Elect. Comm. in Probab. **7** (2002) 55–65

ELECTRONIC COMMUNICATIONS in PROBABILITY

SOME EXTENSIONS OF AN INEQUALITY OF VAPNIK AND CHERVONENKIS

DMITRIY PANCHENKO[1]
Department of Mathematics and Statistics, The University of New Mexico
United States of America
email: `panchenk@math.unm.edu`

submitted August 21, 2001 *Final version accepted* January 17, 2002

AMS 2000 Subject classification: 60E15, 28A35
Concentration of measure, empirical processes

Abstract
The inequality of Vapnik and Chervonenkis controls the expectation of the function by its sample average uniformly over a VC-major class of functions taking into account the size of the expectation. Using Talagrand's kernel method we prove a similar result for the classes of functions for which Dudley's uniform entropy integral or bracketing entropy integral is finite.

1 Introduction and main results.

Let Ω be a measurable space with a probability measure P and Ω^n be a product space with a product measure P^n. Consider a family of measurable functions $\mathcal{F} = \{f : \Omega \to [0,1]\}$. Denote

$$Pf = \int f dP, \quad \bar{f} = n^{-1}\sum_{i=1}^{n} f(x_i), \quad x = (x_1, \ldots, x_n) \in \Omega^n.$$

The main purpose of this paper is to provide probabilistic bounds for Pf in terms of $\bar{f}$ and the complexity assumptions on class $\mathcal{F}$. We are trying to extend the following result of Vapnik and Chervonenkis ([17]). Let $\mathcal{C}$ be a class of sets in Ω. Let

$$S(n) = \max_{x\in\Omega^n}\Big|\Big\{\{x_1, \ldots, x_n\} \cap C : C \in \mathcal{C}\Big\}\Big|$$

The VC dimension d of class $\mathcal{C}$ is defined as

$$d = \inf\{j \geq 1 : S(j) < 2^j\}.$$

$\mathcal{C}$ is called VC if $d < \infty$. The class of functions $\mathcal{F}$ is called VC-major if the class of sets

$$\mathcal{C} = \Big\{\{x \in \Omega : f(x) \leq t\} : f \in \mathcal{F}, t \in R\Big\}$$

is a VC class of sets in Ω, and the VC dimension of $\mathcal{F}$ is defined as the VC dimension of $\mathcal{C}$. The inequality of Vapnik and Chervonenkis states that (see Theorem 5.3 in [18]) if $\mathcal{F}$ is a

VC-major class of $[0,1]$ valued functions with dimension d then for all $\delta > 0$ with probability at least $1-\delta$ for all $f \in \mathcal{F}$

$$\frac{1}{n(Pf)^{1/2}} \sum_{i=1}^{n} (Pf - f(x_i)) \le 2\Bigl(\frac{1}{n}\log S(2n) + \frac{1}{n}\log\frac{4}{\delta}\Bigr)^{1/2}, \tag{1.1}$$

where for $n \ge d$, $S(n)$ can be bounded by

$$S(n) \le \Bigl(\frac{en}{d}\Bigr)^{d}$$

(see [16]) to give

$$\frac{1}{n(Pf)^{1/2}} \sum_{i=1}^{n} (Pf - f(x_i)) \le 2\Bigl(\frac{d}{n}\log\frac{2en}{d} + \frac{1}{n}\log\frac{4}{\delta}\Bigr)^{1/2}. \tag{1.2}$$

The factor $(Pf)^{-1/2}$ allows interpolation between the n^{-1} rate for Pf in the optimistic zero error case $\bar{f} = 0$ and the $n^{-1/2}$ rate in the pessimistic case when $\bar{f}$ is "large". In this paper we will prove a bound of a similar nature under different assumptions on the complexity of the class $\mathcal{F}$. Using Talagrand's abstract concentration inequality in product spaces and the related kernel method for empirical processes [14] we will first prove a general result that interpolates between optimistic and pessimistic cases. Then we will give examples of application of this general result in two situations when it is assumed that either Dudley's uniform entropy integral is finite or the bracketing entropy integral is finite.

