Electronic Journal of Statistics
Vol. 10 (2016) 1709–1728
ISSN: 1935-7524
DOI: 10.1214/15-EJS1055

Bounding the expectation of the supremum of an empirical process over a (weak) VC-major class

Yannick Baraud

Univ. Nice Sophia Antipolis, CNRS
LJAD, UMR 7351
06100 Nice, France
e-mail: baraud@unice.fr

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 when $\mathscr{G}$ 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 $\mathscr{G}$.

MSC 2010 subject classifications: Primary 60E15; secondary 62G05.
Keywords and phrases: Suprema of empirical processes, expectation bounds, VC type classes, concentration inequalities, nonparametric estimation.

Received November 2014.

Contents

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 [12], Birgé and Massart [5], Barron, Birgé and Massart [3] and the connections between empirical process theory and statistics are detailed at length in the book by van der Vaart and Wellner [14]. With the concentration of measure phenomenon and Talagrand's Theorem 1.4 [11] 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 [7]. 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,

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq C(H)\left[\sigma\sqrt{nH\left(2\sigma^{-1}\|F\|_{\mathbb{L}_2(P)}\right)} + H\left(2\sigma^{-1}\|F\|_{\mathbb{L}_2(P)}\right)\right] \tag{1.1}$$

where

$$Z(\mathscr{F}) = \sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\left(f(X_i) - \mathbb{E}\left[f(X_i)\right]\right)\right|, \tag{1.2}$$

$C(H)$ is a positive number depending on H, and $\sigma \in (0,1]$ satisfies

$$\sup_{f\in\mathscr{F}} \mathrm{Var}(f(X_1)) \leq \sigma^2.$$

However, computing the universal entropy of a class of functions $\mathscr{F}$ is not an easy task and inequality (1.1) 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 (1.1) that there exists a universal constant $C > 0$ such that

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leq C\left[\sqrt{n\sigma} + \sigma^{-1}\right]. \tag{1.3}$$

While this inequality provides a satisfactory upper bound for $\mathbb{E}[Z(\mathscr{F})]$ in general, Giné and Koltchinski [7] (Example 3.8 p. 1173) noticed that $\mathbb{E}[Z(\mathscr{F})]$ was actually of smaller order than the right-hand side of (1.3) 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 nonincreasing 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 $\mathbb{E}[Z(\mathscr{F})]$ 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 family of all level sets $\{g > c\}$ with $g \in \mathscr{G}$ and $c \in \mathbb{R}$ form a VC-class of subsets of $\mathbb{R}$. When a function 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 $\mathbb{E}[Z(\mathscr{F})]$ 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 [13] 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 $\mathscr{G} = \{g_\theta,\ \theta \in \Theta\}$ lying within a small ball around g_{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 [2], to be an essential tool for the study of ρ-estimators. Under suitable assumptions on $\mathscr{G}$ and because of the phenomenon we have explained above, one can expect some faster rates of convergence for these estimators toward specific parameters θ_0. An illustration of this fact, which relies on the results of the present paper, can be found in Baraud and Birgé [1]. We show that the ρ-estimator built on a class $\mathscr{F}$ of densities satisfying some shape constraints achieves a rate of convergence toward some specific elements of $\mathscr{F}$ which may be much faster than the minimax rate over the whole class. This phenomenon is actually not specific to ρ-estimators and was already observed for the Grenander estimator of

a monotone density which converges at parametric rate when the target density is piecewise constant, as noticed by Birgé [4], although the minimax rate over the whole set is of order $n^{-1/3}$.

Our paper is organised 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 theorems, namely Theorems 2.1 and 2.2, are postponed to Section 3. We also establish there upper bounds for $\mathbb{E}[Z(\mathscr{F})]$ in the special case where $\mathscr{F}$ consists of indicator functions indexed by a class of sets $\mathscr{C}$ since these bounds may be of independent interest. When $\mathscr{C}$ is VC and the X_i i.i.d., these bounds are compared to those provided by Boucheron *et al.* [6]. Finally Section 4 gathers the proofs of our propositions and that of Corollary 2.2 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. Given two numbers a, b, $a \vee b$ and $a \wedge b$ mean $\max\{a, b\}$ and $\min\{a, b\}$ respectively. 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 (1.2) 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 [14], Lemma 2.3.6):

Lemma 2.1. *For all $a_1, \ldots, a_n \in \mathbb{R}$,*

$$\mathbb{E}\left[\sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}[f(X_i)] \right) \right| \right] \leq 2\mathbb{E}\left[\sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - a_i \right) \right| \right] \tag{2.1}$$

In particular,

$$\mathbb{E}[Z(\mathscr{F})] \leq 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]. \tag{2.2}$$

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

2.1. Basic definitions and properties

We recall the following.

Definition 2.1. 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 an integer $k \in \mathbb{N}$ such that $\mathscr{C}$ cannot shatter any subset of $\mathcal{Z}$ with cardinality larger than k. The dimension $d \in \mathbb{N}$ of $\mathscr{C}$ is then the smallest of these integers k.

Of special interest is the class $\mathscr{C}$ of all intervals of $\mathbb{R}$ which is VC with dimension 2: for $Z = \{0, 1\}$, $\{C \cap Z, C \in \mathscr{C}\} = \mathscr{P}(Z)$ and whatever $Z' = \{x_1, x_2, x_3\}$ with $x_1 < x_2 < x_3$, $\{x_1, x_3\} \notin \{C \cap Z', C \in \mathscr{C}\}$.

