Sample Size Requirements For Feedforward Neural Networks

Michael J. Turmon and Terrence L. Fine
Department of Electrical Engineering
Cornell University
Ithaca, NY 14853
{mjt,tlfine}@ee.cornell.edu

September 1995

Abstract

We address the question of how many training samples are required to ensure that the performance of a neural network of given complexity on its training data matches that obtained when fresh data is applied to the network. This desirable property may be termed "reliable generalization." Well-known results of Vapnik give conditions on the number of training samples sufficient for reliable generalization, but these are higher by orders of magnitude than practice indicates; other results in the mathematical literature involve unknown constants and are useless for our purposes.

This work seeks to narrow the gap between theory and practice by transforming the problem into one of determining the distribution of the supremum of a Gaussian random field in the space of weight vectors, which in turn is attacked by application of a technique called the Poisson clumping heuristic. The idea is that mismatches between training set error and true error occur not for an isolated network but for a group of similar networks. The size of this group of equivalent networks is obtained, and means of computing the size based on the training data are considered. It is shown that in some cases the Poisson clumping technique yields estimates of sample size having the same functional form as earlier ones, but since the new estimates incorporate specific characteristics of the network architecture and data distribution it is felt that more realistic estimates will result. A brief simulation study shows the usefulness of the new sample size estimates.

1 Introduction

1.1 Background

Neural networks have been used to tackle what might be termed 'empirical classification' (or 'empirical regression') problems. Given independent samples of data (x_i, y_i) we wish minimize the error $\mathcal{E}(\eta) = E(y - \eta(x))^2$; that is, to estimate $f(x) = E[Y|X = x]$. The approach taken is to choose a constraint class of networks $\mathcal{N} = \{\eta(x)\}$ and within that class, by an often complex procedure, choose a candidate network η^*. In determining how well this network models f, It is useful to separate the modeling error into two parts. The first is includes approximation error or bias—choosing $\mathcal{N}$ large enough [5] so that $\eta^0 \in \mathcal{N}$, say, models f well—and the second piece is estimation error or variance—how well the chosen η^* performs relative to η^0:

$$\begin{aligned} \mathcal{E}(\eta^*) - \mathcal{E}(f) &= [\mathcal{E}(\eta^0) - \mathcal{E}(f)] + [\mathcal{E}(\eta^*) - \mathcal{E}(\eta^0)] \\ &= E(f - \eta^0)^2 + [\mathcal{E}(\eta^*) - \mathcal{E}(\eta^0)] \quad . \end{aligned}$$

A key question in such empirical regression problems is to choose an appropriate tradeoff [15, 4], and here we examine the second, or estimation error component. One may approximate the estimation error for a particular network by using independent test data, if it is available, or cross-validation, if one is content with its properties. Our work pursues another line, drawing its fundamental outlook from Vapnik [15] and from certain PAC learning results [7] in that we find a bound on estimation error which is uniform across networks. Such a bound provides at once an estimate of the estimation error of the chosen η^*. Furthermore, if we have confidence that the returned η^* is nearly optimal in the sense of emprical error (see below) within $\mathcal{N}$, the uniform bound allows us to assert that *no member* of $\mathcal{N}$ has true performance much better than that of η^*, a conclusion not available otherwise. The primary new feature in this work is that we preserve dependence on the statistics of the data, and its interaction with the architecture $\mathcal{N}$. In this way more realistic estimates of the sample size needed to control estimation error are obtained.

1.2 Problem Setup

We investigate the tradeoffs among *network complexity*, *training set size*, and *statistical performance* of feedforward neural networks so as to allow a reasoned choice of network architecture in the face of limited training data. Nets are functions $\eta(x; w)$, parameterized by their weight vector $w \in \mathcal{W} \subseteq R^d$, which take as input points $x \in R^k$. For classifiers, network output is restricted to $\{0, 1\}$ while for forecasting it may be any real number. The architecture of all nets under consideration is $\mathcal{N}$, whose complexity may be gauged by its Vapnik-Chervonenkis (VC) dimension v, the size of the largest set of inputs the architecture can classify in any desired way ('shatter'). Nets $\eta \in \mathcal{N}$ are chosen on the basis of a training set $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^n$. These n samples are i.i.d. according to an *unknown* probability law P. Performance of a network is measured by the mean-squared error