Let us formulate Talagrand's concentration inequality that is used in the proof of our main Theorem 2 below. Consider a probability measure ν on Ω^n and $x \in \Omega^n$. We will denote by x_i the i^{th} coordinate of x. If $\mathcal{C}_i = \{y \in \Omega^n : y_i \ne x_i\}$, we consider the image of the restriction of ν to $\mathcal{C}_i$ by the map $y \to y_i$, and its Radon-Nikodym derivative d_i with respect to P. As in [14] we assume that Ω is finite and each point is measurable with a positive measure. Let m be a number of atoms in Ω and $p_1, \ldots, p_m$ be their probabilities. By the definition of d_i we have

$$\int_{\mathcal{C}_i} g(y_i)\,d\nu(y) = \int_{\Omega} g(y_i) d_i(y_i)\,dP(y_i).$$

For $\alpha > 0$ we define a function $\psi_\alpha(x)$ by

$$\psi_\alpha(x) = \begin{cases} x^2/(4\alpha), & \text{when } x \le 2\alpha, \\ x - \alpha, & \text{when } x \ge 2\alpha. \end{cases}$$

We set

$$m_\alpha(\nu, x) = \sum_{i \le n} \int \psi_\alpha(d_i)\,dP \quad \text{and} \quad m_\alpha(A, x) = \inf\{m_\alpha(\nu, x) : \nu(A) = 1\}.$$

For each $\alpha > 0$ let L_α be any positive number satisfying the following inequality:

$$\frac{2L_\alpha(e^{1/L_\alpha} - 1)}{1 + 2L_\alpha} \le \alpha. \tag{1.3}$$

The following theorem holds (see [9]).

Theorem 1 *Let $\alpha > 0$ and L_α satisfy (1.3). Then for any n and $A \subseteq \Omega^n$ we have*

$$\int \exp \frac{1}{L_\alpha} m_\alpha(A,x) dP^n(x) \le \frac{1}{P^n(A)}. \tag{1.4}$$

Below we will only use this theorem for $\alpha = 1$ and $L_1 \approx 1.12$. Let us introduce the normalized empirical process as

$$Z(x) = \sup_{\mathcal{F}} \frac{1}{\varphi(f)} \sum_{i=1}^{n} (Pf - f(x_i)), \quad x \in \Omega^n,$$

where $\varphi : \mathcal{F} \to (0, \infty)$ is a function such that Z has a finite median $M = M(Z) < \infty$, i.e.

$$P(Z \ge M) \le \frac{1}{2} \quad \text{and} \quad \forall \varepsilon > 0 \quad P(Z \ge M + \varepsilon) < \frac{1}{2}. \tag{1.5}$$

The factor $\varphi(f)$ will play the same role as $(nPf)^{1/2}$ plays in (1.1) The following theorem holds.

Theorem 2 *Let $L \approx 1.12$. If (1.5) holds then for any $u > 0$,*

$$\mathbb{P}\Big(\exists f \in \mathcal{F} \ \sum_{i \le n} (Pf - f(x_i)) \ge M\varphi(f) + 2\sqrt{LnuPf}\Big) \le 2e^{-u} \tag{1.6}$$

Proof. The proof of the theorem repeats the proof of Theorem 2 in [9] with some minor modifications, but we will give it here for completeness. Let us consider the set $A = \{Z(x) \le M\}$. Clearly, $P^n(A) \ge 1/2$. Let us fix a point $x \in \Omega^n$ and then choose $f \in \mathcal{F}$. For any point $y \in A$ we have

$$\frac{1}{\varphi(f)} \sum_{i=1}^{n} (Pf - f(y_i)) \le M.$$

Therefore, for any probability measure ν such that $\nu(A) = 1$ we will have

$$\frac{1}{\varphi(f)} \sum_{i \le n} (Pf - f(x_i)) - M \le \frac{1}{\varphi(f)} \int \Big(\sum_{i \le n} (Pf - f(x_i)) - \sum_{i \le n} (Pf - f(y_i)) \Big) d\nu(y)$$
$$= \frac{1}{\varphi(f)} \sum_{i \le n} \int (f(y_i) - f(x_i)) d_i(y_i) dP(y_i).$$

