Ann. I. H. Poincaré – PR 39, 1 (2003) 95–119

S0246-0203(02)00010-9/FLA

EXACT RATES IN VAPNIK–CHERVONENKIS BOUNDS

VITESSES EXACTES DANS LES BORNES DE VAPNIK–CHERVONENKIS

Nicolas VAYATIS [1]

Centre de mathématiques et de leurs applications (CMLA), École normale supérieure de Cachan, 61, av. du Président Wilson, 94235 Cachan cedex, France

Received 7 November 2000, revised 4 October 2001

ABSTRACT. – Vapnik–Chervonenkis bounds on rates of uniform convergence of empirical means to their expectations have been continuously improved over the years since the precursory work in [26]. The result obtained by Talagrand in 1994 [21] seems to provide the final word as far as universal bounds are concerned. However, in the case where there are some additional assumptions on the underlying probability distribution, the exponential rate of convergence can be fairly improved. Alexander [1] and Massart [15] have found better exponential rates (similar to those in Bennett–Bernstein inequalities) under the assumption of a control on the variance of the empirical process. In this paper, the case of a particular distribution is considered for the empirical process indexed by a family of sets, and we provide the exact exponential rate based on large deviations theorems, as predicted by Azencott [2].

MSC: 60E15; 60F10

RÉSUMÉ. – Les bornes de Vapnik–Chervonenkis sur les vitesses de convergence uniforme des moyennes empiriques vers leurs espérances ont fait l'objet de nombreuses améliorations depuis leur travail précurseur [26]. Le résultat obtenu par Talagrand en 1994 [21] semble mettre un point final à la question des bornes universelles. Cependant, dans le cas d'hypothèses supplémentaires sur la loi de probabilité sous-jacente, le taux exponentiel de la convergence peut être amélioré. Alexander [1] et Massart [15] ont trouvé de meilleures vitesses exponentielles (similaires à celles des inégalités de type Bennett–Bernstein) sous l'hypothèse d'un contrôle de la variance du processus empirique. Dans cet article, nous étudions le cas d'une loi particulière pour le processus empirique indexé par une famille d'ensembles et nous démontrons un résultat, annoncé par Azencott [2], avec une borne présentant un taux exponentiel exact, conformément aux théorèmes de grandes déviations.

E-mail address: vayatis@ccr.jussieu.fr (N. Vayatis).

[1] Present address: Laboratoire de probabilités et modèles aléatoires, Université Paris 6, 175, rue du Chevaleret, 75013 Paris cedex, France.

1. Introduction and motivations

Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of independent random variables with distribution μ on a Polish space $(\mathcal{X},\mathcal{B})$ where $\mathcal{B}$ is the Borel σ-algebra. We denote by $\mu_n = (\sum_{i=1}^n \delta_{X_i})/n$ the corresponding empirical measure. We consider a countable and totally bounded (for the symmetric difference metric) family Γ of measurable sets of $\mathcal{X}$, with finite Vapnik–Chervonenkis (VC) dimension V.

We recall that a set T in a metric space $\mathcal{T}$ is *totally bounded* if, for every λ, there exists a finite number of closed balls of radius λ covering B (a λ-*covering* of B). The *VC dimension* of a family of sets Γ is defined as the largest integer k such that for some set E of k points, any subset of E is obtained as the intersection between E and some set C of Γ.

In the present paper, we shall examine the uniform deviation over Γ of the empirical measure μ_n from its expectation μ. Our main object of interest will be the following probability tail

$$\rho(\Gamma,\mu,n,\varepsilon) = \Pr\bigl\{\sup_{C\in\Gamma}\bigl|\mu_n(C)-\mu(C)\bigr| > \varepsilon\bigr\}. \tag{1}$$

Such a probability tail is known to tend to zero as the sample size n approaches infinity under the assumption that the family Γ is reasonably small (this is the Glivenko–Cantelli problem, see e.g. [7]). But a challenging issue is also the computation of the rates at which this uniform convergence is achieved. We insist on the fact that finding sharp rates is not mere sophistication and it has a tremendous impact on the applications in the field of machine learning. Indeed, VC bounds are closely related to the error bounds on generalization of learning algorithms like neural networks. Such theoretical results actually provide an important tool in the design of learning structures and the prediction of their performance (see [25]). Pioneering work on is due to Vapnik and Chervonenkis [26,27] who extended classical results of Kiefer [12], and Dvoretzsky, Kiefer and Wolfowitz [8] (see also [15] for a consistent review). Since then, many techniques have been developped in empirical process theory in order to improve these Vapnik–Chervonenkis (VC) inequalities. The general structure for VC bounds is the following: there is an M such that, for $n\varepsilon^2 > M$, we have

$$\rho(\Gamma,\mu,n,\varepsilon) \leqslant K\bigl(n\varepsilon^2\bigr)^{\tau}\exp\bigl\{-n\,\phi(\varepsilon)\bigr\}, \tag{2}$$

where K is a multiplicative constant, τ is the power of the polynomial term which reflects the *capacity* of the family Γ (in some way, it is related to a complexity index, e.g. the VC dimension of the family Γ), and $\phi(\varepsilon)$ is the exponential rate of convergence.

In order to discuss previous results and find out what the "best" K, τ and $\phi(\varepsilon)$ are, we have to point out that there are various assumptions that can be made on the *a priori* knowledge that we have on the underlying distribution μ:

(1) Universal (or distribution-free) case (HVT) – it is assumed that μ can be *any* probability distribution on $\mathcal{M}_1(\mathcal{X})$.
(2) Control-of-the-variance assumption (ALMA) – there exists a constant σ^2 such that

$$\sup_{C\in\Gamma}\mu(C)\bigl(1-\mu(C)\bigr) \leqslant \sigma^2 < \frac{1}{4}. \tag{3}$$

(3) Distribution-dependent case (DD) – we assume that we know exactly the particular distribution μ that underlies the data.

Note that these three types of assumptions correspond to different exponential rates like in the classical case of the deviation of the empirical mean from its expected value. Assumption HVT[2] leads to Hoeffding's exponential rate, Assumption ALMA[3] gives Bennett–Bernstein rates, and Assumption DD corresponds to the exact rate of Chernoff's bound (as confirmed by large deviations theory). We insist on the fact that *the issue of providing sharp exponential rates is prior to the question of getting a "good" polynomial factor*.

The Assumption HVT has been considered by Vapnik and Chervonenkis [26], Vapnik [23], Devroye [5], Pollard [18], Parrondo and van den Broek [17], Lugosi [14], Talagrand [21]. The best exponential rate in that case corresponds to Hoeffding's inequality where $\phi(\varepsilon) = 2\varepsilon^2$ (first obtained by Devroye in [5]), and the best polynomial power of $\tau = V - 1/2$ was obtained, at the cost of significant breakthroughs in empirical processes theory, by Talagrand in [21].

The case of Assumption ALMA has been mainly carried out by Alexander [1] and Massart [15]. Massart establishes a bound with

$$\phi(\varepsilon) = \frac{\varepsilon^2}{2(\sigma^2 + \frac{\sigma}{\sqrt{n}}(3\sigma + \varepsilon\sqrt{n}))} \tag{4}$$

and $\tau = 3V$. Alexander proves a similar result with general exponential rate involving the variance of the empirical process but with a huge capacity term ($\tau = 2^{12}V!$). There are some hints and proof sketches on how to improve these results regarding to the polynomial factor which have been provided by Talagrand in [21].

The purpose of this paper is to investigate the case where Assumption DD is adopted. As pointed out by Azencott [2], the exponential rate which is expected is the one of Sanov's theorem (see e.g. [4]) involving the Kullback information. Indeed, this result has been proved in a large deviations setting (asymptotically on a logarithmic scale) by Wu in [30] for the functional case. We propose to make an accurate statement of the bound in the case of empirical processes indexed by sets and to prove a non-asymptotic result reflecting the general structure of VC bounds.

Indeed, this investigation was motivated by our empirical study on VC bounds and VC dimension in [29]. In this experimental work, our goal was to test the very structure of VC bounds for particular distributions through computer simulations. Our idea (following the general, but incomplete, approach of [28]) was basically to estimate the probability tail $\rho(\Gamma, \mu, n, \varepsilon)$, and then fit the results with the explicit formula given in the bound (2). It is worth noticing that if one has a precise knowledge of the exponential rate $\phi(\varepsilon)$ (this indeed is the most crucial issue for the success of these experiments!), it is then possible to estimate precisely both the complexity index τ and the multiplicative constant K (see [29] for details and examples). This methodology provides an interesting machinery for testing conjectures about the quantities involved in VC bounds.

[2] HVT stands for Hoeffding–Vapnik–Talagrand.

[3] ALMA stands for ALexander–MAssart.

Remark 1.1. – The distribution-dependent VC bounds which are presented in statistical learning literature, in [24] and [25] for instance, refer to different complexity concepts (VC entropy or annealed entropy), but not to the exponential rate.

Remark 1.2. – Except in some very particular cases (see [16]), the multiplicative constants in such bounds are very difficult to control. In this work, we are not concerned about these constants. However, we have proposed in [29] a simulation protocol which leads to sharp empirical estimations of the constant K.

