Distribution-Dependent Vapnik-Chervonenkis Bounds

Nicolas Vayatis[1,2,3] and Robert Azencott[1]

[1] Centre de Mathématiques et de Leurs Applications (CMLA),
Ecole Normale Supérieure de Cachan,
61, av. du Président Wilson - 94 235 Cachan Cedex, France
[2] Centre de Recherche en Epistémologie Appliquée (CREA),
Ecole Polytechnique,
91 128 Palaiseau Cedex, France
[3] Nicolas.Vayatis@cmla.ens-cachan.fr,
WWW home page: http://www.cmla.ens-cachan.fr/Utilisateurs/vayatis

Abstract. Vapnik-Chervonenkis (VC) bounds play an important role in statistical learning theory as they are the fundamental result which explains the generalization ability of learning machines. There have been consequent mathematical works on the improvement of VC rates of convergence of empirical means to their expectations over the years. The result obtained by Talagrand in 1994 seems to provide more or less the final word to this issue as far as universal bounds are concerned. Though for fixed distributions, this bound can be practically outperformed. We show indeed that it is possible to replace the $2\epsilon^2$ under the exponential of the deviation term by the corresponding Cramér transform as shown by large deviations theorems. Then, we formulate rigorous distribution-sensitive VC bounds and we also explain why these theoretical results on such bounds can lead to practical estimates of the effective VC dimension of learning structures.

1 Introduction and motivations

One of the main parts of statistical learning theory in the framework developed by V.N. Vapnik [23], [25] is concerned with non-asymptotic rates of convergence of empirical means to their expectations.

The historical result obtained originally by Vapnik and Chervonenkis (VC) (see [21], [22]) has provided the qualitative form of these rates of convergences and it is a remarkable fact that this result holds with no assumption on the probability distribution underlying the data. Consequently, VC-theory of bounds is considered as a Worst-Case theory.

This observation is the source of most of the criticisms addressed to VC-theory. It has been argued (see e.g. [4], [5], [9], [17]) that VC bounds are loose in general. Indeed, there is an infinite number of situations in which the observed learning curves representing the generalization error of some learning structure are not well described by theoretical VC bounds.

In [17], D. Schuurmans criticizes the *worst-case-argument* by pointing out that there is no practical evidence that pathological probability measures must be taken into account. This is the open problem we want to tackle : **the distribution-sensitivity of VC bounds**.

Another question which motivates our work (Vapnik *et al.* [24]) is the measure of effective VC dimension. The idea to use a VC bound as an estimate of the error probability tail, and to simulate this probability to identify the constants and to estimate the VC dimension "experimentally".

We will show how to improve these results by computing new accurate VC bounds for fixed families of distributions.

It is thus possible to provide a deeper understanding for VC theory and its main concepts. We also want to elaborate a practical method for measuring empirically the VC dimension of a learning problem. This part is still work in progress (see forthcoming [26] for examples and effective simulations).

2 Classical VC bounds

We first present universal VC bounds. For simplicity, we consider the particular case of deterministic pattern recognition with noiseless data. The set-up is standard :

Consider a device T which transforms any input $X \in \mathbb{R}^d$ in some binary output $Y \in \{0,1\}$. Let us denote P the distribution of the random variable (X,Y), μ the distribution of X and R the Borel set in $\mathbb{R}^d$ of all X's associated to the label $Y = 1$.

The goal of learning is to select an appropriate model of the device T among a fixed set Γ of models C on the basis of a sample of empirical data (X_1, Y_1), ..., (X_n, Y_n). Here, Γ is a family[1] of Borel sets of $\mathbb{R}^d$ with finite VC dimension V. The VC dimension is a complexity index which characterizes the capacity of any given family of sets to *shatter* a set of points.

The error probability associated to the selection of C in Γ is :

$$L(C) = \mu(C \Delta R) \qquad (\textit{true} \text{ error})$$

$$\hat{L}_n(C) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{C \Delta R}(X_k) = \mu_n(C \Delta R) \; (\textit{empirical} \text{ error})$$

where μ_n is the empirical measure $\mu_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k}$.

