \ldots, n\}$

$$\mathbb{P}\left(|\langle \Pi_S\xi, e_i\rangle| \geq x\right) \quad \leq \quad 2\exp\left[-\frac{x^2}{2\Lambda_2^2(S)\left(\sigma^2 + cx\right)}\right].$$

We obtain (5.5) by summing up these probabilities for $i = 1, \ldots, n$. □

5.4. Proof of Theorem 4.1

Let us fix some $m \in \mathcal{M}$. It follows from simple algebra and the inequality $\mathrm{crit}(\hat{m}) \leq \mathrm{crit}(m)$ that

$$\left|f - \hat{f}_{\hat{m}}\right|_2^2 \leq \left|f - \hat{f}_m\right|_2^2 + 2\langle \xi, \hat{f}_{\hat{m}} - \hat{f}_m \rangle + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}).$$

Using the elementary inequality $2ab \leq a^2 + b^2$ for all $a, b \in \mathbb{R}$, we have for $K > 1$,

$$\begin{aligned} 2\langle \xi, \hat{f}_{\hat{m}} - \hat{f}_m \rangle &\leq 2\left|\hat{f}_{\hat{m}} - \hat{f}_m\right|_2 |\Pi_{S_m + S_{\hat{m}}}\xi|_2 \\ &\leq K^{-1}\left[\left(1 + \frac{K-1}{K}\right)\left|\hat{f}_{\hat{m}} - f\right|_2^2 + \left(1 + \frac{K}{K-1}\right)\left|f - \hat{f}_m\right|_2^2\right] \\ &\quad + K\,|\Pi_{S_m + S_{\hat{m}}}\xi|_2^2, \end{aligned}$$

and we derive

$$\begin{aligned} \frac{(K-1)^2}{K^2}\left|f - \hat{f}_{\hat{m}}\right|_2^2 &\leq \frac{K^2 + K - 1}{K(K-1)}\left|f - \hat{f}_m\right|_2^2 + K\,|\Pi_{S_m + S_{\hat{m}}}\xi|_2^2 - (\mathrm{pen}(\hat{m}) - \mathrm{pen}(m)) \\ &\leq \frac{K^2 + K - 1}{K(K-1)}\left|f - \hat{f}_m\right|_2^2 + \mathrm{pen}(m) \\ &\quad + K\,|\Pi_{S_m + S_{\hat{m}}}\xi|_2^2 - (\mathrm{pen}(\hat{m}) + \mathrm{pen}(m)). \end{aligned}$$

Setting

$$\begin{aligned} A_1(\hat{m}) &= K\kappa^2\left(\sigma^2 + \frac{2cu}{\kappa}\right)\left(\frac{|\Pi_{S_m + S_{\hat{m}}}\xi|_2^2}{\kappa^2\left(\sigma^2 + \frac{2cu}{\kappa}\right)} - D_{\hat{m}} - D_m - \Delta_{\hat{m}} - \Delta_m\right)_+ \mathbb{1}\left\{|\Pi_{S_m + S_{\hat{m}}}\xi|_\infty \leq u\right\} \\ A_2(\hat{m}) &= K\,|\Pi_{S_m + S_{\hat{m}}}\xi|_2^2\,\mathbb{1}\left\{|\Pi_{S_m + S_{\hat{m}}}\xi|_\infty \geq u\right\} \end{aligned}$$

and using (4.2), we deduce that

$$\frac{(K-1)^2}{K^2}\left|f - \hat{f}_{\hat{m}}\right|_2^2 \leq \frac{K^2 + K - 1}{K(K-1)}\left|f - \hat{f}_m\right|_2^2 + \mathrm{pen}(m) + A_1(\hat{m}) + A_2(\hat{m}),$$

and by taking the expectation on both side we get

$$\frac{(K-1)^2}{K^2}\mathbb{E}\left[\left|f - \hat{f}_{\hat{m}}\right|_2^2\right] \leq \frac{K^2 + K - 1}{K(K-1)}\mathbb{E}\left[\left|f - \hat{f}_m\right|_2^2\right] + \mathrm{pen}(m) + \mathbb{E}\left[A_1(\hat{m})\right] + \mathbb{E}\left[A_2(\hat{m})\right].$$

The index m being arbitrary, it remains to bound $E_1 = \mathbb{E}\left[A_1(\hat{m})\right]$ and $E_2 = \mathbb{E}\left[A_2(\hat{m})\right]$ from above.