$$\begin{aligned} \mathcal{E}(w) &= E(\eta(x; w) - y)^2 & (1) \\ &= P(\eta(x; w) \neq y) \quad \text{(for classifiers)} & (2) \end{aligned}$$

and a good (perhaps not unique) net in the architecture is

$$w^0 = \arg\min_{w \in \mathcal{W}} \mathcal{E}(w).$$

To select a net using the training set we employ the empirical error

$$\nu_{\mathcal{T}}(w) = \frac{1}{n} \sum_{i=1}^{n} (\eta(x_i; w) - y_i)^2 \tag{3}$$

sustained by $\eta(\cdot; w)$ on the training set $\mathcal{T}$. A good choice for a classifier is then

$$w^* = \arg\min_{w \in \mathcal{W}} \nu_{\mathcal{T}}(w).$$

In these terms, the issue raised in the first sentence of the section can be restated as, "How large must n be in order to ensure $\mathcal{E}(w^*) - \mathcal{E}(w^0) \leq \epsilon$ with high probability?"

For purposes of analysis we can avoid dealing directly with the stochastically chosen network w^* by noting

$$\begin{aligned} 0 \leq \mathcal{E}(w^*) - \mathcal{E}(w^0) &\leq |\nu_{\mathcal{T}}(w^*) - \mathcal{E}(w^*)| + |\nu_{\mathcal{T}}(w^0) - \mathcal{E}(w^0)| \\ &\leq 2 \sup_{w \in \mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \quad . \end{aligned}$$

A bound on the last quantity is also useful in its own right, as discussed above.

2 Known Results

Most work has been done in the context of classification so we adopt that setting in this section. The best-known result is due to Vapnik [16, 15], introduced to a wider audience in [6, 8]:

$$P(\sup_{w \in \mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \geq \epsilon) \leq 6 \left(\frac{2en}{v}\right)^v e^{-n\epsilon^2/4} \quad . \tag{4}$$

This remarkable bound not only involves no unknown constant factors, but holds independent of the data distribution P. Analysis shows that sample sizes of about

$$n_c = \frac{9.2v}{\epsilon^2} \log(\frac{8}{\epsilon}) \tag{5}$$

are sufficient to force the bound below unity, after which it drops exponentially to zero. If for purposes of illustration we take $\epsilon = .1$, $v = 50$, we find $n_c = 202\,000$, which disagrees by orders of magnitude with the experience of practitioners who train such low-complexity networks (about 50 connections).

More recently, Talagrand [13] has obtained the bound

$$P(\sup_{w \in \mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \geq \epsilon) \leq K_1 \Big(\frac{K_2 n \epsilon^2}{v}\Big)^v e^{-2n\epsilon^2}, \tag{6}$$

where $n \geq K_3 v/\epsilon^2$, which gives the sufficient condition

$$n_c = \frac{\max((1/2)\log K_2, K_3)v}{\epsilon^2}.$$

However, the values of K_2 and K_3 are inaccessible, and the result in this form is of no practical use. It does, however, illustrate that the general order of dependence is v/ϵ^2, without the extra logarithmic factor.

Related formulations have been examined. Vapnik [15] obtains bounds on

$$P(\sup_{w\in\mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\mathcal{E}(w)^{1/2}} \geq \epsilon) \qquad , \tag{7}$$

(note $\mathcal{E}(w)^{1/2} \approx Var(\nu_{\mathcal{T}}(w))^{1/2}$ when $\mathcal{E}(w) \approx 0$) and Anthony and Biggs [3] work with the equivalent of

$$P(\sup_{w\in\mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|\, 1_{\{0\}}(\nu_{\mathcal{T}}(w)) \geq \epsilon) \qquad . \tag{8}$$

Both provide finer resolution near $\mathcal{E}(w) = 0$. Haussler [9], although not concerned primarily with obtaining tight bounds, considers another type of relative distance in the regression setting. The formulation (8), for example, yields the sufficient condition

$$n_c = \frac{5.8v}{\epsilon} \log \frac{12}{\epsilon} \tag{9}$$

for nets, if any, having $\nu_{\mathcal{T}}(w) = 0$. If one is guaranteed to do reasonably well on the training set, a smaller order of dependence results.