It is easy to observe that for $v \ge 0$, and $-1 \le u \le 1$,

$$uv \le u^2 I(u > 0) + \psi_1(v). \tag{1.7}$$

Therefore, for any $\delta > 1$

$$\sum_{i \le n} (Pf - f(x_i)) - M\varphi(f) \le \delta \sum_{i \le n} \int \frac{f(y_i) - f(x_i)}{\delta} d_i(y_i) dP(y_i)$$
$$\le \frac{1}{\delta} \sum_{i \le n} \int (f(y_i) - f(x_i))^2 I(f(y_i) > f(x_i)) dP(y_i) + \delta \sum_{i \le n} \int \psi_1(d_i) dP$$

Taking the infimum over ν we obtain that for any $\delta > 1$

$$\sum_{i \le n} (Pf - f(x_i)) \le M\varphi(f) + \frac{1}{\delta} \sum_{i \le n} \int (f(y_i) - f(x_i))^2 I(f(y_i) > f(x_i)) dP(y_i) + \delta m_1(A, x).$$

Let us denote the random variable $\xi = f(y_1)$, $F_\xi(t)$ - the distribution function of ξ, and $c_i = f(x_i)$. For $c \in [0,1]$ define the function $h(c)$ as

$$h(c) = \int (f(y_1) - c)^2 I(f(y_1) > c) dP(y_1) = \int_c^1 (t-c)^2 dF_\xi(t).$$

One can check that $h(c)$ is decreasing, convex, $h(0) = Pf^2$ and $h(1) = 0$. Therefore,

$$\frac{1}{n}\sum_{i\le n} h(c_i) \le \Bigl(\frac{1}{n}\sum_{i\le n} c_i\Bigr) h(1) + \Bigl(1 - \frac{1}{n}\sum_{i\le n} c_i\Bigr) h(0) = (1-\bar{f}) P f^2.$$

Hence, we showed that

$$\sum_{i\le n} (Pf - f(x_i) \le M\varphi(f) + \frac{1}{\delta} nPf + \delta m_1(A,x).$$

Theorem 1 then implies via the application of Chebyshev's inequality that with probability at least $1 - 2e^{-u}$, $m_1(A,x) \le Lu$ and, hence

$$\sum_{i\le n} (Pf - f(x_i) \le M\varphi(f) + \inf_{\delta>1} \Bigl(\frac{1}{\delta} nPf + \delta Lu\Bigr).$$

For $u \le nPf/L$ the infimum over $\delta > 1$ equals $2\sqrt{LnuPf}$. On the other hand, for $u \ge nPf/L$ this infimum is greater than $2nPf$ whereas the left-hand side is always less than nPf.

□

We will now give two examples of normalization $\varphi(f)$ where we can prove that (1.5) holds.

1.1 Uniform entropy conditions.

Given a probability distribution Q on Ω we denote

$$d_{Q,2}(f,g) = (Q(f-g)^2)^{1/2}$$

an L_2-distance on $\mathcal{F}$ with respect to Q. Fiven $u > 0$ we say that a subset $\mathcal{F}' \subset \mathcal{F}$ is $u-$separated if for any $f \ne g \in \mathcal{F}'$ we have $d_{Q,2}(f,g) > u$. Let the *packing number* $D(\mathcal{F}, u, L_2(Q))$ be the maximal cardinality of any $u-$separated set. We will say that $\mathcal{F}$ satisfies the uniform entropy condition if

$$\int_0^\infty \sqrt{\log D(\mathcal{F}, u)}\, du < \infty, \tag{1.8}$$

where

$$\sup_Q D(\mathcal{F}, u, L_2(Q)) \le D(\mathcal{F}, u)$$

and the supremum is taken over all discrete probability measures. It is well known (see, for example, [3]) that if one considers the subset $\mathcal{F}_p = \{f \in \mathcal{F} : Pf \le p\}$, then the expectation of $\sup_{\mathcal{F}_p} \sum (Pf - f(x_i))$ can be estimated (in some sense, since the symmetrization argument is required) by

$$\varphi(p) = \sqrt{n} \int_0^{\sqrt{p}} \sqrt{\log D(\mathcal{F}, u)}\, du. \tag{1.9}$$

We will prove that it holds for all $p > 0$ simultaneously.