Let m' be some deterministic index in $\mathcal{M}$. By using Theorem 5.2 with $S = S_m + S_{m'}$ the dimension of which is not larger than $D_m + D_{m'}$ and integrating (5.4) with respect to x we get

$$\mathbb{E}\left[A(m')\right] \leq K\kappa^2 \left(\sigma^2 + \frac{2cu}{\kappa}\right) e^{-\Delta_m - \Delta_{m'}}$$

and thus

$$E_1 \leq \sum_{m' \in \mathcal{M}} \mathbb{E}\left[A(m')\right] \leq K\kappa^2 \left(\sigma^2 + \frac{2cu}{\kappa}\right) \Sigma.$$

Let us now turn to $\mathbb{E}\left[A_2(\hat{m})\right]$. By using that $S_{\hat{m}} + S_m \subset \mathcal{S}_n$, $|\Pi_{S_{\hat{m}}+S_m}\xi|_2^2 \leq |\Pi_{\mathcal{S}_n}\xi|_2^2 \leq n\,|\Pi_{\mathcal{S}_n}\xi|_\infty^2$. Besides, it follows from the definition of $\overline{\Lambda}_\infty$ that

$$|\Pi_{S_{\hat{m}}+S_m}\xi|_\infty = |\Pi_{S_{\hat{m}}+S_m}\Pi_{\mathcal{S}_n}\xi|_\infty \leq \overline{\Lambda}_\infty\,|\Pi_{\mathcal{S}_n}\xi|_\infty\,.$$

and therefore, setting $x_0 = \overline{\Lambda}_\infty^{-1} u$

$$E_2 \quad \leq \quad Kn\mathbb{E}\left[|\Pi_{\mathcal{S}_n}\xi|_\infty^2\, \mathbb{1}\left\{|\Pi_{\mathcal{S}_n}\xi|_\infty \geq x_0\right\}\right].$$

We shall now use the following lemma the proof of which can be found in Baraud (2009).

***Lemma* 5.1.** Let X be some nonnegative random variable satisfying for all $x > 0$,

$$\mathbb{P}\left(X \geq x\right) \leq a \exp\left[-\phi(x)\right] \quad \text{with} \quad \phi(x) = \frac{x^2}{2\left(\alpha + \beta x\right)} \tag{5.7}$$

where $a, \alpha > 0$ and $\beta \geq 0$. For $x_0 > 0$ such that $\phi(x_0) \geq 1$,

$$\mathbb{E}\left[X^p \mathbb{1}\left\{X \geq x_0\right\}\right] \leq a x_0^p e^{-\phi(x_0)} \left(1 + \frac{ep!}{\phi(x_0)}\right), \quad \forall p \geq 1.$$

We apply the lemma with $p = 2$ and $X = |\Pi_{\mathcal{S}_n}\xi|_\infty$ for which we know from (5.5) that (5.7) holds with $a = 2n$, $\alpha = \Lambda_2^2(S)\sigma^2$ and $\beta = \Lambda_2^2(S)c$. Besides, it follows from the definition of x_0 and the fact that $n \geq 2$ that

$$\phi(x_0) = \frac{x_0^2}{2\Lambda_2^2(S)\left(\sigma^2 + cx_0\right)} \geq \log\left(n^2 e^z\right) \geq 1.$$

The assumptions of Lemma 5.1 being checked, we deduce that $E_2 \leq 2Kx_0^2 e^{-z}$ and conclude the proof putting these upper bounds on E_1 and E_2 together.

5.5. Elements of proof for Propositions 3.1, 3.2 and 3.3

The proofs of Propositions 3.1, 3.2 and 3.3 derive from the proposition below which allows to bound $\Lambda_2(S)$ and $\Lambda_\infty(S)$ under suitable assumptions on an orthonormal basis of S. We only give the proof of this proposition and refer the reader to Baraud (2009) for the complete proofs of Propositions 3.1, 3.2 and 3.3.

