BOUNDING THE EXPECTATION OF THE SUPREMUM OF AN EMPIRICAL PROCESS OVER A (WEAK) VC-MAJOR CLASS

Yannick Baraud

BOUNDING THE EXPECTATION OF THE SUPREMUM OF AN EMPIRICAL PROCESS OVER A (WEAK) VC-MAJOR CLASS

Y. BARAUD

Abstract. Given a bounded class of functions $\mathscr{G}$ and independent random variables $X_1, \ldots, X_n$, we provide an upper bound for the expectation of the supremum of the empirical process over elements of $\mathscr{G}$ having a small variance. Our bound applies in the cases where $\mathscr{G}$ is a VC-subgraph or a VC-major class and it is of smaller order than those one could get by using a universal entropy bound over the whole class $\mathscr{G}$. It also involves explicit constants and does not require the knowledge of the entropy of $\mathscr{G}$

1. Introduction

The control of the fluctuations of an empirical process is a central tool in statistics for establishing the rate of convergence over a set of parameters of some specific estimators such as minimum contrast ones for example. These techniques have been used over the years in many papers among which van de Geer (1990), Birgé and Massart (1993), Barron, Birgé and Massart (1999) and the connections between empirical process theory and statistics are detailed at length in the book by van der Vaart and Wellner (1996). With the concentration of measure phenomenon and Talagrand's Theorem 1.4 (1996) relating the control of the supremum of an empirical process over a class of functions $\mathscr{F}$ to the expectation of this supremum, the initial problem reduces to the evaluation of that expectation. This can be done under universal entropy conditions which measure the massiveness of a class $\mathscr{F}$ by bounding from above and uniformly with respect to probability measures Q on $\mathscr{F}$ the number $N(\mathscr{F}, Q, \varepsilon)$ of $\mathbb{L}_2(Q)$-balls of radius ε that are necessary to cover $\mathscr{F}$. A ready to use inequality is given by Theorem 3.1 in Giné and Koltchinski (2006). Roughly speaking their result says the following. Let $\mathscr{F}$ admit an envelop function $F \leq 1$ (which means that $|f| \leq F \leq 1$ for all $f \in \mathscr{F}$) and $\log N(\mathscr{F}, Q, \varepsilon)$ be not larger than $H(\|F\|_{\mathbb{L}_2(Q)}/\varepsilon)$ for some nondecreasing function H independent of Q and satisfying some mild conditions. Then, given n i.i.d. random variables $X_1, \ldots, X_n$ with an arbitrary distribution P,

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq C(H)\left[\sigma\sqrt{nH\left(2\sigma^{-1}\|F\|_{\mathbb{L}_2(P)}\right)} + H\left(2\sigma^{-1}\|F\|_{\mathbb{L}_2(P)}\right)\right] \tag{1}$$

where

$$Z(\mathscr{F}) = \sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\left(f(X_i) - \mathbb{E}\left[f(X_1)\right]\right)\right|, \tag{2}$$

$C(H)$ is a positive number depending on H, and $\sigma \in (0,1]$ satisfies $\sup_{f\in\mathscr{F}} \mathrm{Var}(f(X_1)) \leq \sigma^2$.

Date: November 20, 2014.

However, computing the universal entropy of a class of functions $\mathscr{F}$ is not an easy task in general and inequality (1) might not be so easy to use in general. For illustration, let us consider the case of $\mathscr{F} = \mathscr{G} \cap \mathscr{B}(g_0, r)$ where $\mathscr{G}$ is the set of nonincreasing functions from $[0,1]$ into itself and $\mathscr{B}(g_0, r)$ the $\mathbb{L}_2(P)$-ball centered at $g_0 \in \mathscr{G}$ with radius $r > 0$. The universal entropy of $\mathscr{F}$, which depends on the choice of g_0, is usually unknown. However, one may use that of $\mathscr{G}$, which is of order $1/\varepsilon$, to bound the universal entropy of $\mathscr{F} \subset \mathscr{G}$ from above. Taking for envelope function F the constant function equal to 1, we derive from (1) that there exists a universal constant $C > 0$ such that

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq C\left[\sqrt{n\sigma} + \sigma^{-1}\right]. \tag{3}$$

While this inequality provides a satisfactory upper bound for $\mathbb{E}\left[Z(\mathscr{F})\right]$ in general, Giné and Koltchinski (2006) (Example 3.8 p.1173) noticed that $\mathbb{E}\left[Z(\mathscr{F})\right]$ was actually of smaller order than the right-hand side of (3) when $g_0 = 0$. This phenomenon is actually easy to explain and we shall see that the function $g_0 = 0$ has in fact nothing magic: if g_0 is decreasing very fast on $[0,1]$ then it is quite easy to oscillate around g_0 and still remain non-increasing on $[0,1]$. This implies that $\mathscr{G} \cap \mathscr{B}(g_0, r)$ is actually massive around g_0. It is however impossible to oscillate around a function g_0 which is constant without violating the monotonicity constraint. For a constant function g_0, $\mathscr{G} \cap \mathscr{B}(g_0, r)$ turns out to be less massive and $\mathbb{E}\left[Z(\mathscr{F})\right]$ much smaller than that of the previous set. A general entropy bound on $\mathscr{G}$ which allows to bound the entropies of all sets $\mathscr{G} \cap \mathscr{B}(g_0, r)$ independently of g_0 therefore provides a pessimistic upper bound in the case of a constant function g_0.

