Making Vapnik-Chervonenkis bounds accurate

Léon Bottou

Abstract This chapter shows how returning to the combinatorial nature of the Vapnik-Chervonenkis bounds provides simple ways to increase their accuracy, take into account properties of the data and of the learning algorithm, and provide empirically accurate estimates of the deviation between training error and testing error.

1 Introduction

Although the Vapnik-Chervonenkis (VC) learning theory [18, 19, 20, 15, 16] has been justly acclaimed as a major conceptual breakthrough, applying its essential theorems to practical problems often yields very loose bounds. In the case of the pattern recognition problem, the theorems provide distribution–independent uniform bounds on the deviation between the expected classification error and the empirical classification error. Their derivation reveals many possible causes for their poor quantitative performance:

i) Practical data distributions may lead to smaller deviations than the worst possible data distribution.
ii) Uniform bounds hold for all possible classification functions. Better bounds may hold when one restricts the analysis to functions that perform well on plausible training sets.
iii) A symmetrization lemma translates the main combinatorial result into a bound on the deviation between expected and empirical errors. This lemma is a conservative inequality.
iv) The combinatorial characterization of the Vapnik-Chervonenkis capacity is a conservative upper bound.

Léon Bottou
Microsoft Research, 641 Avenue of the Americas, New York, NY, e-mail: leon@bottou.org

v) The union bound $P(\cup_i A_i) \leq \sum_i P(A_i)$ constitutes a critical step of the proof. This bound could be reasonably accurate if the events were independent events with low probability. Nothing guarantee that this is the case.

An apparently different class of bounds, sometimes called sample compression bounds, often provides much more realistic estimates of the generalization error. Such bounds predate the VC theory: for instance, it was mentioned in Paphos that Chervonenkis knew that the expected error of the generalized portrait algorithm is roughly bounded by the fraction of support vectors found in the training set [21, 17]. This bounds depends on the number of support vectors, an empirical quantities measured a posteriori.

The purpose of this contribution is to explore the gap between these two style of bounds using only simple mathematics and a simple empirical case study. This simplicity results from an apparently bold step: instead of assuming that the examples are independently drawn from an unknown distribution, we will reason on random partitions of an arbitrary data set into equally sized training and testing sets. Deviation estimates then result from the combinatorial argument that forms the central part of the traditional Vapnik-Chervonenkis proofs. Avoiding the symmetrization lemma (see point *iii* above) also provides a simple way to obtain data– and algorithm–dependent bounds (points *i* and *ii*) and to define empirical data– and algorithm–dependent capacities (point *iv*) [3, 4, 24]. The union bound (point *v* above) then remains the main source of quantitative problems.

Although refined bounding techniques have been proposed to address all these issue [6, 8, 12, 7, 5, 13], their sophistication often obscures their connection to the practical reality. We believe that the simple framework described in this contribution provides useful intuitions.

The following discussion is organized as follows. After presenting the random split paradigm, we explain how to easily derive bounds in the style of Vapnik-Chervonenkis and make them take into account the specificities of the data distribution and of the learning algorithm. We then estimate the performance of these bounds on a simple case study and show that more refinements are necessary to obtain a bound with a reasonable amount of computation.

2 Setup

Let $Q(z,w)$ be a loss function that measures the correctness on sample z of the answer produced by a learning machine parameterized by $w \in \mathscr{F}$. In this paper we only consider the case of binary loss functions that take the value one if the answer is wrong and zero if it is correct. For instance, in the case of a pattern recognition system, each sample z is a pair (x,y) composed pattern x and a class label y. Given a classifier $f_w(x)$ parametrized by w, the loss function $Q(z,w)$ is zero when $f_w(x) = y$ and is one otherwise.

Let S be a set of 2ℓ labeled examples $z_1, \cdots, z_{2\ell}$. There are $C_{2\ell}^{\ell}$ ways to split this set into equally sized training and testing sets, S_1 and S_2, containing each ℓ

examples. For each choice of a training set S_1 and a test set S_2, and for each value of the parameter w, we define the training error ν_1, the test error ν_2 and the total error ν as:

$$\nu_1(w) = \frac{1}{\ell} \sum_{z \in S_1} Q(z,w) \ , \qquad \nu_2(w) = \frac{1}{\ell} \sum_{z \in S_2} Q(z,w)$$

$$\nu(w) = \frac{1}{2\ell} \sum_{z \in S} Q(z,w)$$

Consider a deterministic learning algorithm $\mathscr{A}$ that processes the training set S_1 and produces a parameter w^{S_1}. This parameter value usually performs well on the training set S_1 in the hope that it will also perform well on the testing set S_2. For instance, the empirical risk minimization principle suggests to design an algorithm that minimizes $\nu_1(w)$ in the hope to ensure that $\nu_2(w^{S_1})$ is small.