Notations. We need to introduce the *Kullback information function* for Bernoulli distributions which we denote by H.

$$\forall q, p \in (0,1), \quad H(q,p) = q \ln\left(\frac{q}{p}\right) + (1-q)\ln\left(\frac{1-q}{1-p}\right). \tag{5}$$

We recall the standard *Chernoff bound* on large deviations.

$$\forall C \in \Gamma, \quad \Pr\{\mu_n(C) - \mu(C) > \varepsilon\} \leqslant \exp\{-nH(\mu(C)+\varepsilon, \mu(C))\}, \tag{6}$$

$$\forall C \in \Gamma, \quad \Pr\{\mu(C) - \mu_n(C) > \varepsilon\} \leqslant \exp\{-nH(\mu(C)-\varepsilon, \mu(C))\}. \tag{7}$$

As we consider the two-sided probability tail, we will have to consider the "worst" of these two exponential rates, as being the *exact exponential rate*. We shall denote it by $\Lambda_q(\varepsilon)$.

$$\forall q \in (\varepsilon, 1-\varepsilon), \quad \Lambda_q(\varepsilon) = H(q+\varepsilon, q) \wedge H(q-\varepsilon, q). \tag{8}$$

Another important concept is the one of *critical value* of the family Γ. We introduce the range J of values of the mass of the elements C of Γ with respect to the distribution μ.

$$J = \{q = \mu(C) \colon C \in \Gamma\}. \tag{9}$$

DEFINITION 1.3. – *We define the* critical values *p_c of Γ with respect to the distribution μ as the values which minimize the function $q \to \Lambda_q(\varepsilon)$ over the set J.*

The reader should be warned that the constant K is used repeatedly, but, for notational convenience, its value is not fixed. This constant depends indeed on Γ (through its VC dimension V) but we have not captured, in the present work, the type of dependency which is involved. [4]

2. Main results

In this work, we have investigated the two general proof methods which have been developed in proving rates of convergence for empirical processes (cf. [18,6], for methodological inventories). The purpose of both methods is to make the supremum tractable. The *approximation method* allows to reduce the family Γ to a finite approximating family. The *combinatorial method* is based on symmetrization argument allowing to consider the trace of Γ on a fixed sample.

[4] However, our empirical study in [29] provides some insights on this issue.

Our main theorem follows the line of proof of the original Vapnik–Chervonenkis paper [26] and the improved version provided by Devroye [5]. This result indicates the exact exponential rate for empirical processes indexed by classes of sets. In order to keep track of this correct rate, some sophistications were needed and we used abusively some techniques from the work of Talagrand in [21].

THEOREM 2.1. – *Given $\varepsilon > 0$, let p be a critical value of Γ with respect to the distribution μ. We have*

$$p := p(\varepsilon) = \arg\min_{q \in J} \Lambda_q(\varepsilon).$$

There exists some constant K such that, for any n, and ε small enough,

$$\Pr\{\sup_{C \in \Gamma} |\mu_n(C) - \mu(C)| > \varepsilon\} \leqslant K n^{3V+21} \exp\{-n\Lambda_p(\varepsilon)\}. \tag{10}$$

Remark 2.2. – In this result, the polynomial factor is in "n" instead of the typical "$n\varepsilon^2$" (we recall that $M = \varepsilon\sqrt{n}$ is the "natural" variable in the study of such probability tails). This means actually that the bound is trivial in the case of ε being of the order $1/\sqrt{n}$. The VC bound obtained in the theorem becomes active for ε at least of the order $\sqrt{(\log n)/n}$, and, in particular, for ε fixed.

Remark 2.3. – By slightly modifying the end of the proof, we can get a polynomial factor in $n\varepsilon^2$, but we come up with a condition like $n\varepsilon^3$ large enough which seems to be a weaker result. We thus have, with the same assumptions as in the theorem, that there exists some constants K and M such that, for ε small enough, and for $n\varepsilon^3 > M$,

$$\Pr\{\sup_{C \in \Gamma} |\mu_n(C) - \mu(C)| > \varepsilon\} \leqslant K \left(n\varepsilon^2\right)^{3V+21} \exp\{-n\Lambda_p(\varepsilon)\}. \tag{11}$$

Remark 2.4. – A similar result holds for the one-sided probability tail

$$\rho_+(\Gamma, \mu, n, \varepsilon) = \Pr\{\sup_{C \in \Gamma} \left(\mu_n(C) - \mu(C)\right) > \varepsilon\}, \tag{12}$$

except that, in this case, one shall simply have $\Lambda_p(\varepsilon) = H(p + \varepsilon, p)$.

We then formulate a simple characterization of the critical values of Γ (it is based on a detailed study of the exponential rate $\Lambda_q(\varepsilon)$ presented in Section 3).

PROPOSITION 2.5. – *Let J be the range of the values of the mass $\mu(C)$ of the elements of Γ. We assume that there is a neighborhood $\mathcal{V}$ of $\frac{1}{2}$ such that $\mathcal{V} \cap J = \emptyset$. Then, there is an ε_0 such that for $\varepsilon < \varepsilon_0$, the* critical values *$p_c$ of Γ with respect to the distribution μ are the closest value to $1/2$ from the set J. In other words, we have, if ε is small enough,*

$$p_c := \arg\min_{q \in J} \Lambda_q(\varepsilon) = \arg\min_{q \in J} \left| q - \frac{1}{2} \right|. \tag{13}$$

Remark 2.6. – Note that the assumption stated in the proposition can be rephrased by saying that the range J does not contain the value $\frac{1}{2}$. If we drop this assumption, then we have that $p_c \in ((1-\varepsilon)/2, 1/2) \cup (1/2, (1+\varepsilon)/2)$.

Remark 2.7. – In any case, because of the symmetry of the function $q \to \Lambda_q(\varepsilon)$ with respect to the position $1/2$, there are at most two critical values for Γ given the distribution μ. Indeed, if p is a critical value, then $1-p$ is the other possible critical value.

The following sections are dedicated to the proofs of these results. In Section 3, we provide the analysis of the function $q \to \Lambda_q(\varepsilon)$. The approximation method is investigated through Section 4. We establish a partial result (cf. Proposition 4.1) which turns out to be useful in the sequel. The proof of Theorem 2.1, based on the combinatorial method, is eventually presented in Section 5.

3. Proof of Proposition 2.5

We notice that the functions $x \to H(x+\varepsilon, x)$ and $x \to H(x-\varepsilon, x)$ are symmetric with respect to $x = \frac{1}{2}$. We have indeed,

$$\forall x, \quad H(x+\varepsilon, x) = H(1-x-\varepsilon, 1-x).$$

Thus, it suffices to consider the variations of g_ε defined by $g_\varepsilon(x) = H(x+\varepsilon, x)$ (see Fig. 1). A quick study of this function shows that its second derivative is positive as soon as $\varepsilon < \sqrt{3}/2$. Hence, this function is convex. We note $a = \inf J$ and $b = \sup J$. We want to derive the value of $\inf_{x\in J} H(x+\varepsilon, x)$.

- *Case* (1) $a > 1/2$.
  We notice that $g'_\varepsilon(1/2) \geqslant 0$ if $\varepsilon \in (0; 1/2)$, hence g_ε is increasing on J (recall that this function is convex), and we have

  $$\inf_{x\in J} H(x+\varepsilon, x) = H(a+\varepsilon, a).$$

  Similarly, we have

  $$\inf_{x\in J} H(x-\varepsilon, x) = H(a-\varepsilon, a).$$

  Hence,

  $$\inf_{x\in J} \big\{ H(x+\varepsilon, x) \wedge H(x-\varepsilon, x) \big\} = H(a+\varepsilon, a) \wedge H(a-\varepsilon, a).$$

- *Case* (2) $b < 1/2$.
  We notice that $g'_\varepsilon((1-\varepsilon)/2) < 0$ if $\varepsilon > 0$. Thus g_ε decreases on J if we consider $\varepsilon < 1-2b$. We have

  $$\inf_{x\in J} H(x+\varepsilon, x) = H(b+\varepsilon, b).$$

  Thus,

  $$\inf_{x\in J} \big\{ H(x+\varepsilon, x) \wedge H(x-\varepsilon, x) \big\} = H(b+\varepsilon, b) \wedge H(b-\varepsilon, b).$$

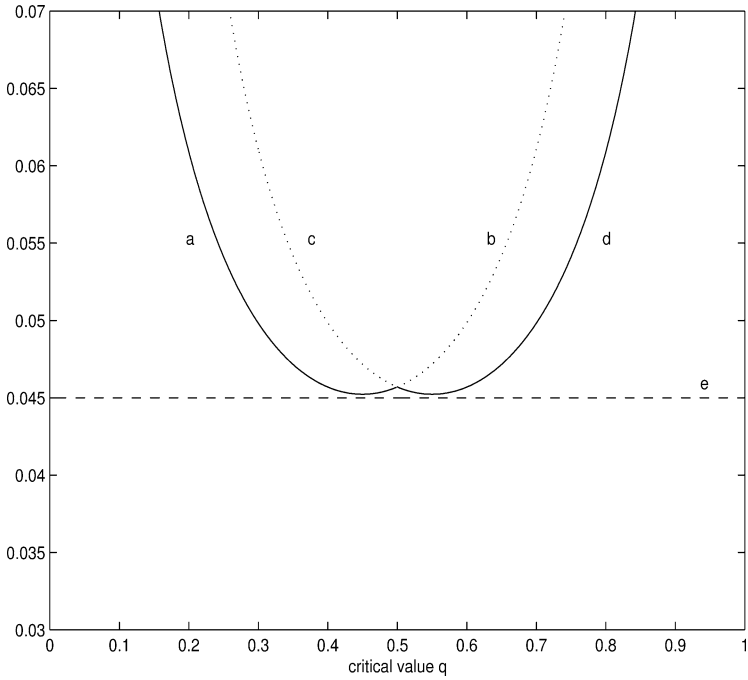

Fig. 1. Plot of the function $q \to \Lambda_q(\varepsilon)$ for $\varepsilon = 0.15$. The branches (a) and (b) represent the function $q \to H(q+\varepsilon, q)$, while (c) and (d) represent the function $q \to H(q-\varepsilon, q)$. The line indexed by (e) corresponds to the universal case (HVT) where the exponential rate is $2\varepsilon^2$.