The problem of model selection consists in minimizing the (unknown) risk functional $L(C) = \mu(C \Delta R)$, problem usually replaced by a tractable one which is the minimization of the empirical risk $\hat{L}_n(C) = \mu_n(C \Delta R)$ (this principle is known as ERM for Empirical Risk Minimization). But then, one has to guarantee that the minimum of the empirical risk is "close" to the theoretical minimum.

[1] Γ satisfies some technical, but unimportant for our purpose, measurability condition. In order to avoid such technicalities, we will assume that Γ is countable.

This is precisely the point where Vapnik-Chervonenkis bound drops in. Their fundamental contribution is the upper bound of the quantity

$$Q(n,\epsilon,\Gamma,\mu) = \mathbf{Pr}\left\{\sup_{C\in\Gamma}|\mu_n(C)-\mu(C)|>\epsilon\right\} \quad .$$

Remark 1. Note that

$$\mathbf{Pr}\left\{\sup_{C\in\Gamma}|\hat{L}_n(C)-L(C)|>\epsilon\right\} = \mathbf{Pr}\left\{\sup_{C\in\Gamma}|\mu_n(C\Delta R)-\mu(C\Delta R)|>\epsilon\right\} \quad ,$$

and by a slight notational abuse without any consequence on the final result[2], we take $C := C\Delta R$ and $\Gamma := \Gamma\Delta R = \{C\Delta R : C \in \Gamma\}$.

We recall here this result :

Theorem 1 (Vapnik-Chervonenkis [21]). *Let Γ be a class of Borel sets of $\mathbb{R}^d$ with finite VC dimension V. Then, for $n\epsilon^2 \geq 2$,*

$$\sup_{\mu\in\mathcal{M}_1(\mathbb{R}^d)} \mathbf{Pr}\left\{\sup_{C\in\Gamma}|\mu_n(C)-\mu(C)|>\epsilon\right\} \leq 4\left(\frac{2en}{V}\right)^V e^{-n\epsilon^2/8} \quad .$$

Remark 2. For a very readable proof, see [7].

This bound actually provides an estimate of the worst rate of convergence of the empirical estimator to the true probability.

To comment on the form of the previous upper bound, we notice that the exponential term quantifies the worst deviation for a single set C and the polynomial term characterizes the richness of the family Γ.

There have been several improvements for this type of bound since the pioneering work of Vapnik and Chervonenkis [21](see Vapnik [23], Devroye[6], Pollard[16] , Alexander[1], Parrondo-Van den Broek [15], Talagrand[19], Lugosi [13]).

Many of these improvements resulted from theory and techniques in empirical processes (see Pollard[16], Alexander[1], Talagrand[19]), and these works indicated that the proper variable is $\epsilon\sqrt{n}$ (or $n\epsilon^2$). Keeping this in mind, we can summarize the qualitative behavior of VC-bounds by the following expression :

$$K(\epsilon,V)\cdot\underbrace{(n\epsilon^2)^{\tau(V)}}_{capacity}\cdot\underbrace{e^{-n\gamma\epsilon^2}}_{deviation} \quad \text{for} \quad n\epsilon^2 \geq M \, ,$$

with M constant, τ an affine function of V, $\gamma \in [0,2]$, and $K(\epsilon,V)$ constant independent of n, possibly depending on ϵ and V (ideally $K(\epsilon,V) \leq K(V)$).

Once we have stated this general form for VC-bounds, we can address the following issues (both theoretically and practically) :

[2] Indeed, for a fixed set R, we have $VCdim(\Gamma) = VCdim(\Gamma\Delta R)$. For a proof, see e.g. [11].

1. What is the best exponent γ in the deviation term ?
2. What is the correct power $\tau(V)$ of n in the capacity term ?
3. What is the order of the constant term $K(V)$ for the bound to be sharp ?

In Table 1, we provide the theoretical answers brought by previous studies, in a distribution-free framework.