All results presented here concern the distribution of the deviation between the training error $\nu_1(w^{S_1})$ and the testing error $\nu_2(w^{S_1})$ when one considers all possible splits $S_1 \cup S_2$ of the dataset S and obtain w^{S_1} by running the learning algorithm $\mathscr{A}$,

$$Pr\left\{ \ \left|\nu_2(w^{S_1}) - \nu_1(w^{S_1})\right| > \varepsilon \ \right\} . \tag{1}$$

The notation $Pr(\mathscr{H})$ denotes the ratio of the number of splits for which condition $\mathscr{H}$ is true over the total number $C_{2\ell}^{\ell}$ of possible splits $S_1 \cup S_2$ of the dataset S. We use this notation instead of the traditional probability notation to emphazise the purely combinatorial nature of this problem.

We argue that the real life behavior of learning algorithms is well characterized by the tail of this distribution. Thousands of machine learning papers are in fact supported by experimental studies that follow the same protocol: randomly holding out testing data, applying the learning algorithm to the remaining data, and assessing its performance on the testing data. A good testing set performance is widely accepted as convincing evidence supporting the use of a specific learning algorithm for a specific learning problem. Bounding the tail of the distribution (1) provides as strong an evidence.

In contrast, traditional statistical approaches of the learning problem assume that the training examples are drawn independently from an unknown distribution. The expected error $\mathbb{E}(Q(z,w^{S_1}))$ then represents the future performance of the system on new examples drawn from this same distribution. Bounding the difference between the training error and the expected error provides a stronger guarantee because the assumed existence of the ground truth distribution provides a means to reason about the algorithm performance on arbitrarily large training sets. Consider for instance a binary classification algorithm that relies on a polynomial discriminant function whose degree grows linearly with the number of training examples. Running such an algorithm on a training set S_1 of a sufficiently small size ℓ could conceivably give acceptable performance on the testing set S_2 of the same size. However this acceptable performance does not guarantee that running the algorithm on all 2ℓ available examples would not overfit.

Avoiding the ground truth assumption is attractive for philosophical reasons. Although epistemology frequently relies on the idea that the world is ruled by simple universal truths waiting to be uncovered, it can be argued that the only thing that is available to us for sure is the finite set of examples. From this point of view, the ground truth distribution is a metaphysical concept because there is no statistical test to determine whether or not our dataset is composed of independent and identically distributed examples and no hope to identify their distribution.

Avoiding the ground truth assumption is also attractive for technical reasons. Working with the combinatorial distribution (1) affords simple ways to derive tail bounds that leverage specific properties of the data or of the learning algorithm.

3 Misclassification Vectors

For each value of the parameter w, the loss function $Q(z,w)$ maps the full set of examples S onto a binary vector $q(w) = (Q(z_1,w),\dots,Q(z_n,w))$ of length 2ℓ that we shall call *misclassification vector*. The ordering of its coefficients does not depend on the dataset split: the i-th component of $q(w)$ indicates whether the learning system parametrized by w processes the example z_i incorrectly, regardless of its appartenance to either the training set or the testing set.

The misclassification vector $q(w)$ encapsulates all that we need to know about the performance of the system parametrized by vector w. Parameter values that lead to the same misclassification vector will also lead to the same total error, training error, and the testing error. Therefore we often write them as $\nu(q)$, $\nu_1(q)$ and $\nu_2(q)$ instead of $\nu(w)$, $\nu_1(w)$ and $\nu_2(w)$.

The function $\eta(q,\varepsilon) = Pr\{|\nu_2(q)-\nu_1(q)| > \varepsilon\}$ does not depend on the ordering of the coefficients in the misclassification vector q because all possible splits are considered and because the quantities $\nu_1(q)$ and $\nu_2(q)$ do not depend on the ordering of the coefficients within each subset. We therefore write $\eta(q,\varepsilon) = \eta(\ell,\nu(q),\varepsilon)$.

Consider an urn containing $2\nu\ell$ red marbles and $2(1-\nu)\ell$ white marbles. Out of the $C_{2\ell}^{\ell}$ possible ways to draw ℓ marbles without replacement, there are exactly $C_{2\nu\ell}^{k} C_{2(1-\nu)\ell}^{\ell-k}$ ways to draw exactly k red marbles. The analytic expression of $\eta(\ell,\nu,\varepsilon)$ is obtained by summing this quantity for all values of k that ensure that the difference between the number k of red marbles drawn from the urn and the number $2\nu\ell - k$ of red marbles left in the urn exceeds $\ell\varepsilon$:

$$\eta(\ell,\nu,\varepsilon) = \frac{1}{C_{2\ell}^{\ell}} \sum_{2|\nu\ell-k|>\ell\varepsilon} C_{2\nu\ell}^{k} C_{2(1-\nu)\ell}^{\ell-k} \tag{2}$$

There are efficient numerical methods for computing this *hypergeometric tail* [14].

