The Role of Critical Sets in Vapnik-Chervonenkis Theory

Nicolas Vayatis[*]
vayatis@cmla.ens-cachan.fr

Centre de Mathématiques
et de Leurs Applications (CMLA)
Ecole Normale Supérieure de Cachan
61, avenue du Président Wilson
94 235 Cachan Cedex, France.

Equipe de Modélisation
Aléatoire (MODAL'X)
Université Paris X - Nanterre
200, avenue de la République
92 100 Nanterre Cedex, France.

Abstract

In the present paper, we present the theoretical basis, as well as an empirical validation, of a protocol designed to obtain effective VC dimension estimations in the case of a simple pattern recognition issue. We first formulate particular (distribution-dependent) VC bounds in which a special attention has been given to the exact exponential rate of convergence. We show indeed that the most significant contribution in such bounds is due to the "worst" elements of the model class (designated as *the critical sets*). We then explain how these results can lead to a rigorous framework for computer simulations involving speed-up techniques for rare event simulation (importance sampling) as well as parameter estimation (linear regression). We thus obtain accurate empirical estimates of the complexity measure and of the multiplicative constant in VC bounds. In particular, we develop the idea of a *local complexity* characterization associated to every critical set.

1 Introduction

One of the main issues in Statistical Learning Theory is the improvement of bounds on the generalization error of learning machines in model selection problems. Those bounds rely on rates of convergence in uniform laws of large numbers. Seminal results are due to Vapnik and Chervonenkis [8] and they have witnessed continuous improvements over the years (see [4] or [11] for a review). In this paper, we appeal to large deviations techniques (see [1] or [2]) to show that there still is some room for improving classical VC bounds by introducing some additional considerations both on the model class and the underlying distribution. A key concept in the present study is the one of *critical element* of the model class. An element is said to be *critical* if it achieves the *worst*, in some sense, expected risk. We will show that a proper characterization of the critical elements in a particular learning problem leads to a significant improvement of the rate of uniform convergence of empirical risk towards expected risk. First, we shall study the simple case of a finite class of models in order to provide some intuitive arguments on the behavior of the probability tail of worst deviation. We then state our main result providing an exact distribution-dependent exponential rate for the typical VC bound. We also suggest a precise formula which should be valid for data samples of sufficiently large size. The final section is devoted to a brief overview of a simulation protocol aiming at the experimental validation of such theoretical results. As an application, we obtain effective VC dimension estimations with a proper treatment of calibration issues.

2 Setting

In order to keep things as simple as possible, we restrict ourselves to the simplest setting. Thus, we shall consider the example of pattern recognition, in the particular case of binary deterministic classification. Thus, the model class shall be thought as a family of sets.

2.1 Notations

Let Γ be a family of measurable sets of $\mathbb{R}^d$ with finite VC dimension V. The data are represented by a sample $X(n) = \{X_1, ..., X_n\}$ of i.i.d. random variables with probability distribution μ over $\mathbb{R}^d$. We denote by $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, the empirical measure over $X(n)$. We recall some basic definitions from Vapnik-Chervonenkis (VC) theory.

Definition 2.1 (trace)

$$\mathbf{Tr}(\Gamma, x(n)) = \{C \cap x(n),\, C \in \Gamma\}\,.$$

Definition 2.2 (theoretical VC dimension)

$$\begin{aligned} V_{th}(\Gamma) :=& V(\Gamma) \\ &= \max\{k \in \mathbb{N} \,:\, \sup_{x(k)} |\mathbf{Tr}(\Gamma, x(k))| = 2^k\}\,. \end{aligned}$$

Definition 2.3 (empirical VC dimension)

$$V_{emp}(\Gamma, x(n)) := \\ \max\{k, k \le n \,:\, \sup_{x(k)\subset x(n)} |\mathbf{Tr}(\Gamma, x(k))| = 2^k\}\,.$$

In the present study, we will focus on the estimation of the one-sided probability tail of the worst deviation of the empirical mean from its expectation on the family Γ described by

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

[*] This work has been supervised by Pr. Robert Azencott (CMLA, ENS Cachan) as a part of the author's PhD thesis.