The above argument is not only valid when $\mathscr{G}$ consists of monotone functions but more generally when $\mathscr{G}$ is a bounded VC-major class on $\mathbb{R}$ for instance. For such a class, the level sets $\{g > c\}$ with $g \in \mathscr{G}$ and $c \in \mathbb{R}$ form a VC-class of subsets of $\mathbb{R}$. When g oscillates around c, the level set $\{g > c\}$ is a union of disjoint intervals and since the class of all unions of disjoint intervals is not VC, the elements of $\mathscr{G}$ cannot oscillate arbitrarily around the constant function $g_0 = c$.

The aim of this paper is to provide an upper bound for $\mathbb{E}\left[Z(\mathscr{F})\right]$ when $\mathscr{F}$ consists of the elements of a class $\mathscr{G}$ (including the cases of VC-major and VC-subgraph classes) which satisfy some suitable control of their $\mathbb{L}_2$-norms or variances. The bounds we get are non-asymptotic, involve explicit numerical constants and are true as long as the random variables $X_1, \ldots, X_n$ are independent but not necessarily i.i.d.. They allow to improve the bounds one could obtain by using a naive upper bound on the entropy of the whole class $\mathscr{G}$.

As already mentioned, the expectations of suprema of empirical processes play a central role in statistics and it is well known (we refer the reader to Theorem 5.52 in the book of van der Vaart (1998) and to the historical references therein) that, given a sampling model indexed by a metric space Θ, the rate of convergence of a minimum contrast estimator toward a parameter $\theta_0 \in \Theta$ is governed by the expectation of the supremum of an empirical process over the elements g_θ of a class $\mathscr{G} = \{g_\theta,\ \theta \in \Theta\}$ for which the parameters θ lie in a small ball around θ_0. Such connections between suprema of empirical processes and rates of convergence (or more generally risk bounds) of an estimator are not restricted to minimum contrast estimators and have also recently proved, in Baraud, Birgé and Sart (2014), to be an essential tool for the study of ρ-estimators. In all these cases, the distance between the parameters θ and θ_0 in the metric space Θ controls the $\mathbb{L}_2$-distance between the functions g_θ and g_{θ_0} so that what we need to control is in fact the supremum of the empirical

process over the intersection of $\mathscr{G}$ with a small $\mathbb{L}_2$-ball around g_{θ_0}. Under suitable assumptions on $\mathscr{G}$ and because of the phenomenon we have explained above, one can expect some faster rates of convergence for minimum contrast estimators (as well as ρ-estimators) toward specific parameters θ_0. For example, the Grenander estimator of a monotone density converges at parametric rate when the target density is piecewise constant, as noticed by Birgé (1989), while the minimax rate is of order $n^{-1/3}$. The statistical implications of the results established in the present paper will be detailed in a forthcoming one.

Our paper is organized as follows. The main definitions, including those of VC-classes, VC-major and weak VC-major classes, as well as some basic properties relative to these classes are given in Section 2.1. The main results are presented in Section 2.2. The proof of our main theorem, namely Theorem 1, is postponed to Section 3 where we also establish upper bounds for $\mathbb{E}[Z(\mathscr{F})]$ in the special case of a class $\mathscr{F}$ consisting of indicator functions since these bounds may be of independent interest. Finally Section 4 gathers the proofs of our propositions and that of Corollary 2 which is specific to the case of $\mathscr{F}$ being a VC-major class and $X_1, \ldots, X_n$ i.i.d.

In the sequel, we shall use the following conventions and notations. The word *countable* will always mean finite or countable and, given a set A, $|A|$ and $\mathscr{P}(A)$ will respectively denote the cardinality of A and the class of all its subsets. By convention, $\sum_{\varnothing} = 0$.

2. The setting and the main result

Throughout the paper, $X_1, \ldots, X_n$ are independent random variables defined on a probability space $(\Omega, \mathcal{W}, \mathbb{P})$ with values in a measurable space $(\mathscr{X}, \mathscr{A})$, $\mathscr{F}$ is a class of real-valued measurable functions on $(\mathscr{X}, \mathscr{A})$ and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables (which means that ε_i takes the values ± 1 with probability $1/2$) independent of the X_i. We recall that $Z(\mathscr{F})$ is defined by (2) and set

$$\overline{Z}(\mathscr{F}) = \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|.$$

In order to avoid measurability issues, $\mathbb{E}[Z(\mathscr{F})]$ and $\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]$ mean $\sup_{\mathscr{F}'} \mathbb{E}[Z(\mathscr{F}')]$ and $\sup_{\mathscr{F}'} \mathbb{E}\left[\overline{Z}(\mathscr{F}')\right]$, respectively, where the suprema run among all countable subsets $\mathscr{F}'$ of $\mathscr{F}$. The relevance of the random variable $\overline{Z}(\mathscr{F})$ is due to the following classical symmetrization argument (see van der Vaart and Wellner (1996), Lemma 2.3.6) :

Lemma 1. *For all $a_1, \ldots, a_n \in \mathbb{R}$,*

$$\mathbb{E}\left[\sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}[f(X_i)]\right) \right|\right] \leq 2\mathbb{E}\left[\sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - a_i\right) \right|\right] \tag{4}$$

In particular,

$$\mathbb{E}[Z(\mathscr{F})] \leq 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]. \tag{5}$$

For the sake of completeness, we provide a proof in Section 3 below.