Since the function $\eta(\ell,\nu,\varepsilon)$ is monotonically decreasing with ε, we define the inverse function $\varepsilon(\ell,\nu,\eta)$ and write

$$\forall q \quad Pr\{\ |\nu_2(q)-\nu_1(q)| > \varepsilon(\ell,\nu(q),\eta)\ \} \ = \ \eta\,. \tag{3}$$

Although there is no known analytic form for the inverse function $\varepsilon(\ell,\nu,\eta)$, its exact values can be directly read from a table of its inverse $\eta(\ell,\nu,\varepsilon)$. This function is also well described by relatively accurate bounds and approximations such as those derived by Vapnik and Chervonenkis [15, inequality A5, page 176].

$$\varepsilon(\ell,\nu,\eta) \leq \sqrt{4\left(\nu+\frac{1}{2\ell}\right)\left(1-\nu+\frac{1}{2\ell}\right)\frac{\log(2/\eta)}{\ell+1}} \tag{4}$$

$$\approx \sqrt{\frac{4\,\nu(1-\nu)\log(2/\eta)}{\ell}}\,. \tag{5}$$

4 Data– and Algorithm–Independent Bounds

Let $\Delta_{\mathscr{F}}(S) = \{q(w) : w \in \mathscr{F}\}$ be the set of misclassification vectors associated with all potential values of the parameter w. Bounds on the deviation (1) are then derived from the following chain of inequalities.

$$\begin{aligned}
&Pr\left\{\ \left|\nu_2(w^{S_1})-\nu_1(w^{S_1})\right| > \varepsilon(\ell,\nu(w^{S_1}),\eta)\ \right\}\\
&\quad = Pr\left\{\ \left|\nu_2(q^{S_1})-\nu_1(q^{S_1})\right| > \varepsilon(\ell,\nu(q^{S_1}),\eta)\ \right\}\\
&\quad \leq Pr\left\{\ \exists q\in\Delta_{\mathscr{F}}(S)\ :\ |\nu_2(q)-\nu_1(q)| > \varepsilon(\ell,\nu(q),\eta)|\ \right\}\\
&\quad \leq \sum_{q\in\Delta_{\mathscr{F}}(S)} Pr\left\{\ |\nu_2(q)-\nu_1(q)| > \varepsilon(\ell,\nu(q),\eta)\ \right\}\ =\ \eta\,\mathrm{Card}\,\Delta_{\mathscr{F}}(S)\,.
\end{aligned} \tag{6}$$

The first inequality above majorizes (1) by a uniform bound. The second inequality is an application of the union bound $Pr(A\cup B) \leq Pr(A)+Pr(B)$, and the final result is obtained by applying equation (3).

Traditional data– and algorithm–independent deviation bounds control $\varepsilon(\ell,\nu,\eta)$ by the more convenient expression (4) and then invoke the landmark combinatorial lemma of Vapnik and Chervonenkis [18, theorem 1], which states that $\mathrm{Card}\,\Delta_{\mathscr{F}}(S)$ is either equal to $2^{2\ell}$ or smaller than $(2\ell e/h)^h$ for some positive integer h that does not depend on the data S and is now called the VC-dimension of the family of indicator functions $\{\,z \mapsto Q(w,z)\ :\ w\in\mathscr{F}\,\}$. Simple algebraic manipulations then yield data– and algorithm–independent bounds for both the absolute and the relative deviation.

$$Pr\left\{\ \left|\nu_2(w^{S_1})-\nu_1(w^{S_1})\right| > \sqrt{\frac{h(1+\log\frac{\ell}{h})-\log\frac{\eta}{2}}{\ell-1}}\ \right\} \leq \eta\ ,$$

$$Pr\left\{\ \frac{\left|\nu_2(w^{S_1})-\nu_1(w^{S_1})\right|}{\sqrt{\nu(w^{S_1})+\frac{1}{2\ell}}} > 2\sqrt{\frac{h(1+\log\frac{\ell}{h})-\log\frac{\eta}{2}}{\ell}}\ \right\} \leq \eta\ .$$

5 Data– and Algorithm–Dependent Bounds

There are several obvious ways to make these bounds tighter. Instead of using the bound (4), we can tabulate the exact values of $\varepsilon(\ell,\nu,\eta)$ as suggested in section 3. Instead of bounding $\mathrm{Card}\,\Delta_{\mathscr{F}}(S)$, we can design empirical procedures to measure its value [22, 3]. The only remaining causes of inaccuracy are then the two inequalities appearing in the derivation (6), namely the uniform bound and the union bound.

The first source of concern is the majorization of the error deviation by a uniform bound. Many elements of $\Delta_{\mathscr{F}}(S)$ are misclassification vectors that no reasonable learning algorithm would produce. Realistic learning algorithms tend to produce solutions that perform well on the training examples and also contain critical capacity control aspects. For instance one can argue that multilayer network training often achieve good performance because their poor optimization algorithm is unable to find solutions far away from the initialial point. All these aspects severely restricts the set of misclassification vectors.

