Vapnik-Chervonenkis bounds for generalization

1993 J. Phys. A: Math. Gen. 26 2211

(http://iopscience.iop.org/0305-4470/26/9/016)

J. Phys. A: Math. Gen. 26 (1993) 2211–2223. Printed in the UK

Vapnik–Chervonenkis bounds for generalization

J M R Parrondo[†] and C Van den Broeck[‡]

Department of Chemistry, 0340, University of California, San Diego, La Jolla, CA-92093, USA

Received 25 August 1992

Abstract. We review the Vapnik and Chervonenkis theorem as applied to the problem of generalization. By combining some of the technical modifications proposed in the literature we derive tighter bounds and a new version of the theorem bounding the accuracy in the estimation of generalization probabilities from finite samples. A critical discussion and comparison with the results from statistical mechanics is given.

1. Introduction

Over the past few years, there has been a rising interest in the statistical mechanics of generalization and rule extraction [1–5]. A major reason for this new development is the success with which techniques borrowed from the fields of spin glasses and associative memories can be used to evaluate analytically training and generalization curves for the perceptron and its variants. On the other hand, statistics is the home turf of what generalization is all about: predicting data on the basis of a training set. Traditionally, most of the focus was on parametric statistics which involves assumptions about the form or prior knowledge of the underlying probability distributions. More recently, major progress has been achieved in the context of non-parametric statistics. One of the beautiful results is the so-called Vapnik–Chervonenkis (VC) theorem, establishing bounds for the uniform convergence of frequencies to probabilities for a whole class of events, and this independently of the underlying probability distributions. The theorem was recently applied in the context of generalization for neural network architectures by Baum and Haussler [6].

Even though insightful and didactical introductions to the VC theorem are available in the literature [3, 7, 8], we believe that the more powerful variants of the theorem are not well known, especially in the statistical physics community. The purpose of this paper is to fill this gap, while at the same time adding significant technical refinements to the theorem itself. We also present a streamlined proof, which we believe is clearer and allows technical improvements to be easily incorporated and the derivation of variants of the theorem. In section 2, we give a self-contained presentation of an improved version of the VC theorem, with the proof given in appendix A. In section 3, we introduce a variant of the theorem related, on the one hand, to the teacher–student scenario which has been studied in detail in the statistical mechanics literature and, on the other hand, to the *probably approximately correct* (PAC) learning model [9], the usual approach for 'learnable' problems in the computational science literature. In section 4, we introduce another variant of the

[†] Permanent address: Departamento Física Aplicada I, Universidad Complutense, 28040-Madrid, Spain.

[‡] Permanent address: Limburgs Universitair Centrum, B 3590 Diepenbeek, Belgium.

theorem which combines the features of both previous versions. Finally, in section 5, we conclude with a critical discussion of the practical implementation of the VC theorem, and make a comparison with the results from the statistical physics literature.

2. Review of the Vapnik–Chervonenkis theorem

We first reformulate the VC theorem, in a way that is more convenient to discuss the problem of generalization. Consider the functions f, defined on a domain Ω, and taking values $+1$ or -1

$$\begin{aligned} f : \Omega &\rightarrow \{+1, -1\} \\ \xi &\rightarrow f(\xi). \end{aligned} \tag{1}$$

These functions provide all possible binary classifications of the input or question space Ω.

We select one of these functions as the so-called target function, say the function $\bar{f}$. A quantity of interest is the 'error' probability e_f, defined as the probability that f and $\bar{f}$ give a different classification for a randomly chosen example, $\xi \in \Omega$. By random choice, we mean that the example is chosen according to a probability distribution defined on Ω. Note that e_f depends on the latter probability distribution and on the target function $\bar{f}$ (but we omit explicit reference to this dependence for simplicity of notation). The interest of the quantity e_f is that it tells us how well or rather how badly the function f predicts the outcomes of the target function $\bar{f}$†.

One expects the frequency of errors observed on a finite sample of test examples to be close to the error probability e_f when the size of the test set is large enough. This is the basic link between generalization and statistics: quantifying the convergence of frequencies to probabilities which is, in the context of generalization, nothing but the convergence of the performance over the training set to the actual performance of the net. For example, to evaluate the quantity e_f for a particular function f without an exhaustive comparison over all the possible input patterns, we can invoke Bernouilli's theorem stating that the observed frequency ν_f^m of error, defined as the fraction of times that f and $\bar{f}$ give a different answer on m independently chosen examples, converges to the true probability e_f in the limit of an infinite number of independent trials $m \rightarrow \infty$. The Hoeffding inequality [10, 11] gives an idea of how fast this convergence takes place with increasing test size m‡

$$\mathrm{Prob}[|\nu_f^m - e_f| > \epsilon] \leqslant 2\mathrm{e}^{-2\epsilon^2 m}. \tag{2}$$

† A sometimes useful view of these error probabilities arises when they are interpreted as *distances*. In fact, in the particular case that the input variables are binary, $\Omega = \{-1, +1\}^n$ and if, moreover, the probability distribution on Ω is uniform, i.e. every single of the 2^n input patterns is equally likely, e_f is proportional to the so-called Hamming distance between f and $\bar{f}$:

$$e_f = d(f)/2^n$$

where the Hamming distance between the functions f and $\bar{f}$ is denoted by $d(f)$, $0 \leqslant d(f) \leqslant 2^n$ and is equal to the number of differences in their respective truth table.