2.1. Basic definitions and properties.

We recall the following.

Definition 1. *A class $\mathscr{C}$ of subsets of some set $\mathcal{Z}$ is said to shatter a finite subset Z of $\mathcal{Z}$ if $\{C \cap Z, C \in \mathscr{C}\} = \mathscr{P}(Z)$ or, equivalently, $|\{C \cap Z, C \in \mathscr{C}\}| = 2^{|Z|}$. A non-empty class $\mathscr{C}$ of subsets of $\mathcal{Z}$ is a VC-class if there exists a finite subset Z of $\mathcal{Z}$ which is not shattered*

by $\mathscr{C}$. The dimension $d \geq 0$ of this VC-class is the smallest integer $k \in \mathbb{N}$ for which there exists Z of cardinality $k+1$ not shattered by $\mathscr{C}$.

Of special interest is the case of $(\mathscr{X}, \mathscr{A}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ for which the class $\mathscr{C}$ of all intervals is a VC-class with dimension 2.

We extend this definition from classes of sets to classes of functions in the following way.

Definition 2. *Let $\mathscr{F}$ be a non-empty class of functions on a set $\mathscr{X}$. We shall say that $\mathscr{F}$ is weak VC-major with dimension $d \in \mathbb{N}$ if d is the smallest integer $k \in \mathbb{N}$ such that, for all $u \in \mathbb{R}$, the class*

$$\mathscr{C}_u(\mathscr{F}) = \big\{\{x \in \mathscr{X} \text{ such that } f(x) > u\},\ f \in \mathscr{F}\big\}$$

is a VC-class of subsets of $\mathscr{X}$ with dimension not larger than k.

If $\mathscr{F}$ consists of monotone functions on $(\mathscr{X}, \mathscr{A}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, $\mathscr{C}_u(\mathscr{F})$ consists of intervals of $\mathbb{R}$ and $\mathscr{F}$ is therefore weak VC-major with dimension not larger than 2. For the same reasons, this is also true for the class $\mathscr{F}$ of nonnegative functions f on $\mathbb{R}$ which are monotone on an interval of $\mathbb{R}$ (depending on f) and vanish elsewhere.

There exist other ways of extending to classes of functions the concept of a VC-class of sets. The two main ones encountered in the literature are the following:

Definition 3. *Let $\mathscr{F}$ be a non-empty class of functions on a set $\mathscr{X}$.*

- *The class $\mathscr{F}$ is VC-major with dimension $d \in \mathbb{N}$ if*

  $$\mathscr{C}(\mathscr{F}) = \big\{\{x \in \mathscr{X} \text{ such that } f(x) > u\},\ f \in \mathscr{F},\ u \in \mathbb{R}\big\}$$

  is a VC-class of subsets of $\mathscr{X}$ with dimension d.
- *The class $\mathscr{F}$ is VC-subgraph with dimension d if*

  $$\mathscr{C}_\times(\mathscr{F}) = \big\{\{(x,u) \in \mathscr{X} \times \mathbb{R} \text{ such that } f(x) > u\},\ f \in \mathscr{F}\big\}$$

  is a VC-class of subsets of $\mathscr{X} \times \mathbb{R}$ with dimension d.

These two notions are related to that of weak VC-major class in the following way.

Proposition 1. *If $\mathscr{F}$ is either VC-major or VC-subgraph with dimension d then $\mathscr{F}$ is weak VC-major with dimension not larger than d.*

An alternative definition for a weak VC-major class can be obtained from the following proposition.

Proposition 2. *The class $\mathscr{F}$ is weak VC-major with dimension d if and only if d is the smallest integer $k \in \mathbb{N}$ such that, for all $u \in \mathbb{R}$, the class*

$$\mathscr{C}_u^+(\mathscr{F}) = \big\{\{x \in \mathscr{X} \text{ such that } f(x) \geq u\},\ f \in \mathscr{F}\big\}$$

is a VC-class of subsets of $\mathscr{X}$ with dimension not larger than k.

The following permanence properties can be established for weak VC-major classes.