Table 1. Universal bounds

	M	$K(\epsilon, V)$	$\tau(V)$	γ
Pollard (1984)	2	$8\left(\frac{e}{V}\frac{1}{\epsilon^2}\right)^V$	V	1/32
Vapnik-Chervonenkis (1971)	2	$4\left(\frac{2e}{V}\frac{1}{\epsilon^2}\right)^V$	V	1/8
Vapnik (1982)	2	$6\left(\frac{2e}{V}\frac{1}{\epsilon^2}\right)^V$	V	1/4
Parrondo-Van den Broeck (1993)	2	$6e^{2\epsilon}\left(\frac{2e}{V}\frac{1}{\epsilon^2}\right)^V$	V	1
Devroye (1982)	1	$4e^{4\epsilon+4\epsilon^2}\left(\frac{e}{V}\frac{1}{\epsilon^2}\right)^V$	$2V$	2
Lugosi (1995)	$\frac{V}{2}$	$4e(V+1)\left(\frac{32e^5}{V^2}\frac{1}{\epsilon}\right)^V$	$2V$	2
Alexander (1984)	64	16	$2048V$	2
Talagrand (1994)	0	$K(V)$	$V-\frac{1}{2}$	2

to conclude this brief review, we point out that in the above distribution-free results, the optimal value for the exponent γ is 2 (which actually is the value in Hoeffding's inequality), and the best power achieved for the capacity term is the one obtained by Talagrand $V - \frac{1}{2}$ (see also the discussion about this point in [19]). In most of the results, the function $K(\epsilon, V)$ is not bounded as ϵ goes to zero, and only Alexander's and Talagrand's bounds satisfy the requirement $K(\epsilon, V) \leq K(V)$.

Our point in the remainder of this paper is that the $2\epsilon^2$ term under the exponential can be larger in particular situations.

3 Rigorous distribution-dependent results

In the continuity of the results evoked in the previous section, one issue of interest is the construction of bounds taking into account the characteristics of the underlying probability measure μ.

There are some works tackling this problem but with very different perspectives (see Vapnik [25], Bartlett-Lugosi [3], in a learning theory framework; Schuurmans [17], in a PAC-learning framework; Pollard [16], Alexander [1], Massart [14], who provide the most significant results in empirical processes).

We note that :

– in learning theory, the idea of distribution-dependent VC-bounds led to other expressions for the capacity term, involving different concepts of entropy as

VC-entropy, annealed entropy or metric entropy, depending on the probability measure.
- while in the theory of empirical processes, a special attention was given to refined exponential rates for restricted families of probability distributions (see [1], [14]).

Our purpose is to formulate a distribution-dependent result preserving the structure of universal VC bounds with an optimal exponential rate and with some nearly optimal power $\tau(V)$, though we will keep the concept of VC dimension unchanged[3].

Indeed, we would like to point out that if we consider a particular case where the probability measure μ underlying the data belongs to a restricted set $\mathcal{P} \subset \mathcal{M}_1(\mathbb{R}^d)$, then the deviation term can be fairly improved. Our argument is borrowed from large deviations results which provide asymptotically exact estimates of probability tails on a logarithmic scale. A close look at the proof of the main theorem in the case of real random variables (Cramér's theorem, for a review, see [2] or [18]) will reveal that the result holds as a non-asymptotical upper bound. Thanks to this result, we obtain the *exact* term under the exponential quantifying the worst deviation.

In order to formulate our result, we need to introduce the Cramér transform (see the appendix) of a Bernoulli law with parameter p given by : $\Lambda_p(x) = x \log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right)$, for x in $[0,1]$.

Then, the uniform deviation of the empirical error from its expectation, for a fixed family of probability distributions, can be estimated according to the following theorem (a sketch of its proof is given in Sect. 6) :

Theorem 2. *Let Γ be a family of measurable sets C of $\mathbb{R}^d$ with finite VC dimension V, and $\mathcal{P} \subset \mathcal{M}_1(\mathbb{R}^d)$ a fixed family of probability distributions μ.*

Let Λ_p be the Cramér transform of a Bernoulli law with parameter p, let $J = \{q : q = \mu(C), (\mu, C) \in \mathcal{P} \times \Gamma\}$ and set $p = \arg\min_{q \in J} | q - \frac{1}{2} |$. For every $\beta > 0$, there exists $M(\beta, p, V)$ and $\epsilon_0(\beta, p, V) > 0$ such that if $\epsilon < \epsilon_0(\beta, p, V)$ and $n\epsilon^2 > M(\beta, p, V)$, we have :

$$\sup_{\mu \in \mathcal{P}} \mathbf{Pr}\left\{ \sup_{C \in \Gamma} |\mu_n(C) - \mu(C)| > \epsilon \right\} \leq K(V)(n\epsilon^2)^V \, e^{-n \cdot (1-\beta) \cdot \Lambda_p(\epsilon + p)} \quad .$$

Remark 3. The corrective term β can be chosen to be as small as possible at the cost of increasing $M(\beta, p, V)$.