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

Definition 2.2. Let $\mathscr{F}$ be a non-empty class of functions on a set $\mathcal{X}$. 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

$$\mathscr{C}_u(\mathscr{F}) = \big\{\{x \in \mathcal{X} \text{ such that } \ f(x) > u\},\ f \in \mathscr{F}\big\} \tag{2.3}$$

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

If $\mathscr{F}$ consists of monotone functions on $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, $\mathscr{C}_u(\mathscr{F})$ 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 the concept of a VC-class of sets to classes of functions. The two main ones encountered in the literature are the following:

Definition 2.3. Let $\mathscr{F}$ be a non-empty class of functions on a set $\mathcal{X}$.

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

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

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

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

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

These two properties are stronger than that of being weak VC-major:

Proposition 2.1. *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 2.2. *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 2.3. *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 results

Let us first introduce some combinatoric quantities. For $u \in (0,1)$, $\mathscr{C}_u(\mathscr{F})$ defined by (2.3) and $\mathbf{X} = (X_1, \ldots, X_n)$ let

$$\mathscr{E}_u(\mathbf{X}) = \{\{i,\ X_i \in C\},\ C \in \mathscr{C}_u(\mathscr{F})\} \quad \text{and} \quad \Gamma_u = \mathbb{E}\left[\log(2\,|\mathscr{E}_u(\mathbf{X})|)\right]. \tag{2.4}$$

Since $\mathscr{C}_u(\mathscr{F}) \neq \varnothing$ and $\mathscr{E}_u(\mathbf{X}) \subset \mathscr{P}(\{1,\ldots,n\})$, $1 \leq |\mathscr{E}_u(\mathbf{X})| \leq 2^n$. Hence, Γ_u is well defined and satisfies $\log 2 \leq \Gamma_u \leq (n+1)\log 2$ for all $u \in (0,1)$. The upper bound $(n+1)\log 2$ can be improved as follows when $\mathscr{F}$ is weak VC-major with dimension d. For $u \in (0,1)$, the class $\mathscr{C}_u(\mathscr{F})$ being VC with dimension not larger than d, a classical lemma of Sauer [10] (see also van der Vaart and Wellner [14], Section 2.6.3 p. 136) asserts that $|\mathscr{E}_u(\mathbf{X})| \leq \sum_{j=0}^{d\wedge n}\binom{n}{j}$ for all $n \geq 1$, therefore $\Gamma_u \leq \overline{\Gamma}_n(d)$ for all $u \in (0,1)$ with

$$\overline{\Gamma}_n(d) = \log\left[2\sum_{j=0}^{d\wedge n}\binom{n}{j}\right]. \tag{2.5}$$

Using the classical inequality $\sum_{j=0}^{k}\binom{n}{j} \leq (en/k)^k$ for $k \leq n$ (see Barron, Birgé and Massart [3], Lemma 6), a convenient upper bound for $\overline{\Gamma}_n(d)$ when $d \geq 1$ is given by

$$\overline{\Gamma}_n(d) \leq \log 2 + (d\wedge n)\log\left(\frac{en}{d\wedge n}\right) \leq (d\wedge n)\log\left(\frac{2en}{d\wedge n}\right).$$

Since for $d \leq n$, $\overline{\Gamma}_n(d) \geq \log\binom{n}{d}$, it is not difficult to see that

$$\overline{\Gamma}_n(d) = d\log n(1+o(1)) \quad \text{when} \quad n \to +\infty.$$

The following result holds.

Theorem 2.1. *If $\mathscr{F}$ is a class of functions with values in $[0,1]$ and*

$$\sigma = \sup_{f\in\mathscr{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[f^2(X_i)\right]\right]^{1/2}, \tag{2.6}$$

then,

$$\mathbb{E}\left[Z(\mathscr{F})\right] \le 2\sqrt{2n}\,\sigma\left[\frac{1}{\sigma}\int_0^\sigma \sqrt{\Gamma_u}du + \int_\sigma^1 \frac{\sqrt{\Gamma_u}}{u}du\right] + 8\int_0^1 \Gamma_u du, \tag{2.7}$$

with Γ_u *defined by (2.4). In particular, if* $\mathscr{F}$ *is weak VC-major with dimension* d*,*

$$\mathbb{E}\left[Z(\mathscr{F})\right] \le 2\sqrt{\overline{\Gamma}_n(d)}\left[\sigma\log\left(\frac{e}{\sigma}\right)\sqrt{2n} + 4\sqrt{\overline{\Gamma}_n(d)}\right] \tag{2.8}$$

with $\overline{\Gamma}_n(d)$ *given by (2.5).*

In view of analysing (2.8), let $\mathscr{G}$ be a weak VC-major class with dimension $d \ge 1$ consisting of functions with values in $[0,1]$, $\sigma \in [0,1]$ and

$$\mathscr{F} = \mathscr{G}(\sigma) = \left\{ f \in \mathscr{G},\ \sum_{i=1}^n \mathbb{E}\left[f^2(X_i)\right] \le n\sigma^2 \right\}. \tag{2.9}$$

As a subset of $\mathscr{G}$, $\mathscr{F}$ is weak VC-major with dimension not larger than d and we may therefore apply our Theorem 2.1 to bound $\mathbb{E}\left[Z(\mathscr{F})\right]$ from above. When n is large enough, the right-hand side of (2.8) is of order $\sigma\log(e/\sigma)\sqrt{nd\log n}$ for $\sigma \ge \sqrt{d/(n\log n)}$ and is equivalent to $2\sigma\log(e/\sigma)\sqrt{2nd\log n}$ when σ is fixed and n tends to infinity. In the opposite situation where $\sigma < \sqrt{d/(n\log n)}$, (2.8) is of order $d\log n$.