Proposition 3. *Let $\mathscr{F}$ be weak VC-major with dimension d. Then for any monotone function F, $F \circ \mathscr{F} = \{F \circ f,\ f \in \mathscr{F}\}$ is weak VC-major with dimension not larger than d. In particular $\{-f,\ f \in \mathscr{F}\}$ and $\{f \vee 0,\ f \in \mathscr{F}\}$ are weak VC-major with respective dimensions not larger than d.*

2.2. **The main result.** Let us first introduce some combinatoric quantities. For $u \in (0,1)$, let

$$\mathscr{E}_u(\mathbf{X}) = \{\{i,\ X_i \in C\},\ C \in \mathscr{C}_u(\mathscr{F})\} \quad \text{and} \quad \Gamma_u = \mathbb{E}\left[\log(2\,|\mathscr{E}_u(\mathbf{X})|\right]. \tag{6}$$

When $\mathscr{F}$ is weak VC-major with dimension d, the quantity Γ_u can be bounded independently of u as follows. For $u \in (0,1)$, the class $\mathscr{C}_u(\mathscr{F})$ being VC with dimension not larger than d, Sauer's lemma (see van der Vaart and Wellner (1996), Section 2.6.3 p.136) asserts that for all $n \geq 1$

$$|\mathscr{E}_u(\mathbf{X})| \leq \sum_{j=0}^{d\wedge n} \binom{n}{j}$$

and therefore for all $u \in (0,1)$, $\Gamma_u \leq \overline{\Gamma}(d)$ with

$$\overline{\Gamma}(d) = \log\left(2\sum_{j=0}^{d\wedge n}\binom{n}{j}\right) \leq \log 2 + (d\wedge n)\log\left(\frac{en}{d\wedge n}\right) \tag{7}$$

and the convention $0\times\log(+\infty) = 0$. The bound on $\overline{\Gamma}(d)$ comes from the classical inequality $\sum_{j=0}^{k}\binom{n}{j} \leq (en/k)^k$ for $k \leq n$ (see Barron, Birgé and Massart (1999), Lemma 6). The following result holds.

Theorem 1. *If $\mathscr{F}$ is a class of functions with values in $[0,1]$ and*

$$\sigma = \sup_{f\in\mathscr{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[f^2(X_i)\right]\right]^{1/2}, \tag{8}$$

then,

$$\frac{1}{2}\mathbb{E}\left[Z(\mathscr{F})\right] \leq \mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq \sqrt{2n}\sigma\left[\frac{1}{\sigma}\int_0^\sigma\sqrt{\Gamma_u}du + \int_\sigma^1\frac{\sqrt{\Gamma_u}}{u}du\right] + 4\int_0^1\Gamma_u du \tag{9}$$

with Γ_u defined by (6). In particular, if $\mathscr{F}$ is weak VC-major with dimension d,

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq 2\left[\sigma\log(e/\sigma)\sqrt{2n\overline{\Gamma}(d)} + 4\overline{\Gamma}(d)\right]$$

with $\overline{\Gamma}(d)$ given by (7).

For the sake of comparison, consider the case where the X_i are i.i.d. with values in $[0,1]$ and $\mathscr{F}$ consists of nondecreasing functions from $[0,1]$ into $[0,1]$ which satisfy $\mathbb{E}[f^2(X_1)] \leq \sigma^2$. The class $\mathscr{F}$ is weak VC-major with dimension 1 and, since $\overline{\Gamma}(1) = \log(2(n+1))$, Theorem 1 gives

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq 2\sigma\log(e/\sigma)\sqrt{2n\log(2(n+1))} + 8\log(2(n+1)). \tag{10}$$

For $\sigma < e^{-e}$, Giné and Koltchinskii (2006) (Example 3.8 p.1173) obtained an upper bound for $\mathbb{E}\left[Z(\mathscr{F})\right]$ of order

$$B(n,\sigma) = \sigma\sqrt{nL(\sigma)} + L(\sigma) + \sqrt{\log n} \quad \text{with} \quad L(\sigma) = \left[\log\left(\sigma^{-1}\right)\right]^{3/2}\log\log\left(\sigma^{-1}\right). \tag{11}$$

If $\sigma \geq \sqrt{\log n/n}$, then $B(n,\sigma) \geq \sqrt{n}\sigma$ while $B(n,\sigma) \geq \sqrt{\log n}$ for $\sigma \leq \sqrt{\log n/n}$. In any case, $B(n,\sigma) \geq \max\{\sqrt{n}\sigma, \sqrt{\log n}\}$, which shows that the bound (11) can only improve ours by some power of $\log n$.

Giné and Koltchinskii's bound is based on the fact that the envelop function $F = \sup_{f\in\mathscr{F}} f$ is smaller than the crude bound 1 for this specific class of functions. This is no longer true for the class $\mathscr{F}' = \{f(\cdot - t)\mathbb{1}_{[0,1]},\ t \in \mathbb{R},\ f \in \mathscr{F}\}$ the elements of which also satisfy $\mathbb{E}[f^2(X_1)] \leq \sigma^2$ when the X_i are uniformly distributed on $[0,1]$ for instance. While their trick fails for the class $\mathscr{F}'$, our Theorem 1 still applies: since $\mathscr{F}'$ is weak-VC major with dimension not larger than 2 and $\overline{\Gamma}(2) \leq 2\overline{\Gamma}(1)$, $\mathbb{E}\left[Z(\mathscr{F}')\right]$ is actually not larger than twice the right-hand side of (10).

When the elements of $\mathscr{F}$ take their values in $[-b,b]$ for some $b > 0$, one should rather use the following result.