Remark 4. Here we achieved $\tau(V) = V$ instead of the optimal $V - \frac{1}{2}$ found by Talagrand in [19]. However, refining the proof by using a smart partitioning of the family Γ should lead to this value.

Remark 5. Note that the result above can be extended to the other fundamental problems of statistics as regression or density estimation.

[3] However, we could use alternatively effective VC dimension which is a distribution-dependent index (see [26] for details).

4 Comparison with Universal VC Bounds

To appreciate the gain in considering distribution-dependent rates of convergence instead of universal rates, we provide a brief discussion in which we compare the $\Lambda_p(\epsilon + p)$ in our result with the universal $\gamma\epsilon^2$.

First, we point out that even in the worst-case situation (take $\mathcal{P} = \mathcal{M}_1(\mathbb{R}^d)$) where $p = \frac{1}{2}$, we have a better result since $\Lambda = \Lambda_{\frac{1}{2}}(\epsilon + \frac{1}{2}) \geq 2\epsilon^2$ (see Fig. 1).

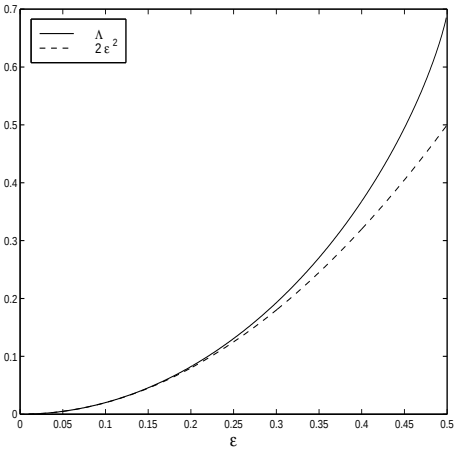

Fig. 1. Comparison between $\Lambda = \Lambda_{\frac{1}{2}}(\epsilon + \frac{1}{2})$ and $2\epsilon^2$.

In the general case when $p \neq \frac{1}{2}$, we claim that the distribution-dependent VC bound obtained in Theorem 2 is of the same type of universal bounds listed in Sect. 2. In order to make the comparison, we recall a result proved by W. Hoeffding :

Proposition 1 (Hoeffding [10]). *For any $p \in [0,1]$, the following inequality holds :*

$$\frac{\Lambda_p(\epsilon + p)}{\epsilon^2} \geq g(p) \geq 2 \ ,$$

where the function g is defined by :

$$g(p) = \begin{cases} \dfrac{1}{1-2p} \log\left(\dfrac{1-p}{p}\right) & , \text{ if } p < \frac{1}{2} \\ \dfrac{1}{2p(1-p)} & , \text{ if } p \geq \frac{1}{2} \ . \end{cases}$$

With the help of Fig. 2, the comparison between $g(p)$ and the values of γ becomes quite explicit. Indeed, it is clear that, as soon as $p \neq 1/2$, we have a better bound than in the universal case.

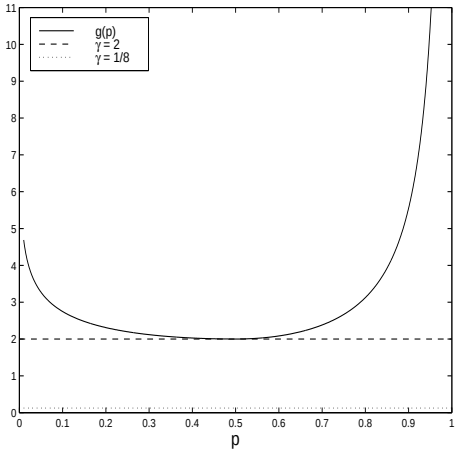

Fig. 2. Comparison between *distribution-dependent* $g(p)$ and *universal* γ's.