- *Case* (3) $a < 1/2 < b$.
  We write $J = J_1 \cup J_2$, where

$$J_1 = J \cap [0, 1/2[, \quad \text{and} \quad J_2 = J \cap [1/2, 1].$$

  We note $u = \sup J_1$ and $v = \inf J_2$. Since we have assumed that there exists a neighborhood $\mathcal{V}$ of $\frac{1}{2}$ such that $\mathcal{V} \cap J = \emptyset$, we have $u < 1/2$. Thus, we have, from Case (2), that, if $\varepsilon < 1 - 2u$,

$$\inf_{x \in J_1} H(x+\varepsilon, x) = H(u+\varepsilon, u),$$

  and, from Case (1)

$$\inf_{x \in J_2} H(x+\varepsilon, x) = H(v+\varepsilon, v)$$

  to obtain, eventually,

$$\inf_{x \in J} H(x+\varepsilon, x) = H(u+\varepsilon, u) \wedge H(v+\varepsilon, v).$$

  We also have

$$\inf_{x \in J} H(x-\varepsilon, x) = H(u-\varepsilon, u) \wedge H(v-\varepsilon, v).$$

Assume that u is the closest value to $1/2$. Then, since $1/2 < 1-u < v$, we have

$$H(u-\varepsilon,u) = H(1-u+\varepsilon,1-u) < H(v+\varepsilon,v),$$

and, similarly,

$$H(u+\varepsilon,u) = H(1-u-\varepsilon,1-u) < H(v-\varepsilon,v).$$

Thus, if we assume that $u = \arg\min_{q:\ q=\mu(C),\, C\in\Gamma} |q-\frac{1}{2}|$, we then have

$$\inf_{x\in J}\big\{H(x+\varepsilon,x)\wedge H(x-\varepsilon,x)\big\} = H(u+\varepsilon,u)\wedge H(u-\varepsilon,u).$$

The same argument in the case where $v = \arg\min_{q:\ q=\mu(C),\, C\in\Gamma} |q-\frac{1}{2}|$ leads to

$$\inf_{x\in J}\big\{H(x+\varepsilon,x)\wedge H(x-\varepsilon,x)\big\} = H(v+\varepsilon,v)\wedge H(v-\varepsilon,v).$$

Therefore, we have proved that, given J, then, for ε small enough, we have

$$p := \operatorname*{arg\,min}_{q:\ q=\mu(C),\, C\in\Gamma} \big(H(q+\varepsilon,q)\wedge H(q-\varepsilon,q)\big) = \operatorname*{arg\,min}_{q:\ q=\mu(C),\, C\in\Gamma} \left|q-\frac{1}{2}\right|.$$

Remark 3.1. – Note that the closer p is to $1/2$, the smaller ε has to be.

4. Approximation method

The bound we have obtained through the approximation method (Proposition 4.1) is simply a preliminary step since it involves a disturbing corrective term. However, this step turns out to play a key role in the proof of our main theorem (Theorem 2.1).

PROPOSITION 4.1. – *For every $\beta > 0$, there exist $M(\beta, p, V)$ and $\varepsilon_0(\beta, p, V) > 0$ such that if $\varepsilon < \varepsilon_0(\beta, p, V)$ and $n\varepsilon^2 > M(\beta, p, V)$, we have*

$$\Pr\big\{\sup_{C\in\Gamma}\big|\mu_n(C)-\mu(C)\big| > \varepsilon\big\} \leqslant \exp\big\{-n\,(1-\beta)\,\Lambda_p(\varepsilon)\big\}. \tag{14}$$

Remark 4.2. – We mention that the constant M is of the order $\mathcal{O}(\frac{1}{\beta^4}\exp\frac{1}{\beta^2})$.

We now turn to the proof of this proposition.

4.1. Proof of Proposition 4.1

We define a λ-*net* Γ_λ which is a finite approximation of Γ such that, for any $C\in\Gamma$, there is a $C^*\in\Gamma_\lambda$ such that $\mu(C\Delta C^*) < \lambda$. The element C^* is called the *projection* of C on Γ_λ. We denote by $J_\lambda = \{q = \mu(C)\colon\ C\in\Gamma_\lambda\}$. The cardinality of Γ_λ is denoted by $\mathcal{N}(\lambda)$. We shall consider that $\lambda = \frac{1}{n\varepsilon^2}$. We denote by $G_n = \mu_n - \mu$ the centered empirical process. We then have

$$\Pr\Big\{\sup_{C\in\Gamma}\big|G_n(C)\big|>\varepsilon\Big\}\leqslant\Pr\Big\{\sup_{C\in\Gamma}\big|G_n(C^*)\big|>\varepsilon_1\Big\}+\Pr\Big\{\sup_{C\in\Gamma}\big|G_n(C)-G_n(C^*)\big|>\varepsilon_2\Big\}\tag{15}$$

where $\varepsilon_1+\varepsilon_2=\varepsilon$.

The first term corresponds to the same problem as the initial one but for a finite family of sets. It is easy to control it with a straightforward application of Chernoff's inequality. Hence, we have, for some β,

$$\begin{aligned}\Pr\Big\{\sup_{C\in\Gamma}\big|G_n(C^*)\big|>\varepsilon_1\Big\}&=\Pr\Big\{\sup_{C\in\Gamma_\lambda}\big|G_n(C)\big|>\varepsilon_1\Big\}\\&\leqslant 2\mathcal{N}(\lambda)\exp\Big\{-n\inf_{q\in J_\lambda}\Lambda_q(\varepsilon_1)\Big\}&(16)\\&\leqslant 2\mathcal{N}(\lambda)\exp\Big\{-n\inf_{q\in J}\Lambda_q(\varepsilon_1)\Big\}&(17)\\&\leqslant 2\mathcal{N}(\lambda)\exp\big\{-n\Lambda_p(\varepsilon_1)\big\}&(18)\\&\leqslant 2\mathcal{N}(\lambda)\exp\big\{-n(1-\beta)\Lambda_p(\varepsilon)\big\}.&(19)\end{aligned}$$

Inequality (16) comes from the union-of-events bound and an application of the Chernoff bound (see the inequalities (6) and (7) in Section 1). The sum is bounded by the worst exponential rate over the range of possible values of $\mu(C)$. Inequality (17) simply uses the fact that $J_\lambda\subset J$, and inequality (18) is a notational transformation thanks to the definition of the critical value p (see Definition 1.3). We now explain inequality (19). We note that (see Fig. 1), for p fixed, we have

$$H(p+\varepsilon,p)<H(p-\varepsilon,p),\quad\text{for } p<\frac{1}{2},\quad\text{and}$$

$$H(p+\varepsilon,p)>H(p-\varepsilon,p),\quad\text{for } p>\frac{1}{2}.$$

Thus, we have that either $\Lambda_p(\varepsilon)=H(p+\varepsilon,p)$, either $\Lambda_p(\varepsilon)=H(p-\varepsilon,p)$ when the parameter p is fixed. Hence, the function Λ_p is a convex function of ε. We can write that

$$\Lambda_p(\varepsilon_1)\geqslant\Lambda_p(\varepsilon)+(\varepsilon_1-\varepsilon)\Lambda'_p(\varepsilon_1).$$

As we want to obtain the correct exponential rate $\Lambda_p(\varepsilon)$ with possibly some corrective term, we set β such as $\Lambda_p(\varepsilon_1)=(1-\beta)\Lambda_p(\varepsilon)$. We also set $\varepsilon_1=(1-\theta)\varepsilon$. Then, we have that β and θ are related through

$$\theta\geqslant\beta\left(\frac{\Lambda_p(\varepsilon)}{\varepsilon\Lambda'_p(\varepsilon)}\right),$$

and the factor between the two is a bounded non-zero quantity. We can keep in mind that $\theta\simeq\beta/2$.

The difficult part of the work is to control efficiently the second term which is due to the approximation. We have

$$\Pr\Big\{\sup_{C\in\Gamma}\big|G_n(C)-G_n(C^*)\big|>\varepsilon_2\Big\}$$

$$\leqslant \mathcal{N}(\lambda) \max_{C^* \in \Gamma_\lambda} \Pr\Big\{ \sup_{C \in \mathcal{B}(C^*,\lambda)} \big|G_n(C) - G_n(C^*)\big| > \varepsilon_2 \Big\}. \tag{20}$$

The essential part of the proof is dedicated to the control of the localized empirical process. Indeed we need to obtain a tractable exponential bound on the quantity

$$\Pr\Big\{ \sup_{C \in \mathcal{B}(C^*,\lambda)} \big|G_n(C) - G_n(C^*)\big| > \varepsilon_2 \Big\}. \tag{21}$$

The classical way to deal with suprema in empirical processes theory is to use the *chaining* trick (cf. [19,22,13]). However, this cannot be done straightforwardly since the process involved here does not satisfy a subgaussian inequality. The argument developed here is due to Talagrand in [21] and it can also be found in [22]. In the following Section 4.2, we shall prove

PROPOSITION 4.3. – *With $\lambda = 1/(n\varepsilon^2)$, and the same β as before, we have, for $n\varepsilon^2$ larger than a constant $M(\beta)$,*

$$\Pr\Big\{ \sup_{C \in \mathcal{B}(C^*,\lambda)} \big|G_n(C) - G_n(C^*)\big| > \varepsilon_2 \Big\} \leqslant 16 \exp\big\{-n(1-\beta)\Lambda_p(\varepsilon)\big\}. \tag{22}$$

We have $M(\beta) = \mathcal{O}(\frac{1}{\beta^4} \exp \frac{1}{\beta^2})$.