Corollary 1. *Assume that $\mathscr{F}$ is a weak VC-major class with dimension d consisting of functions with values in $[-b,b]$ for some $b > 0$. Then,*

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq 4\left[\sigma\log(eb/\sigma)\sqrt{2n\overline{\Gamma}(d)} + 4b\overline{\Gamma}(d)\right]$$

with σ given by (8).

Proof. By homogeneity, we may assume that $b = 1$. Since $\mathscr{F}$ is weak VC-major with dimension d, $\mathscr{F}_+ = \{f \vee 0,\ f \in \mathscr{F}\}$ and $\mathscr{F}_- = \{(-f) \vee 0,\ f \in \mathscr{F}\}$ are both weak VC-major with dimension not larger than d by Proposition 3. The elements of $\mathscr{F}_+$ and $\mathscr{F}_-$ take their values in $[0,1]$ and

$$\max_{\epsilon\in\{-,+\}} \sup_{f\in\mathscr{F}_\epsilon} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f^2(X_i)\right] \leq \sigma^2.$$

We may therefore bound $\mathbb{E}\left[\sup_{f\in\mathscr{F}_\epsilon} |\sum_{i=1}^n \varepsilon_i f(X_i)|\right]$ from above for $\epsilon \in \{-,+\}$ by applying Theorem 1. To conclude we use that $f = f \vee 0 - (-f) \vee 0$ for all $f \in \mathscr{F}$ so that

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \leq \mathbb{E}\left[\sup_{f\in\mathscr{F}_+}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + \mathbb{E}\left[\sup_{f\in\mathscr{F}_-}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right].$$

□

Finally, we end this section with the particular situation when the X_i are i.i.d. and $\mathscr{F}$ is VC-major. In this case, it is possible to replace the control of the $\mathbb{L}_2(P)$-norm of the elements of $\mathscr{F}$ by a control of their variances. More precisely, the following holds.

Corollary 2. *Assume that $X_1,\dots,X_n$ are i.i.d and $\mathscr{F}$ is VC-major with dimension d. Let $\overline{\Gamma}(d)$ be given by (7). Then,*

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq 2\left[\sigma\log(2eb/\sigma)\sqrt{2n\overline{\Gamma}(d)} + 8b\overline{\Gamma}(d)\right] \quad \textit{with} \quad \sigma = \sup_{f\in\mathscr{F}}\sqrt{\mathrm{Var}[f(X_1)]} \in (0,b].$$

3. Proof of Theorem 1

3.1. **Proof of Lemma 1.** Let $(X_1', \ldots, X_n')$ be an independent copy of $\mathbf{X} = (X_1, \ldots, X_n)$. Then

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \left(f(X_i) - \mathbb{E}\left[f(X_i)\right]\right)\right|\right] &= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \left(f(X_i) - \mathbb{E}\left[f(X_i')\middle|\mathbf{X}\right]\right)\right|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\mathbb{E}\left[\sum_{i=1}^n \left(f(X_i) - f(X_i')\right)\middle|\mathbf{X}\right]\right|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \left(f(X_i) - f(X_i')\right)\right|\right].
\end{aligned}$$

By symmetry $\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \left(f(X_i) - f(X_i')\right)\right|$ and $\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \varepsilon_i \left(f(X_i) - f(X_i')\right)\right|$ have the same distribution. Therefore

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \left(f(X_i) - f(X_i')\right)\right|\right] &= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \varepsilon_i \left(f(X_i) - a_i - \left[f(X_i') - a_i\right]\right)\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \varepsilon_i \left(f(X_i) - a_i\right)\right|\right].
\end{aligned}$$

3.2. **The particular case of a class $\mathscr{F}$ of indicator functions.** Let us first consider the situation of $\mathscr{F}$ being a class of indicator functions and start with the following elementary situation.

Lemma 2. *Let $\mathscr{E}$ be a non-empty subset of $\mathscr{P}(\{1, \ldots, n\})$ with elements E of cardinality bounded by $m \leq n$. Then*

$$\mathbb{E}\left[\max_{E\in\mathscr{E}}\left|\sum_{i\in E}\varepsilon_i\right|\right] \leq \sqrt{2\log(2\left|\mathscr{E}\right|)m}. \tag{12}$$

Proof. The result is clear when $\mathscr{E} = \{\varnothing\}$ in view of our convention $\sum_{\varnothing} = 0$. Therefore we may restrict ourselves to the case of $\max_{E\in\mathscr{E}}|E| \geq 1$. Let us now fix some $\lambda > 0$. For all $E \in \mathscr{E}$,

$$\begin{aligned}
\exp\left[\lambda\left|\sum_{i\in E}\varepsilon_i\right|\right] &= \max\left\{\exp\left[\lambda\sum_{i\in E}\varepsilon_i\right], \exp\left[-\lambda\sum_{i\in E}\varepsilon_i\right]\right\} \\
&\leq \exp\left[\lambda\sum_{i\in E}\varepsilon_i\right] + \exp\left[-\lambda\sum_{i\in E}\varepsilon_i\right].
\end{aligned}$$