5 PAC-Learning Application of the Result

A PAC-learning formulation of distribution-dependent VC bounds in terms of sample complexity can easily be deduced from the main result :

Corollary 1. *Under the same assumptions as in Theorem 2. The* sample complexity $N(\epsilon, \delta)$*, that guarantees :*

$$\mathbf{Pr}\left\{\sup_{C \in \Gamma} |\mu_n(C) - \mu(C)| > \epsilon\right\} \leq \delta$$

for $n \geq N(\epsilon, \delta)$*, is bounded by :*

$$N(\epsilon, \delta) \leq max\left(\frac{2V}{\Lambda}\log\left(\frac{2V\epsilon^2}{\Lambda}\right), \frac{2}{\Lambda}\log\left(\frac{K(V)}{\delta}\right)\right)$$

where $\Lambda = (1-\beta)\cdot\Lambda_p(\epsilon + p)$*.*

Remark 6. In order to appreciate this result, one should consider that $\Lambda_p(\epsilon+p) \simeq g(p)\epsilon^2$.

Proof. Consider n such that : $(n\epsilon^2)^V \leq e^{n\Lambda/2}$. Then , taking the log and multiplying by ϵ^2, we obtain : $n\epsilon^2 \geq \frac{2V\epsilon^2}{\Lambda}\log(n\epsilon^2)$. Thus, taking the log again, we have $\log(n\epsilon^2) \geq \log(\frac{2V\epsilon^2}{\Lambda})$ which we inject in the last inequality. We get : $n \geq \frac{2V}{\Lambda}\log\left(\frac{2V\epsilon^2}{\Lambda}\right)$. If n satisfies the previous condition, we have : $(n\epsilon^2)^V e^{-n\Lambda} \leq e^{-n\Lambda/2}$, and we want $K(V)e^{-n\Lambda/2}$ to be smaller than δ. Hence, n should also satisfy : $n \geq \frac{2}{\Lambda}\log\left(\frac{K(V)}{\delta}\right)$. □

As a matter of fact, Theorem 2 provides an appropriate theoretical foundation for computer simulations. Indeed, in practical situations, *a priori* informations about the underlying distribution and about realistic elements C of the family Γ turn distribution-dependent VC bounds in an operational tool for obtaining estimates of the effective VC dimension V and of the constant $K(V)$ as well (see [26] for examples).

6 Elements of proof for Theorem 2

In this section, we provide a sketch of the proof of Theorem 2 (for a complete and general proof, see [26]). It relies on some results from empirical processes theory. The line of proof is inspired from the direct approximation method exposed by D. Pollard [16] while most of the techniques and intermediate results used in this proof are due to M. Talagrand and come from [19], [20].

First, note that if the family Γ is finite, the proof is a straightforward consequence of Chernoff's bound (see the appendix) together with the union-of-events bound. In the case of a countable family, we introduce a finite approximation Γ_λ which is a λ-net[4] for the symmetric difference associated to the measure μ, with cardinality $N(\Gamma, \mu, \lambda) = N(\lambda)$. We shall take $\lambda = \frac{1}{n\epsilon^2}$.

The first step of the proof is to turn the global supremum of the empirical process $G_n(C) = \mu_n(C) - \mu(C)$ into a more tractable expression like the sum of a maximum over a finite set and some local supremum. Then, the tail $Q(n, \epsilon, \Gamma, \mu)$ is bounded by $A + B$, where A is the tail of the maximum of a set of random variables which can be bounded by :

$$A \leq N(\lambda) \max_{C^* \in \Gamma_\lambda} \mathbf{Pr}\left\{ |G_n(C^*)| > (1 - \frac{\beta}{2})\epsilon \right\} \quad , \tag{1}$$

and B is the tail of the local supremum of a family of random variables bounded as follows :

$$B \leq N(\lambda) \max_{C^* \in \Gamma_\lambda} \mathbf{Pr}\left\{ \sup_{C \in \mathcal{B}(C^*, \lambda)} |G_n(C) - G_n(C^*)| > \frac{\beta\epsilon}{2} \right\} \quad , \tag{2}$$

where $\mathcal{B}(C^*, \lambda) = \{C \in \Gamma : \mu(C \Delta C^*) \leq \lambda\}$.