Combining inequalities (19) and (22), we finally obtain a global bound on $\rho(\Gamma, \mu, n, \varepsilon)$.

$$\Pr\Big\{ \sup_{C \in \Gamma} \big|G_n(C)\big| > \varepsilon \Big\} \leqslant 18 \mathcal{N}(\lambda) \exp\big\{-n(1-\beta)\Lambda_p(\varepsilon)\big\}. \tag{23}$$

We conclude the proof of this proposition by using the relationship between metric entropy and the VC dimension. Indeed, we recall from [10] and [22] that there exists a constant K such that

$$\mathcal{N}(\lambda) \leqslant K \left(\frac{1}{\lambda}\right)^V. \tag{24}$$

We eventually set λ being equal to $1/(n\varepsilon^2)$. We also consider that

$$18K \big(n\varepsilon^2\big)^V \leqslant \exp\big\{n\beta\Lambda_p(\varepsilon)\big\} \tag{25}$$

as soon as $n\varepsilon^2 = \mathcal{O}(\frac{1}{\beta} \log \frac{1}{\beta})$.

Then, at the cost of modifying β up to a multiplicative constant, we obtain Proposition 4.1.

We now turn to the proof of Proposition 4.3.

4.2. Proof of Proposition 4.3

4.2.1. Symmetrization

Consider the centered stochastic process $\{Z_n(C)\}_{C \in \mathcal{B}(C^*,\lambda)}$ where we have set $Z_n(C) = G_n(C) - G_n(C^*)$ (we suppose here that C^* is fixed). We introduce the independent Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_n$ (ε_i takes values 1 and -1 with

probability 1/2 each). Thanks to a result from [21] (Lemma 3.1, p. 44 – original result due to Giné and Zinn [9], Lemma 2.7, pp. 936–937), we have, for $n\varepsilon_2{}^2 \geqslant 8$ (here, the measurable functions f are of the form $(\mathbf{1}_C - \mathbf{1}_{C^*})$),

$$\Pr\Big\{ \sup_{C\in\mathcal{B}(C^*,\lambda)} \big|Z_n(C)\big| > \varepsilon_2 \Big\} \leqslant 4\Pr\left\{ \sup_{C\in\mathcal{B}(C^*,\lambda)} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i (\mathbf{1}_C - \mathbf{1}_{C^*})(X_i) \right| > \frac{\varepsilon_2}{4} \right\}, \tag{26}$$

and, since the random variables

$$\sup_{C\in\mathcal{B}(C^*,\lambda)} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i (\mathbf{1}_C - \mathbf{1}_{C^*})(X_i) \right| \quad \text{and} \quad \sup_{C\in\mathcal{B}(C^*,\lambda)} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{C\Delta C^*}(X_i) \right|$$

have the same distribution, we have

$$\Pr\Big\{ \sup_{C\in\mathcal{B}(C^*,\lambda)} \big|Z_n(C)\big| > \varepsilon_2 \Big\} \leqslant 4\Pr\left\{ \sup_{C\in\mathcal{B}(C^*,\lambda)} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{C\Delta C^*}(X_i) \right| > \frac{\varepsilon_2}{4} \right\}. \tag{27}$$

4.2.2. Conditioning and decomposition using the median

We set some notations

$$X_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{C\Delta C^*}(x_i), \tag{28}$$

$$\|X_n\| = \sup_{C\in\mathcal{B}(C^*,\lambda)} |X_n|, \tag{29}$$

$$\sigma^2 = \sup_{C\in\mathcal{B}(C^*,\lambda)} \left(\frac{1}{n^2}\sum_{i=1}^n \mathbf{1}_{C\Delta C^*}(X_i) \right) \tag{30}$$

and $\Pr_X$, $\mathbb{E}_X$ (respectively $\Pr_\varepsilon$, $\mathbb{E}_\varepsilon$) are the conditional distributions and expectations given (ε_i) (respectively (x_i)).

We attempt to control the the following probability tail.

$$\Pr\left\{ \sup_{C\in\mathcal{B}(C^*,\lambda)} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{1}_{C\Delta C^*}(X_i) \right| > \frac{\varepsilon_2}{4} \right\} = \Pr\left\{ \|X_n\| > \frac{\varepsilon_2}{4} \right\}. \tag{31}$$

We fix $X_i = x_i$ for $i = 1, \ldots, n$ and we consider the conditional probability of the previous tail. We observe that this is the tail of the supremum norm over the set of indicator functions $\{\mathbf{1}_{C\Delta C^*};\ C \in \Gamma\}$ of a Rademacher process over the Banach space $\mathbb{R}^d$ of the form $X = \sum_{i=1}^n \varepsilon_i x_i$ which is known to be subgaussian. This fact is guaranteed by a concentration inequality from [13] on the deviations from the median. This result involves the median $M(X_n)$ of the process (conditionally on the (x_i)'s) which is defined by the following inequalities

$$\Pr_\varepsilon\big\{\|X\| > M_X\big\} \leqslant \frac{1}{2} \leqslant \Pr_\varepsilon\big\{\|X\| \geqslant M_X\big\}. \tag{32}$$

Then, using the right inequality, we obtain with the help of Markov inequality

$$M_X \leqslant 2\mathbb{E}_\varepsilon \|X\|. \tag{33}$$

According to this concentration result from [13] (Inequality 4.10, p. 100), we have that

$$\forall t > 0, \quad \Pr_\varepsilon\{\|X_n\| - M(X_n) > t\} \leqslant 2\mathrm{e}^{-t^2/8n^2\sigma^2}.$$

We can then write (computation follows essentially [21] recomposed in [22]), for all $u > 0$,

$$\Pr\left\{\|X_n\| > \frac{\varepsilon_2}{4}\right\} = \mathbb{E}_X \Pr_\varepsilon\left\{\|X_n\| > \frac{\varepsilon_2}{4}\right\} \tag{34}$$

$$\leqslant \mathbb{E}_X \Pr_\varepsilon\left\{\|X_n\| - M(X_n) > \frac{\varepsilon_2}{8}\right\} + \Pr_X\left\{M(X_n) > \frac{\varepsilon_2}{8}\right\} \tag{35}$$

$$\leqslant \mathbb{E}_X\left(2\exp\left\{-\frac{{\varepsilon_2}^2}{512\sigma^2}\right\}\right) + \Pr_X\left\{\mathbb{E}_\varepsilon\|X_n\| > \frac{\varepsilon_2}{16}\right\} \tag{36}$$

$$\leqslant 2\exp\left\{-\frac{{\varepsilon_2}^2}{512u}\right\} + \Pr_X\{\sigma^2 > u\} + \Pr_X\left\{\mathbb{E}_\varepsilon\|X_n\| > \frac{\varepsilon_2}{16}\right\}. \tag{37}$$

We adopt the following notation for this bound,

$$H = 2\exp\left\{-\frac{{\varepsilon_2}^2}{512u}\right\} + \Pr_X\{\sigma^2 > u\} + \Pr_X\left\{\mathbb{E}_\varepsilon\|X_n\| > \frac{\varepsilon_2}{16}\right\} = D + F + G. \tag{38}$$

4.2.3. Estimating the terms F and G

Now we use the following result from [22] (Lemma A.4.3, p. 455) in order to bound F and G in the previous inequality. Indeed, if S_n is a permutation-symmetric map such that

$$\text{(i)} \quad S_n(x) \leqslant S_{n+m}(x, y),$$

$$\text{(ii)} \quad S_{n+m}(x, y) \leqslant S_n(x) + S_m(y).$$

Then, we have, for any strictly positive t, and for any integer n,

$$\Pr\{S_n > t\} \leqslant \exp\left(-\frac{t}{2}\log\left(\frac{t}{4(2\mathbb{E}S_n + 1)}\right)\right). \tag{39}$$

Remark 4.4. – We have slightly modified the original result by keeping the term $4(2\mathbb{E}S_n + 1)$ in the denominator under the logarithm rather than $12(\mathbb{E}S_n \vee 1)$ which is given in [22].

We now apply this result to get the following upper bounds.

- On the one hand, by applying the lemma with $S_n(X) = n^2\sigma^2$, and $t := n^2u$, we get

$$F = \Pr_X\{\sigma^2 > u\} \leqslant \exp\left(-\frac{1}{2}n^2u\log\left(\frac{nu}{4(2n\mathbb{E}\sigma^2 + \frac{1}{n})}\right)\right). \tag{40}$$

- On the other hand, by setting $S_n(X) = n\mathbf{E}_\varepsilon \|X_n\|$ and $t := n\varepsilon_2/16$, we have [5]

$$G = \Pr_X\left\{\mathbb{E}_\varepsilon\|X_n\| > \frac{\varepsilon_2}{16}\right\} \leqslant \exp\left(-\frac{1}{2}\cdot\frac{n\varepsilon_2}{16}\log\left(\frac{\varepsilon_2}{64(2\mathbb{E}\|X_n\| + \frac{1}{n})}\right)\right). \quad (41)$$

Now we appeal to some result from [21] (Corollary 3.2) to bound $\mathbb{E}\sigma^2$. Applying this inequality for indicator functions of the form $f = \mathbf{1}_{C\Delta C^*}$ (we have $\mathbb{E}f = \mu(C\Delta C^*)$), we obtain

$$n\mathbb{E}\sigma^2 \leqslant \lambda + 2\mathbb{E}\|X_n\|. \quad (42)$$

Thus, we have

$$4\left(2n\mathbb{E}\sigma^2 + \frac{1}{n}\right) \leqslant 4\left(2\lambda + 4\mathbb{E}\|X_n\| + \frac{1}{n}\right), \quad (43)$$

and we can use the following bound for F.