For the sake of comparison with the results of Giné and Koltchinskii [7], consider the case where the X_i are i.i.d. with a nonatomic distribution P on $[0,1]$, $\mathscr{G}$ is the set of nondecreasing functions f from $[0,1]$ into $[0,1]$ and $\mathscr{F} = \mathscr{G}(\sigma)$ is given by (2.9). The class $\mathscr{F}$ is weak VC-major with dimension 1 because the elements of $\mathscr{C}_u(\mathscr{F})$ are all of the form $(a,1]$ or $[a,1]$ with $a \in [0,1]$ for all u and such classes of intervals cannot shatter a set of two elements $\{x_1, x_2\}$ with $0 \le x_1 < x_2 \le 1$ (the subset $\{x_1\}$ cannot be picked up). Besides, $\overline{\Gamma}_n(1) = \log(2(n+1))$ and Theorem 2.1 gives

$$\mathbb{E}\left[Z(\mathscr{F})\right] \le 2\sigma\log(e/\sigma)\sqrt{2n\log(2(n+1))} + 8\log(2(n+1)). \tag{2.10}$$

For $\sigma < e^{-e}$, Giné and Koltchinskii [7] (Example 3.8 p. 1173) obtained an upper bound for $\mathbb{E}\left[Z(\mathscr{F})\right]$ of order

$$B(n,\sigma) = \sigma\sqrt{nL(\sigma)} + L(\sigma) + \sqrt{\log n} \quad \text{with}$$
$$L(\sigma) = \left[\log\left(\sigma^{-1}\right)\right]^{3/2}\log\log\left(\sigma^{-1}\right). \tag{2.11}$$

If $\sigma \ge \sqrt{\log n/n}$, then $B(n,\sigma) \ge \sqrt{n}\sigma$ while $B(n,\sigma) \ge \sqrt{\log n}$ for $\sigma \le \sqrt{\log n/n}$. In any case, $B(n,\sigma) \ge \max\{\sqrt{n}\sigma, \sqrt{\log n}\}$, which shows that the bound (2.11) can only improve ours by some power of $\log n$.

Giné and Koltchinskii's bound is based on the fact that the class $\mathscr{F}$ possesses an envelop function $F = \sup_{f\in\mathscr{F}} f$ whose $\mathbb{L}_2(P)$-norm equals $\sigma[\log(e/\sigma^2)]^{1/2}$ and is therefore small when σ is small. This property is no longer satisfied for the class $\mathscr{F}' = \{f(\cdot - t)\mathbb{1}_{[0,1]}(\cdot),\ t \in \mathbb{R},\ f \in \mathscr{F}\}$ for which $\sup_{f\in\mathscr{F}'} f = 1$.

The elements of $\mathscr{F}'$ also satisfy $\mathbb{E}[f^2(X_1)] \leq \sigma^2$ when the X_i are uniformly distributed on $[0,1]$ for instance, however, while Giné and Koltchinskii's trick fails for the class $\mathscr{F}'$, our Theorem 2.1 still applies: since $\mathscr{F}'$ is weak-VC major with dimension not larger than 2 and $\overline{\Gamma}_n(2) \leq 2\overline{\Gamma}_n(1)$, $\mathbb{E}[Z(\mathscr{F}')]$ is actually not larger than twice the right-hand side of (2.10).

When σ^2 is large enough compared to $\overline{\Gamma}_n(d)/n$, inequality (2.8) can be further improved as we shall see below. Let

$$\overline{H}(x) = x\sqrt{d\left[5 + \log\left(\frac{1}{x}\right)\right]} \quad \text{for } x \in (0,1] \text{ and } a = \left(32\sqrt{\frac{\overline{\Gamma}_n(d)}{n}}\right) \wedge 1. \tag{2.12}$$

Note that $a = 32\sqrt{(d\log n)/n}(1+o(1))$ when n tends to infinity.

Theorem 2.2. *If $\mathscr{F}$ is a weak VC-major class with dimension not larger than $d \geq 1$, of functions with values in $[0,1]$,*

$$\mathbb{E}[Z(\mathscr{F})] \leq 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq 10\sqrt{n}B(\sigma) \tag{2.13}$$

where σ is given by (2.6) and

$$B(\sigma) = \begin{cases} \overline{H}\left[\sigma\log(1/\sigma) + \sigma\right] & \text{for } \sigma \geq a \\ \\ \overline{H}\left[\sigma\log(1/a) + a\right] & \text{for } \sigma < a \end{cases}. \tag{2.14}$$

In both cases, we may note that

$$B(\sigma) \leq \overline{H}\left[(\sigma \vee a)\log\left(\frac{e}{\sigma \vee a}\right)\right] \quad \text{for all } \sigma \in [0,1].$$

When $\mathscr{F} = \mathscr{G}(\sigma)$ is given by (2.9) and n is large, the right-hand side of (2.13) is of order $\sigma\log^{3/2}(e/\sigma)\sqrt{nd}$ when $\sigma \geq a$ and improves (2.8) when $\log(1/\sigma)$ is small enough compared to $\log n$. When $\sigma < a$, two situations may occur. Either $\sigma \geq \sqrt{d/(n\log n)}$ and the right-hand sides of (2.13) and (2.8) are both of order $\sigma\log(e/\sigma)\sqrt{nd\log n}$, or $\sigma < \sqrt{d/(n\log n)}$ and the right-hand side of (2.8), which is of order $d\log n$ improves that of (2.13) which is of order $d\log^{3/2} n$.