The probability tail in (1) can be bounded by large deviations estimates according to Chernoff's bound :

$$\mathbf{Pr}\left\{ |G_n(C^*)| > (1 - \frac{\beta}{2})\epsilon \right\} \leq 2e^{-n \cdot A_p((1-\frac{\beta}{2})\epsilon + p)} \quad ,$$

where $p = \arg\min_{q\,:\,q=\mu(C),\,(\mu,C)\in\mathcal{P}\times\Gamma} |\, q - \frac{1}{2}\,|$.

[4] If Γ is totally bounded, by definition, it is possible, for every $\lambda > 0$, to cover Γ by a finite number of balls of radius λ centered in Γ. Consider a minimal cover of Γ, then a λ-net will be the set of all the centers of the balls composing this cover.

The estimation of (2) requires the use of technical results on empirical processes mainly from [19] and [20] : symmetrization of the empirical processes with Rademacher random variables, decomposition of the conditional probability tail using the median, application of the chaining technique. In the end, we introduce the parameter u to obtain the bound :

$$\begin{aligned} B &\leq 4N(\lambda)\left(2e^{-\frac{\beta^2\epsilon^2}{1024u}} + e^{-\frac{1}{2}n^2u\log\left(\frac{nu}{64m_1}\right)} + e^{-\frac{n\beta\epsilon}{64}\log\left(\frac{\beta\epsilon}{128m_2}\right)}\right) \\ &= 4N(\lambda)\,(\,D + F + G\,) \quad , \end{aligned} \tag{3}$$

where $m_1 = k_1(V)\cdot(1/n\epsilon^2)\cdot\log(k_2\,n\epsilon^2))$, and $m_2 = k_3(V)\cdot(1/n\epsilon)\cdot\log(k_2\,n\epsilon^2))$. The meaning of each of the terms in (3) is the following : D measures the deviation of the symmetric process from the median, F controls its variance and G bounds the tail of the median which can be controlled thanks to the chaining technique.

To get the proper bound from (3), one has to consider the constraint on u :

$$u \in \mathcal{I} = \left[\frac{k_5(\beta,p,V)}{n\log(n\epsilon^2)}\,,\ k_4(\beta,p)\cdot\frac{1}{n}\right] \quad ,$$

which leads to the condition : $n\epsilon^2 > M(\beta,p,V)$.

To get the desired form of the bound, we eventually apply a result due to D. Haussler [8]:

$$N(\lambda) \leq e(V+1)\left(\frac{2e}{\lambda}\right)^V \quad ,$$

and set $\lambda = \frac{1}{n\epsilon^2}$, $u \in \mathcal{I}$, which ends the proof.

7 Appendix - Chernoff's bound on large deviations

We remind the setting for Chernoff's bound (see [2] for further results).

Consider ν a probability measure over $\mathbb{R}$. $\hat{\nu} : \mathbb{R} \longrightarrow]0,+\infty]$ is the Laplace transform of ν, defined by $\hat{\nu}(t) = \int_{\mathbb{R}} e^{tx}\,\nu(dx)$.

The Cramér transform $\Lambda : \mathbb{R} \longrightarrow [0,+\infty]$ of the measure ν is defined, for $x \in \mathbb{R}$, by

$$\Lambda(x) = \sup_{t\in\mathbb{R}}\left(tx - \log\hat{\nu}(t)\right) \quad .$$

If we go through the optimization of the function of t inside the *sup* (it is a simple fact that this function is infinitely differentiable, cf. e.g. [18]), we can compute exactly the optimal value of t. Let $t(x)$ be that value. Then, we write

$$\Lambda(x) = t(x)x - \log\hat{\nu}(t(x)) \quad .$$

Proposition 2 (Chernoff's bound). *Let $U_1, ..., U_n$ be real i.i.d. random variables. Denote their sum by $S_n = \sum_{i=1}^n U_i$. Then, for every $\epsilon > 0$, we have :*

$$\mathbf{Pr}\left\{|S_n - \mathbf{E}S_n| > \epsilon\right\} \leq 2e^{-n\Lambda(\epsilon+\mathbf{E}U_1)}$$

where Λ is the Cramér Transform of the random variable U_1.