‡ Note that $m\nu_f^m$ is a binomial random variable since it is the sum of m independent random variables which take the value 1 with probability e_f and 0 with probability $1 - e_f$. From this point of view the Hoeffding inequality (2) is just a bound for the tail of the binomial distribution, whose main virtue is that it is independent of the value of e_f. For the special case that $e_f = 1/2$, one can derive a better bound, namely the one implicit in equation (6).

Roughly speaking, one concludes that the convergence of frequency to probability goes as $1/\sqrt{m}$, which is actually the familiar type of convergence associated with a central limit theorem.

In the context of generalization, one tries to match the outcomes of an unknown function $\bar{f}$. A natural way to do so is to work with a whole class of candidate functions $\Im$, rather than with a single function, and to check how each of them is performing on the training set. Hopefully, the one function that is selected because it has the smallest observed error rate will have a true probability for error which is equally small. The class of functions that is considered is sometimes called the rule or hypothesis space. For example, $\Im$ could correspond to the class of functions that can be implemented by a specific neural network architecture. It is tempting to invoke the Hoeffding inequality to claim that the true and observed probabilities of error for the selected function will be close to each other. This, however, is incorrect, for the same reason that, for example, the largest of a number of identically distributed random variables has a probability density which is no longer equal to the original density. To illustrate this point further, consider a rule space which contains all the binary classifications. It obviously contains all the functions that match $\bar{f}$ perfectly on the training set, so their test error is zero. However, these functions realize all the possible classifications on the other examples, so that for the latter examples there is no correlation between test error and true error. The correct way to proceed is to bound the maximum deviation between frequencies and corresponding probabilities for the whole class $\Im$ of functions or, more precisely, a bound for the probability

$$\mathrm{Prob}[\max_{f\in\Im} |\nu_f^m - e_f| > \epsilon]. \tag{3}$$

This bound then *a fortiori* applies to any one of the functions that one decides to select.

If the class $\Im$ is finite, say it contains N elements, such a bound can easily be derived from the Hoeffding inequality since the probability of observing a deviation larger than ϵ for at least one of the functions $f \in \Im$ is smaller than the sum of the probabilities of observing such a deviation for every single one of the functions. Hence,

$$\mathrm{Prob}[\max_{f\in\Im} |\nu_f^m - e_f| > \epsilon] \leqslant 2N\mathrm{e}^{-2\epsilon^2 m} \tag{4}$$

and one says that the error frequencies converge 'uniformly' to their corresponding error probabilities. As an example, consider the class of functions that can be implemented by a network with n binary weights. A special case is the class of binary or Ising perceptrons (the weight vector $\boldsymbol{J}$ has components $J_i = +1$ or $-1, i = 1, \ldots, n$). One can apply the above result with $N = 2^n$. In the limit m and $n \to \infty$, with a fixed value of the ratio $\alpha = m/n$, one concludes that with probability one, none of the observed frequencies differs from their corresponding probability by more than the accuracy threshold $\epsilon_{\mathrm{th}} = \sqrt{\ln 2/(2\alpha)}$. Note that this result is independent of the target function $\bar{f}$.

From the above derivation, it would appear that not much can be said about the 'uniform convergence' of frequencies to probabilities for a class $\Im$ of functions with an infinite number of elements. On the other hand, one expects that there exist such classes of functions with a very limited 'classification diversity'. To illustrate this point, consider the class of functions $\Im$ for which at most n input patterns are classified as $+1$. If the number of elements in Ω is infinite, there are an infinite number of such functions. To quantify the 'classification diversity' of this class of functions we introduce the quantity $\Delta(m)$ defined as the maximum number of different classifications which can be induced by the functions $f \in \Im$ on m

examples. Clearly, every possible classification can be induced when $m \leqslant n$. On the other hand, for $m > n$, one can only perform all those classifications in which the number of $+1$ is not larger than n. Consequently one finds

$$\Delta(m) \begin{cases} = 2^m & \text{for } m \leqslant n \\ = \sum_{l=0}^{n} \binom{m}{l} & \text{for } m \geqslant n. \end{cases} \tag{5}$$

An estimate of how the number of classifications is limited for $m > n > 1$ is provided by the following bounds [12–14]

$$\sum_{l=0}^{n} \binom{m}{l} \leqslant 1.5 \frac{m^n}{n!} \leqslant \left(\frac{em}{n}\right)^n. \tag{6}$$

Therefore, instead of the exponentially large total number of classifications 2^m, only a polynomial large number of them is realized for $m > n$.

It was proven by Vapnik and Chervonenkis [15] and independently by Sauer [13] that the above described behaviour of $\Delta(m)$ is, in fact, very general. For every class of functions $\Im$, there exists a unique integer number d_{VC}, called the VC dimension (which is possibly equal to ∞), such that for $m \leqslant d_{\mathrm{VC}}$, all the 2^m classifications can be implemented (for at least one choice of the m examples), while for $m > d_{\mathrm{VC}}$, this is no longer the case. Moreover the above example, with the growth function (5), gives the largest number of classifications that can arise for all the classes with a VC dimension $d_{\mathrm{VC}} = n$. Consequently, given knowledge of the VC dimension, one concludes that (cf equation (6))

$$\Delta(m) \begin{cases} = 2^m & \text{for } m \leqslant d_{\mathrm{VC}} \\ \leqslant [em/d_{\mathrm{VC}}]^{d_{\mathrm{VC}}} & \text{for } m > d_{\mathrm{VC}}. \end{cases} \tag{7}$$

The VC dimension is thus a way to identify classes of functions with a limited scheme of classifications. This limitation in the classification capacity is obviously a necessary characteristic for generalization. As we explained before, for a class that provides all possible classifications there is no correlation between the error frequencies on the training set and the error probabilities.