Theorem 3 *Assume that $D(\mathcal{F},1) \geq 2$ and (1.8) holds. If φ is defined by (1.9) then the median*

$$M = M\Big(\sup_{\mathcal{F}} \frac{1}{\varphi(Pf)} \sum_{i=1}^{n} (Pf - f(x_i))\Big) \leq K < \infty,$$

is finite, where K is an absolute constant.

Proof. The proof is based on standard symmetrization and chaining techniques. We will first prove that

$$\mathbb{P}\Big(\sup_{\mathcal{F}} \frac{\sum(Pf - f(x_i))}{\varphi(Pf)} \geq u\Big) \leq 2\mathbb{P}\Big(\sup_{\mathcal{F}} \frac{\sum(f(y_i) - f(x_i))}{\varphi(\bar{f}(x,y))} \geq u - \Big(\frac{2}{\log 2}\Big)^{1/2}\Big). \tag{1.10}$$

where

$$\bar{f}(x,y) = \frac{1}{2n}\sum(f(y_i) + f(x_i)).$$

Let

$$A = \Big\{x : \sup_{\mathcal{F}} \frac{\sum(Pf - f(x_i))}{\varphi(Pf)} \geq u\Big\}.$$

Let $x \in A$ and $f \in \mathcal{F}$ be such that $\sum(Pf - f(x_i))/\varphi(Pf) \geq u$. Chebyshev's inequality implies

$$\mathbb{P}\Big(|\sum_{i=1}^{n}(Pf - f(y_i))| \geq \sqrt{2nPf}\Big) \leq \frac{n\mathrm{Var}f}{2nPf} \leq \frac{1}{2},$$

where $y = (y_1, \ldots, y_n)$ lives on an independent copy of (Ω^n, P^n). We will show that the inequalities

$$nPf \leq \sum f(y_i) + \sqrt{2nPf}, \quad u \leq \frac{\sum(Pf - f(x_i))}{\varphi(Pf)}$$

imply that

$$\frac{\sum(f(y_i) - f(x_i))}{\varphi(\bar{f}(x,y))} \geq u - \Big(\frac{2}{\log 2}\Big)^{1/2}.$$

If we define by $\mathbb{P}_y$ the probability measure on the space of y, it would mean that

$$\frac{1}{2}I(x \in A) \leq \mathbb{P}_y\Big(|\sum_{i=1}^{n}(Pf - f(y_i))| \geq \sqrt{2nPf}\Big) \leq \mathbb{P}_y\Big(\frac{\sum(f(y_i) - f(x_i))}{\varphi(\bar{f}(x,y))} \geq u - \Big(\frac{2}{\log 2}\Big)^{1/2}\Big)$$
$$\leq \mathbb{P}_y\Big(\sup_{\mathcal{F}} \frac{\sum(f(y_i) - f(x_i))}{\varphi(\bar{f}(x,y))} \geq u - \Big(\frac{2}{\log 2}\Big)^{1/2}\Big)$$

and taking expectation of both sides with respect to x would prove (1.10). To show the remaining implication we consider two cases when $nPf \leq \sum f(y_i)$ and $nPf \geq \sum f(y_i)$. First assume that $nPf \leq \sum f(y_i)$. Since, as easily checked, both $\varphi(p)$ and $p/\varphi(p)$ are increasing we get

$$\frac{\sum(Pf - f(x_i))}{\varphi(Pf)} \leq \frac{\sum(f(y_i) - f(x_i))}{\varphi(n^{-1}\sum f(y_i))} \leq \frac{\sum(f(y_i) - f(x_i))}{\varphi(\bar{f}(x,y))}.$$

In the case $nPf \geq \sum f(y_i)$ we have

$$\frac{\sum(Pf - f(x_i))}{\varphi(Pf)} \leq \frac{\sum(f(y_i) - f(x_i))}{\varphi(Pf)} + \frac{\sqrt{2nPf}}{\varphi(Pf)} \leq \frac{\sum(f(y_i) - f(x_i))}{\varphi(\bar{f}(x,y))} + \frac{\sqrt{2nPf}}{\varphi(Pf)}.$$