These are sufficient conditions on n to force $P(\sup_{w\in\mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \geq \epsilon)$ low. In [6] Baum and Haussler find the *necessary* condition $n \geq \frac{v}{32\epsilon}$ which differs by an important factor of $1/\epsilon$ from the general sufficient condition.

3 Perceptron Example

By examining in detail the case of a perceptron classifier under multivariate normal input we have derived another necessary condition, in that any VC-type bound must include this as a special case. Suppose the observed data $x \in R^d$ has equal prior probability of being $N(\mu_0, I_d)$ or $N(\mu_1, I_d)$, and that $n/2$ correctly classified samples are gathered from each prior. The classifier $\eta(\cdot; w^0)$ minimizing true error probability simply compares $x^T(\mu_0 - \mu_1)$ to a threshold. The empirically chosen classifier η^* is formed by substituting the sample means under each hypothesis, $\bar{x}_0$ and $\bar{x}_1$, into η^0; this is *Fisher's linear discriminant.* The error $\mathcal{E}(w^0)$ is easily written down, and after approximating $\mathcal{E}(w^*)$, we enforce the condition

$$\mathcal{E}(w^*) - \mathcal{E}(w^0) < 2\epsilon \quad \text{(with high prob.)}$$

which must hold for all values of μ_0 and μ_1. The key to forcing a sample size of order v/ϵ^2 is to let $\mu_0 - \mu_1$ shrink with n. Analysis reveals that meeting the above condition requires

$$n \geq \frac{v}{280\epsilon^2} \qquad , \tag{10}$$

lower than the VC sufficient condition by a factor of order just $\log(1/\epsilon)$, and of the same order as the Talagrand condition. This demonstrates that the v/ϵ^2 behavior need not arise due to a strange “worst-case” input distribution or network architecture. Rather, such a sample size occurs simply by letting the data distributions under each pattern class approach one another as n increases.

4 Applying the Poisson Clumping Heuristic

We now adopt a different approach to the problem. For the moderately large values of n we anticipate, the central limit theorem informs us that

$$\sqrt{n}\,[\nu_{\mathcal{T}}(w) - \mathcal{E}(w)]$$

has nearly the distribution of a zero-mean Gaussian random variable. It is therefore reasonable[1] to suppose that

$$P(\sup_{w\in\mathcal{W}} |\,[\nu_{\mathcal{T}}(w) - \mathcal{E}(w)]\,| \geq \epsilon) \simeq P(\sup_{w\in\mathcal{W}} |Z(w)| \geq \epsilon\sqrt{n}) \leq 2P(\sup_{w\in\mathcal{W}} Z(w) \geq \epsilon\sqrt{n})$$

where $Z(w)$ is a Gaussian process with mean zero and covariance

$$R(w,v) = EZ(w)Z(v) = Cov\big((y-\eta(x;w))^2,\, (y-\eta(x;v))^2\big) \qquad .$$

Further, by symmetry of the zero-mean Gaussian process,

$$P(\sup_{w\in\mathcal{W}} Z(w) \geq \epsilon\sqrt{n}) \leq P(\sup_{w\in\mathcal{W}} |Z(w)| \geq \epsilon\sqrt{n}) \leq 2P(\sup_{w\in\mathcal{W}} Z(w) \geq \epsilon\sqrt{n}) \quad (11)$$

The factor of two makes no significant contribution to the sample size estimate because of the exponential nature of the probability bounds to be developed, so the absolute value may be ignored. The problem about extrema of the original empirical process is equivalent to one about extrema of a corresponding Gaussian process.