Classical VC theory is devoted to the control of such small probabilities by means of some complexity concept (mainly VC entropy, or VC dimension) *independently of the underlying distribution μ*. These probability tails are known to behave like

$$K\,(n\epsilon^2)^\beta \exp\left\{-An\epsilon^2\right\} \;,$$

where K is some constant, β is a complexity exponent linked to the VC dimension V, and A is the exponential rate (see [11] for an overview of universal VC bounds). The best distribution-free result is the one obtained by Talagrand [5] in which $A = 2$ and $\beta = V - 1/2$. We suggest here some directions in computing refined bounds taking into account particular distributions. In the following formulations, we shall make use of the information function

$$H(q,p) = q\log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$$

in order to express the exact exponential rate in VC bounds.

Remark 2.4 *Note that for ϵ small, we have*

$$H(q+\epsilon,q) \sim \frac{\epsilon^2}{2q(1-q)} \;.$$

2.2 Basic definitions

We now define the notion of *critical value* and *critical set* of Γ relatively to the fixed distribution μ.

Definition 2.5 *We introduce*

- *the range of admissible values of $\mu(C)$ denoted by*

$$J := \{\mu(C) \;:\; C \in \Gamma\} \;,$$

- *the μ-critical value $p := \arg\min_{q\in J} H(q+\epsilon,q)$,*
- *the μ-critical subfamily*

$$\Gamma_p := \{C \in \Gamma \;:\; \mu(C) = p\} \;,$$

- *and the μ-critical sets which are the elements of the subfamily Γ_p.*

Remark 2.6 *The value p minimizing $H(q+\epsilon,q)$ depends on ϵ but, since we have $H(q+\epsilon,q) \underset{\epsilon\to 0}{\sim} \frac{\epsilon^2}{2q(1-q)}$, actually it is very close to 1/2. One could retain, for simplicity, that p is the closest admissible value to 1/2.*

In several applications such as concept learning, the knowledge of *a priori* information on the target set (or concept) allows to consider a restricted part of the whole family of candidates. This knowldege could induce serious restrictions on the range J of admissible values. Our point is that the constraints on the possible values of $\mu(C)$, for C in Γ, lead to a consequent improvement of the bound on generalization. In particular, a small part of Γ actually contributes to the probability of worst deviation.

3 Heuristics - The Finite Case

The theoretical point on which we want to insist is the significant improvement that can be obtained in classical VC bounds in a distribution-dependent setting. Indeed universal bounds are based on Hoeffding's inequality, but as soon as the critical value is different from 1/2, then it would much more appropriate to use Chernoff's inequality. Indeed, for a single element C, it gives

$$\mathbf{Pr}\left\{\mu_n(C) - \mu(C) > \epsilon\right\} \le \exp\left\{-nH(\mu(C)+\epsilon,\mu(C))\right\} \;.$$

Moreover, we know (see [1]) by Cramér-Chernoff theorem on large deviations that this rate is asymptotically exact and thus it cannot be improved. In practical learning problems where "bad" models can be eliminated, one can consider that the range J can be significantly constrained. Let us assume that the family Γ is finite. Then, we obtain an improved exponential rate on the upper bound of the probability tail $\rho(\mu,n,\epsilon)$ thanks to the union-of-events bound jointly with Chernoff's bound,

$$\rho(\mu,n,\epsilon) \le |\Gamma| \exp\left\{-nH(p+\epsilon,p)\right\} \qquad (1)$$

where p is the μ-critical value of Γ. More precisely, we shall have, for n large, a behaviour like

$$\rho(\mu,n,\epsilon) \sim |\Gamma_p| \exp\left\{-nH(p+\epsilon,p)\right\} \qquad (2)$$

and we see that the main contributions in the value of the probability tail are due to the critical sets of Γ.

Our first step shall be to extend such a reasoning to infinite families Γ.