$$\Pr_X\{\sigma^2 > u\} \leqslant \exp\left(-\frac{1}{2}n^2u\log\left(\frac{nu}{8(\lambda + 2\mathbb{E}\|X_n\| + \frac{1}{n})}\right)\right). \quad (44)$$

Let m_1, m_2 be such that

$$m_1 \geqslant \lambda + 2\mathbb{E}\|X_n\| + \frac{1}{n}, \quad (45)$$

$$m_2 \geqslant 2\mathbb{E}\|X_n\| + \frac{1}{n}. \quad (46)$$

We have obtained the following control of the tail

$$L = 2\exp\left\{-\frac{{\varepsilon_2}^2}{512u}\right\} + \exp\left\{-\frac{1}{2}n^2u\log\left(\frac{nu}{8m_1}\right)\right\} + \exp\left\{-\frac{1}{2}\frac{n\varepsilon_2}{16}\log\left(\frac{\varepsilon_2}{64m_2}\right)\right\}. \quad (47)$$

4.2.4. Bounding $\mathbb{E}\|X_n\|$

To get an estimate for m_1, m_2, we need to compute $\mathbb{E}\|X_n\|$. We recall the technical result obtained by Talagrand [21] (Proposition 6.2). If we set

$$\lambda = \sup_{C\in\mathcal{B}(C^*,\lambda)} \mu(C\Delta C^*),$$

then

$$\mathbb{E}\|X_n\| \leqslant \frac{K}{\sqrt{n}}\left(\left(\lambda + K^2\cdot\frac{v}{n}\log\frac{U}{4\lambda}\right)v\log\frac{U}{4\lambda}\right)^{1/2}, \quad (48)$$

where K, U, and v are some constants.

As we have set $\lambda = 1/(n\varepsilon^2)$, we get

$$\mathbb{E}\|X_n\| \leqslant \frac{K}{n\varepsilon}\log(n\varepsilon^2). \quad (49)$$

[5] Note that $\mathbb{E}S_n = n\mathbb{E}_X\mathbb{E}_\varepsilon\|X_n\| = n\mathbb{E}\|X_n\|$.

Hence, we can take

$$m_1 = \frac{K}{n\varepsilon^2}\log\bigl(n\varepsilon^2\bigr), \tag{50}$$

$$m_2 = \frac{K}{n\varepsilon}\log\bigl(n\varepsilon^2\bigr). \tag{51}$$

4.2.5. Optimization

Now, we adjust the various parameters in such a way that each exponential term in the bound becomes smaller than $\exp\{-n(1-\beta)\Lambda_p(\varepsilon)\}$. We set

$$\frac{(\theta\varepsilon)^2}{512u} \geqslant n(1-\beta)\Lambda_p(\varepsilon), \tag{52}$$

$$\frac{n^2u}{2}\log\left(\frac{nu}{64m_1}\right) \geqslant n(1-\beta)\Lambda_p(\varepsilon), \tag{53}$$

$$\frac{n\theta\varepsilon}{32}\log\left(\frac{\theta\varepsilon}{64m_2}\right) \geqslant n(1-\beta)\Lambda_p(\varepsilon). \tag{54}$$

From the first two conditions (52) and (53), we can get a single inequality and there is no need to provide the proper choice for u, but only guarantee that such a choice is possible. We have

$$\frac{\beta^2}{2048(1-\beta)}\frac{\varepsilon^2}{\Lambda_p(\varepsilon)} \geqslant nu, \tag{55}$$

$$\log\left(\frac{nu}{64K}\frac{n\varepsilon^2}{\log(n\varepsilon^2)}\right) \geqslant 2(1-\beta)\Lambda_p(\varepsilon). \tag{56}$$

Using the fact that the term $\frac{\varepsilon^2}{\Lambda_p(\varepsilon)}$ can assumed to be constant (possibly depending on p) and taking into account the fact that we have, up to a constant,

$$x \geqslant A\log A \quad \Rightarrow \quad \frac{x}{\log x} \geqslant A,$$

we deduce a condition of the form

$$n\varepsilon^2 \geqslant M_1(\beta, p, V) = \frac{K}{\beta^4}\exp\left\{\frac{C}{\beta^2}\right\},$$

where K and C are some constants.

Now the third condition (54) leads with a similar argument to

$$n\varepsilon^2 \geqslant M_2(\beta, p, V) = \frac{K'}{\beta^2}\exp\left\{\frac{C'}{\beta}\right\},$$

with K', C' being some constants.

5. Combinatorial method

We now turn to the proof of Theorem 2.1. We introduce a new notation for the empirical mean which makes a more explicit reference to the sample,

$$\mu\big(C, U_1^n\big) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_C(U_i). \tag{57}$$

We introduce a "ghost" i.i.d. sample $Y_1, \ldots, Y_m$ from the same probability distribution μ, and we use the symmetrization lemma due to Devroye [5] which states that, for any $\varepsilon > 0$, $m \geqslant 1$ and $0 < \alpha < 1$, we have

$$\begin{aligned}&\Pr\Big\{\sup_{C\in\Gamma}\big|\mu\big(C, X_1^n\big) - \mu(C)\big| > \varepsilon\Big\}\\ &\quad\leqslant \left(1 - \frac{1}{4m\alpha^2\varepsilon^2}\right)^{-1} \Pr\Big\{\sup_{C\in\Gamma}\big|\mu\big(C, X_1^n\big) - \mu\big(C, Y_1^m\big)\big| > (1-\alpha)\varepsilon\Big\}.\end{aligned} \tag{58}$$

We set $\varepsilon^* = (1-\alpha)\varepsilon$ and

$$\tau(n, m, \varepsilon, \mu) = \Pr\Big\{\sup_{C\in\Gamma}\big|\mu\big(C, X_1^n\big) - \mu\big(C, Y_1^m\big)\big| > \varepsilon^*\Big\}. \tag{59}$$

With these notations, we rephrase the previous lemma in the following form by

$$\rho(\Gamma, \mu, n, \varepsilon) \leqslant \left(1 - \frac{1}{4m\alpha^2\varepsilon^2}\right)^{-1} \tau(n, m, \varepsilon, \mu). \tag{60}$$

We introduce the notation: $N = n + m$. Now, we work on the symmetrized probability tail.

$$\tau(n, m, \varepsilon, \mu) = \int \mathbf{1}_{\{\sup_{C\in\Gamma}|\mu(C,X_1^n)-\mu(C,Y_1^m)|>\varepsilon^*\}}\, d\mu^{\otimes N}\big(X_1^n \times Y_1^m\big) \tag{61}$$

$$= \int \sup_{C\in\Gamma} \mathbf{1}_{\{|\mu(C,X_1^n)-\mu(C,Y_1^m)|>\varepsilon^*\}}\, d\mu^{\otimes N}\big(X_1^n \times Y_1^m\big). \tag{62}$$

Following the line of proof of [26], we consider $\mathcal{T} = \{T_i\}_{i=1,\ldots,N!}$ the set of all permutations of the set $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ and we notice that there is a finite number $\mathcal{N}(\Gamma, X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ of equivalence classes of sets in Γ achieving different values for the quantity $|\mu(C, X_1^n) - \mu(C, Y_1^m)|$. Hence, it is possible to replace the supremum over Γ by a supremum over an approximation Γ^* of the family Γ, and Γ^* has a finite number of elements equal to $\mathcal{N}(\Gamma, X_1, \ldots, X_n, Y_1, \ldots, Y_m)$. Thus, we can write

$$\begin{aligned}&\tau(n, m, \varepsilon, \mu)\\ &\quad\leqslant \int \sum_{C^*\in\Gamma^*} \frac{1}{N!}\sum_{i=1}^{N!} \mathbf{1}_{\{|\mu(C^*,T_iX_1^n)-\mu(C^*,T_iY_1^m)|>\varepsilon^*\}}\, d\mu^{\otimes N}\big(X_1^n \times Y_1^m\big)\end{aligned} \tag{63}$$

$$= \int \sum_{C^*\in\Gamma^*} \Pr\Big\{\big|\mu\big(C^*, X'^n_1\big) - \mu\big(C^*, Y'^m_1\big)\big| > \varepsilon^*\Big\}\, d\mu^{\otimes N}\big(X_1^n \times Y_1^m\big), \tag{64}$$

where $X'^n_1 \times Y'^m_1$ is obtained by a sampling without replacement draw from $X_1^n \times Y_1^m$. We bound each term of the sum separately thanks to the following proposition. Thus, we will consider that the element C^* is fixed.