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

Corollary 2.1. *Assume that $\mathscr{F}$ is a weak VC-major class with dimension not larger than $d \geq 1$ consisting of functions with values in $[-b,b]$ for some $b > 0$. Then,*

$$4^{-1}\mathbb{E}[Z(\mathscr{F})] \leq \left[\sigma\log\left(\frac{eb}{\sigma}\right)\sqrt{2n\overline{\Gamma}_n(d)} + 4b\overline{\Gamma}_n(d)\right] \wedge \left[5\sqrt{n}bB(\sigma b^{-1})\right].$$

with $\overline{\Gamma}_n(d)$ given by (2.5), σ by (2.6) and $B(\cdot)$ by (2.14).

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 2.3. The elements of $\mathscr{F}_+$ and $\mathscr{F}_-$ take their values in $[0,1]$ and

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

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 Theorems 2.1 and 2.2. To conclude we use that $f = f\vee 0-(-f)\vee 0$ for all $f\in\mathscr{F}$ so that

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

□

Finally, we conclude this section with the special case of i.i.d. X_i and a VC-major class $\mathscr{F}$. It is then 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 2.2. *Let $X_1,\dots,X_n$ be i.i.d random variables, $\mathscr{F}$ a VC-major class of functions with values in $[-b,b]$ and*

$$\sigma = \sup_{f\in\mathscr{F}}\sqrt{\mathrm{Var}[f(X_1)]} \in (0,b].$$

If $\mathscr{F}$ is a VC-major class with dimension not larger than $d\ge 1$,

$$\mathbb{E}\left[Z(\mathscr{F})\right] \le \left[2\sigma\log\left(\frac{2eb}{\sigma}\right)\sqrt{2n\overline{\Gamma}_n(d)} + 16b\overline{\Gamma}_n(d)\right] \wedge \left[20\sqrt{n}bB\left(\frac{b}{\sigma}\right)\right]$$

where $\overline{\Gamma}_n(d)$ is given by (2.5) and $B(\cdot)$ by (2.14).

3. Proofs of Theorem 2.1 and 2.2

3.1. Proof of Lemma 2.1

Let $(X_1',\dots,X_n')$ be an independent copy of $\mathbf{X}=(X_1,\dots,X_n)$. Then

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

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

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

3.2. The particular case of a class $\mathscr{F}$ of indicator functions

We start with the following elementary situation.

Lemma 3.1. *For a finite and non-empty subset T of $\mathbb{R}^n$ and $v^2=\max_{t\in T}\sum_{i=1}^n t_i^2$,*

$$\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n\varepsilon_i t_i\right|\right]\leq\sqrt{2\log(2|T|)v^2}. \tag{3.1}$$

Proof. For $\overline{T}=T\cup\{-t,\ t\in T\}$,

$$\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n\varepsilon_i t_i\right|\right]=\mathbb{E}\left[\sup_{t\in\overline{T}}\sum_{i=1}^n\varepsilon_i t_i\right]$$

and the result follows from inequality (6.3) in Massart [9]. □

Let us now prove an analogue of Theorem 2.1 when $\mathscr{F}$ is a family of indicator functions.

Theorem 3.1. *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 $\mathscr{X}$. For $\mathscr{F}=\{\mathbb{1}_C,\ C\in\mathscr{C}\}$, $\mathscr{E}(\mathbf{X})=\big\{\{i,\ X_i\in C\},\ C\in\mathscr{C}\big\}$,*

$$\sigma=\sup_{C\in\mathscr{C}}\left[\frac{1}{n}\sum_{i=1}^n\mathbb{P}(X_i\in C)\right]^{1/2}\quad\text{and}\quad\Gamma=\mathbb{E}\left[\log(2|\mathscr{E}(\mathbf{X})|)\right]$$

the following holds,

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

This result is of the same flavour as the one Pascal Massart established in Massart [9] (see his Lemma 6.4). Massart's result involves an inexplicit constant, is established under the assumption that the X_i are i.i.d. and for σ satisfying an inequality while our bound is true for all σ. Nevertheless, the proof of our Theorem 3.1 is essentially included in that provided by Massart for his Lemma 6.4. We provide a proof below to assess the constants.