4 Distribution-Dependent VC Bound - Result and Conjecture

Indeed, we have already tackled the question of distribution--dependent VC bounds in a previous paper [11], but here we provide a finer result together with some conjectures on exact asymptotics of VC probability tails. Moreover, we stress the role of critical sets in these results and we point out the fact that theoretical VC dimensions actually mask the presence of local complexity exponents attached to each of those critical sets. The main theorem actually improves our previous result from [11] since we have managed to get rid of the disturbing corrective term. We formulate a VC bound with an exact exponential term which exceeds the universal rate $A = 2$ as soon as the range J does not contain a neighbourhood around the value 1/2. Moreover, the general form of the capacity term has been recovered. A sketch of proof is provided in the **Appendix** while the detailed proof can be found in [10].

Theorem 4.1 *Let $\Gamma \subset \left\{ C \subset \mathbb{R}^d \;:\; C \text{ measurable } \right\}$ a totally bounded family of sets and let p be the μ-critical value of Γ. We shall assume that $p \neq \frac{1}{2}$. There exist some constants M and $K(V)$ such that, if $\epsilon < \min(1-p,p)$ and $n \ge M$,*

$$\rho(\mu,n,\epsilon) \le K(V)\, n^{5V+21} \exp\left\{-nH(p+\epsilon,p)\right\} \;. \qquad (3)$$

However, this result can still be improved in the sense that the capacity term has to be computed more precisely. Combining the previous theorem with a result by Talagrand [5], we formulate the following conjecture.

Conjecture 4.2 (Azencott-Talagrand) *Under the same assumptions, with $n\epsilon^2$ sufficiently large,*

$$\rho(\mu, n, \epsilon) \leq K(V)\,(n\epsilon^2)^{V-1/2} \exp\{-nH(p+\epsilon, p)\} \,. \tag{4}$$

Remark 4.3 *In some particular cases (for instance, if $\Gamma = \{halfspaces\}$), we shall have $V-1$ instead of $V-1/2$ (indeed this is the case for Smirnov statistics).*

Now let us try to refine this conjecture. As we have seen it in the finite case, the worst deviation over the family is expected to be achieved at one of the critical sets. We denote by

$$C_n = \arg\sup_{C\in\Gamma}(\mu_n(C) - \mu(C)) \,, \tag{5}$$

the empirical critical set. We propose to partition the event

$$\Omega(\mu, n, \epsilon) = \left\{ \sup_{C\in\Gamma}(\mu_n(C) - \mu(C)) > \epsilon \right\}$$

with respect to the location of C_n, and particularly, regarding its proximity to some critical set. We assume, for simplicity, that the critical subfamily Γ_p is finite. Thus, for α small enough, we have the following decomposition

$$\begin{aligned}\rho(\mu, n, \epsilon) = &\\ &\sum_{\gamma\in\Gamma_p} \mathbf{Pr}\left\{\Omega(\mu, n, \epsilon) \bigcap (\mu(C_n\Delta\gamma) \leq \alpha)\right\}\\ &\qquad + \mathbf{Pr}\left\{\Omega(\mu, n, \epsilon) \bigcap K(\alpha) \right\} \,,\end{aligned}$$

where $K(\alpha) = \bigcap_{\gamma\in\Gamma_p}\{\mu(C_n\Delta\gamma) > \alpha\}$. As the second term is a residue, we can thus conjecture the following refined estimate of the probability tail, for n sufficiently large,

$$\rho(\mu, n, \epsilon) \sim \left(\sum_{i=1}^{L} K_i\,(n\epsilon^2)^{\beta_i}\right) \exp\{-nH(p+\epsilon, p)\} \tag{6}$$

where the β_i's are complexity exponents related to each critical set and $L = |\Gamma_p|$. These exponents could certainly be linked to some concept of *local VC dimension* characterizing the capacity of the subfamily of sets around a critical element C_i.

5 Validation Through Simulations

One of our motivations in the present study was to obtain effective VC dimension estimations in the spirit of [9]. Indeed, we found that the issue of critical sets needed to be considered and that, moreover, one should provide error computations and confidence intervals on the estimations in order to draw further conclusions. We show that the previous theoretical considerations are necessary in order to face efficiently the simulation part. Our first step is to specify the nature of the concepts we shall be able to measure.