Taking expectations on both sides and using that $\mathbb{E}\left[e^{\lambda\varepsilon}\right] = \cosh(\lambda) \leq e^{\lambda^2/2}$, we get

$$\mathbb{E}\left[\exp\left[\lambda\left|\sum_{i\in E}\varepsilon_i\right|\right]\right] \leq 2\mathbb{E}\left[\exp\left[\lambda\sum_{i\in E}\varepsilon_i\right]\right] = 2\prod_{i\in E}\mathbb{E}\left[\exp\left[\lambda\varepsilon_i\right]\right] \leq 2\exp\left[\frac{\lambda^2|E|}{2}\right].$$

It follows from Jensen's Inequality that

$$\begin{aligned}
&\mathbb{E}\left[\max_{E\in\mathscr{E}}\left|\sum_{i\in E}\varepsilon_i\right|\right] \\
&\quad= \frac{1}{\lambda}\log\left(\exp\left[\mathbb{E}\left[\lambda\max_{E\in\mathscr{E}}\left|\sum_{i\in E}\varepsilon_i\right|\right]\right]\right) \;\leq\; \frac{1}{\lambda}\log\mathbb{E}\left[\exp\left[\lambda\max_{E\in\mathscr{E}}\left|\sum_{i\in E}\varepsilon_i\right|\right]\right] \\
&\quad= \frac{1}{\lambda}\log\mathbb{E}\left[\max_{E\in\mathscr{E}}\exp\left[\lambda\left|\sum_{i\in E}\varepsilon_i\right|\right]\right] \;\leq\; \frac{1}{\lambda}\log\left(\sum_{E\in\mathscr{E}}\mathbb{E}\left[\exp\left[\lambda\left|\sum_{i\in E}\varepsilon_i\right|\right]\right]\right) \\
&\quad\leq \frac{1}{\lambda}\log\left(2\sum_{E\in\mathscr{E}}\exp\left[\frac{\lambda^2|E|}{2}\right]\right) \;\leq\; \frac{\log(2\,|\mathscr{E}|)}{\lambda}+\frac{\lambda}{2}\max_{E\in\mathscr{E}}|E|\,.
\end{aligned}$$

We then conclude by minimizing this upper bound with respect to λ. □

Let us now prove an analogue of Theorem 1 when $\mathscr{F}$ is a family of indicator functions.

Theorem 2. *Let $\mathbf{X}=(X_1,\ldots,X_n)$ be a random vector with independent components taking their values in the measurable space $(\mathscr{X},\mathscr{A})$ and let $\mathscr{C}$ be a countable family of measurable subsets of $\mathscr{X}$. Let $\mathscr{F}=\{\mathbb{1}_C,\ C\in\mathscr{C}\}$, $\mathscr{E}(\mathbf{X})=\big\{\{i,\ X_i\in C\},\ C\in\mathscr{C}\big\}$ and*

$$\sigma=\sup_{C\in\mathscr{C}}\left[\frac{1}{n}\sum_{i=1}^n\mathbb{P}(X_i\in C)\right]^{1/2}.$$

Then

$$\mathbb{E}\left[Z(\mathscr{F})\right]\leq 2\left[\sigma\sqrt{2n\Gamma}+4\Gamma\right]\qquad\textit{with}\qquad\Gamma=\mathbb{E}\left[\log(2\,|\mathscr{E}(\mathbf{X})|)\right].$$

This result is of the same flavour as the one Pascal Massart established in Massart (2007) (see his Lemma 6.4). The bound he gets can be smaller than ours when σ is not too small. Unlike his, our bound contains explicit constants.

Proof. By the symmetrization argument (4),

$$\begin{aligned}
\mathbb{E}\left[\sup_{C\in\mathscr{C}}\sum_{i=1}^n\mathbb{1}_C(X_i)\right] &\leq \mathbb{E}\left[\sup_{C\in\mathscr{C}}\sum_{i=1}^n\left(\mathbb{1}_C(X_i)-\mathbb{P}(X_i\in C)\right)\right]+n\sigma^2 \\
&\leq 2\mathbb{E}\left[\sup_{C\in\mathscr{C}}\left|\sum_{i=1}^n\varepsilon_i\mathbb{1}_C(X_i)\right|\right]+n\sigma^2=2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]+n\sigma^2. \qquad (13)
\end{aligned}$$

Let us denote by $\mathbb{E}_\varepsilon$ the conditional expectation given $\mathbf{X}=(X_1,\ldots,X_n)$. By (12),

$$\mathbb{E}_\varepsilon\left[\sup_{C\in\mathscr{C}}\left|\sum_{i=1}^n\varepsilon_i\mathbb{1}_C(X_i)\right|\right]=\mathbb{E}_\varepsilon\left[\max_{E\in\mathscr{E}(\mathbf{X})}\left|\sum_{i\in E}\varepsilon_i\right|\right]\leq\sqrt{2\log(2\,|\mathscr{E}(\mathbf{X})|)\sup_{C\in\mathscr{C}}\sum_{i=1}^n\mathbb{1}_C(X_i)}.$$

Taking expectations with respect to $\mathbf{X}$ on both sides of this inequality, we derive from Cauchy-Schwarz's Inequality and (13) that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]\leq\sqrt{2\Gamma\mathbb{E}\left[\sup_{C\in\mathscr{C}}\sum_{i=1}^n\mathbb{1}_C(X_i)\right]}\leq\sqrt{2\Gamma\left(2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]+n\sigma^2\right)}.$$