Several classes of interest with a finite VC dimension have been identified: rectangles and half-planes (perceptrons with no threshold) in $\mathbb{R}^n$ [16], Boolean functions [8], general neural networks [6] and so on.

For such classes, Vapnik and Chervonenkis [15] were able to derive an upper bound for the probability that any of the frequencies ν_m^f differs by more than ϵ from its corresponding 'true' frequency e_f, $\forall f \in \Im$:

$$\mathrm{Prob}[\max_{f \in \Im} |\nu_f^m - e_f| > \epsilon] \leqslant 4\Delta(2m)\mathrm{e}^{-m\epsilon^2/8}. \tag{8}$$

This original bound has been subsequently refined by Vapnik himself [12] who obtained $-m\epsilon^2/4$ in the exponent. Looking for a faster convergence in ϵ for a fixed number of examples m, Devroye [17] succeeded in obtaining the same exponent as in the Hoeffding inequality, i.e. $-2m\epsilon^2$, but at the expense of evaluating the growth function in m^2 instead of m. By a combination of the ideas present in both proofs, we have obtained the following improved VC result (see appendix A):

$$\mathrm{Prob}[\max_{f \in \Im} |\nu_f^m - e_f| > \epsilon] \leqslant c_1 \Delta(2m)\mathrm{e}^{-m\epsilon^2} \tag{9}$$

with $c_1 = 6e^{2\epsilon}$, a constant slightly larger than six. By comparing the VC bounds (8) and (9) and the Hoeffding inequality (2) we note that the proportionality factor $\Delta(2m)$ plays the role of the effective number of elements in the class. On the other hand, we have lost a factor of 2 for the convergence rate in the exponent (but this factor can be restored at the cost of increasing the prefactor, cf [17] and see below).

In the case of a class of functions with a finite VC dimension, the prefactor $\Delta(m)$ only grows like a power law for $m > d_{\mathrm{VC}}$, and we again conclude that the differences between observed frequencies ν_f^m and true probabilities e_f will become uniformly small as $m \to \infty$. This is nicely illustrated by applying the inequality (7) for the growth function and introducing the variable $\alpha = m/d_{\mathrm{VC}}$

$$\mathrm{Prob}[\max_{f\in\Im} |\nu_f^m - e_f| > \epsilon] \leqslant c_1 \exp[-d_{\mathrm{VC}}(\alpha\epsilon^2 - \ln(2\alpha) - 1)]. \tag{10}$$

In many applications, and particularly those considered in the statistical mechanics literature, the regime of interest is the analogue of a *thermodynamic limit* in which $d_{\mathrm{VC}} \to \infty$. In that case the right-hand side of (10) has a sharp behaviour as a function of the accuracy ϵ. There is an accuracy threshold

$$\epsilon_{\mathrm{th}}(\alpha) = \sqrt{(\ln(2\alpha) + 1)/\alpha} \tag{11}$$

above which the right-hand side of (10) vanishes in that limit. Therefore, *with probability one* all the error probabilities e_f lie in an interval of radius $\epsilon_{\mathrm{th}}(\alpha)$ and centred at the corresponding frequency ν_f^m over a test sample of size $m = \alpha d_{\mathrm{VC}}$.

A stronger version of the VC theorem can be obtained using the more general result, valid for any integer m' (cf appendix A)

$$\mathrm{Prob}[\max_{f\in\Im} |\nu_f^m - e_f| > \epsilon] \leqslant c_2 \Delta(m + m') \exp\left[-2m\left(\frac{m'\epsilon}{m+m'}\right)^2\right] \tag{12}$$

with $c_2 = 4e^{4\epsilon mm'/(m+m')^2}$ a constant slightly larger than four. In the limit $m' \gg m$ one recovers the Hoeffding exponent but the argument of the growth function $m + m'$ is also increased. In the above introduced thermodynamic limit it is convenient to set $m' = x d_{\mathrm{VC}}$, so that we can rewrite (12) as

$$\mathrm{Prob}[\max_{f\in\Im} |\nu_f^m - e_f| > \epsilon] \leqslant c_2 \exp\left[-d_{\mathrm{VC}}\left(2\alpha\left(\frac{x\epsilon}{\alpha+x}\right)^2 - \ln(\alpha + x) - 1\right)\right] \tag{13}$$

which is valid for any value of x. In the limit $d_{\mathrm{VC}} \to \infty$, the accuracy threshold becomes

$$\epsilon_{\mathrm{th}} = \min_x \frac{\alpha + x}{x}\sqrt{(\ln(\alpha + x) + 1)/2\alpha}.$$

For α large one can choose $x = r\alpha$ with $1 \ll r \ll \alpha$, finding the following asymptotic result

$$\epsilon_{\mathrm{th}} \simeq \sqrt{\ln\alpha/2\alpha} \tag{14}$$

which is a factor $\sqrt{2}$ sharper than the result given by (11).