The assumption $D(\mathcal{F},\sqrt{Pf})\geq D(\mathcal{F},1)\geq 2$ garantees that $\varphi(Pf)\geq\sqrt{nPf\log 2}$ and, finally,

$$u\leq\frac{\sum(f(y_i)-f(x_i))}{\varphi(\bar{f}(x,y))}+\Big(\frac{2}{\log 2}\Big)^{1/2}$$

which completes the proof of (1.10). We have

$$\mathbb{P}\Big(\sup_{\mathcal{F}}\frac{\sum(f(y_i)-f(x_i))}{\varphi(\bar{f}(x,y))}\geq u\Big)=\mathbb{E}\mathbb{P}_{\varepsilon}\Big(\sup_{\mathcal{F}}\frac{\sum\varepsilon_i(f(y_i)-f(x_i))}{\varphi(\bar{f}(x,y))}\geq u\Big),$$

where (ε_i) is a sequence of Rademacher random variables. We will show that there exists u independent of n such that for any $x,y\in\Omega^n$

$$\mathbb{P}_{\varepsilon}\Big(\sup_{\mathcal{F}}\frac{\sum\varepsilon_i(f(y_i)-f(x_i))}{\varphi(\bar{f}(x,y))}\geq u\Big)<\frac{1}{2}.$$

Clearly, this will prove the statement of the theorem. For a fixed $x,y\in\Omega^n$ let

$$F=\{(f(x_1),\ldots,f(x_n),f(y_1),\ldots,f(y_n)):f\in\mathcal{F}\}\subset R^{2n}$$

and

$$d(f,g)=\Big(\frac{1}{2n}\sum_{i=1}^{2n}(f_i-g_i)^2\Big)^{1/2},\qquad f,g\in F.$$

The packing number of F with respect to d can be bounded by $D(F,u,d)\leq D(\mathcal{F},u)$. Consider an increasing sequence of sets

$$\{0\}=F_0\subseteq F_1\subseteq F_2\subseteq\ldots$$

such that for any $g\neq h\in F_j$, $d(g,h)>2^{-j}$ and for all $f\in F$ there exists $g\in F_j$ such that $d(f,g)\leq 2^{-j}$. The cardinality of F_j can be bounded by

$$|F_j|\leq D(F,2^{-j},d)\leq D(\mathcal{F},2^{-j}).$$

For simplicity of notations we will write $D(u):=D(\mathcal{F},u)$. If $D(2^{-j})=D(2^{-j-1})$ then in the construction of the sequence (F_j) we will set F_j equal to F_{j+1}. We will now define the sequence of projections $\pi_j:F\to F_j$, $j\geq 0$ in the following way. If $f\in F$ is such that $d(f,0)\in(2^{-j-1},2^{-j}]$ then set $\pi_0(f)=\ldots=\pi_j(f)=0$ and for $k\geq j+1$ choose $\pi_k(f)\in F_k$ such that $d(f,\pi_k(f))\leq 2^{-k}$. In the case when $F_k=F_{k+1}$ we will choose $\pi_k(f)=\pi_{k+1}(f)$. This construction implies that $d(\pi_{k-1}(f),\pi_k(f))\leq 2^{-k+2}$. Let us introduce a sequence of sets

$$\Delta_j=\{g-h:g\in F_j,h\in F_{j-1},d(g,h)\leq 2^{-j+2}\},\quad j\geq 1,$$

and let $\Delta_j=\{0\}$ if $D(2^{-j})=D(2^{-j+1})$. The cardinality of Δ_j does not exceed

$$|\Delta_j|\leq|F_j|^2\leq D(2^{-j})^2.$$

By construction any $f\in F$ can be represented as a sum of elements from Δ_j

$$f=\sum_{j\geq 1}(\pi_j(f)-\pi_{j-1}(f)),\quad \pi_j(f)-\pi_{j-1}(f)\in\Delta_j.$$

Let

$$I_j = \sqrt{n} \int\limits_{2^{-j-1}}^{2^{-j}} \sqrt{\log D(u)}du$$

and define the event

$$A = \bigcup_{j=1}^{\infty} \{ \sup_{f\in\Delta_j} \sum_{i=1}^{n} \varepsilon_i (f_{i+n} - f_i) \geq uI_j \}.$$