Therefore, instead of considering the set $\Delta_{\mathscr{F}}(S)$ of the misclassification vectors associated with all potential parameter $w \in \mathscr{F}$, we can consider the set $\Delta_{\mathscr{A}}(S)$ of the misclassification vectors associated with the parameters produced by applying algorithm $\mathscr{A}$ to all training set S_1 extracted from data set S:

$$\Delta_{\mathscr{A}}(S) = \left\{\ q(\mathscr{A}(S_1))\ \forall S_1 \subset S \text{ s.t.} \operatorname{Card}(S_1) = \ell\ \right\} .$$

Replicating the derivation (6) leads to a data– and algorithm–dependent deviation bound,

$$Pr\left\{\ \left|\nu_2(w^{S_1}) - \nu_1(w^{S_1})\right| > \varepsilon(\ell, \nu(w^{S_1}), \eta)\ \right\} \leq \eta\ \operatorname{Card}\Delta_{\mathscr{A}}(S)\ . \qquad (7)$$

The second source of concern is the union bound which, in (6), majorizes the probability of the union of K events $A_1 \ldots A_K$ of probability η by the sum $K\eta$ of their probabilities. Let us tentatively assume that the events A_i can be considered pairwise independent. We can then write

$$K\eta - Pr(\cup_k A_k)\ \leq\ \sum_{i<j} Pr(A_i \cap A_j)\ \approx\ \frac{K^2}{2}\eta^2 \qquad (8)$$

and show that the majorization error is much smaller than $K\eta$. The deviation bound (7) is likely to be quite accurate if this assumption holds. Whether this is true will be clarified in section 7.

6 Empirical Study

In order to illustrate the performance of bound (7), we report on a simple experimental study using 1000 examples of MNIST handwritten digit recognition dataset [2].

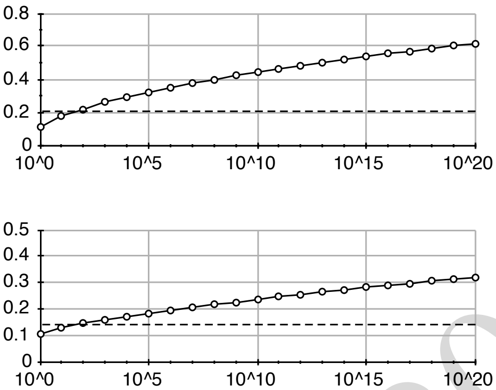

Fig. 1 Bounds on the median relative deviation (top) and median testing error ν_2 (bottom) as a function of $\mathrm{Card}\,\Delta_{\mathscr{A}}(S)$. The dotted line indicates the observed values.

The classifier is a medium-size convolutional network Lenet5 described in [10] with 60,000 adjustable parameters. Training is performed using mean square error back-propagation with learning rates periodically adjusted by estimating the diagonal of the Hessian matrix [11]. This case study should be viewed as a typical example of multilayer neural network training technology using a proven implementation. In particular, this learning algorithm should not be seen as a simple empirical risk minimization algorithm because the cost function is nonconvex and because the first-order nature of the algorithm favors solutions that are relatively close to the initial conditions.

We train this classifier on 1000 random splits of the examples into equally sized training and testing sets containing $\ell = 500$ examples each. We always use the same weight initialization. The observed median training error, median testing error and median relative deviation are, respectively,

$$\mathrm{Median}\left[\nu_1(w^{S_1})\right] \approx 0.075\,, \qquad \mathrm{Median}\left[\nu_2(w^{S_1})\right] \approx 0.14\,,$$

$$\mathrm{Median}\left[\frac{|\nu_2(w^{S_1}) - \nu_1(w^{S_1})|}{\sqrt{\nu(w^{S_1})(1-\nu(w^{S_1})}}\right] \approx 0.21\ .$$

The median deviation can also be estimated by setting the right hand side of (7) to 0.5 and using the approximation (5),

$$\mathrm{Median}\left[\frac{|\nu_2(w^{S_1}) - \nu_1(w^{S_1})|}{\sqrt{\nu(w^{S_1})(1-\nu(w^{S_1}))}} - 2\sqrt{\frac{\log(4\,\mathrm{Card}\,\Delta_{\mathscr{A}}(S))}{l}}\right] \overset{?}{\approx} 0 \qquad (9)$$

Figure 1 (top plot) shows how the bound on the relative deviation (9) depends on the value $\mathrm{Card}\,\Delta_{\mathscr{A}}(S)$. Figure 1 (bottom) plots a corresponding bound on the median testing error ν_2, obtained by setting the training error $\nu_1 = 0.075$ and numerically solving (9) for ν_2 with $\nu = (\nu_1+\nu_2)/2$. Both plots show that $\mathrm{Card}\,\Delta_{\mathscr{A}}(S)$ must be as low as 62 for the bounds to match empirical observations. However these plots also show that values as large as 10^8 still provide reasonable estimates.