Solving the last inequality with respect to $\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]$ leads to

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq \sqrt{2\Gamma n\sigma^2 + (2\Gamma)^2} + 2\Gamma \leq \sqrt{2\Gamma n\sigma^2} + 4\Gamma$$

and the conclusion follows from (5). □

Of particular interest is the situation when $\mathscr{C}$ is VC with dimension d. In this case, we derive from Sauer's lemma that for all $n \geq 1$

$$|\mathscr{E}(\mathbf{X})| \leq \sum_{j=0}^{d\wedge n} \binom{n}{j}.$$

This shows that, if $\mathscr{C}$ is a VC-class with dimension not larger than d, $\log(2\,|\mathscr{E}(\mathbf{X})|) \leq \overline{\Gamma}(d)$ where $\overline{\Gamma}(d)$ is given by (7). We immediately deduce from Theorem 2 the following corollary.

Corollary 3. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector with independent components taking their values in the measurable space $(\mathscr{X}, \mathscr{A})$ and let $\mathscr{C}$ be a countable family of measurable subsets of $\mathscr{X}$ which is a VC-class with dimension d. If $\mathscr{F} = \{\mathbb{1}_C,\ C \in \mathscr{C}\}$ and $\overline{\Gamma}(d)$ is given by (7), then*

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq \sigma\sqrt{2n\overline{\Gamma}(d)} + 4\overline{\Gamma}(d) \quad \textit{with} \quad \sigma = \sup_{C\in\mathscr{C}} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(X_i \in C)\right]^{1/2}.$$

3.3. **End of the proof of Theorem 1.** In view of our convention about the definition of $\mathbb{E}\left[Z(\mathscr{F})\right]$ we may assume with no loss of generality that $\mathscr{F}$ is countable. Let us fix $u \in]0,1[$ and write for simplicity, $\mathscr{C}_u(\mathscr{F}) = \mathscr{C}_u$. Since $\mathscr{F}$ is weak VC-major with dimension d, $\mathscr{C}_u$ is a VC-class with dimension not larger than d. Besides, $\mathscr{C}_u$ is countable since $\mathscr{F}$ is and, by Markov's Inequality and concavity,

$$\sup_{C\in\mathscr{C}_u} \sum_{i=1}^{n} \mathbb{P}(X_i \in C) = \sup_{f\in\mathscr{F}} \sum_{i=1}^{n} \mathbb{P}(f(X_i) > u) \leq \sup_{f\in\mathscr{F}} \sum_{i=1}^{n} \left[\frac{\mathbb{E}\left(f^2(X_i)\right)}{u^2} \wedge 1\right] \leq n\left[\frac{\sigma^2}{u^2} \wedge 1\right].$$

Applying Corollary 3 to the class of sets $\mathscr{C}_u$ leads to

$$\mathbb{E}\left[\sup_{C\in\mathscr{C}_u} \left|\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_C(X_i)\right|\right] \leq \left(\frac{\sigma}{u} \wedge 1\right)\sqrt{2n\Gamma_u} + 4\Gamma_u. \tag{14}$$

Since the elements $f \in \mathscr{F}$ take their values in $[0,1]$,

$$\left|\sum_{i=1}^{n} \varepsilon_i f(X_i)\right| = \left|\int_0^1 \sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{f(X_i)>u}\, du\right| \leq \int_0^1 \left|\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{f(X_i)>u}\right| du.$$

Moreover,

$$\sup_{f\in\mathscr{F}} \left|\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{f(X_i)>u}\right| = \sup_{C\in\mathscr{C}_u} \left|\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_C(X_i)\right|$$

and it follows that

$$\sup_{f\in\mathscr{F}} \left|\sum_{i=1}^{n} \varepsilon_i f(X_i)\right| \quad \leq \quad \int_0^1 \sup_{C\in\mathscr{C}_u} \left|\sum_{i=1}^{n} \varepsilon_i \mathbb{1}_C(X_i)\right| du.$$

Using (14) and taking expectations on both sides lead to

$$\begin{aligned}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] &= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right] \leq \int_0^1\left[\left(\frac{\sigma}{u}\wedge 1\right)\sqrt{2n\Gamma_u}+4\Gamma_u\right]du \\ &= \sqrt{2n}\sigma\left[\frac{1}{\sigma}\int_0^\sigma\sqrt{\Gamma_u}du+\int_\sigma^1\frac{\sqrt{\Gamma_u}}{u}\right]+4\int_0^1\Gamma_u du\end{aligned}$$

and the conclusion follows from (5).

4. Additional proofs

4.1. **Proof of Proposition 1.** If $\mathscr{F}$ is VC-major with dimension d, $\mathscr{C}(\mathscr{F})$ is a VC-class with dimension d therefore, whatever $u\in\mathbb{R}$, its subset $\mathscr{C}_u(\mathscr{F})$ is also a VC-class with dimension not larger than d. Let us now turn to the case where $\mathscr{F}$ is VC-subgraph with dimension d. Let $u\in\mathbb{R}$, if $\mathscr{C}_u$ shatters $\{x_1,\ldots,x_k\}$, for any subset E of $\{1,\ldots,k\}$ one can find a function $f\in\mathscr{F}$, such that

$$E=\big\{i\in\{1,\ldots,k\}\text{ such that }f(x_i)>u\big\}$$

which exactly means that $\mathscr{C}_\times(\mathscr{F})$ shatters $\{(x_1,u),\ldots,(x_k,u)\}$ and implies that $k\leq d$.