3. A VC theorem for learnable rule

Our purpose here is to draw attention to a variant of the VC theorem [6, 8, 18], which allows one to obtain a much stronger bound on the generalization performance. The point is that the convergence of frequencies to probabilities is much faster when the observed frequencies are close to 0, which is precisely the situation of interest in the context of generalization. For example, the probability that the observed frequency ν_f^m for a single function f is found to be equal to 0 while the true probability is larger then ϵ, can be bound as follows:

$$\mathrm{Prob}[\nu_f^m = 0 \text{ and } e_f > \epsilon] = \int_\epsilon^1 \mathrm{d}x\, P_{e_f}(x)(1-x)^m \leqslant (1-\epsilon)^m \leqslant \mathrm{e}^{-\epsilon m} \tag{15}$$

and this inequality is valid for any probability density of the error $P_{e_f}(x)$, i.e. either for a fixed target function $\bar{f}$ and a fixed f or for random choice of both according to some probability distribution on the class $\Im$. We conclude that the error probability e_f decreases as $1/m$ to be compared with the $1/\sqrt{m}$ behaviour observed in the Hoeffding inequality (2).

Consider now the more complicated situation of a class of functions $\Im$ from which one selects all the functions f that score perfectly $\nu_f^m = 0$ on m training examples. These are the so-called compatible functions. The class of these functions will be denoted by $\Im_m^* = \{f | f \in \Im \text{ and } \nu_f^m = 0\}$ and is sometimes called the *version space*. We will assume that this class is not empty; in other words, we are considering the case of a 'learnable' rule. In the statistical physics literature, this student–teacher scenario was introduced by Gardner and Derrida [19] in the context of the binary perceptron. The related PAC approach was introduced by Valiant in computational science literature [9]. The following improved VC bound can be derived following essentially the same steps as those used in the original VC proof (see appendix B for details):

$$\mathrm{Prob}[\max_{f\in\Im_m^*} e_f > \epsilon] \leqslant 2\Delta(m')\left(1+\frac{m}{m'}\right)^{1-\epsilon m'} \tag{16}$$

and it is valid for any integer m'. In particular, for $m' = m$, one finds (compare with (15))

$$\mathrm{Prob}[\max_{f\in\Im_m^*} e_f > \epsilon] \leqslant 4\Delta(m)2^{-\epsilon m}. \tag{17}$$

Note the important improvement over the VC theorem (12) with a factor ϵ rather than ϵ^2 in the exponent in the right-hand side of the inequality.

In order to study the thermodynamic limit m and $d_{\mathrm{VC}} \to \infty$, with a fixed value of the ratio $\alpha = m/d_{\mathrm{VC}}$, we set $m' = x d_{\mathrm{VC}}$ and rewrite equation (16) for $x > 1$ as follows

$$\mathrm{Prob}[\max_{f\in\Im_m^*} e_f > \epsilon] \leqslant 2\left(1+\frac{\alpha}{x}\right)\exp[-d_{\mathrm{VC}}(\epsilon x \ln(1+\alpha/x) - \ln x - 1)]. \tag{18}$$

We conclude that the accuracy threshold is given by

$$\epsilon_{\mathrm{th}}(\alpha) = \min_{x>1} \frac{\ln x + 1}{x \ln(1+\alpha/x)} \tag{19}$$

and all the error probabilities e_f of the compatible functions are smaller than $\epsilon_{\mathrm{th}}(\alpha)$ with probability one in the thermodynamic limit. Note that for α large, one can choose $x = r\alpha$

with $1 \ll r \ll \alpha$ and the error probabilities converge to zero at least as $\epsilon_{\text{th}}(\alpha) \simeq \ln\alpha/\alpha$. This improves the result given in [16] by a factor of $2/\ln 2$ (see also [20], section 10).

Another version of the theorem due to Blumer *et al* [18] shows that the behaviour of ϵ_{th} proportional to $\ln\alpha/\alpha$ persists to a certain extend for unlearnable rules. They derive the following inequality, valid for $0 \leqslant \gamma < 1$,

$$\text{Prob}[\max_{f\in\Im} \theta(\epsilon\gamma - \nu_f^m)e_f > \epsilon] \leqslant 2\Delta(2m)\exp\left[-\frac{\epsilon(1-\gamma)^2 m}{4}\right] \tag{20}$$

where the error probability of the worst function among those which have error frequencies less than $\gamma\epsilon$ is compared with ϵ. Introducing in (20) the bound for $\Delta(m)$ (7), the accuracy threshold ϵ_{th} is found to be proportional to $\ln\alpha/\alpha$.

4. VC theorem for frequency sampling in the version space

In this section we present a version of the VC theorem that provides the convergence conditions for the frequency of errors to the error probability for those functions that result after training, i.e. functions in a given *version space*. The setup is: we first select from the hypotheses space those functions that make no error on a set of m_1 training examples, and determine the error frequencies of the functions in the resulting version space over a sample of m_2 examples. The question that we address here is how these error frequencies converge to their respective error probabilities by bounding the following probability

$$P(m_1, m_2) = \text{Prob}[\sup_{f\in\Im} \delta(\nu_f^{m_1}, 0)|\nu_f^{m_2} - e_f| > \epsilon]. \tag{21}$$

Following identical steps to those in the preceding sections (see appendix C for details), one finds

$$P(m_1, m_2) \leqslant 2\Delta(2m_2)\max_k \exp\left[-\frac{(m_2+1)(\epsilon m_2 - 1)^2}{(k+1)(2m_2-k+1)} - \frac{m_1 k}{2m_2}\right] \tag{22}$$

where k is an integer that runs from 0 to $2m_2$.