5.1 Distribution-dependent concepts

Classical definitions of VC dimension, growth function and VC entropies are known to be *worst-case* since they are distribution-free quantities. Indeed, one single configuration of points is enough to increase the VC dimension. However, it has been noticed ([6], [7]) that on a particular sample, the *empirical VC dimension* can be much smaller. In our experiments, we assume that the probability distribution underlying the data is known and we want to observe if the predictions of VC theory are confirmed. As we have to adapt the theory taking into account the particular exponential rate, we also have to specify the corresponding particular concepts. Thus, we introduce the notion of *effective VC dimension* in a different manner than in [9].

Definition 5.1 (Effective VC dimension)

$$V_{eff}(\Gamma, \mu) = \frac{\mathbb{E}\log|\mathbf{Tr}(\Gamma, X(n))|}{\log 2} \tag{7}$$

Note that this definition allows non-integer values for the VC dimension.

5.2 Rare-event simulation

In order to validate our conjectures on the analytical expression of the probability tail through simulation, we face the issue of effectively simulating the event

$$\Omega(\mu, n, \epsilon) = \left\{ \sup_{C\in\Gamma}(\mu_n(C) - \mu(C)) > \epsilon \right\} \,.$$

But as a rare event, it hardly never can be observed. Thus, we propose to use a speed-up technique, known as *importance sampling*, in order to carry out reasonable computer experiments. The basic idea is

- to change the distribution underlying the data such that the event can frequently be observed, and then,
- to put some weights on the empirical estimator in order to obtain the proper correction on the numerical value.

First we fix an element C_0 of Γ which is a μ-critical set (*i.e.* $\mu(C_0) = p$). We assume in the sequel that μ is the *uniform* distribution with support $\mathcal{K}$. Consider the distribution ν defined by the exponential change of measure[1]. This change of measure will force most of the trajectory paths to be "near" to the critical element C_0. We denote by U the random variable with distribution ν. We obtain, in our case,

$$\frac{d\mu}{d\nu}(u) = \begin{cases} \dfrac{p}{p+\epsilon} & \text{if} \quad u \in C_0 \,, \\[2ex] \dfrac{1-p}{1-p-\epsilon} & \text{if} \quad u \in \mathcal{K} - C_0 \,. \end{cases}$$

We notice that ν is absolutely continuous with respect to μ and that it puts a mass of $p+\epsilon$ on C_0 (while μ put p). By

[1] This is the same change of measure as the one used in the proof of the Cramér-Chernoff theorems on large deviations of the empirical mean of independent real random variables (see [2]). The distribution ν is also known as the *twisted* distribution.

$U(n)$ we denote the i.i.d. sample of size n with distribution ν. Then the event

$$\Omega(\nu, n, \epsilon) = \left\{ \sup_{C \in \Gamma} (\nu_n(C) - \mu(C)) > \epsilon \right\}$$

is not a rare event. We use the Importance Sampling (IS) estimator,

$$\hat{\rho}(\nu, n, \epsilon) = \frac{1}{M} \sum_{m=1}^{M} Z_m \ , \tag{8}$$

where the Z_m's are independent copies of the same random variable

$$Z = \mathbb{1}_{\Omega(\nu,n,\epsilon)} \cdot W(\nu_n(C_0), p, \epsilon)^n \ , \tag{9}$$

and the weights due to the change of measure

$$W(x, p, \epsilon) = \left(\frac{p}{p+\epsilon} \right)^{nx} \left(\frac{1-p}{1-p-\epsilon} \right)^{n(1-x)} \tag{10}$$

correspond to Radon-Nykodym derivatives. The IS estimator satisfies two important properties:

- $\hat{\rho}(\nu, n, \epsilon)$ is an unbiased estimator of $\rho(\mu, n, \epsilon)$, meaning that $\mathbb{E} Z = \rho(\mu, n, \epsilon)$
- Z possesses the variance reduction property

  $$\frac{\mathbf{Var}(Z)}{\mathbf{Var}(T)} \sim \exp\{-n \, H(p+\epsilon, p)\} \ ,$$

  where $T = \mathbb{1}_{\Omega(\mu,n,\epsilon)}$ is the standard estimator.