Proof. By the symmetrization argument (2.1),

$$\begin{aligned}\mathbb{E}\left[\sup_{C\in\mathscr{C}}\sum_{i=1}^{n}\mathbb{1}_C(X_i)\right] &\leq \mathbb{E}\left[\sup_{C\in\mathscr{C}}\sum_{i=1}^{n}(\mathbb{1}_C(X_i)-\mathbb{P}(X_i\in C))\right]+n\sigma^2\\ &\leq 2\mathbb{E}\left[\sup_{C\in\mathscr{C}}\left|\sum_{i=1}^{n}\varepsilon_i\mathbb{1}_C(X_i)\right|\right]+n\sigma^2\\ &= 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]+n\sigma^2. \qquad (3.2)\end{aligned}$$

Let us denote by $\mathbb{E}_\varepsilon$ the conditional expectation given $\mathbf{X}=(X_1,\ldots,X_n)$. Applying Lemma 3.1 with $T=\{(1_E(1),\ldots,1_E(n)),\ E\in\mathscr{E}(\mathbf{X})\}$ we get

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

Taking expectations with respect to $\mathbf{X}$ on both sides of this inequality, we derive from Cauchy-Schwarz's inequality and (3.2) 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 (2.2). □

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 for a VC-class $\mathscr{C}$ with dimension not larger than d, $\log(2\,|\mathscr{E}(\mathbf{X})|)\leq\overline{\Gamma}_n(d)$ where $\overline{\Gamma}_n(d)$ is given by (2.5). We immediately deduce from Theorem 3.1 the following corollary.

Corollary 3.1. *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 VC with dimension d. For $\mathscr{F}=\{\mathbb{1}_C,\ C\in\mathscr{C}\}$*

$$\mathbb{E}\left[Z(\mathscr{F})\right]\leq 2\left[\sigma\sqrt{2n\overline{\Gamma}_n(d)}+4\overline{\Gamma}_n(d)\right]\quad\text{with}\quad\sigma^2=\sup_{C\in\mathscr{C}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(X_i\in C)\quad(3.3)$$

and $\overline{\Gamma}_n(d)$ given by (2.5).

To analyse this bound, let us consider the situation where $\mathscr{G}$ is the family of indicators $\{\mathbb{1}_C,\ C \in \mathscr{D}\}$ indexed by a VC-class $\mathscr{D}$ of subsets of $\mathscr{X}$ with dimension $d \geq 1$ and $\mathscr{F} = \mathscr{G}(\sigma)$ given by (2.9). The bound we get on $\mathbb{E}[Z(\mathscr{F})]$ writes as

$$2\sqrt{2n}\sigma\sqrt{d\log n}(1+o(1)) \ \text{ when } n \to +\infty.$$

It can be used to bound from above the smaller quantity

$$E = \max\left\{\mathbb{E}\left[\sup_{C\in\mathscr{C}}\sum_{i=1}^{n}(\mathbb{1}_C(X_i) - \mathbb{P}(X_i \in C))\right]; \right.$$
$$\left.\mathbb{E}\left[\sup_{C\in\mathscr{C}}\sum_{i=1}^{n}(\mathbb{P}(X_i \in C) - \mathbb{1}_C(X_i))\right]\right\}.$$

When the X_i are i.i.d., an alternative bound on E is given in Theorem 13.7 of Boucheron *et al.* [6]. This bound, that we recall below, is based on the control of the universal entropy of a VC-class of sets which is due to Haussler [8].

$$E \leq 72\sqrt{n}\sigma\sqrt{d\log\left(\frac{4e^2}{\sigma}\right)} \ \text{ provided that } \ \sigma \geq 24\sqrt{\frac{d}{5n}\log\left(\frac{4e^2}{\sigma}\right)}. \tag{3.4}$$

This constraint on σ can be reformulated as $\sigma \geq \sigma_n$ where

$$\sigma_n = \frac{24}{\sqrt{10}}\sqrt{\frac{d\log n}{n}}(1+o(1)) \ \text{ when } n \to +\infty.$$

In the case $\sigma = \sigma_n$, inequality (3.3) improves their bound in terms of constants at least when n is large enough. However in the situation where σ is fixed and n is large, their bound improves ours by a $\sqrt{\log n}$ factor. We provide below an improvement of Boucheron *et al.*'s bound (and hence of (3.3)) in terms of constants at least when σ is large enough compared to σ_n.

Proposition 3.1. *Under the assumptions of Corollary 3.1 and provided that the dimension of $\mathscr{C}$ is not larger than $d \geq 1$,*

$$\mathbb{E}[Z(\mathscr{F})] \leq 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq 10\sqrt{n}\ \overline{H}\,(\sigma \vee a) \tag{3.5}$$

where $\overline{H}$ and a are given by (2.12).

Proof. Throughout this proof d stands for $d \wedge n$. Given $\mathbf{X} = (X_1, \ldots, X_n)$, let $P_{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\delta_{X_i}$ be the empirical distribution based on the X_i and for $\eta > 0$ let $\mathscr{C}_\eta = \mathscr{C}_\eta(\mathbf{X})$ be a maximal η-separated subset of $\mathscr{C}$ for the $\mathbb{L}_1(P_{\mathbf{X}})$-norm, that is, $\mathscr{C}_\eta$ is a (random) subset of $\mathscr{C}$ satisfying the following properties: for all $C, C' \in \mathscr{C}_\eta$ with $C \neq C'$, $|C\Delta C'|_{1,\mathbf{X}} = \sum_{i=1}^{n}|\mathbb{1}_{X_i\in C} - \mathbb{1}_{X_i\in C'}| > n\eta$ and for all $C \in \mathscr{C}$, their exists $\Pi_\eta C \in \mathscr{C}_\eta$ such that $|C\Delta\Pi_\eta C|_{1,\mathbf{X}} \leq n\eta$. Note that for $\eta < 1/n$, we necessarily have that $|C\Delta\Pi_\eta C|_{1,\mathbf{X}} = 0$ which means that

$$\mathbb{1}_C(X_i) = \mathbb{1}_{\Pi_\eta C}(X_i) \ \text{ for all } C \in \mathscr{C} \text{ and } 1 \leq i \leq n. \tag{3.6}$$

The proof is decomposed into three steps.