On the complement A^c of the event A we have for any $f \in F$ such that $d(f,0) \in (2^{-j-1}, 2^{-j}]$

$$\sum_{i=1}^{n} \varepsilon_i (f_{i+n} - f_i) = \sum_{k\geq j+1} \sum_{i=1}^{n} \varepsilon_i \big((\pi_k(f) - \pi_{k-1}(f))_{i+n} - (\pi_k(f) - \pi_{k-1}(f))_i \big)$$

$$\leq \sum_{k\geq j+1} uI_k \leq u\sqrt{n} \int\limits_{0}^{2^{-j-1}} \sqrt{\log D(u)}du \leq u\sqrt{n} \int\limits_{0}^{(\bar{f})^{1/2}} \sqrt{\log D(u)}du,$$

where $\bar{f} = (2n)^{-1} \sum_{i\leq 2n} f_i$, since $2^{-j-1} < d(f,0) \leq (\bar{f})^{1/2}$. It remains to prove that for some absolute constant u, $\bar{P}(A) < 1/2$. Indeed,

$$P(A) \leq \sum_{j=1}^{\infty} P\big(\sup_{f\in\Delta_j} \sum_{i=1}^{n} \varepsilon_i (f_{i+n} - f_i) \geq uI_j \big)$$

$$\leq \sum_{j=1}^{\infty} |\Delta_j| \exp\Big\{ -\frac{u^2 I_j^2}{n 2^{-2j+6}} \Big\} I\big(D(2^{-j}) > D(2^{-j+1}) \big)$$

$$\leq \sum_{j=1}^{\infty} \exp\Big\{ 2\log D(2^{-j}) - \frac{u^2 I_j^2}{n 2^{-2j+6}} \Big\} I\big(D(2^{-j}) > D(2^{-j+1}) \big),$$

since for $f \in \Delta_j$

$$\sum_{i=1}^{n} (f_{i+n} - f_i)^2 \leq 2 \sum_{i=1}^{2n} f_i^2 \leq n4 \cdot 2^{-2j+4}.$$

The fact that $D(u)$ is decreasing implies

$$\frac{I_j}{\sqrt{n}\; 2^{-(j+1)}} \geq \sqrt{\log D(2^{-j})}$$

and, therefore,

$$P(A) \leq \sum_{j=1}^{\infty} \exp\{ -\log D(2^{-j})(u^2 2^{-8} - 2) \} I\big(D(2^{-j}) > D(2^{-j+1}) \big)$$

$$\leq \sum_{j=1}^{\infty} \frac{1}{D(2^{-j})^{\alpha}} I\big(D(2^{-j}) > D(2^{-j+1}) \big) \leq \sum_{j=2}^{\infty} \frac{1}{j^{\alpha}} < \frac{1}{2},$$

for $\alpha = u^2/2^8 - 2$ big enough.

□

Combining Theorem 2 and Theorem 3 we get

Corollary 1 *If (1.8) holds then there exists an absolute constant $K > 0$ such that for any $u > 0$ with probability at least $1 - 2e^{-u}$ for all $f \in \mathcal{F}$*

$$\sum_{i=1}^{n} (Pf - f(x_i)) \le K\Big(\sqrt{n} \int_0^{(Pf)^{1/2}} \sqrt{\log D(\mathcal{F}, u)}du + \sqrt{nuPf}\Big).$$

1.2 Bracketing entropy conditions.

Given two functions $g, h : \Omega \to [0,1]$ such that $g \le h$ and $(P(h-g)^2)^{1/2} \le u$ we will call a set of all functions f such that $g \le f \le h$ a $u-$bracket with respect to $L_2(P)$. The $u-$bracketing number $N_{[]}(\mathcal{F}, u, L_2(P))$ is the minimum number of $u-$brackets needed to cover $\mathcal{F}$. Assume that

$$\int_0^{\infty} \sqrt{\log N_{[]}(\mathcal{F}, u, L_2(P))}du < \infty \tag{1.11}$$

and denote

$$\varphi(p) = \sqrt{n} \int_0^{\sqrt{p}} \sqrt{\log N_{[]}(\mathcal{F}, u, L_2(P))}du.$$

Then the following theorem holds.