The Poisson clumping heuristic (PCH), introduced in a remarkable book [2] by D. Aldous, provides a tool of wide applicability for estimating such exceedance probabilities. Consider the excursions above level $b\,(=\epsilon\sqrt{n} \gg 1)$ of a sample path of a stochastic process $Z(w)$. As in figure 1a, the set $\{w : Z(w) \geq b\}$ can be visualized as a group of “clumps” scattered in weight space $\mathcal{W}$. The PCH says that, provided Z has no long-range dependence and the level b is large, the centers of the clumps fall according to the points of a Poisson process on $\mathcal{W}$, and the clump shapes themselves are independent. Figure 1b illustrates this clump process. The vertical arrows illustrate two clump centers (points of the Poisson process); the clumps themselves are the bars centered about the arrows.

In fact, with $p_b(w) = P(Z(w) \geq b)$, $C_b(w)$ the size of a clump located at w, and $\lambda_b(w)$ the rate of occurrence of clump centers, the fundamental equation is

$$p_b(w) \simeq \lambda_b(w) EC_b(w). \quad (12)$$

[1] There are some technical details, of no interest here, in this passage to the limiting normal process; see chapter 7 of [11].

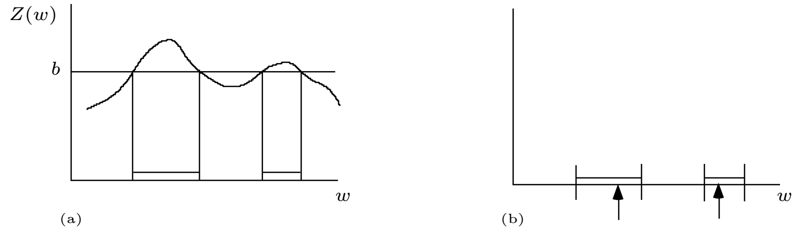

Figure 1: The Poisson clumping heuristic

Since clump centers form a Poisson process, the number of clumps in $\mathcal{W}$ is a Poisson random variable N_b with parameter $\int_{\mathcal{W}} \lambda_b(w)\, dw$. The probability of a clump, which we wish to make small since it corresponds to existence of a bad estimate of $\mathcal{E}(w)$ by $\nu_{\mathcal{T}}(w)$, is

$$P(N_b > 0) = 1 - \exp\bigl(-\int_{\mathcal{W}} \lambda_b(w)\, dw\bigr) \simeq \int_{\mathcal{W}} \lambda_b(w)\, dw$$

where the last approximation holds because our goal is to operate in a regime where this probability is near zero. Letting $\bar{\Phi}(b) = P(N(0,1) > b)$ and $\sigma^2(w) = R(w,w)$, we have $p_b(w) = \bar{\Phi}(b/\sigma(w))$. The fundamental equation becomes

$$P(\sup_{w \in \mathcal{W}} Z(w) \geq b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma(w))}{EC_b(w)}\, dw \quad . \tag{13}$$

It remains only to find the mean clump size $EC_b(w)$ in terms of the network architecture and the statistics of (x, y).

5 Poisson Clumping for Smooth Processes

Assume $Z(w)$ has two mean-square derivatives in w. (If the network activation functions have two derivatives in w, for example, $Z(w)$ will have two almost sure derivatives.) We can then write a quadratic approximation to Z in the vicinity of a point w_0:

$$Z(w) \simeq Z_0 + (w - w_0)^T G + \frac{1}{2}(w - w_0)^T \mathbf{H}(w - w_0) \tag{14}$$

where the gradient $G = \nabla Z(w)$ and Hessian matrix $\mathbf{H} = \nabla\nabla Z(w)$ are both evaluated at w_0. One pictures a downward-turning parabola peaking near w_0 which attains height at least b; the clump size is the volume V of the ellipsoid in R^d formed by the intersection of the parabola with the level b. Provided $Z_0 \geq b$, that is that there is a clump at w_0, simple computations reveal

$$V \simeq \kappa_d \frac{(2(Z_0 - b) - G^T \mathbf{H}^{-1} G)^{d/2}}{|\mathbf{H}|^{1/2}} \tag{15}$$

where κ_d is the volume of the unit ball in R^d and $|\cdot|$ is the determinant. The mean clump size is approximately[2] the expectation of V conditioned on $Z(w_0) \geq b$.