Step 1: An entropy bound In the sequel, we provide an upper bound for the quantities $\log|\mathscr{C}_\eta|$ with $\eta > 0$. We first note that given two distinct sets $C, C' \in \mathscr{C}_\eta$, $|C\Delta C'|_{1,\mathbf{X}} > n\eta > 0$, hence

$$C \cap \{X_1, \ldots, X_n\} \neq C' \cap \{X_1, \ldots, X_n\},$$

and since the number of such subsets of $\{X_1, \ldots, X_n\}$ is not larger than $\sum_{k=0}^{d} \binom{n}{k}$ by Sauer's lemma, we necessarily have

$$\log|\mathscr{C}_\eta| \leq \log\left[\sum_{j=0}^{d}\binom{n}{j}\right] = \overline{\Gamma}_n(d) - \log 2 \quad \text{for all } \eta > 0.$$

Since two arbitrary subsets $C, C' \in \mathscr{C}$ satisfy $|C\Delta C'|_{1,\mathbf{X}} \leq n$, if $\eta \geq 1$ one should take $\mathscr{C}_\eta = \mathscr{C}_1 = \{C_0\}$ for some arbitrary $C_0 \in \mathscr{C}$ so that $\log|\mathscr{C}_\eta| = 0$ for all $\eta \geq 1$.

When $\eta \in (0,1)$ there exists $k \in \{1, \ldots, n\}$ such that $(k-1)/n \leq \eta < k/n$ and for all $C, C' \in \mathscr{C}_\eta$, $|C\Delta C'|_{1,\mathbf{X}} > k-1$, hence $|C\Delta C'|_{1,\mathbf{X}} \geq k$, and it follows from Haussler [8] Theorem 1 that

$$\log\left(|\mathscr{C}_\eta|\right) \leq \log\left[e(d+1)\left(\frac{2e}{\eta}\right)^d\right].$$

Putting these bounds on $\log|\mathscr{C}_\eta|$ together we obtain that, for all $\eta > 0$, $\log|\mathscr{C}_\eta| \leq h(\eta)$ with

$$h(\eta) = \left\{\left[\log\left(e(d+1)(2e)^d\right) + d\log\frac{1}{\eta}\right] \wedge \left[\overline{\Gamma}_n(d) - \log 2\right]\right\} \mathbb{1}_{(0,1)}(\eta).$$

Note that h is a nonnegative, right-continuous and nonincreasing function which is bounded from above by $\overline{\Gamma}_n(d) - \log 2$ and satisfies for $d \geq 1$, $n \geq 1$ and $\eta \in (0,1)$,

$$h(\eta) \geq \min\{2\log(2e), \log(n+1)\} \geq \log 2. \tag{3.7}$$

Step 2: Preliminary calculations For $q = 2^{5/2}e^{-6} \in (0,1)$, the function H defined by

$$H(x) = \int_0^x \sqrt{\log 2 + h(u^2) + h(q^2u^2)}du \quad \text{for } x > 0$$

is nondecreasing and concave. It is also differentiable from the right on $(0, +\infty)$ and its right-hand derivative at $x > 0$ is given by

$$H'(x) = \sqrt{\log 2 + h(x^2) + h(q^2x^2)} \leq \sqrt{2\overline{\Gamma}_n(d)}. \tag{3.8}$$

Besides, for $x \in (0,1)$ H is differentiable and

$$H'(x) \leq \sqrt{c_d + 4d\log\frac{1}{x}}$$

with

$$c_d = \log 2 + 2\log\left(e(d+1)(2e)^d\right) + 2d\log(1/q) \leq 16d \quad \text{for } d \geq 1.$$

In particular, we deduce from Jensen's inequality that for $x \in (0,1]$,

$$\begin{aligned} H(x) &\leq x \times \frac{1}{x}\int_0^x \sqrt{c_d + 4d\log\frac{1}{u}}du \leq x\left[\frac{1}{x}\int_0^x\left(c_d+4d\log\frac{1}{u}\right)du\right]^{1/2} \\ &= x\left[c_d+4d\log\frac{e}{x}\right]^{1/2} \leq 2x\left[d\log\frac{e^5}{x}\right]^{1/2} = 2\overline{H}(x). \end{aligned} \tag{3.9}$$

Let

$$\eta_0 = \eta_0(\mathbf{X}) = \sup_{C\in\mathscr{C}}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\in C}\right] \in [0,1].$$

By the symmetrization argument (2.1),

$$\begin{aligned} n\mathbb{E}\left[\eta_0(\mathbf{X})\right] &\leq \mathbb{E}\left[\sup_{C\in\mathscr{C}}\sum_{i=1}^n\left(\mathbb{1}_C(X_i)-\mathbb{P}(X_i\in C)\right)\right]+n\sigma^2 \\ &\leq 2\mathbb{E}\left[\sup_{C\in\mathscr{C}}\left|\sum_{i=1}^n\varepsilon_i\mathbb{1}_C(X_i)\right|\right]+n\sigma^2 \\ &= 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]+n\sigma^2. \end{aligned} \tag{3.10}$$