PROPOSITION 5.1. – *Let $X_1^n \times Y_1^m$ be a fixed sample. For a fixed C^*, we set*

$$r = r(C^*) = \sum_{l=1}^{n} \mathbf{1}_{C^*}(X_l) + \sum_{l=1}^{m} \mathbf{1}_{C^*}(Y_l). \tag{65}$$

We denote by $X'^n_1 \times Y'^m_1$ the random variables obtained through sampling without replacement from the original sample $X_1^n \times Y_1^m$. We have

$$\begin{aligned} &\Pr\left\{\left|\mu\left(C^*, X'^n_1\right) - \mu\left(C^*, Y'^m_1\right)\right| > \varepsilon^* \mid X_1^n \times Y_1^m\right\} \\ &\quad \leqslant 2(n+m+1)^7 \exp\left\{-n\Lambda_{\frac{r}{n+m}}\left(\left(\frac{m}{n+m}\right)\varepsilon^*\right)\right\}. \end{aligned} \tag{66}$$

Proof. – We notice that the actual distribution of the random variable r is binomial with parameters $(N, \mu(C^*))$. Now we consider that the sample $X_1^n \times Y_1^m$ is fixed, so that all the probabilities involved in this proof are conditional probabilities given the sample. Suppose that $\mu(C^*, X'^n_1) = \frac{k}{n}$, then we have $\mu(C^*, Y'^m_1) = \frac{r-k}{m}$. Thus,

$$\mu\left(C^*, X'^n_1\right) - \mu\left(C^*, Y'^m_1\right) = \frac{k}{n} - \frac{r-k}{m} = \frac{N}{m}\left(\frac{k}{n} - \frac{r}{N}\right),$$

and $\frac{r}{N} = \mu(C^*, X'^n_1 \times Y'^m_1)$. Hence, we have

$$\begin{aligned} &\Pr\left\{\left|\mu\left(C^*, X'^n_1\right) - \mu\left(C^*, Y'^m_1\right)\right| > \varepsilon^*\right\} \\ &\quad = \Pr\left\{\left|\mu\left(C^*, X'^n_1\right) - \mu\left(C^*, X'^n_1 \times Y'^m_1\right)\right| > \left(\frac{m}{N}\right)\varepsilon^*\right\}, \end{aligned} \tag{67}$$

where $\mu(C^*, X'^n_1 \times Y'^m_1) = \frac{r}{N}$ is non-random if C^* is fixed as well as the sample $X_1^n \times Y_1^m$ obtained from the distribution μ. Now we have to state an upper bound of Cramér–Chernoff type for the probability

$$\Pr\left\{\left|\mu\left(C^*, X'^n_1\right) - \frac{r}{N}\right| > \left(\frac{m}{N}\right)\varepsilon^*\right\}, \tag{68}$$

where X'^n_1 is a sampling without replacement draw from the set of points $X_1^n \times Y_1^m$. We write

$$\Pr\left\{\left|\mu\left(C^*, X'^n_1\right) - \frac{r}{N}\right| > \left(\frac{m}{N}\right)\varepsilon^*\right\} = \sum_{k:\ |\frac{k}{n} - \frac{r}{N}| > (\frac{m}{N})\varepsilon^*} \frac{C_r^k C_{N-r}^{n-k}}{C_N^n}. \tag{69}$$

Then, by Stirling's formula, one gets straightforwardly [6]

$$\frac{1}{n}\log\left(\frac{C_r^k C_{N-r}^{n-k}}{C_N^n}\right) \leqslant -\left\{H\left(\frac{k}{n}, \frac{r}{N}\right) + \frac{m}{n}H\left(\frac{r-k}{m}, \frac{r}{N}\right) + \frac{6}{n}\log(N+1)\right\}, \tag{70}$$

[6] One could also check some neat large deviations formulations for sampling without replacement for a binary alphabet in [4] pp. 20–22 and pp. 318–323.

where $H(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$.

We recall that the function $x \to H(x,q)$ is decreasing for $x < q$ and increasing for $x > q$. Let us consider k such that $\frac{k}{n} - \frac{r}{N} > (\frac{m}{N})\varepsilon^*$. We notice that, in this case, $k \to H(\frac{k}{n}, \frac{r}{N})$ increases, and we have

$$H\left(\frac{k}{n}, \frac{r}{N}\right) > H\left(\frac{r}{N} + \frac{m}{N}\varepsilon^*, \frac{r}{N}\right). \tag{71}$$

Moreover, we have that $\frac{r-k}{m} < \frac{r}{N}$ and $\to H(\frac{r-k}{m}, \frac{r}{N})$ is also increasing in k. Thus,

$$H\left(\frac{r-k}{m}, \frac{r}{N}\right) > H\left(\frac{r}{N} - \frac{n}{N}\varepsilon^*, \frac{r}{N}\right). \tag{72}$$

We then have, by brutally bounding the second term under the exponential (by zero!),

$$\sum_{k:\ \frac{k}{n} - \frac{r}{N} > (\frac{m}{N})\varepsilon^*} \frac{C_r^k C_{N-r}^{n-k}}{C_N^n} \leqslant r(N+1)^6 \exp\left\{-n\left(H\left(\frac{r}{N} + \frac{m}{N}\varepsilon^*, \frac{r}{N}\right) + \frac{m}{n}H\left(\frac{r}{N} - \frac{n}{N}\varepsilon^*, \frac{r}{N}\right)\right)\right\} \tag{73}$$

$$\leqslant r(N+1)^6 \exp\left\{-nH\left(\frac{r}{N} + \frac{m}{N}\varepsilon^*, \frac{r}{N}\right)\right\}. \tag{74}$$

Similarly, for k such that $\frac{k}{n} - \frac{r}{N} < -(\frac{m}{N})\varepsilon^*$, we obtain

$$\sum_{k:\ \frac{k}{n} - \frac{r}{N} < -(\frac{m}{N})\varepsilon^*} \frac{C_r^k C_{N-r}^{n-k}}{C_N^n} \leqslant r(N+1)^6 \exp\left\{-nH\left(\frac{r}{N} - \frac{m}{N}\varepsilon^*, \frac{r}{N}\right)\right\}. \tag{75}$$

We finally use the fact that $r \leqslant N$ to end the proof. □

In the sequel, we will consider only the case of a critical value p smaller than $1/2$ (this is indeed just a matter of notations since $\Lambda_p = \Lambda_{1-p}$). The other part shall be treated in the same way. Now consider the random variable r which has a binomial distribution with parameters $(N, \mu(C^*))$. Its rate function is known to be $x \to H(x, \mu(C^*))$.

Intuitive argument. Suppose we can directly proceed as in the proof of Varadhan's lemma (see e.g. in [4] the remark p. 137). Then, we would have the following upper bound on $\tau(n,m,\varepsilon,\mu)$ (which is not rigorously correct).

$$\int_0^1 \sum_{C^* \in \Gamma^*} 2(N+1)^7 \exp\left\{-n\,\Lambda_u\left(\frac{m}{N}\varepsilon^*\right) - NH\left(u, \mu(C^*)\right)\right\} du.$$

Intuitively, we can see that it suffices to show that the value of the integral is given essentially on the neighborhoods around $\mu(C^*)$ for each C^*. Then taking N large enough compared to n shall end the proof.

Rigorous argument. Let us go through the details. In the sequel, x will denote a point in the product space $(\mathbb{R}^d)^N$ (a sample of N points of $\mathbb{R}^d$). We start from the bound on $\tau(n, m, \varepsilon, \mu)$,

$$2\int \sum_{C\in\Gamma^*(x)} \phi\big(r(C)\big)\, d\mu^{\otimes N}(x) \tag{76}$$

where

$$\phi(r) := \phi(r, m, n, \varepsilon^*) = (N+1)^7 \exp\left\{-n\Lambda_{\frac{r}{N}}\left(\frac{m}{N}\varepsilon^*\right)\right\}.$$

First, we fix δ such that $p + 4\delta < 1/2$, and we decompose, for any x, the finite family $\Gamma^*(x)$,

$$\Gamma^*(x) = \Gamma_1^*(x) \cup \Gamma_2^*(x) \cup \Gamma_3^*(x) \tag{77}$$

where

$$\Gamma_1^*(x) = \left\{C \in \Gamma^*(x):\ \frac{r(C)}{N} \leqslant p + 4\delta\right\},$$

$$\Gamma_2^*(x) = \left\{C \in \Gamma^*(x):\ p + 4\delta < \frac{r(C)}{N} < 1 - p - 4\delta\right\},$$

$$\Gamma_3^*(x) = \left\{C \in \Gamma^*(x):\ \frac{r(C)}{N} \geqslant 1 - p - 4\delta\right\}.$$

We have that, for $r/N \leqslant p + 4\delta < 1/2$, ϕ is non-decreasing, and for $r/N \geqslant 1 - p - 4\delta > 1/2$, ϕ is non-increasing. Therefore, we obtain the following inequalities

$$\forall C \in \Gamma_1^*, \quad \phi\big(r(C)\big) \leqslant (N+1)^7 \exp\left\{-nH\left(p + 4\delta + \left(\frac{m}{N}\right)\varepsilon^*, p + 4\delta\right)\right\}, \tag{78}$$

$$\forall C \in \Gamma_3^*, \quad \phi\big(r(C)\big) \leqslant (N+1)^7 \times \exp\left\{-nH\left(1 - p - 4\delta - \left(\frac{m}{N}\right)\varepsilon^*, 1 - p - 4\delta\right)\right\}, \tag{79}$$

and then we can bound uniformly the corresponding parts of the sum using the fact that $|\Gamma^*(x)| = \mathcal{N}(\Gamma, X_1, \ldots, X_n, Y_1, \ldots, Y_m) \leqslant s(\Gamma, N)$. Using the symmetry properties of the function H, we finally obtain

$$\sum_{C\in\Gamma_1^*(x)\cup\Gamma_3^*(x)} \phi(r(C)) \leqslant s(\Gamma, N)(N+1)^7 \exp\left\{-nH\left(p + \delta + \left(\frac{m}{N}\right)\varepsilon^*, p + \delta\right)\right\}. \tag{80}$$

For the remaining sets (in Γ_2^*), we detect the "worst set" which shall be denoted by

$$\widetilde{C}(x) = \operatorname*{arg\,max}_{C\in\Gamma_2^*(x)} \phi\left(\frac{r(C)}{N}\right) \tag{81}$$

and we use a uniform bound for the sum

$$\sum_{C\in\Gamma_2^*(x)} \phi\big(r(C)\big) \leqslant \big|\Gamma_2^*(x)\big|\phi\left(\frac{r(\widetilde{C}(x))}{N}\right). \tag{82}$$

Now we introduce the event

$$\Omega = \left\{ x\colon\ p + 4\delta < \frac{r(\widetilde{C}(x))}{N} < 1 - p - 4\delta \right\} \tag{83}$$

and we attempt to control its probability by applying Varadhan's lemma uniformly over the set Ω. We recall indeed that the rate function of the random variable $\frac{r(C)}{N}$ is $u \to H(u, \mu(C))$. However, we cannot apply Varadhan's lemma straightforwardly on the random variable $\frac{r(\widetilde{C}(x))}{N}$ because its expected value $\mu(\widetilde{C}(x))$ depends on the sample x. First, we notice that

$$\left|\Gamma_2^*(x)\right| \phi\left(\frac{r(\widetilde{C}(x))}{N} \right) \leqslant s(\Gamma, N)(N+1)^7, \tag{84}$$

and we then focus on the estimation of the integral $\int \mathbf{1}_\Omega(x)\, d\mu^{\otimes N}(x)$.