In order to control the experimental results we need to go through calibration considerations. Our approach here adopts some rough simplifications but turns out to be efficient in simulation. Denote by

$$\hat{\sigma}_Z^2 = \frac{1}{M-1} \sum_{m=1}^{M} (Z_m - \hat{\rho})^2 \tag{11}$$

the empirical variance estimator, where M is the number of trials. Then, by an approximate use of the Central Limit Theorem, we obtain the order of relative error with high confidence (95%).

$$\frac{|\rho - \hat{\rho}|}{\rho} \simeq \frac{1.96}{\sqrt{M}} \frac{\hat{\sigma}_Z}{\hat{\rho}} \ . \tag{12}$$

In simulations, we achieve a relative error on $\hat{\rho}$ smaller than 10%.

5.3 Application - the effective VC dimension

Our idea in introducing these ideas was to validate through simulations the following exact form of VC bounds

$$\rho(\mu, n, \epsilon) \sim K \, (n\epsilon^2)^{V-1} \, \exp\{-nH(p+\epsilon, p)\} \ , \tag{13}$$

where V shall denote, from now on, the effective VC dimension of the family Γ. As an application, it is then possible to get effective VC dimension estimations. The procedure we develop consists in treating every critical set separately. Indeed, even though the IS estimator is unbiased and should provide a good estimation for a sufficicnetly large number of iterations, a rough application can generate unexpected fluctuations due to the behaviour of the weights.

Simulation protocol

Given Γ, μ, α, we go through the following steps.

1. localization of critical sets

   Practical issue : provide a detailed description of the submanifold of critical sets which is obtained as the set of minimizers of the functional

   $$C \in \Gamma \longrightarrow H(\mu(C) + \epsilon, \mu(C)) \in \mathbb{R} \, .$$

2. for each critical set γ, simulate the event

   $$\left\{ \Omega(\mu, n, \epsilon) \bigcap (\mu(C_n \Delta \gamma) \leq \alpha) \right\}$$

   and compute its measure $\hat{\rho}(\gamma)$.

   Practical issues : compute the supremum over an infinite family Γ of sets and achieve the estimation of very small probabilities.

3. we use the following fit

   $$\hat{\rho}(\gamma) \simeq K_\gamma (n\epsilon^2)^{\beta_\gamma} \exp\{-nH(p+\epsilon, p)\}$$

   to estimate the multiplicative constant K_γ and the complexity exponent β_γ.

   Practical issue : achieve parameter estimation.

Taking care of practical and algorithmic issues

In this exploratory work, we have decided to limitate as much as possible the practical problems by choosing the simplest examples. At this point, we aim at validating the protocol and its basic components which are rare event simulation and parameter estimation. Thus, we shall not consider the problem of localizing critical sets in general. We basically construct the example by choosing the critical sets in the first place. The second and very significant issue is how to compute the supremum over an infinite family of sets. In the one-dimensional and two-dimensional case (at least for the uniform distribution), it is possible to show that there is only a finite number of sets to consider. The main problem is then to control the algorithmic complexity of the procedure and some solutions are explored in [10] using some ideas from the field of computational geometry. The question of estimating very small probabilities has been developped in the subsection **5.2** where the rare event simulation procedure has been explained. The last issue is rather standard since parameter estimation can be achieved by simple linear regression. Indeed, we set

$$\begin{aligned} X &= \log(n\epsilon^2) \\ Y &= nH(p+\epsilon, p) + \log \hat{\rho} \, , \end{aligned}$$

and we use the linear model $Y = \beta X + \log K$. The complexity parameter $\beta = V - 1$ is then obtained as the slope of the linear regression model.