Step 3: Completion of the proof Let us now define for all positive integers k, $\eta_k = q^{2k}\eta_0$, $C_k = \Pi_{\eta_k}C$ for $C\in\mathscr{C}$ and T_k as the subset of $\mathbb{R}^n$ gathering those vectors of the form $(\mathbb{1}_{X_1\in C_{k+1}}-\mathbb{1}_{X_1\in C_k},\ldots,\mathbb{1}_{X_n\in C_{k+1}}-\mathbb{1}_{X_n\in C_k})$ as C varies along $\mathscr{C}$. For all $i\in\{1,\ldots,n\}$,

$$\mathbb{1}_{X_i\in C} = \mathbb{1}_{X_i\in C_0}+\sum_{k=0}^{+\infty}\left(\mathbb{1}_{X_i\in C_{k+1}}-\mathbb{1}_{X_i\in C_k}\right)$$

where the sum is actually finite because of (3.6). Hence,

$$\left|\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in C}\right| \leq \left|\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in C_0}\right|+\sum_{k=0}^{+\infty}\left|\sum_{i=1}^n\varepsilon_i\left(\mathbb{1}_{X_i\in C_{k+1}}-\mathbb{1}_{X_i\in C_k}\right)\right|$$

and

$$\overline{Z}(\mathscr{F}) \leq \left|\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in C_0}\right|+\sum_{k=0}^{+\infty}\sup_{t\in T_k}\left|\sum_{i=1}^n\varepsilon_i t_i\right|.$$

Denoting by $\mathbb{E}_\varepsilon$ the conditional expectation given $\mathbf{X}$, the quantities $\mathbb{E}_\varepsilon\left[\left|\sum_{i=1}^n\varepsilon_i\mathbb{1}_{X_i\in C_0}\right|\right]$ and $\mathbb{E}_\varepsilon\left[\sup_{t\in T_k}\left|\sum_{i=1}^n\varepsilon_i t_i\right|\right]$ can be bounded from above by means of Lemma 3.1 using the facts that $\sum_{i=1}^n\mathbb{1}_{X_i\in C_0}\leq n\eta_0$, $|T_k|\leq|\mathscr{C}_{\eta_k}||\mathscr{C}_{\eta_{k+1}}|\leq e^{h(\eta_k)+h(q^2\eta_k)}$ for all $k\geq 1$ and for all $C\in\mathscr{C}$

$$\begin{aligned} \sum_{i=1}^n\left(\mathbb{1}_{X_i\in C_{k+1}}-\mathbb{1}_{X_i\in C_k}\right)^2 &= n\left|C_{k+1}\Delta C_k\right|_{1,\mathbf{X}} \leq n\left[\left|C_{k+1}\Delta C\right|_{1,\mathbf{X}}+\left|C_k\Delta C\right|_{1,\mathbf{X}}\right] \\ &\leq n(1+q^2)\eta_k = n\frac{1+q^2}{(1-q)^2}(\sqrt{\eta_k}-\sqrt{\eta_{k+1}})^2. \end{aligned}$$

We get,

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left[\overline{Z}(\mathscr{F})\right] \\
&\quad\le \sqrt{2n}\left[\sqrt{\eta_0\log 2}+\frac{\sqrt{1+q^2}}{1-q}\sum_{k=0}^{+\infty}\left(\sqrt{\eta_k}-\sqrt{\eta_{k+1}}\right)\sqrt{\log 2+h(\eta_k)+h(q^2\eta_k)}\right] \\
&\quad\le \sqrt{2n}\left[\sqrt{\eta_0\log 2}+\frac{\sqrt{1+q^2}}{1-q}\sum_{k=0}^{+\infty}\int_{\sqrt{\eta_{k+1}}}^{\sqrt{\eta_k}}\sqrt{\log 2+h(u^2)+h(q^2u^2)}du\right] \\
&\quad\le \sqrt{2n}\left[\sqrt{\eta_0\log 2}+\frac{\sqrt{1+q^2}}{1-q}\int_0^{\sqrt{\eta_0}}\sqrt{\log 2+h(u^2)+h(q^2u^2)}du\right].
\end{aligned}$$

Using (3.7),

$$\sqrt{\eta_0\log 2}\le\sqrt{\frac{\log 2}{3\log 3}}\int_0^{\sqrt{\eta_0}}\sqrt{\log 2+h(u^2)+h(q^2u^2)}du$$

and hence,

$$\mathbb{E}_\varepsilon\left[\overline{Z}(\mathscr{F})\right]\le\sqrt{n}b_qH\left[\sqrt{\eta_0(\mathbf{X})}\right]\quad\text{with}\quad b_q=\sqrt{2}\left(\frac{\sqrt{1+q^2}}{1-q}+\sqrt{\frac{1}{3}}\right)<2.5.$$

Taking the expectation with respect to $\mathbf{X}$ on both sides and using Jensen's inequality yield to

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]\le\sqrt{n}b_q\mathbb{E}\left[H\left(\sqrt{\eta_0(\mathbf{X})}\right)\right]\le\sqrt{n}b_qH\left[\sqrt{\mathbb{E}\left[\eta_0(\mathbf{X})\right]}\right]. \tag{3.11}$$

If $\overline{a}=32(\overline{\Gamma}_n(d)/n)^{1/2}\ge 1$, $a=\overline{a}\wedge 1=1$ and

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]\le\sqrt{n}b_qH\left[\sqrt{\mathbb{E}\left[\eta_0(\mathbf{X})\right]}\right]\le 2.5\sqrt{n}H(1)=2.5\sqrt{n}H(\sigma\vee 1). \tag{3.12}$$