In the limits $d_{\text{VC}}, m_1, m_2 \to \infty$ with $\alpha_i = m_i/d_{\text{VC}}$ finite and using the bound (7) for the growth function, the inequality (22) can be rewritten as ($\alpha_1 > 1$ and $\alpha_2 > 1$)

$$P(\alpha_1, \alpha_2) \leqslant 2\exp[-d_{\text{VC}}(f(\alpha_1, \alpha_2) - \ln(2\alpha_2) - 1)] \tag{23}$$

with

$$f(\alpha_1, \alpha_2) = \min_{x\in[0,1]} \left\{\frac{\alpha_2\epsilon^2}{4x(1-x)} + \alpha_1 x\right\}. \tag{24}$$

For $\alpha_1 = 0$ the minimum is attained for $x = 1/2$, and we recover the VC theorem (10). On the other hand for $\epsilon\sqrt{\alpha_2/\alpha_1} \ll 1$, one finds the minimum at $x = \frac{1}{2}\epsilon\sqrt{\alpha_2/\alpha_1}$ and f becomes

$$f(\alpha_1, \alpha_2) \simeq \epsilon\sqrt{\alpha_1\alpha_2}. \tag{25}$$

Combining (23) and (25) one finds the accuracy threshold ($\alpha_1 > 1$ and $\alpha_2 > 1$)

$$\epsilon_{\text{th}} = \ln(2\alpha_2)/\sqrt{\alpha_1\alpha_2}. \tag{26}$$

As in the previous section, it can be seen that the threshold decreases with respect to the one given by the original VC theorem as a consequence of restricting the supremum to the version space induced by the m_1 first examples. When α_1 and α_2 are both large and of the same order of magnitude the threshold value (26) is still valid. Assuming, for example, that $\alpha_1 = \alpha_2 = \alpha$, i.e. that we use the first half of a set of examples to train the network and the second half to test its performance, we recover the familiar asymptotic behaviour $\ln\alpha/\alpha$, but having in mind that this accuracy threshold bounds the distance between the true error probability and the frequency measured in the second half of the examples.

5. Discussion

Explicit analytic results have been obtained in the statistical physics literature for the error curves in function of α for several variants of the perceptron and for various training schemes [2,4,5,19,21–28]. For a learnable rule (i.e. the target function is an element of the hypothesis space) and perfect training, one typically observes a $1/\alpha$ asymptotic decay of the error. Such behaviour is also predicted on the basis of an approximate theory [29,30] (provided the *a priori* distribution has no gap at zero distance from the target [3]). Furthermore, when the *a priori* distribution of the target function is known, one can define a Bayesian and Gibbsian strategy (which classifies a new question following the majority vote or a random vote from the version space respectively). In this case, it can be proven [20] that the average error is not larger than $1/\alpha$ and $2/\alpha$ for Bayes and Gibbs respectively (with the conjecture that the real bounds are a factor 2 sharper). Finally, it is also worth mentioning that for systems with a continuous valued output (rather than the $\{1, 0\}$-valued output that we have considered here) trained by gradient descent on a smooth error function, a replica calculation again predicts a $1/\alpha$ asymptotic decay of the error [4]. The $1/\alpha$ behaviour has to be compared with the $\ln\alpha/\alpha$ decay predicted by the VC theorem for perfect learning (section 3). The difference between both results is usually explained by pointing out that the VC theorem corresponds to a 'worst case scenario', both with respect to the choice of the target function $\bar{f}$ and of the selection mechanism. We do not believe that this explanation settles the issue. First, one can wonder whether or not the VC proof can be improved to come in line with the $1/\alpha$ behaviour. The answer seems to be no, since an explicit example has been constructed that gives rise to a $(1 - 1/e)\ln\alpha/\alpha$ worst case behaviour [16]. In this example, however, the $\ln\alpha/\alpha$ behaviour is observed for a number of training examples much larger then the number of elements in the question space Ω. This certainly is a very artificial situation, which cannot occur when the question space has an infinite number of elements. Second, under which circumstances can a 'truly' worst case scenario arise? For example, those perceptron training schemes which have been studied in the statistical physics literature and give rise to a $1/\alpha$ behaviour, are completely symmetric with respect to the choice of the target function, and consequently have error curves which are independent of the target function. In this case, there is not really a 'worst' choice of a target function (this is certainly not a typical situation, cf [31]). One can thus raise the question whether a $1/\alpha$ behaviour arises in all cases for which there is such a symmetry. More generally speaking, it remains an open question whether one can recover the ubiquitous $1/\alpha$ behaviour from a modified VC theorem by introducing some additional assumptions of a general nature (such as the specification that the question space Ω has an infinite number of elements), or by including some additional properties of the VC class (such as symmetry with respect to the target function).