The same argument used to show that $Z(w)$ is approximately normal shows that G and $\mathbf{H}$ are approximately normal too. In fact,

$$\begin{aligned} E[\mathbf{H}|Z(w_0) = z] &= \frac{z}{\sigma^2(w_0)}\Lambda(w_0) \\ \Lambda(w_0) &= -EZ(w_0)\mathbf{H} = -\nabla_w \nabla_w R(w_0, w)|_{w=w_0} \end{aligned}$$

so that, since b (and hence z) is large, the second term in the numerator of (15) may be neglected. The expectation is then easily computed, resulting in

Lemma 1 (Smooth process clump size) *Let the network activation functions be twice continuously differentiable, and let $b \gg \sigma(w)$. Then the process clump size is*

$$EC_b(w) \simeq (2\pi)^{d/2} \left|\frac{\Lambda(w)}{\sigma^2(w)}\right|^{-1/2} \left(\frac{\sigma(w)}{b}\right)^d \qquad .$$

Substituting into (13) yields

$$\begin{aligned} P(\sup_{w\in\mathcal{W}} Z(w) \geq b) &\simeq (2\pi)^{-d/2} \int_{\mathcal{W}} \left|\frac{\Lambda(w)}{\sigma^2(w)}\right|^{1/2} \left(\frac{b}{\sigma(w)}\right)^d \bar{\Phi}(b/\sigma(w))\, dw \qquad (16) \\ &\simeq (2\pi)^{-(d+1)/2} \int_{\mathcal{W}} \left|\frac{\Lambda(w)}{\sigma^2(w)}\right|^{1/2} \left(\frac{b}{\sigma(w)}\right)^{d-1} e^{-b^2/2\sigma^2(w)}\, dw, \end{aligned}$$

where use of the asymptotic expansion $\bar{\Phi}(z) \simeq (z\sqrt{2\pi})^{-1}\exp(-z^2/2)$ is justified since $(\forall w) b \gg \sigma(w)$ is necessary to have each individual probability $P(Z(w) \geq b)$ low—let alone the supremum. To proceed farther, we need some information about the variance $\sigma^2(w)$ of $(y - \eta(x; w))^2$. In general this must come from the problem at hand, but suppose for example the process has a unique variance maximum $\bar{\sigma}^2$ at $\bar{w}$. Then, since the level b is large, we can use Laplace's method to approximate the d-dimensional integral.

Laplace's method is a way to find asymptotic expansions for integrals of the form

$$\int_{\mathcal{W}} g(w) \exp(-f(w)^2/2)\, dw$$

when $f(w)$ has two continuous derivatives and a unique positive minimum at w_0 in the interior of $\mathcal{W} \subseteq R^d$, and $g(w)$ is continuous at w_0. Suppose $f(w_0) \gg 1$ so

[2] The subtle distinction between $EC_b(w_0)$ and $E[V \mid Z(w_0) > b]$ is that the condition $Z(w_0) > b$ is not precisely equivalent to occurrence of a *clump center* at b, which implicitly conditions the mean clump size; in fact, the latter implies the former. However, it is apparent that the two events are closely related so that the approximation is reasonable. We shall have more to say on the tightness of this approximation in section 6.

that the exponential factor is decreasing much faster than the slowly varying g. Expanding f about w_0, substituting into the exponential, and ignoring terms of more than second order yields

$$\int_{\mathcal{W}} g(w)\exp(-f(w_0)^2/2)\exp(-(w-w_0)^T[f(w_0)M](w-w_0)/2)\,dw$$
$$\simeq g(w_0)\exp(-f(w_0)^2/2)\int_{\mathcal{W}}\exp(-(w-w_0)^T[f(w_0)M](w-w_0)/2)\,dw$$

where $M = \nabla\nabla f(w)|_{w_0}$, the Hessian of f. In the latter equation we have used that g is changing slowly relative to the exponential. The integral is expanded to all of R^d—it is negligible away from w_0—and is easily performed. The Laplace asymptotic expansion is

$$\int_{\mathcal{W}} g(w)\exp(-f(w)^2/2)\,dw \simeq (2\pi)^{d/2}|f(w_0)M|^{-1/2}g(w_0)\exp(-f(w_0)^2/2)$$