In contrast, since it is clear that the VC dimension of such a large multilayer neural network exceeds the total number of examples, $\mathrm{Card}\,\Delta_{\mathscr{F}}(S) = 2^{2\ell} \approx 10^{301}$, leading to a vacuous bound on the median testing error, $\nu_2 \leq 1.25$.

We can attempt to directly measure $\mathrm{Card}\,\Delta_{\mathscr{A}}(S)$ by counting the number $\mathscr{N}_0(t)$ of distinct misclassification vectors seen after training the classifier on t random splits. Such an attempt was unsuccessful because we lack the computing resources to process a large enough number of splits. We stopped after processing $18,000$ random splits and producing $18,000$ distinct misclassification vectors. Birthday problem considerations [1] show that $\mathrm{Card}\,\Delta_{\mathscr{A}}(S) > 10^8$ with confidence greater than 80%. As illustrated in Figure 1, even such large values of $\mathrm{Card}\,\Delta_{\mathscr{A}}(S)$ can still lead to reasonable estimates, within a factor two of the observed deviations.

Since directly counting $\mathrm{Card}\,\Delta_{\mathscr{A}}(S)$ is computationally too expensive, we must design simpler empirical procedures to characterize the properties of the set $\Delta_{\mathscr{A}}(S)$ of misclassification vectors.

7 Coverings

The solid curve in figure 2 shows the histogram of the Hamming distances measured between the misclassification vectors associated with pairs of random splits. This histogram shows a very peaky distribution. We can accurately determine the location of this peak by processing a moderate number of pairs. All our misclassification vectors appear to be located at or around Hamming distance 75 of each other.

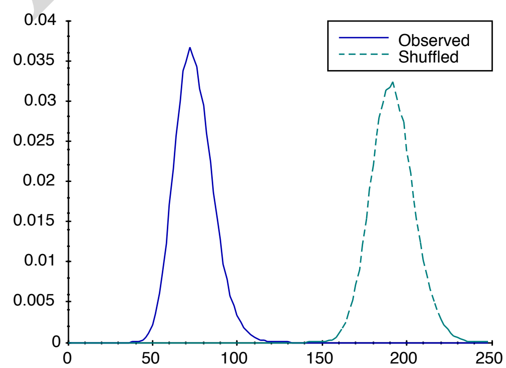

Fig. 2 Histogram of Hamming distances between misclassification vectors. The solid curve shows the histogram of the Hamming distances separating random pairs of misclassification vectors. The dashed curve shows what this histogram would have been if the coefficient of the misclassification vectors were independently sampled from a Bernoulli distribution.

It is well known that the distribution of the Hamming distance separating two d-dimensional binary vectors follows a very peaky distribution centered on $2dp(1-p)$

where p is the probability of nonzero coefficients [9]. The dotted curve figure 2 represents the histogram obtained by randomly shuffling the coefficient of each misclassification vectors before computing the Hamming distances. This curve verifies the theoretical prediction with a peak centered at $4\ell\nu(1-\nu) \approx 180$. The actual misclassification vectors $q(w^{S_1})$ therefore appear considerably less dispersed than random binary vectors. This observation invalidates the independence assumption that could have given us confidence in the accuracy of the union bound (8).

This peaky Hamming distance distribution suggests to characterize the set $\Delta_{\mathscr{A}}(S)$ of misclassification vectors using covering numbers. Let $B_c(q)$ represent a Hamming ball of radius c centered on q. The covering number $\mathscr{N}_c(\Delta)$ is the smallest number of Hamming balls of radius c necessary to cover the set Δ:

$$\mathscr{N}_c(\Delta) \;=\; \min_{C \subseteq \Delta} \; \mathrm{Card}(C) \quad \text{such that} \quad \Delta \subseteq \bigcup_{q \in C} B_c(q)\;.$$

Let us consider an arbitrary split of the data set into training and testing sets and assume that there exists $q' \in B_c(q)$ such that $|\nu_2(q') - \nu_1(q')| > \varepsilon$. A simple derivation then establishes that $|\nu_2(q) - \nu_1(q)| > \varepsilon - c/\ell$.

Combining this observation with (3) gives

$$\forall q \quad Pr\left\{ \exists q' \in B_c(q) \;:\; \left|\nu_2(q') - \nu_1(q')\right| > \frac{c}{\ell} + \varepsilon(\ell, \nu(q), \eta) \right\} \;=\; \eta\,,$$

and a chain of inequality similar to (6) gives

$$Pr\left\{ \left|\nu_2(w^{S_1}) - \nu_1(w^{S_1})\right| > \frac{c}{\ell} + \varepsilon(\ell, \nu(w^{S_1}), \eta) \right\} \;\leq\; \eta\, \mathscr{N}_c(\Delta_{\mathscr{A}}(S))\,. \qquad (10)$$

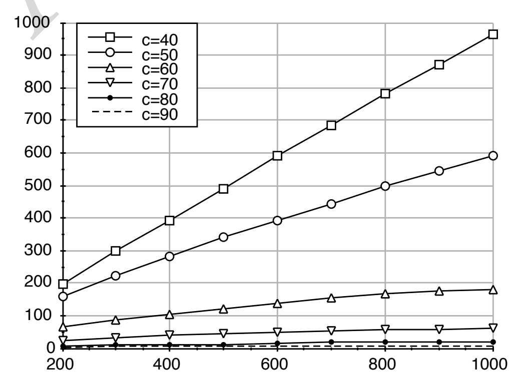

Fig. 3 Empirical covering sizes. Each curve shows how many Hamming balls (of size 40 to 100) are needed to cover the misclassification vectors obtained using the number of splits specified on the X axis. These curves should reach the corresponding convering number when the number of splits increases to infinity.