Example

We have experimented this protocol on constrained Smirnov statistics. We consider real-valued random variables and we take

$$\Gamma = \{ \, [0, x] \, : \, x < q \, \} \ ,$$

q	V	δV	K	$\delta K/K$
1.0	1.00	0.22	0.95	23%
0.9	0.99	0.20	0.97	23%
0.8	0.99	0.14	0.95	15%
0.7	1.03	0.14	0.90	16%
0.6	1.06	0.18	0.80	20%
0.5	0.98	0.14	0.67	16%
0.4	0.78	0.09	0.53	10%
0.3	0.68	0.06	0.42	7%
0.2	0.62	0.04	0.33	5%
0.1	0.57	0.04	0.23	4%

Table 1: Effective VC dimension and constant K with confidence intervals

where $q \in [0,1]$. We have $J = [0,q]$. Hence, the critical subset is here reduced to the only element $[0,p]$ where

$$p = \arg\min_{x \in J} H(x+\epsilon, x) \ ,$$

and we have for the theoretical VC dimension $V(\Gamma) = 1$. We have checked the formula

$$\rho(\mu, n, \epsilon) \underset{n\to\infty}{\sim} K\left(n\epsilon^2\right)^{V-1} \exp\left\{-nH(p+\epsilon, p)\right\} \ , \qquad (14)$$

where V designates the effective VC dimension. Note also that, if $q = 1$, this is exactly the one-sided Smirnov statistics, because we then have $K = 1$ and $H(p+\epsilon, p) \simeq 2\epsilon^2$. We have provided an experimental validation of this result and an extension of the formula in the case where $q < 1$. The regression model behaves well on our simulations and we provide a synthesis of experimental results in **Table 1** and **Figures 1, 2**.

Other examples

Our machinery has also been tested on other families (see [10] for further examples). The idea is to consider the cases where there are many critical sets, as for example the family of intervals of size smaller than some fixed p.

6 Conclusions and Open Questions

From the experimental part of the present study, we can draw the following conclusions:

- The VC bound we have conjectured provides an exact formula for the constrained Smirnov statistic.
- Effective VC dimension can take non-integer values strictly smaller than theoretical VC dimension.
- As soon as J contains the value $1/2$, effective and theoretical VC dimensions are the same.

Moreover, we have introduced the concept of local VC dimension which is unavoidable in the design of simulations whenever there are many critical sets. However, the theoretical relevance of such a notion still has to be investigated.

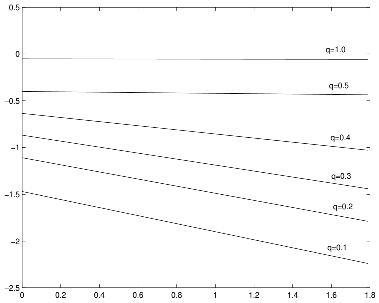

Figure 1: Comparison of the regression lines for various values of q

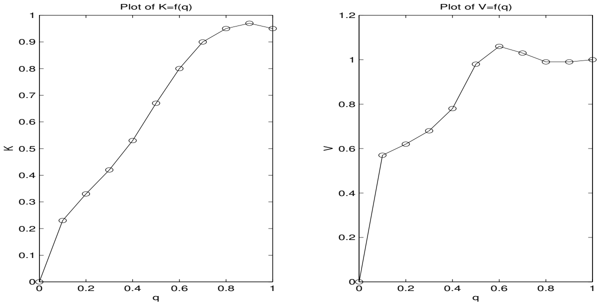

Figure 2: Graph of the experimental functions $K(q)$ and $V(q)$

Appendix - Sketch of proof for Theorem 4.1

Basically, we follow the combinatorial scheme ([8]) as it has been stated by Devroye ([3]) using an adaptive size for the symmetrized sample. We denote by $X(n)$ and $Y(m)$ two i.i.d. samples with distribution μ, respectively of size n and m, and by $\mu_n^{(X)}$, $\mu_m^{(Y)}$ the corresponding empirical measures. We shall set $m \sim n^3$. By symmetrization, we are led to the following probability tail

$$\tau(\mu, n, \epsilon) = \mathbf{Pr}\left\{\sup_{C \in \Gamma}\left(\mu_n^{(X)}(C) - \mu_m^{(Y)}(C)\right) > \epsilon\right\} \ .$$