Applying this method to the integral (16) results in

Theorem 1 *Let the network activation functions be twice continuously differentiable. Let the variance have a unique maximum $\bar\sigma$ at $\bar w$ in the interior of $\mathcal{W}$ and the level $b \gg \bar\sigma$. Then the PCH estimate of exceedance probability is given by*

$$P(\sup_{w\in\mathcal{W}} Z(w) \geq b) \simeq \frac{|\Lambda(\bar w)|^{1/2}}{|\Lambda(\bar w)-\Gamma(\bar w)|^{1/2}}\frac{\bar\sigma/b}{\sqrt{2\pi}}e^{-b^2/2\bar\sigma^2} \quad (17)$$
$$\simeq \frac{|\Lambda(\bar w)|^{1/2}}{|\Lambda(\bar w)-\Gamma(\bar w)|^{1/2}}\,\bar\Phi(b/\bar\sigma) \quad (18)$$

where $\Gamma(\bar w) = \nabla_w\nabla_v R(w,v)|_{w=v=\bar w}$. Furthermore, $\Lambda - \Gamma$ is positive-definite at $\bar w$; it is $-1/2$ the Hessian of $\sigma^2(w)$. The leading constant thus strictly exceeds unity.

The above probability is just $P(Z(\bar w) \geq b)$ multiplied by a factor to account for the other random variables in the supremum. Letting $b = \epsilon\sqrt{n}$ in the above reveals that

$$n = \frac{d\bar\sigma^2\log K}{\epsilon^2} \quad (19)$$
$$K^d = |\Lambda(\bar w)|/|\Lambda(\bar w)-\Gamma(\bar w)| = |I_d - \Lambda(\bar w)^{-1}\Gamma(\bar w)| \quad (20)$$

samples will force $P(\sup_w|\nu_{\mathcal{T}}(w)-\mathcal{E}(w)| \geq \epsilon)$ low. With d playing the role of VC dimension v, this is similar to Vapnik's bound although we have retained dependence on the underlying P and $\mathcal{N}$. Similar results (but in which the constant is explicitly determined) are available for different versions of this problem, and certain other related problems [14].

We note that the above probability has the property of being determined by behavior near the maximum-variance point, which for example in classification is where $\mathcal{E}(w) = 1/2$. (This 'concentration' property is in fact quite common [12]). Such nets are not very interesting as classifiers, and certainly it is not desirable for them to determine the entire probability. This problem can be avoided by focusing instead on

$$P\Big(\sup_{w \in \mathcal{W}} \frac{\nu_{\mathcal{T}}(w) - \mathcal{E}(w)}{\sigma(w)} \geq \epsilon\Big) \simeq P\Big(\sup_{w \in \mathcal{W}} \frac{Z(w)}{\sigma(w)} \geq \epsilon\sqrt{n}\Big) \qquad , \tag{21}$$

which has the added benefit of allowing a finer resolution to be used where $\mathcal{E}(w)$ is near zero. In classification for example, if n is such that with high probability

$$\sup_{w \in \mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sigma(w)} = \sup_{w \in \mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sqrt{\mathcal{E}(w)(1 - \mathcal{E}(w))}} < \epsilon \quad , \tag{22}$$

then $\nu_{\mathcal{T}}(w^*) = 0$ implies $\mathcal{E}(w^*) < \epsilon^2(1 + \epsilon^2)^{-1} \simeq \epsilon^2 \ll \epsilon$. We see that around $\nu_{\mathcal{T}}(w^*) = 0$ the condition (22) is much more powerful than the corresponding unnormalized one. Sample size estimates using this formulation give results having a functional form similar to (9).