On the other hand, it is remarkable that similar considerations can be made on the VC theorem for non-perfect learning (section 2). One example where such a version of the theorem is applicable and explicit calculations have been performed is given by the perceptron trained by Hebbian rule when the teacher is also a perceptron [21]. We find again a similar situation: the exact training error and generalization error both decrease as $1/\sqrt{\alpha}$ for α large, so does their difference, whereas the VC theorem predicts a $\sqrt{\ln\alpha/\alpha}$ behaviour (cf equations (11) or (14)).

We conclude with a word of caution concerning the practical implementation of the VC theorem. It is often stated that the VC theorem guarantees good generalization if one is able to load onto a given network a number of training examples much larger than the corresponding VC dimension. This statement can be misleading for the following reason.

Just as the Hoeffding inequality cannot be applied to the best function selected from a *whole class of functions, the VC theorem cannot be used for the hypothesis class that* is selected for its best performance amongst a whole set of classes. In fact, trying to load the training examples on various hypothesis classes (corresponding, for example, to various neural network architectures) is tantamount to enlarging the hypothesis class with a corresponding increase in the VC dimension. In such a case it is plainly wrong to apply the VC theorem using the VC dimension of the class that is selected because of its best performance.

Acknowledgments

We would like to thank one of the referees for pointing out some errors in the original version of the paper. This work was supported by DOE grant DE-FG03-86ER13606. J M R Parrondo also acknowledges support from a fellowship *Del Amo* from Universidad Complutense de Madrid and Grants PB91-0378 and PB91-0222 from the DGICYT (Spain) and C Van den Broeck from the Programme on Inter-University Attraction Poles, Belgian Government, and from the NFWO Belgium.

Appendix A

Defining the step function as $\theta(x) = 1$ for $x > 0$ and 0 otherwise, one can write probability (3) as a mean value

$$
\begin{aligned}
\mathrm{Prob}[\sup_{f\in\mathfrak{F}} |\nu_f^m - e_f| > \epsilon] &= \langle \theta(\sup_{f\in\mathfrak{F}} |\nu_f^m - e_f| - \epsilon)\rangle \\
&= \langle \sup_{f\in\mathfrak{F}} \theta(|\nu_f^m - e_f| - \epsilon)\rangle \\
&= \langle \sup_{f\in\mathfrak{F}} [\theta(\nu_f^m - e_f - \epsilon) + \theta(e_f - \nu_f^m - \epsilon)]\rangle .
\end{aligned}
\tag{27}
$$

The bracket $\langle\ldots\rangle$ denotes an average over the choice of the m examples determining the value of ν_f^m. One of the main steps in the theorem is to construct an inequality such that one can eliminate the e_f dependence in the mean value. This can be achieved by introducing a new training set with m' examples for which the function f gives a frequency of errors $\nu_f^{m'}$. Since $\theta(x)\theta(x') \leqslant \theta(x + x')$, one has

$$
\theta(\nu_f^m - e_f - \epsilon)\theta\left(e_f - \nu_f^{m'} + \frac{1}{m'}\right) \leqslant \theta\left[\nu_f^m - \nu_f^{m'} - \left(\epsilon - \frac{1}{m'}\right)\right]. \tag{28}
$$

The inequality (28) is valid for any value of $\nu_f^{m'}$ and hence also for the average $\langle\ldots\rangle'$ over the choices of the m' examples, i.e.

$$
\theta(\nu_f^m - e_f - \epsilon)\left\langle\theta\left(e_f - \nu_f^{m'} + \frac{1}{m'}\right)\right\rangle' \leqslant \left\langle\theta\left[\nu_f^m - \nu_f^{m'} - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle'. \tag{29}
$$

Recalling that $m'\nu_f^{m'}$ is a random variable with binomial distribution and mean value $m'e_f$ *is not hard to prove [18] that*

$$
\left\langle\theta\left(e_f - \nu_f^{m'} + \frac{1}{m'}\right)\right\rangle' = \mathrm{Prob}\left[0 \leqslant \nu_f^{m'} < e_f + \frac{1}{m'}\right] \geqslant \frac{1}{2} \tag{30}
$$

and from (30) we conclude that

$$\theta(\nu_f^m - e_f - \epsilon) \leqslant 2\left\langle \theta\left[\nu_f^m - \nu_f^{m'} - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle \tag{31}$$

In a similar way one can easily prove that

$$\theta(e_f - \nu_f^m - \epsilon) \leqslant 2\left\langle \theta\left[\nu_f^{m'} - \nu_f^m - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle . \tag{32}$$

Finally, combining (27), (31) and (32) yields

$$\begin{aligned}\mathrm{Prob}[\sup_{f\in\Im} |\nu_f^m - e_f| > \epsilon] &\leqslant 2\left\langle \sup_{f\in\Im}\left\langle \theta\left[|\nu_f^m - \nu_f^{m'}| - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle' \right\rangle \\ &\leqslant 2\left\langle\left\langle \sup_{f\in\Im} \theta\left[|\nu_f^m - \nu_f^{m'}| - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle\right\rangle'\end{aligned}$$

or

$$\mathrm{Prob}[\sup_{f\in\Im} |\nu_f^m - e_f| > \epsilon] \leqslant 2\mathrm{Prob}\left[\sup_{f\in\Im} |\nu_f^m - \nu_f^{m'}| > \left(\epsilon - \frac{1}{m'}\right)\right] . \tag{33}$$

This result is called the *basic lemma* by Vapnik and Chervonenkis [15]. Note that the result given here is an improvement on the original VC result in two ways. First, they considered only the case $m = m'$. On the other hand, instead of the lower limit $\epsilon - 1/m$ they have $\epsilon/2$ with the extra condition that $1/m < \epsilon/2$. Both ideas are present in the proof given by Devroye [17] but he did not split the step function of the absolute value obtaining a much weaker inequality valid only for $m' > m^2/4$.