Theorem 4 *Assume that $N_{[]}(\mathcal{F}, 1, L_2(P)) \ge 2$ and (1.11) holds. If φ is defined by (1.9) then the median*

$$M = M\Big(\sup_{\mathcal{F}} \frac{1}{\varphi(Pf)} \sum_{i=1}^{n} (Pf - f(x_i))\Big) \le K(\mathcal{F}) < \infty,$$

where $K(\mathcal{F})$ does not depend on n.

We omit the proof of this theorem since it is a modification of a standard bracketing entropy bound (see Theorem 2.5.6 and 2.14.2 in [15]) similar to what Theorem 3 is to the standard uniform entropy bound. The argument is more subtle as it involves a truncation argument required by the application of Bernstein's inequality but otherwise it repeats Theorem 3. Combining Theorem 2 and Theorem 4 we get

Corollary 2 *If (1.11) holds then there exists an absolute constant $K > 0$ such that for any $u > 0$ with probability at least $1 - 2e^{-u}$ for all $f \in \mathcal{F}$*

$$\sum_{i=1}^{n} (Pf - f(x_i)) \le K\Big(\sqrt{n} \int_0^{(Pf)^{1/2}} \sqrt{\log N_{[]}(\mathcal{F}, u, L_2(P))}du + \sqrt{nuPf}\Big).$$

2 Examples of application.

Example 1 (VC-subgraph classes of functions). A class of functions $\mathcal{F}$ is called VC-subgraph if the class of sets

$$\mathcal{C} = \Big\{\{(\omega, t) : \omega \in \Omega, t \in R, t \le f(\omega)\} : f \in \mathcal{F}\Big\}$$

is a VC-class of sets in $\Omega \times R$. The VC dimension of $\mathcal{F}$ is equal to the VC dimension d of $\mathcal{C}$. On can use Corollary 3 in [4] to show that

$$D(\mathcal{F}, u) \le e(d+1)\Big(\frac{2e}{u^2}\Big)^d.$$

Corollary 1 implies in this case that for any $\delta > 0$ with probability at least $1-\delta$ for all $f \in \mathcal{F}$

$$\frac{1}{n(Pf)^{1/2}} \sum_{i=1}^{n} (Pf - f(x_i)) \le K\Big(\Big(\frac{d}{n}\log n\Big)^{1/2} + \Big(\frac{1}{n}\log\frac{1}{\delta}\Big)^{1/2}\Big), \tag{2.1}$$

where $K > 0$ is an absolute constant. Instead of the $\log n$ on the right-hand side of (2.1) one could also write $\log(1/Pf)$, but we simplify the bound to eliminate this dependence on Pf. Note that the bound is similar to the bound (1.2) for VC classes of set and VC-major classes. Unfortunately, our proof does not allow us to recover the same small value of $K = 2$ as for VC classes of sets.

(2.1) improves the main result in [7], where it was shown that for any fixed $\nu > 0$ for any $\delta > 0$ with probability at least $1-\delta$ for all $f \in \mathcal{F}$

$$\frac{\sum(Pf - f(x_i))}{\sum(Pf + f(x_i)) + n\nu} \le K\Big(\frac{1}{n\nu}\Big(d\log\frac{1}{\nu} + \log\frac{1}{\delta}\Big)\Big)^{1/2}. \tag{2.2}$$

It is easy to see that, in a sense, one would get (2.1) from (2.2) only after optimizing over ν. Indeed, for $Pf \lesssim \nu$, (2.2) gives

$$\frac{1}{n}\sum(Pf - f(x_i)) \lesssim \Big(\frac{\nu}{n}\Big(d\log\frac{1}{\nu} + \log\frac{1}{\delta}\Big)\Big)^{1/2},$$

which is implied by (2.1) as well. For $Pf \gtrsim \nu$, (2.2) gives

$$\frac{1}{n}\sum(Pf - f(x_i)) \lesssim \Big(\frac{Pf}{\nu}\Big)^{1/2}\Big(\frac{Pf}{n}\Big(d\log\frac{1}{\nu} + \log\frac{1}{\delta}\Big)\Big)^{1/2},$$

which compared to (2.1) contains an additional factor of $(Pf/\nu)^{1/2}$. In the situation when ν is small (this is the only interesting case) this factor introduces an unnecessary penalty for any function f such that $Pf \gg \nu$. Hence, for a fixed ν (2.2) improves the bound for $Pf \le \nu$ at cost of f with $Pf \ge \nu$.