6 Lower Bounds via Approximate Clump Size

In this section we explore some approximations to clump size that also figure into accurate lower bounds for the exceedance probability. These approximations have the advantage of being easier to express for arbitrary processes than the clump size. The starting point is the total exceedance volume

$$D_b \equiv \int_{\mathcal{W}} U(Z(w') - b)\, dw' = \mathrm{vol}(\{w' \in \mathcal{W} \,:\, Z(w') > b\}) \tag{23}$$

where $U(z) = 1$ iff $z > 0$. At levels b of interest to us, $D_b = 0$ with high probability, motivating the introduction of the *mean bundle size*

$$E[D_b \,|\, Z(w) > b] \quad . \tag{24}$$

As the name suggests, the mean bundle size is different from the clump size partly because the former includes all exceedances of the level b, not just the region corresponding to a given clump center. The bound is an overestimate when the number N_b of clumps exceeds one, but recall that we are in a regime where b (equivalently n) is large enough so that

$$P(N_b > 1)/P(N_b = 1) \simeq \int_{\mathcal{W}} \lambda_b(w)\, dw \ll 1 \qquad .$$

Thus error in (23) due to this source is negligible.

The principal source of difference between $EC_b(w)$ and $E[D_b \,|\, Z(w) > b]$ has to do with biased sampling. The event conditioning bundle size is an exceedance at w, or equivalently occurrence of a clump capturing w. By virtue of its having been large enough to overlap a point, such a clump is on average larger than a clump centered at w, and in fact one can show that

$$\frac{E[D_b \,|\, Z(w)\!>\!b]}{EC_b(w)} \simeq \frac{EC_b(w)^2}{(EC_b(w))^2} \geq 1 \quad . \tag{25}$$

In the examples we have studied, the amount of overestimation is not enough to significantly affect the sample size estimates.

The usefulness of the bundle size is due to the following results, which are an extension (see also [1]) of the union bound. For clarity we make explicit the dependence on the experimental outcome $\underline{\omega} \in \Omega$ (further distinguished from $w \in \mathcal{W}$ by the underbar).

Lemma 2 *Let $S_w \subseteq \Omega$ be measurable sets for each $w \in \mathcal{W}$, and let θ be a measure on $\mathcal{W}$. Assume that*

$$D = D(\underline{\omega}) := \theta(\{w \in \mathcal{W} \,:\, \underline{\omega} \in S_w\}) \tag{26}$$

is a well-defined random variable. If the regularity conditions $D(\underline{\omega}) < \infty$ a.s. and $\underline{\omega} \in S_w \Rightarrow D(\underline{\omega}) > 0$ are met, then

$$P\big(\bigcup_{w \in \mathcal{W}} S_w\big) = \int_{\mathcal{W}} P(S_w)\, E[D^{-1} \,|\, S_w]\, \theta(dw) \quad .$$

In order for $D(\underline{\omega})$ to make sense, for each fixed $\underline{\omega}$ the θ-measure must be defined. Then, the resulting function $\Omega \rightarrow R$ must be a random variable.
Proof. The regularity conditions on $D(\underline{\omega})$ allow us to define $D_w(\underline{\omega}) = 1/D(\underline{\omega})$ for $\underline{\omega} \in S_w$ and zero otherwise. Rewriting the union and iterating the expectation give

$$\begin{aligned}
1_{\bigcup S_w} \quad \text{a.s.} &= \int_{\mathcal{W}} D_w 1_{S_w}\, \theta(dw) \\
\Rightarrow P\big(\bigcup_{w \in \mathcal{W}} S_w\big) &= \int_{\mathcal{W}} E\, D_w 1_{S_w}\, \theta(dw) \\
&= \int_{\mathcal{W}} E\, E[D_w 1_{S_w} \,|\, 1_{S_w}]\, \theta(dw) \\
&= \int_{\mathcal{W}} E\, 1_{S_w} E[D_w \,|\, 1_{S_w}]\, \theta(dw) \\
&= \int_{\mathcal{W}} E\, 1_{S_w} E[D_w \,|\, S_w]\, \theta(dw) \\
&= \int_{\mathcal{W}} P(S_w)\, E[D^{-1} \,|\, S_w]\, \theta(dw)
\end{aligned}$$

since $D_w(\underline{\omega}) = D^{-1}(\underline{\omega})$ if $\underline{\omega} \in S_w$.