We can now introduce the classification diversity of our class of functions $\Im$: for a given choice of these examples the functions $f \in \Im$ can be collected in equivalence classes $\hat{f}$, such that all functions within a given equivalence class $\hat{f}$ have an identical classification of the $m + m'$ examples. Since the functions in $\Im$ induce at most $\Delta(m + m')$ different classifications of any $m + m'$ training examples, the total number of equivalence classes is bounded by $\Delta(m + m')$ for any choice of the examples.

Since the outcome on the examples is the same for all the elements belonging to the same class $\hat{f}$ one has

$$\left\langle\left\langle \sup_{f\in\Im} \theta\left[|\nu_f^m - \nu_f^{m'}| - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle\right\rangle' \leqslant \left\langle\left\langle \sum_{\hat{f}} \theta\left[|\nu_{\hat{f}}^m - \nu_{\hat{f}}^{m'}| - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle\right\rangle' \tag{34}$$

where the sum in the right-hand side runs over all the equivalence classes.

Note that these equivalence classes depend on the choice of the examples but not on the order they appear in the total sample. On the other hand, the total mean value is also invariant under permutations of the $m + m'$ examples. Then it is possible to write

$$\begin{aligned}&\left\langle\left\langle \sum_{\hat{f}} \theta\left[|\nu_{\hat{f}}^m - \nu_{\hat{f}}^{m'}| - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle\right\rangle' \\ &\qquad = \left\langle\left\langle \sum_{\hat{f}} \frac{1}{(m+m')!} \sum_{\sigma} \theta\left[|\sigma(\nu_{\hat{f}}^m) - \sigma(\nu_{\hat{f}}^{m'})| - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle\right\rangle'\end{aligned} \tag{35}$$

where the sum runs over all the possible permutations σ of the $m+m'$ examples and $\sigma(\nu_f^m)$ and $\sigma(\nu_f^{m'})$ are the frequencies resulting in the two considered subsamples after permuting the examples of the whole sample (note that the composition of each subsample can be modified by the permutation). It turns out that the quantity

$$\Gamma_\epsilon = \frac{1}{(m+m')!}\sum_\sigma \theta[|\sigma(\nu_f^m) - \sigma(\nu_f^{m'})| - \epsilon] \tag{36}$$

can be bound for all the possible outcomes. We will consider two different bounds for Γ. The first one is valid for $m = m'$ and can be found in [12]

$$\Gamma_\epsilon < 3\mathrm{e}^{-m\epsilon^2}. \tag{37}$$

The second one, derived in [11] and [17], is weaker for the special case $m = m'$ but gives stronger results when $m' \gg m$. It reads

$$\Gamma_\epsilon \leqslant 2\exp\left[-2m\left(\frac{m'}{m+m'}\epsilon\right)^2\right]. \tag{38}$$

Finally, combining the basic lemma (33), (35) and the bound for Γ (37), one immediately obtains equation (9) of the main text. If (38) is used instead of (37) then (12) is recovered.

Appendix B

The proof runs along lines similar to those of appendix A. One has

$$\mathrm{Prob}(\sup_{f\in\Im^*} |\nu_f^m - e_f| > \epsilon) = \langle \sup_{f\in\Im} \delta(\nu_f^m, 0)\theta(e_f - \epsilon)\rangle \tag{39}$$

where $\delta(i, j)$ is the Kronecker's delta.

Using again that

$$\theta(e_f - \epsilon)\theta\left(\nu_f^{m'} - e_f + \frac{1}{m'}\right) \leqslant \theta\left(\nu_f^{m'} - \left(\epsilon - \frac{1}{m'}\right)\right) \tag{40}$$

one finds

$$\theta(e_f - \epsilon) \leqslant 2\left\langle \nu_f^{m'} - \left(\epsilon - \frac{1}{m'}\right)\right\rangle'. \tag{41}$$

This leads to the following *basic lemma* for the frequency one case (cf equation (33))

$$\mathrm{Prob}[\sup_{f\in\Im^*} |\nu_f^m - e_f| > \epsilon] \leqslant 2\ \mathrm{Prob}\left[\sup_{f\in\Im} \delta(\nu_f^m, 0)\nu_f^{m'} > \left(\epsilon - \frac{1}{m'}\right)\right]. \tag{42}$$

The basic lemma derived in [18] follows steps similar to those of the original proof by Vapnik and Chervonenkis [15]. Therefore their basic lemma is a special case of ours with $m = m'$ and $\epsilon/2$ instead of $\epsilon - 1/m'$ in the right-hand side of the inequality (42). Anthony and Biggs [8] also derive a basic lemma which is, however, again slightly weaker and more restrictive than (42).