***Proposition* 5.2.** Let P be some partition of $\{1,\dots,n\}$, J some nonempty index set and

$$\{\phi_{j,I},\ (j,I)\in J\times P\}$$

an orthonormal system such that for some $\Phi>0$ and all $I\in P$

$$\sup_{j\in J}|\phi_{j,I}|_\infty\le\frac{\Phi}{\sqrt{|I|}}\ \text{ and }\ \langle\phi_{j,I},e_i\rangle=0\ \forall i\notin I.$$

If S is the linear span of the $\phi_{j,I}$ with $(j,I)\in J\times P$,

$$\Lambda_2^2(S)\le\left(\frac{|J|\Phi^2}{\min_{I\in P}|I|}\right)\wedge 1\ \text{ and }\ \Lambda_\infty(S)\le\left(|J|\Phi^2\right)\wedge\left(\sqrt{n}\Lambda_2(S)\right).$$

Proof of Proposition 5.2. We have already seen that $\Lambda_2(S)\le 1$ and $\Lambda_\infty(S)\le\sqrt{n}\Lambda_2(S)$, so it remains to show that

$$\Lambda_2^2(S)\le\frac{|J|\Phi^2}{\min_{I\in P}|I|}\ \text{ and }\ \Lambda_\infty(S)\le|J|\Phi^2.$$

Let $i=1,\dots,n$. There exists some unique $I\in P$ such that $i\in I$ and since $\langle\phi_{j,I'},e_i\rangle=0$ for all $I'\neq I$, $\Pi_S e_i=\sum_{j\in J}\langle e_i,\phi_{j,I}\rangle\phi_{j,I}$. Consequently,

$$|\Pi_S e_i|_2^2=\sum_{j\in J}\langle e_i,\phi_{j,I}\rangle^2\le\frac{|J|\Phi^2}{|I|}\le\frac{|J|\Phi^2}{\min_{I\in P}|I|}$$

and

$$|\Pi_S e_i|_1\ =\ \sum_{i'\in I}\left|\sum_{j\in J}\langle e_i,\phi_{j,I}\rangle\langle e_{i'},\phi_{j,I}\rangle\right|\le|I|\frac{|J|\Phi^2}{|I|}\le|J|\Phi^2.$$

We conclude since i is arbitrary. □

Acknowledgements

We thank Jonas Kahn for pointing out this counter-example in Subsection 2.2 and to Lucien Birgé for his useful comments and for making us aware of the book of Talagrand which has been the starting point of this paper.

References

Baraud, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493.

Baraud, Y. (2009). A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression. Technical report, arXiv:0909.1863v1.

Baraud, Y., Comte, F., and Viennet, G. (2001). Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.*, 5:33–49 (electronic).

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

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500.

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 213–247. Birkhäuser, Basel.

Klartag, B. and Mendelson, S. (2005). Empirical processes and random projections. *J. Funct. Anal.*, 225(1):229–245.

Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077.

Ledoux, M. (1996). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87 (electronic).

Mason, J. C. and Handscomb, D. C. (2003). *Chebyshev polynomials*. Chapman & Hall/CRC, Boca Raton, FL.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. With a foreword by Jean Picard.

Mendelson, S. (2008). On weakly bounded empirical processes. *Math. Ann.*, 340(2):293–314.

Mendelson, S., Pajor, A., and Tomczak-Jaegermann, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282.

Rio, E. (2002). Une inégalité de Bennett pour les maxima de processus empiriques. *Ann. Inst. H. Poincaré Probab. Statist.*, 38(6):1053–1057. En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.

Sauvé, M. (2008). Histogram selection in non gaussian regression. *ESAIM Probab. Statist.*, to appear.

Sudakov, V. N. and Cirel'son, B. S. (1974). Extremal properties of half-spaces for spherically invariant measures. *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.*

(LOMI), 41:14–24, 165. Problems in the theory of probability distributions, II.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, (81):73–205.

Talagrand, M. (2005). *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin. Upper and lower bounds of stochastic processes.

van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.*, 18:907–924.