We then notice that, for a fixed sample $X(n) \times Y(m)$, there is only a finite number of sets C in Γ to be considered. We denote by Γ^* this finite (and random) subfamily of Γ and, using the union-of-events bound, we have

$$\tau(\mu, n, \epsilon) \leq \int \sum_{C \in \Gamma^*} \mathbf{Pr}\left\{\mu_n^{(X')}(C) - \mu_m^{(Y')}(C) > \epsilon\right\} \, d\mu^{\otimes(n+m)} \ ,$$

where the probability under the integral is the probability of a sampling without replacement draw of $X'(n) \times Y'(m)$ out of $X(n) \times Y(m)$. For a fixed set C in Γ^*, we set $r = r(C) = \sum_{i=1}^{n} \mathbb{1}_C(X_i) + \sum_{i=1}^{m} \mathbb{1}_C(Y_i)$. Thus, we have

$$\mathbf{Pr}\left\{\mu_n^{(X')}(C) - \mu_m^{(Y')}(C) > \epsilon\right\} \leq \mathbf{Pr}\left\{\mu_n^{(X')}(C) - \frac{r}{n+m} > \left(\frac{m}{n+m}\right)\epsilon\right\} \ ,$$

and we can then use some combinatorics in order to control the deviations of a sample without replacment draw from its expectation. Thanks to Stirling's formula and some monotonicity arguments concerning the function $H\left(\,\cdot\,, \frac{r}{n+m}\right)$, we can bound the last probability tail by

$$(n+m)^7 \exp\left\{-n\, H\left(\frac{r}{N} + \left(\frac{m}{N}\right)\epsilon, \frac{r}{N}\right)\right\} ,$$

after setting $N = n + m$. Integrating this quantity requires some technical steps which aim at formulating a uniform Varadhan-Laplace estimate. Note that $\frac{m}{N} \simeq 1$ and the dominating term corresponds to $\frac{r}{N} \simeq p$. The basic problem is to control the integral over the rare event which is to observe an empirical frequency $r/(n+m)$ far above the critical value p. Our idea was to indroduce a fixed and finite approximation $\tilde{\Gamma} = \{C_1, ..., C_I\}$ of the family Γ and then to decompose the event $A_i = \{\mu(C\Delta C_i) < \lambda\}$ into $A_i = (A_i \cap K_{\beta,i}) \cup (A_i \cap \overline{K}_{\beta,i})$, where $K_{\beta,i}$ is the open ball of center C_i and radius β for the empirical measure. Note that we have $I \sim \left(\frac{1}{\lambda}\right)^V$. On the fist set, Varadhan's lemma (see [1]) applies, while on the second set, a uniform control of local deviations of empirical means is required. A weaker version of **Theorem 4.1** we have established in [10], based on the work of Talagrand [5], conducts to proper estimations. A final optimization leads to the optimal choice of λ which shall be like $\frac{1}{n}$.

References

[1] R. Azencott. Grandes déviations. In P.L. Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour VIII-1978*, volume 774 of *Lecture Notes in Mathematics*. Springer-Verlag, 1978.

[2] J.A. Bucklew. *Large Deviations Techniques in Decision, Simulation and Estimation*. Wiley, 1990.

[3] L. Devroye. Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, 12:72–79, 1982.

[4] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[5] M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22(1):28–76, 1994.

[6] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[7] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[8] V. N. Vapnik and A. Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

[9] V.N. Vapnik, E. Levin, and Y. Le Cun. Measuring the VC-Dimension of a Learning Machine. *Neural Computation*, 6:851–876, 1994.

[10] N. Vayatis. *Inégalités de Vapnik-Chervonenkis et mesures de complexité*. PhD thesis, Ecole Polytechnique, 2000. In English.

[11] N. Vayatis and R. Azencott. Distribution-Dependent Vapnik-Chervonenkis Bounds. In P. Fischer and H.U. Simon, editors, *Computational Learning Theory*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 230–240, 1999.