We construct coverings with the following greedy algorithm. Let $q_1, q_2, \ldots$ be the misclassification vectors associated with successive random splits of our dataset. We construct a covering C_t of the first t vectors using the following recursive procedure:

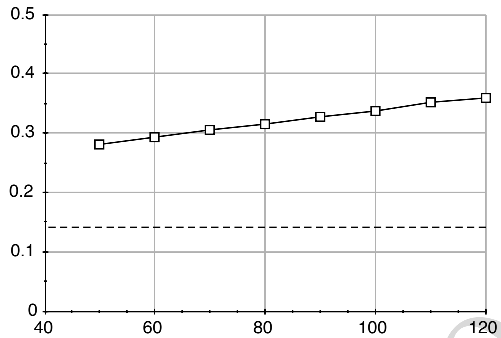

Fig. 4 Covering-based bounds on the median testing error $\nu_2(q^{S_1})$ as a function of the Hamming ball radius c. The dotted line indicates the observed median testing error.

if q_t belongs to one of the Hamming balls centered on an element of C_{t-1}, we set $C_t = C_{t-1}$, otherwise we set $C_t = C_{t-1} \cup \{q_t\}$.

This empirical covering size $N_c(t) = \mathrm{Card}(C_t)$ should converge to an upper bound on $\mathcal{N}(\Delta_{\mathscr{A}}(S))$ when t increases. Figure 3 plots the empirical covering sizes for several values of the Hamming ball radius c. When the radius is smaller than peak of the Hamming distance histogram, this convergence cannot be observed in practice. When the radius is larger than the peak, $\mathcal{N}_c(t)$ converges to a small value.

In the intermediate regime, the empirical covering size appears to converge but its limit is hard to determine. We can work around this difficulty by writing

$$Pr\left\{ \ \left|\nu_2(w^{S_1}) - \nu_1(w^{S_1})\right| > \frac{c}{\ell} + \varepsilon(\ell, \nu(w^{S_1}), \eta) \ \right\} \ \le \ \eta\, \mathcal{N}_c(T) + Pr(R_T)\ , \quad (11)$$

where $R_t \subseteq \Delta_{\mathscr{A}}(S)$ denotes the set of misclassification vectors that are not covered by any of the Hamming balls centered on the elements of C_T. Let $q_{t+1}, \dots, q_{t+s}$ denote the longest sequence of misclassification vectors such that $C_{t+s} = C_t$. None of these vectors belongs to R_t. Since they are obtained by considering random splits independent from the t previous random splits, the probability that none of this vectors belongs to R_t is $(1 - Pr(R_t))^s$. We can therefore write with confidence $1 - \varepsilon$ that $Pr(R_T) \le Pr(R_t) \le 1 - \sqrt[s]{\varepsilon} \le -\log(\varepsilon)/s$. Empirical covering sizes $\mathcal{N}_c(T)$ were collected for $T = 10{,}000$. They range from $N_{120}(10000) = 1$ to $N_{50}(10000) = 3317$. We cannot ensure that $Pr(R_T)$ is small enough when $c < 50$.

Setting the right-hand side of (11) to 0.5, using approximation (5), and solving for $\nu_2(w^{S_1})$ yields a bound on the median testing error. Figure 4 plots this bound as a function of the Hamming ball radius c. Although their empirical accuracy is far from ideal, these covering-based bounds are within a factor of two of the observed testing error. This is clearly better than the vacuous bounds usually afforded by the data– and algorithm–independent bounding technique.

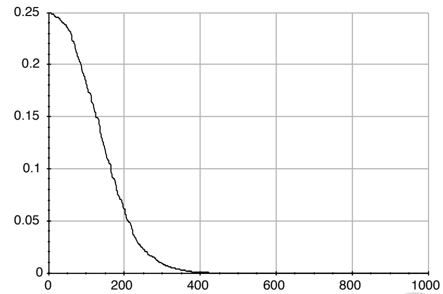

Fig. 5 Empirical variance of the loss function. Only a fraction of the examples z_i have losses $Q(z_i, w^{S_1})$ that vary from one split to the next. The other examples are either always correctly classified or always misclassified.