Otherwise $a=\overline{a}<1$ and let us set $G(u)=H(\sqrt{u})$ for $u>0$. The function G is nondecreasing, concave, differentiable from the right on $(0,+\infty)$ and its right-hand derivative at $x>0$ is given by $G'(x)=H'(\sqrt{x})/(2\sqrt{x})$. In particular, using (3.10) and the fact that the graph of a concave function lies below its tangents, we obtain that

$$\begin{aligned}
H\left[\sqrt{\mathbb{E}\left[\eta_0(\mathbf{X})\right]}\right] &= G(\mathbb{E}\left[\eta_0(\mathbf{X})\right])\le G\left(\sigma^2+2n^{-1}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]\right) \\
&\le G\left(\sigma^2\vee a^2+2n^{-1}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]\right) \\
&\le G(\sigma^2\vee a^2)+2n^{-1}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]G'(\sigma^2\vee a^2) \\
&= H(\sigma\vee a)+\frac{H'(\sigma\vee a)}{an}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \\
&\le H(\sigma\vee a)+\frac{H'(a)}{an}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right].
\end{aligned}$$

This inequality together with (3.11), leads to

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq \sqrt{n} b_q H(\sigma \vee a) + \frac{b_q H'(a)}{a\sqrt{n}} \mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \tag{3.13}$$

and, since by (3.8) and our choice of $\overline{a}$ (that is $\overline{a} > b_q \sqrt{2n^{-1}\overline{\Gamma}_n(d)}/(1-b_q/2.5)$),

$$\frac{b_q H'(a)}{a\sqrt{n}} \leq \frac{b_q}{a}\sqrt{\frac{2\overline{\Gamma}_n(d)}{n}} \leq 1 - \frac{b_q}{2.5},$$

we obtain that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq 2.5\sqrt{n} H(\sigma \vee a). \tag{3.14}$$

Putting (3.12) and (3.14) together and using (3.9), we obtain that in both cases

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq 2.5\sqrt{n} H(\sigma \vee a) \leq 5\sqrt{n}\ \overline{H}(\sigma \vee a)$$

and we conclude by (2.2). □

3.3. *Completion of the proofs of Theorems 2.1 and 2.2*

We start with the proof of Theorem 2.1. In view of our convention about the definition of $\mathbb{E}[Z(\mathscr{F})]$ 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 not larger than d, $\mathscr{C}_u$ is VC with dimension not larger than d as well. Besides, $\mathscr{C}_u$ is countable since $\mathscr{F}$ is and by Markov's inequality

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

Applying Theorem 3.1 to the class of sets $\mathscr{C}_u$ leads to

$$\mathbb{E}\left[\sup_{C\in\mathscr{C}_u}\left|\sum_{i=1}^n \varepsilon_i \mathbb{1}_C(X_i)\right|\right] \leq \left(\frac{\sigma}{u} \wedge 1\right)\sqrt{2n\Gamma_u} + 4\Gamma_u. \tag{3.15}$$

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

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

and taking expectations on both sides gives

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leq \int_0^1 \mathbb{E}\left[\sup_{C\in\mathscr{C}_u}\left|\sum_{i=1}^n \varepsilon_i \mathbb{1}_C(X_i)\right|\right] du. \tag{3.16}$$

Using (3.15),

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

and the conclusion follows from (2.2).

The proof of Theorem 2.2 is quite similar except that we now bound the right-hand side of (3.16) using Proposition 3.1. Since $u\mapsto \overline{H}(u)$ is concave and nondecreasing on $[0,1]$, we get

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] &\leq 5\sqrt{n}\int_0^1 \overline{H}\left[(u^{-1}\sigma)\wedge 1)\vee a\right] du \\ &\leq 5\sqrt{n}\,\overline{H}\left[\int_0^1 [(u^{-1}\sigma)\wedge 1)\vee a]du\right] \\ &= 5\sqrt{n}\,\overline{H}\left[\sigma\vee a - \sigma\log(\sigma\vee a)\right]\end{aligned}$$

which leads to the result.

4. Additional proofs

4.1. Proof of Proposition 2.1

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

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

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 2.2

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 [14] (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 2.3

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 2.2

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 2.2 will follow from Corollary 2.1 if we can prove that $\mathscr{G}$ is weak VC-major. This is a consequence of the next lemma.

Lemma 4.1. *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

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

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

Acknowledgements

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

References

[1] Baraud, Y. and Birgé, L. (2015). Rates of convergence of rho-estimators for sets of densities satisfying shape constraints. http://arxiv.org/abs/1503.04427.

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

[3] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413. MR1679028

[4] Birgé, L. (1989). The Grenander estimator: A nonasymptotic approach. *Ann. Statist.*, 17(4):1532–1549. MR1026298

[5] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1–2):113–150. MR1240719

[6] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities*. Oxford University Press, Oxford. MR3185193

[7] Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216. MR2243881

[8] Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A*, 69(2):217–232. MR1313896

[9] 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. MR2319879

[10] Sauer, N. (1972). On the density of families of sets. *J. Combinatorial Theory Ser. A*, 13:145–147. MR0307902

[11] Talagrand, M. (1996). New concentration inequalities in product space. *Invent. Math.*, 126:505–563. MR1419006

[12] VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.*, 18:907–924. MR1056343

[13] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. MR1652247

[14] 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. MR1385671