One can find alternative extensions of (2.2) in [5]. For some other applications of Corollary 1 see [10].

Example 2 (Bracketing entropy). Assume that either

$$D(\mathcal{F}, u) \le cu^{-\gamma} \text{ or } N_{[]}(\mathcal{F}, u, L_2(P)) \le cu^{-\gamma}, \ \ \gamma \in (0, 2).$$

Then Corollary 1 or Corollary 2 imply that for $u > 0$ with probability at least $1 - 2e^{-u}$ for all $f \in \mathcal{F}$

$$Pf - \bar{f} \le \frac{c_\gamma}{\sqrt{n}}\big((Pf)^{\frac{1}{2}-\frac{\gamma}{4}} + (uPf)^{\frac{1}{2}}\big).$$

If $\bar{f} = 0$ then it is easy to see that for $u \le n^{\frac{\gamma}{2+\gamma}}$ we have

$$Pf \le K_\gamma n^{-\frac{2}{2+\gamma}}.$$

As an example, if $\mathcal{F}$ is a class of indicator functions for sets with $\alpha-$smooth boundary in $[0,1]^l$ and P is Lebesgue absolutely continuous with bounded density then well known bounds on the bracketing entropy due to Dudley (see [3]) imply that $\gamma = 2(l-1)/\alpha$ and $Pf \leq K_\alpha n^{-\frac{\alpha}{l-1+\alpha}}$. Even though $\gamma = 2(l-1)/\alpha$ may be greater than 2 and Corollary 2 is not immediately applicable, one can generalize Theorem 4 to different choices of $\varphi(x)$, using the standard truncation in the chaining argument, to obtain the above rates even for $\gamma \geq 2$.

Acknowledgments. We would like to thank the referee for several very helpful comments and suggestions.

References

[1] Boucheron, S., Lugosi, G., Massart, P., A sharp concentration inequality with applications, *Random Structures Algorithms,* **16** (2000), 277 - 292.

[2] Dembo, A., Information inequalities and concentration of measure, *Ann. Probab.,* **25** (1997), 527 - 539.

[3] Dudley, R.M., Uniform Central Limit Theorems, Cambridge University Press, (1999).

[4] Haussler, D., Sphere packing numbers for subsets of the boolean $n-$cube with bounded Vapnik-Chervonenkis dimension, *J. Combin. Theory Ser. A,* **69** (1995), 217 - 232.

[5] Kohler, M., Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *J. Statist. Plann. Inference,* **89** (2000), no. 1-2, 1 - 23.

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

[7] Li, Y., Long, P.M., Srinivasan,A., Improved bounds on the sample complexity of learning, *Journal of Computer and System Sciences,* **62** (2001), 516 - 527.

[8] Massart, P., About the constants in Talagrand's concentration inequalities for empirical processes, *Ann. Probab.,* **28** (2000), 863 - 885.

[9] Panchenko, D., A note on Talagrand's concentration inequality, *Elect. Comm. in Probab.,* **6** (2001), 55 - 65.

[10] Panchenko, D., New zero-error bounds for voting algorithms, (2001), preprint.

[11] Rio E., Inégalités exponentielles pour les processus empiriques, *C.R. Acad. Sci. Paris,* t.330, Série I (2000), 597-600.

[12] Rio E., Inégalités de concentration pour les processus empiriques de classes de parties, *Probab. Theory Relat. Fields,* 119 (2001), 163-175.

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

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

[15] van der Vaart, A., Wellner, J., Weak Convergence and Empirical Processes: With Applications to Statistics, John Wiley & Sons, New York, (1996).

[16] Vapnik, V.N., Chervonenkis, A.Ya., On the uniform convergence of relative frequencies of events to their probabilities, *Soviet Math. Dokl.*,**9**, (1968), 915 - 918.

[17] Vapnik, V., Chervonenkis, A., Theory of Pattern Recognition. Nauka, Moscow (1974).

[18] Vapnik, V.N., Statistical Learning Theory, Wiley, New York (1998) .