The main difference between the general case comes from the bound for the fraction of permutations Γ_ϵ that satisfy the inequality appearing as the argument of the probability in the right-hand side in the basic lemma (42). Now Γ_ϵ reads

$$\Gamma_\epsilon = \frac{1}{(m+m')!}\sum_\sigma \delta(\sigma(\nu_f^m), 0)\theta(\sigma(\nu_f^{m'}) - \epsilon) \tag{43}$$

and, by a direct combinatorial analysis, can be bound by

$$\Gamma_\epsilon \leqslant \left(1+\frac{m}{m'}\right)^{-\epsilon m'} . \tag{44}$$

Finally, from inequalities (42), (44) and by means of an argument similar to the one used in appendix A, the final result (16) is obtained.

Appendix C

The proof of the inequality (22) again runs along similar lines. The main difference lies in the evaluation of the combinatorial factor Γ that counts the fraction of permutations over the whole sample with $m_1 + m_2 + m'$ examples which contributes to the mean value

$$\left\langle\!\!\left\langle \delta\left(\nu_f^{m_1}, 0\right)\theta\left[|\nu_f^{m_2} - \nu_f^{m'}| - \left(\epsilon - \frac{1}{m'}\right)\right]\right\rangle\!\!\right\rangle' . \tag{45}$$

This fraction $\Gamma_\epsilon(m_1, m_2, , k)$, as a function of the number k of errors in the whole sample, can be calculated combining arguments of the preceding cases

$$\Gamma_\epsilon(m_1, m_2, k) \leqslant \Gamma_\epsilon(0, m_2, k)\left[1+\frac{m_1}{m_2+m'}\right]^{-k} . \tag{46}$$

Finally, using the bound for $\Gamma_\epsilon(0, m_2, k)$ derived in [12] for the special case $m' = m_2$, one finds

$$\Gamma_\epsilon(m_1, m_2, k) \leqslant 2\exp\left[-\frac{(m_2+1)(\epsilon m_2 - 1)^2}{(k+1)(2m_2 - k + 1)}\right]\left[1+\frac{m_1}{m_2+m'}\right]^{-k} \tag{47}$$

which immediately leads to (22).

References

[1] Denker J, Schwartz D, Wittner B, Solla S, Howard R, Jackel L and Hopfield J 1987 *Complex Systems* **1** 877–922

[2] Gyorgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed K Theumann and R Köeberle (Singapore: World Scientific) p 3

[3] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)

[4] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 9056–91

[5] Watkin T L H, Rau A and Biehl M 1993 The statistical mechanics of learning a rule *Rev. Mod. Phys.* in press

[6] Baum E B and Haussler D 1989 *Neur. Comput.* **1** 151–60

[7] Abu-Mustafa Y S 1989 *Neur. Comput.* **1** 312–17

[8] Anthony M and Biggs N 1992 *Computational Learning Theory* (Cambridge: Cambridge University Press)

[9] Valiant L G 1984 *Comm. ACM* **27** 1134–42

[10] Hoeffding W 1956 *Ann. Math. Statist.* **27** 713–21

[11] Shorak G R and Wellner J A 1986 *Empirical Processes with Applications to Statistics* (New York: Wiley)

[12] Vapnik V N 1982 *Estimation of Dependences Based on Empirical Data* (Berlin: Springer)

[13] Sauer N 1972 *J. Comb. Th.* A **13** 145–7

[14] Dudley R M 1984 *École d'Été de Probabilités de Saint-Flour XII (1982) (Lecture Notes in Mathematics 1097)* (Berlin: Springer)

[15] Vapnik V N and Chervonenkis A Y 1971 *Theory of Probability and its Applications* **16** 264–80

[16] Haussler D, Littlestone N and Warmuth M K 1990 Predicting {0, 1}-functions on randomly drawn points *Technical Report* University of California, Santa Cruz

[17] Devroye L 1982 *J. Multivariate Anal.* **12** 72–9

[18] Blumer A, Ehrenfeucht A, Haussler D and Warmuth M K 1989 *J. Association for Computer Machinery* **36** 929–65

[19] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983–94

[20] Haussler D, Kearns M and Schapire R 1991 *Computational Learning Theory: Proc. Fourth Annual Workshop* (San Mateo, CA: Morgan Kaufmann)

[21] Vallet F 1989 *Europhys. Lett.* **8** 747–51

[22] Sompolinsky H, Tishby N and Seung H S 1990 *Phys. Rev. Lett.* **65** 1683–6

[23] Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** L581–L586

[24] Gyorgyi G 1990 *Phys. Rev.* A **41** 7097–100

[25] Opper M and Haussler D 1991 *Computational Learning Theory: Proc. Fourth Annual Workshop* (San Mateo, CA: Morgan Kaufmann)

[26] Opper M and Haussler D 1991 *Phys. Rev. Lett.* **66** 2677–80

[27] Meir R and Fontanari J F 1992 *Phys. Rev.* A **45** 8874–84

[28] Van den Broeck C and Bouten M 1993 Clipped Hebbian training of the perceptron *Europhys. Lett.* in press

[29] Schwartz D B, Samalam V K, Solla S A and Denker J S 1990 *Neur. Comput.* **2** 374–85

[30] Van den Broeck C and Kawai R 1990 *Phys. Rev.* A **42** 6210–8

[31] Van den Broeck C and Kawai R 1991 *Int. AMSE Conf. Neural Networks: Methodologies and Applications* ed G Mesnard and R Swiniarski (Tassin-de-la-demi-lune: AMSE)