Estimation of $\int \mathbf{1}_\Omega(x)\, d\mu^{\otimes N}(x)$. We first introduce a finite, and fixed, δ-approximation $\widetilde{\Gamma}$ of Γ such that $\widetilde{\Gamma} = \{\widetilde{C}_1, \ldots, \widetilde{C}_I\}$, and

$$\forall C \in \Gamma, \quad \exists\, i\colon\ \mu(C \Delta \widetilde{C}_i) < \delta.$$

From inequality (24), we have that

$$I = |\widetilde{\Gamma}| \sim \left(\frac{1}{\delta}\right)^V. \tag{85}$$

We introduce the sets

$$A_i = \left\{ x\colon\ \mu\left(\widetilde{C}(x) \Delta \widetilde{C}_i\right) < \delta \right\}, \tag{86}$$

$$K_i = \left\{ x\colon\ \left| \frac{r(\widetilde{C}(x))}{N} - \frac{r(\widetilde{C}_i)}{N} \right| < 2\delta \right\}, \tag{87}$$

and we use the following decomposition

$$\left(\mathbb{R}^d\right)^N \subset \bigcup_{i=1}^{I} A_i = \bigcup_{i=1}^{I} \left((A_i \cap K_i) \cup (A_i \cap \overline{K}_i) \right), \tag{88}$$

which holds because Γ is assumed to be a totally bounded family.

On the one hand, we have

$$\int\limits_{A_i \cap K_i} \mathbf{1}_\Omega(x)\, d\mu^{\otimes N}(x) \leqslant \int \mathbf{1}_{\{p+2\delta < \frac{r(\widetilde{C}_i)}{N} < 1-p-2\delta\}}(x)\, d\mu^{\otimes N}(x), \tag{89}$$

and this last integral can be controlled thanks to Chernoff's inequality. We have indeed

$$\int \mathbf{1}_{\{p+2\delta < \frac{r(\widetilde{C}_i)}{N} < 1-p-2\delta\}}(x)\, d\mu^{\otimes N}(x) = \Pr\left\{ p + 2\delta < \frac{r(\widetilde{C}_i)}{N} < 1 - p - 2\delta \right\}. \tag{90}$$

Then, we have, if $\mu(\widetilde{C}_i) < p$,

$$\Pr\left\{ p + 2\delta < \frac{r(\widetilde{C}_i)}{N} < 1 - p - 2\delta \right\} \leqslant \Pr\left\{ \frac{r(\widetilde{C}_i)}{N} > p + 2\delta \right\}, \tag{91}$$

$$\leqslant \exp\{-NH(p + 2\delta, \mu(\widetilde{C}_i))\}, \tag{92}$$

$$\leqslant \exp\{-NH(p + 2\delta, p)\}, \tag{93}$$

where the inequality (92) is the straightforward application of Chernoff inequality (6), and the inequality (93) is due to the fact that the function $q \to \exp\{-NH(u, q)\}$ is non-decreasing for $u > q$.

In a similar way, if $\mu(\widetilde{C}_i) > 1 - p$, we obtain

$$\Pr\left\{ p + 2\delta < \frac{r(\widetilde{C}_i)}{N} < 1 - p - 2\delta \right\} \leqslant \exp\{-NH(1 - p - 2\delta, 1 - p)\}, \tag{94}$$

thanks to the monotonicity of the function $q \to \exp\{-NH(u, q)\}$, which is non-increasing for $u > q$, and also to the second Chernoff inequality (7).

Hence, we have obtained the following bound for any index i (because $H(x + y, x) = H(1 - x - y, 1 - x)$),

$$\int\limits_{A_i \cap K_i} \mathbf{1}_\Omega(x)\, d\mu^{\otimes N}(x) \leqslant \exp\{-NH(p + 2\delta, p)\}. \tag{95}$$

On the other hand, we have

$$\int\limits_{A_i \cap \overline{K}_i} \mathbf{1}_\Omega(x)\, d\mu^{\otimes N}(x) \leqslant \mu^{\otimes N}(A_i \cap \overline{K}_i). \tag{96}$$

We introduce the following notation

$$(Z_1, \ldots, Z_N) := x = (X_1, \ldots, X_n, Y_1, \ldots, Y_m),$$

and we notice that

$$\left| r\big(\widetilde{C}(x)\big) - r(\widetilde{C}_i) \right| \leqslant \sum_{k=1}^{N} \left| (\mathbf{1}_{\widetilde{C}(x)} - \mathbf{1}_{\widetilde{C}_i})(Z_k) \right| \tag{97}$$

$$= \sum_{k=1}^{N} \mathbf{1}_{\widetilde{C}(x) \Delta \widetilde{C}_i}(Z_k). \tag{98}$$

Hence,

$$x \in A_i \cap \overline{K}_i \quad \Rightarrow \quad \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\widetilde{C}(x) \Delta \widetilde{C}_i}(Z_k) - \mu\big(\widetilde{C}(x) \Delta \widetilde{C}_i\big) \geqslant \delta \quad \text{and} \quad \mu\big(\widetilde{C}(x) \Delta \widetilde{C}_i\big) < \delta$$

$$\Rightarrow \quad \sup_{C \in \mathcal{B}(\widetilde{C}_i, \delta)} \left(\frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{C \Delta \widetilde{C}_i}(Z_k) - \mu(C \Delta \widetilde{C}_i) \right) \geqslant \delta,$$

where $\mathcal{B}(\widetilde{C}_i, \delta) = \{C\colon\ \mu(C \Delta \widetilde{C}_i) < \delta\}$. Thus, we have

$$\mu^{\otimes N}(A_i \cap \overline{K}_i) \leqslant \Pr\left\{ \sup_{C \in \mathcal{B}(\widetilde{C}_i, \delta)} \left(\frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{C \Delta \widetilde{C}_i}(Z_k) - \mu(C \Delta \widetilde{C}_i) \right) > \delta \right\}. \tag{99}$$

At this point, we use the proposition proved with the approximation method in the previous section (Proposition 4.1) which states that, for $N\delta^2$ large enough, there exists a constant K and a corrective term β such that

$$\begin{aligned} &\Pr\left\{ \sup_{C \in \mathcal{B}(\widetilde{C}_i, \lambda)} \left(\frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{C \Delta \widetilde{C}_i}(Z_k) - \mu(C \Delta \widetilde{C}_i) \right) > \delta \right\} \\ &\quad \leqslant K \exp\bigl\{ -N(1-\beta) H(2\delta, \delta) \bigr\}. \end{aligned} \tag{100}$$

We can fix $\beta = 1/2$. We have proved that, for $N\delta^2$ large enough, the following bound holds,

$$\int_{A_i \cap \overline{K}_i} \mathbf{1}_{\Omega}(x)\, d\mu^{\otimes N}(x) \leqslant K \exp\left\{ -\frac{N}{2} H(2\delta, \delta) \right\}. \tag{101}$$

Thus, we have obtained a uniform version of Varadhan's lemma, if $N\delta^2$ is large enough, for the integral

$$\begin{aligned} &\int \mathbf{1}_{\Omega}(x)\, d\mu^{\otimes N}(x) \\ &\quad \leqslant K \left(\left(\frac{1}{\delta} \right)^{V} \exp\bigl\{ -N H(p + 2\delta, p) \bigr\} + \exp\left\{ -\frac{N}{2} H(2\delta, \delta) \right\} \right). \end{aligned} \tag{102}$$

We will now show that the ratio

$$\frac{H(p + 2\delta, p)}{\frac{1}{2} H(2\delta, \delta)} \tag{103}$$

is smaller than 1 for δ small enough. We will need a simple inequality by Hoeffding [11]. Indeed, if $x < 1/2$, we have

$$H(x + y, x) \geqslant \frac{1}{1 - 2x} \log\left(\frac{1 - x}{x} \right) y^2. \tag{104}$$

Moreover, we have, for fixed x, when y tends to zero,

$$H(x + y, x) \sim \frac{y^2}{2x(1 - x)}. \tag{105}$$

Hence, as δ comes closer to zero, we have

$$\frac{H(p + 2\delta, p)}{\frac{1}{2} H(2\delta, \delta)} \leqslant \frac{\frac{4\delta^2}{2p(1-p)}}{\frac{1}{2} \frac{1}{1-2\delta} \log(\frac{1-\delta}{\delta}) \delta^2} \sim \frac{4}{\log(1/\delta)} < 1. \tag{106}$$

We can then neglect the second term at the cost of increasing the multiplicative constant K. Thus, we shall use the bound

$$\int \mathbf{1}_{\Omega}(x)\, d\mu^{\otimes N}(x) \leqslant K\left(\frac{1}{\delta}\right)^{V} \exp\{-NH(p+2\delta,\, p)\}. \tag{107}$$