References

1. Alexander, K.: Probability Inequalities for Empirical Processes and a Law of the Iterated Logarithm. Annals of Probability **4** (1984) 1041-1067
2. Azencott, R.: Grandes Déviations, in Hennequin, P.L. (ed.): Ecole d'Eté de Probabilités de Saint-Flour VIII-1978. Lecture Notes in Mathematics, Vol. **774**. Springer-Verlag, Berlin Heidelberg New York (1978)
3. Bartlett, P., Lugosi, G.: An Inequality for Uniform Deviations of Sample Averages from their Means. To appear (1998)
4. Cohn, D., Tesauro, G.: How Tight Are the Vapnik-Chervonenkis Bounds ? Neural Computation **4** (1992) 249-269
5. Cohn, D.: Separating Formal Bounds from Practical Performance in Learning Systems. PhD thesis, University of Washington (1992)
6. Devroye, L.: Bounds for the Uniform Deviation of Empirical Measures. Journal of Multivariate Analysis **12** (1982) 72-79
7. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer-Verlag, Berlin Heidelberg New York (1996)
8. Haussler, D.: Sphere Packing Numbers for Subsets of the Boolean n-Cube with Bounded Vapnik-Chervonenkis Dimension. Journal of Combinatorial Theory, Series A **69** (1995) 217-232
9. Haussler, D., Kearns, M., Seung, H.S., Tishby, N.: Rigorous Learning Curve Bounds from Statistical Mechanics. Machine Learning (1996) 195-236
10. Hoeffding, W.: Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association **58** (1963) 13-30
11. Kearns, M.J., Vazirani, U.V.: An Introduction to Computational Learning Theory. MIT Press, Cambridge Massachussets (1994)
12. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer-Verlag, Berlin Heidelberg New York (1992)
13. Lugosi, G.: Improved Upper Bounds for Probabilities of Uniform Deviations. Statistics and Probability Letters **25** (1995) 71-77
14. Massart, P.: Rates of Convergence in the Central Limit Theorem for Empirical Processes. Annales de l'Institut Henri Poincaré, Vol. 22, No. 4 (1986) 381-423
15. Parrondo, J.M.R., Van den Broeck, C.: Vapnik-Chervonenkis Bounds for Generalization. J. Phys. A : Math. Gen. **26** (1993) 2211-2223
16. Pollard, D.: Convergence of Stochastic Processes. Springer-Verlag, Berlin Heidelberg New York (1984)
17. Schuurmans, D. E.: Effective Classification Learning. PhD thesis, University of Toronto (1996)
18. Stroock, D.W.: Probability Theory, an Analytic View. Cambridge University Press (1993)
19. Talagrand, M.: Sharper Bounds for Gaussian and Empirical Processes. The Annals of Probability, Vol. 22, No. 1 (1994) 28-76
20. van der Vaart, A. W., Wellner, J. A.: Weak Convergence and Empirical Processes. Springer-Verlag, Berlin Heidelberg New York (1996)
21. Vapnik, V. N., Chervonenkis, A. Ya.: On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. Theory of Probability and its Applications, Vol. XVI, No. 2 (1971) 264-280
22. Vapnik, V. N., Chervonenkis, A. Ya.: Necessary and Sufficient Conditions for the Uniform Convergence of Means to their Expectations. Theory of Probability and its Applications, Vol. XXVI, No. 3 (1981) 532-553

23. Vapnik, V. N.: Estimation of Dependences Based on Empirical Data. Springer-Verlag, Berlin Heidelberg New York (1982)
24. Vapnik, V. N., Levin, E., Le Cun, Y.: Measuring the VC Dimension of a Learning Machine. Neural Computation **6** (1994) 851-876
25. Vapnik, V. N.: The Nature of Statistical Learning Theory. Springer-Verlag, Berlin Heidelberg New York (1995)
26. Vayatis, N.: Learning Complexity and Pattern Recognition. PhD thesis, Ecole Polytechnique. To appear (1999)