8 Discussion

There is still a significant margin to improve the accuracy of these empirical bounds. The most interesting effect revealed by our empirical study certainly is the low dispersion of the misclassification vectors (figure 2) because it implies that the union bound is very inaccurate. Although relying on empirical covering numbers should in principle reduce the negative impact of this low dispersion, Dudley's chaining technique [6, 13] is a much more refined way to improve on the union bound. Vorontsov's recent work [23] is therefore very interesting because it leverages a more refined characterization of the distribution of misclassification vectors in a manner related to Dudley's chaining.

It is also interesting to investigate the cause of the low dispersion of the misclassification vectors. The observed Hamming distance histogram (figure 2) looks strikingly like the Hamming distance histogram separating random binary vectors of lower dimensionality. Could it be that only a subset of the examples are responsible for the misclassification vector variations? This would mean that most of the examples are always correctly recognized (or misrecognized when their label is incorrect) regardless of the dataset split. This hypothesis is confirmed by figure 5 which plots the observed variance of the loss $Q(z_i, w^{S_1})$ for all examples z_i ordered by decreasing variance. This observation is interesting because it established a connection with sample compression bounds: the only examples that matter are those that switch from being correctly classified to being misclassified when one changes how the data is split into training and testing sets. The connection between capacity and compression therefore appears to be a manifestation of the subtleties of the union bound.

Finally, one of the main criticisms against the approach outlined in this paper is its computational requirement. Why spend time characterizing the set of misclassification vectors to produce a mediocre bound on the testing error while a fraction of this time is sufficient to compute the testing error itself? This is a valid criticism of this work as an empirical measuring technique. However this work also has value because it helps us understand the subtleties of the learning mathematics. Measuring and understanding are two equally important aspects of the scientific approach.

Acknowledgements This work originates in long discussions held in the mid nineties with my AT&T Labs colleagues Olivier Bousquet, Corinna Cortes, John Denker, Isabelle Guyon, Yann LeCun, Sara Solla, and Vladimir Vapnik. My interest was revived in Paphos by Konstantin Vorontsov and Vladimir Vovk. I would like to thank Vladimir Vovk for convincing me to write it up and Matus Tegarlsky for suggesting to use the birthday problem to lower bound $\mathrm{Card}\,\Delta_{\mathscr{A}}$ using empirical evidence.

References

1. Bloom, D.: A birthday problem. American Mathematical Monthly **80**, 1141–1142 (1973)
2. Bottou, L., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Jackel, L.D., Le Cun, Y., Muller, U.A., Säckinger, E., Simard, P., Vapnik, V.N.: Comparison of classifier methods: a case study in handwritten digit recognition. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, vol. 2, pp. 77–82. IEEE, Jerusalem (1994)
3. Bottou, L., Cortes, C., Vapnik, V.N.: On the effective VC dimension. Tech. Rep. bottou-effvc.ps.Z, Neuroprose (ftp://archive.cis.ohio-state.edu/pub/neuroprose) (1994). URL http://leon.bottou.org/papers/bottou-cortes-vapnik-94
4. Bottou, L., Le Cun, Y., Vapnik, V.N.: Report: Predicting learning curves without the ground truth hypothesis (1999). URL http://leon.bottou.org/papers/bottou-lecun-vapnik-1999
5. Bousquet, O.: Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. Ph.D. thesis, Ecole Polytechnique (2002)
6. Dudley, R.M.: The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. Journal of Functional Analysis **1**(3), 290–330 (1967)
7. Dudley, R.M.: Uniform Central Limit Theorems. Cambridge University Press, Cambridge, UK (1999)
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**, 217–232 (1995)
9. Kanerva, P.: Sparse Distributed Memory. The MIT Press (1988)
10. Le Cun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient based learning applied to document recognition. Proceedings of IEEE **86**(11), 2278–2324 (1998)
11. Le Cun, Y., Bottou, L., Orr, G.B., Müller, K.R.: Efficient backprop. In: Neural Networks, Tricks of the Trade, Lecture Notes in Computer Science LNCS 1524. Springer Verlag (1998)
12. Shawe-Taylor, J., Bartlett, P., Williamson, R., Anthony, M.: Structural risk minimization over data-dependent hierarchies. IEEE Transactions on Information Theory **44**(5), 1926–1940 (1998)
13. Talagrand, M.: The Generic Chaining: Upper and Lower Bounds of Stochastic Processes. Springer (2005)
14. Trong Wu: An accurate computation of the hypergeometric distribution function. ACM Transactions on Mathematical Software **19**(1), 33–43 (1993)
15. Vapnik, V.N.: Estimation of Dependences based on Empirical Data. Springer Series in Statistics. Springer Verlag, Berlin, New York (1982)
16. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
17. Vapnik, V.N., Chervonenkis, A.Y.: A note on one class of perceptrons. Automation and Remote Control **25** (1964)
18. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Proceedings of the USSR Academy of Sciences **181**(4), 781–783 (1968). English translation: Soviet Math. Dokl., 9:915-918, 1968
19. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16**(2), 264–281 (1971)
20. Vapnik, V.N., Chervonenkis, A.Y.: Theory of Pattern Recognition. Nauka (1974). German translation: Akademie–Verlag, Berlin, 1979