Global bound on $\tau(n, m, \varepsilon, \mu)$. Hence, we have obtained an explicit bound on $\tau(n, m, \varepsilon, \mu)$ for δ small enough, which is the following

$$\tau(n, m, \varepsilon, \mu) \leqslant Ks(\Gamma, N)(N+1)^{7}\left(\exp\{-nH_1\} + \left(\frac{1}{\delta}\right)^{V} \exp\{-NH_2\}\right), \tag{108}$$

where we have used the following notations

$$H_1 = H\left(p + 4\delta + \frac{m}{N}\varepsilon^*,\, p + 4\delta\right), \tag{109}$$

$$H_2 = H(p + 2\delta,\, p). \tag{110}$$

Now consider the following functions,

$$f_x(q) = H(q + x, q) \quad (x \text{ fixed}), \tag{111}$$

$$g_p(x) = H(p + x,\, p) \quad (p \text{ fixed}). \tag{112}$$

Thanks to the convexity of both functions f_x and g_p, we have

$$f_x(p + 4\delta) \geqslant f_x(p) + 4\delta f_x'(p), \tag{113}$$

$$g_p(2\delta) \geqslant c(p)\delta^2, \tag{114}$$

where $c(q)$ is some constant depending on q. Hence, if we set $x = \frac{m}{N}\varepsilon^*$, we have

$$\begin{aligned}\exp\{-nH_1\} &= \exp\{-nf_x(p + 4\delta)\} \\ &= \exp\left\{-nH\left(p + 4\delta + \frac{m}{N}\varepsilon^*,\, p + 4\delta\right)\right\} \qquad (115)\\ &\leqslant \exp\left\{-nH\left(p + \frac{m}{N}\varepsilon^*,\, p\right)\right\} \cdot \exp\{-n4\delta f_x'(p)\}. \qquad (116)\end{aligned}$$

Thus, δ should be at most of the order $\frac{1}{n}$ since $f_x'(p)$ is usually negative (and bounded since $m \sim N$) and behaves like ε^2. Moreover, to control the second term, we shall take N such that

$$NH_2 \geqslant nH_1. \tag{117}$$

Thus, we impose N to satisfy

$$Ng_p(2\delta) \geqslant nH\left(p + \frac{m}{N}\varepsilon^*,\, p\right), \tag{118}$$

while we are assuming that δ is of the order $\frac{1}{n}$. Hence, we can choose δ and N verifying

$$Nc(p)\delta^2 \geqslant nH\left(p + \frac{m}{N}\varepsilon^*,\, p\right), \tag{119}$$

which leads to a δ of the order $\sqrt{\frac{n}{N}}$. Eventually, when δ goes to zero as $\frac{1}{n}$, choosing $N = n^3$ gives a bound on $\tau(n, m, \varepsilon, \mu)$, for some constant K, like

$$\tau(n, m, \varepsilon, \mu) \leqslant K n^{3V+21} \exp\{-n\, H(p+\varepsilon, p)\}, \tag{120}$$

where we have used Sauer's lemma [7] to bound $s(\Gamma, N)$.

To obtain, the global bound on $\rho(\Gamma, n, m, \varepsilon, \mu)$, we set $\alpha = \frac{1}{n}$ in the inequality (60).

Remark 5.2. – Note that we have obtained an exponential rate in $H(p+\varepsilon, p)$ because we assumed that the critical value p is smaller than $1/2$. If we had taken $p > 1/2$, we would have found a rate in $H(p-\varepsilon, p)$. The exponential rate $\Lambda_p(\varepsilon)$ given in the formulation of Theorem 2.1 covers both cases.

6. Open issues

The result presented in this paper certainly appeals to several improvements. Talagrand has suggested in [21] that the proof technique used to prove his universal bound could be adapted in order to lead to tight distribution-dependent results (see [21], p. 63, last paragraph of Section 6). Indeed, the issue is to obtain the same exponential rate $\phi(\varepsilon) = \Lambda_p(\varepsilon)$ as in Theorem 2.1 with a fairly tight capacity term (like $\tau = V - 1/2$ instead of our $\tau = 3V + 21$).

There are other related issues which could be explored in the same spirit.

- Formulate and prove similar bounds in a functional setting.
- Compute tight bounds on the expected value of the maximal deviation, which is a question of growing interest since the impressive recent results on concentration inequalities (see [3,20], and their references).

Moreover, we point out that theoretical analysis on VC bounds and VC dimension could benefit of an empirical study. Indeed, we have proposed in [29] to use computer simulations to estimate the probability of the event $\{\sup_{C\in\Gamma} |\mu_n(C) - \mu(C)| > \varepsilon\}$ for particular distributions μ. Through this experimental approach, there are several conjectures which can be tested on particular examples.

- Validation of the general structure of VC bounds, and control of asymptotical corrections (existence of polynomial terms smaller than $(n\varepsilon^2)^{\tau}$).
- Numerical values for the multiplicative constant K and the capacity index τ.
- Test of the relationship between the index τ and the VC dimension V (for instance, we have checked, that, for halfspaces, the formula $\tau = V - 1$ holds true).
- Dependence of the effective VC dimension on the underlying distribution μ (on a simple example, we have observed that the estimated values of the VC dimension V for a fixed family Γ depend strongly on the distribution μ).

We find this experimental work very stimulating and complementary to the theoretical analysis. Indeed, the simulation part appears as a very promising means to develop intuition and state conjectures about issues such as the distribution-sensitivity of combinatorial capacities.

[7] This combinatorial result gives a polynomial bound on the shattering coefficient in case of a finite VC dimension: $s(\Gamma, N) \leqslant (\mathrm{e}N/V)^V$.

Acknowledgements

The origin of this work is due to Professor Robert Azencott. For his guidance and support, the author is deeply grateful to him. The author also thanks Professors Gábor Lugosi and Alain Trouvé for their meticulous reading of the previous version of this paper (in [29]) and their insightful comments.

REFERENCES

[1] K. Alexander, Probability inequalities for empirical processes and a law of the iterated logarithm, Ann. Probab. 4 (1984) 1041–1067.
[2] R. Azencott, Communication for the fifty years of the Department of Mathematics at Brown University (USA), 1996.
[3] S. Boucheron, G. Lugosi, P. Massart, A sharp concentration inequality with applications, Random Structures and Algorithms 16 (3) (2000) 277–292.
[4] A. Dembo, O. Zeitouni, Large Deviations Techniques and Applications, 2nd Edition, Springer, 1998.
[5] L. Devroye, Bounds for the uniform deviation of empirical measures, J. Multivariate Anal. 12 (1982) 72–79.
[6] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, 1996.
[7] R.M. Dudley, A course on empirical processes, in: P.L. Hennequin (Ed.), Ecole d'Eté de Probabilités de Saint-Flour XII – 1982, in: Lecture Notes in Mathematics, Vol. 1097, Springer-Verlag, 1982, pp. 1–142.
[8] A. Dvoretzky, J. Kiefer, J. Wolfowitz, Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator, Ann. Math. Statist. 27 (1956) 642–669.
[9] E. Giné, J. Zinn, On the central limit theorem for empirical processes, Ann. Probab. 12 (1984) 929–989.
[10] D. Haussler, Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension, J. Combin. Theory Series A 69 (1995) 217–232.
[11] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.
[12] J. Kiefer, On large deviations of the empirical d.f. of vector chance variables and a law of the iterated logarithm, Pacific J. Math. 11 (1961) 649–660.
[13] M. Ledoux, M. Talagrand, Probability in Banach Spaces, Springer-Verlag, 1992.
[14] G. Lugosi, Improved upper bounds for probabilities of uniform deviations, Statist. Probab. Lett. 25 (1995) 71–77.
[15] P. Massart, Rates of convergence in the central limit theorem for empirical processes, Annales de l'Institut Henri Poincaré 22 (4) (1986) 381–423.
[16] P. Massart, The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality, Ann. Probab. 18 (1990) 1269–1283.
[17] J.M.R. Parrondo, C. van den Broeck, Vapnik–Chervonenkis bounds for generalization, J. Phys. A: Math. Gen. 26 (1993) 2211–2223.
[18] D. Pollard, Convergence of Stochastic Processes, Springer-Verlag, 1984.
[19] D. Pollard, Empirical Processes: Theory and Applications, in: NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2, Institute of Mathematical Statistics, 1991.

[20] E. Rio, Inégalités de concentration pour les processus empiriques de classes de parties, Probab. Theory Related Fields (2000), to appear.
[21] M. Talagrand, Sharper bounds for Gaussian and empirical processes, Ann. Probab. 22 (1) (1994) 28–76.
[22] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, 1996.
[23] V.N. Vapnik, Estimation of Dependencies on the Basis of Empirical Data, Springer, 1982.
[24] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[25] V.N. Vapnik, Statistical Learning Theory, Wiley-Interscience, 1998.
[26] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (1971) 264–280.
[27] V.N. Vapnik, A.Y. Chervonenkis, Necessary and sufficient conditions for the uniform convergence of the means to their expectations, Theory Probab. Appl. 26 (1981) 532–555.
[28] V.N. Vapnik, E. Levin, Y. Le Cun, Measuring the VC-dimension of a learning machine, Neural Comput. 6 (1994) 851–876.
[29] N. Vayatis, Inégalités de Vapnik–Chervonenkis et mesures de complexité, Ph.D. thesis, Ecole Polytechnique, 2000, in English.
[30] L. Wu, Large deviations, moderate deviations and LIL for empirical processes, Ann. Probab. 22 (1) (1994) 17–27.