4.2. **Proof of Proposition 2.** For all $f\in\mathscr{F}$ and $u\in\mathbb{R}$, we can write

$$\mathbb{1}_{\{f\geq u\}}(x)=\lim_{m\to+\infty}\mathbb{1}_{\{f>u-(1/m)\}}(x)\quad\text{for all }x\in\mathscr{X}.$$

This means that $\mathscr{C}_u^+$ is the sequential closure of $\mathscr{C}_u$ for the pointwise convergence of indicator functions. Lemma 2.6.17 (vi) in van der Vaart and Wellner (1996) (and its proof) asserts that $\mathscr{C}_u^+(\mathscr{F})$ is a VC-class with dimension not larger than that of $\mathscr{C}_u$. For the reciprocal, note that for all $f\in\mathscr{F}$ and $u\in\mathbb{R}$,

$$\mathbb{1}_{\{f>u\}}(x)=\lim_{m\to+\infty}\mathbb{1}_{\{f\geq u+(1/m)\}}(x)\quad\text{for all }x\in\mathscr{X}$$

and conclude in the same way.

4.3. **Proof of Proposition 3.** Let $u\in\mathbb{R}$. If $\mathscr{C}_u(F\circ\mathscr{F})$ cannot shatter at least one point, its dimension is 0 and there is nothing to prove since $d\geq 0$. Otherwise, there exist $k\geq 1$ points $x_1,\ldots,x_k$ in $\mathscr{X}$ and m functions $f_1,\ldots,f_m\in\mathscr{F}$ such that the set $\big\{\{F\circ f_j>u\},\ j=1,\ldots,m\big\}$ shatters $\{x_1,\ldots,x_k\}$. In particular, there exists a point x_i and a function f_j such that $F\circ f_j(x_i)\leq u$ so that

$$s=\max_{i,j}\{f_j(x_i)\ \text{ such that }\ F\circ f_j(x_i)\leq u\}$$

is well-defined. Clearly, for all $i=1,\ldots,k$ and $j=1,\ldots,m$,

$$F\circ f_j(x_i)>u\quad\text{if and only if}\quad f_j(x_i)>s$$

and $\mathscr{C}_s(\mathscr{F})$ therefore shatters $\{x_1,\ldots,x_k\}$, which implies that $k\leq d$.

4.4. **Proof of Corollary 2.** Let $\mathscr{G}$ be the class of all functions g_f, $f \in \mathscr{F}$, defined on $\mathscr{X}$ and with values in $[-b, b]$ given by

$$g_f(x) = \frac{1}{2}\left(f(x) - \mathbb{E}\left[f(X_1)\right]\right).$$

Since

$$\sup_{g\in\mathscr{G}} \mathbb{E}\left[g_f^2(X_1)\right] = \frac{1}{4}\sup_{f\in\mathscr{F}} \mathrm{Var}(f(X_1)) \leq \frac{\sigma^2}{4},$$

Corollary 2 will follow from Corollary 1 if we can prove that $\mathscr{G}$ is weak VC-major, which is a consequence of the next lemma.

Lemma 3. *If $\mathscr{F}$ is VC-major with dimension d, $\mathscr{G}$ is weak VC-major with dimension not larger than d.*

Proof. Let $u \in \mathbb{R}$ and $\{x_1, \ldots, x_k\}$ be a nonempty subset of $\mathscr{X}$ which is shattered by $\mathscr{C}_u(\mathscr{G})$ (if no such set exists then the dimension of $\mathscr{C}_u(\mathscr{G})$ is 0 and there is nothing to prove). For any $E \subset \{1, \ldots, k\}$, there exists $f \in \mathscr{F}$ such that

$$E = \left\{i \in \{1,\ldots,k\} \text{ such that } g_f(x_i) > u\right\} = \left\{i \in \{1,\ldots,k\} \text{ such that } f(x_i) > t\right\}$$

with $t = 2(u + \mathbb{E}[f(X_1)])$. Consequently, the class of sets $\mathscr{C}(\mathscr{F}) = \left\{\{f > t\},\ f \in \mathscr{F}, t \in \mathbb{R}\right\}$ shatters $\{x_1, \ldots, x_k\}$ which implies that $k \leq d$. □

Acknowledgement. The author would like to thank Lucien Birgé for his numerous comments that have led to an improved version of the present paper.

References

Baraud, Y., Birgé, L., and Sart, M. (2014). A new method for estimation and model selection: ρ-estimation. *http://arxiv.org/abs/1403.6057*.

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

Birgé, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.*, 17(4):1532–1549.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216.

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.

Talagrand, M. (1996). New concentration inequalities in product space. *Invent. Math.*, 126:505–563.

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

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York.

Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351, 06100 Nice, France.

E-mail address: baraud@unice.fr