21. Vapnik, V.N., Lerner, A.Y.: Pattern recognition using generalized portrait method. Automation and Remote Control **24**, 774–780 (1963)
22. Vapnik, V.N., Levin, E., LeCun, Y.: Measuring the VC-dimension of a learning machine. Neural Computation **6**(5), 851–876 (1994)
23. Vonrontsov, K.V.: Exact combinatorial bounds on the probability of overfitting for empirical risk minimization. Pattern Recognition and Image Analysis. (Advances in Mathematical Theory and Applications **20**(3), 269–285 (2010)
24. Vorontsov, K.V.: Combinatorial substantiation of learning algorithms. Computational Mathematics and Mathematical Physics **44**(11), 1997–2009 (2004)

Rejoinder: Making V.-C. bounds accurate

Léon Bottou

I am very grateful to my colleagues Olivier Catoni and Vladimir Vovk because their insightful comments add considerable value to my article.

Olivier elegantly points out out how similar conclusions can be achieved with a PAC-Bayesian approach. He convinced me to try filling my knowledge gap by reading parts of his excellent monograph [2]. The introductory material of [1] also provides a broad overview of the connections between PAC-Bayesian bounds and more traditional empirical process bounds. I find instructive to observe how the same fundamental phenomena can be discussed from a purely combinatorial viewpoint (as in my text) or from a purely probabilistic approach (as in Olivier's comment.)

Besides providing a beautiful connection between sample compression bounds and conformal prediction, Vladimir raises two issues that I should have discussed much more precisely in the first place. The first issue focuses on the level of data dependence for learning bounds. Four successive data dependence levels make the bounds potentially more accurate and also less useful for predicting the risk because they depend on quantities that have not been observed at the time of the prediction. Since combinatorial bounds belong to the last category ("*data super-dependence*"), they are not very useful to predict the expected risk. The second issue raises questions about the exact difference between the exchangeability assumption and the i.i.d. assumption. These two issues are in fact intimately connected.

De Finetti's theorem characterizes exchangeable sample distributions as *mixtures* of i.i.d. distributions. Such mixtures are usually not i.i.d. distributions themselves. Consider for instance a sample of k real numbers drawn from a equal mixture of normal distributions centered in two distinct points $x, y \in \mathbb{R}$. The expected sample mean is of course $(x+y)/2$. However, regardless of k, one half of the samples has an empirical mean close to x and the other half has an empirical mean close to y. We have exchangeability but the law of large numbers does not apply.

Such a situation is far from unrealistic. Every data collection campaign is in practice corrupted by uncontrolled variables that can be viewed as latent mixture

---

Léon Bottou
Microsoft Research, 641 Avenue of the Americas, New York, NY, e-mail: leon@bottou.org

variables. Despite this, the combinatorial error bounds accurately describe what can be observed when one splits the data into training set and testing set. One cannot expect these same bounds to predict the expected error because it is impossible to construct such a prediction without additional assumption (such as independence assumptions). This is why, in practice, gathering representative data consistently remains the hardest part of building a machine learning application.

Finally, I find instructive to question whether predicting the expected risk is the true purpose of learning bounds. Under i.i.d. assumptions, the most accurate "*inductively data-dependent*" way to estimate the expected risk almost always consists of holding out testing data. Held out data affords considerably better confidence intervals; they easily compensate what is lost by reducing the training set size. In fact, it is easy to see that one can match the best learning bounds by holding out a fraction of examples inversely proportional to $\log \mathrm{Card}\, \Omega_{\mathscr{A}}(S)$.

Let us nevertheless imagine a training set so small that we cannot afford to save a few testing examples, and let us also ignore the fact that the resulting learning system will probably perform too poorly to be of any use. Rather than using a learning bound, the practitioner would be wise to use a k-fold cross-validation approach and average the predictions of the k learning systems. Under the appropriate convexity conditions, this ensemble should perform at least as well as the average of the errors estimated on each fold.

Why then are we devoting considerable efforts to construct more accurate learning bounds? The history of our field provides an easy answer: building more accurate learning bounds forces us to describe new phenomena and acquire new insights. These insights are often useful to inspire and to characterize new learning algorithms. Consider for instance the under-dispersion of the error vectors (figure 7.2). If our data super-dependent learning bound cannot be accurate without taking this effect into account, we can expect that accurate risk bounds or efficient learning algorithms should somehow to take this phenomenon into account.

References

1. Audibert, J.Y., Bousquet, O.: PAC-Bayesian generic chaining. In: S. Thrun, L. Saul, B. Schölkopf (eds.) Advances in Neural Information Processing Systems 16, pp. 1125–1132. MIT Press (2004)
2. Catoni, O.: PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, *IMS Lecture Note Monograph Series*, vol. 56. Institute of Mathematical Statistics (2007)