As a simple corollary we have

Theorem 2 *If $Z(w)$ is continuous and $D_b < \infty$ a.s.,*

$$\begin{aligned} P(\sup_{w\in\mathcal{W}} Z(w) > b) &= \int_{\mathcal{W}} \frac{\bar\Phi(b/\sigma)}{E[D_b^{-1}\,|\,Z(w)\!>\!b]^{-1}}\,dw \\ &\geq \int_{\mathcal{W}} \frac{\bar\Phi(b/\sigma)}{E[D_b\,|\,Z(w)\!>\!b]}\,dw \end{aligned}$$

Proof. Take θ as Lebesgue measure and $S_w = \{\underline{\omega} \in \Omega \,:\, Z(w) > b\}$ in the proposition. Then $D(\underline{\omega}) = D_b$. Continuity of $Z(w)$ as a function $\mathcal{W} \to R$ ensures that D_b is a well-defined random variable. In fact, continuity tells us that the preimage $Z^{-1}\big((b,\infty)\big) \subseteq \mathcal{W}$ is open a.s., so if $Z(w_0) > b$ the preimage is also nonempty and its Lebesgue measure is positive. The second assertion is a consequence of the harmonic mean inequality: $f > 0 \Rightarrow (Ef^{-1})^{-1} \leq Ef$.

The bundle size is easier to compute than the clump size because

$$E[D_b\,|\,Z(w) > b] = \int_{\mathcal{W}} P(Z(w') \geq b|Z(w) \geq b)\,dw' \quad . \tag{27}$$

Since $Z(w)$ and $Z(w')$ are jointly normal, abbreviate $\sigma = \sigma(w)$, $\sigma' = \sigma(w')$, $\rho = \rho(w,w') = R(w,w')/(\sigma\sigma')$, and let

$$\zeta = \zeta(w,w') = (\sigma/\sigma')\frac{1-\rho\sigma'/\sigma}{\sqrt{1-\rho^2}} \tag{28}$$

$$= \sqrt{\frac{1-\rho}{1+\rho}} \quad \text{(constant variance case)} \quad . \tag{29}$$

Evaluating the conditional probability above presents no problem, and we obtain

Lemma 3 (Clump size estimate) *For $b \gg \sigma$ the mean bundle size is*

$$ED_b(w) \simeq \int_{\mathcal{W}} \bar\Phi\left((b/\sigma)\zeta\right)\,dw' \quad . \tag{30}$$

Remark 1. This integral will be used in (13) to find

$$P(\sup_w Z(w) > b) \simeq \int_{\mathcal{W}} \frac{\bar\Phi(b/\sigma)}{\int_{\mathcal{W}} \bar\Phi\left((b/\sigma)\zeta\right)\,dw'}dw \quad . \tag{31}$$

Since b is large, the main contribution to the outer integral occurs for w near a variance maximum, i.e. for $\sigma'/\sigma \leq 1$. If the variance is constant then all $w \in \mathcal{W}$ contribute. In either case ζ is nonnegative. By comparison with the results in lemma 1, we expect the estimate (30) to be, as a function of b, of the form (const $\sigma/b)^p$ for, say, $p = d$. In particular, we do not anticipate the exponentially small bundle sizes resulting if $(\forall w')\zeta(w,w') \geq M \gg 0$. To achieve such polynomial sizes, ζ must come close to zero over some range of w', which evidently can happen only when $\rho \approx 1$, that is, for w' in a neighborhood of

w. The behavior of the covariance locally in such neighborhoods is the key to finding the bundle size.

Remark 2. There is a simple interpretation of the bundle size; it represents the volume of $w' \in \mathcal{W}$ for which $Z(w')$ is highly correlated with $Z(w)$. The exceedance probability is a sum of the point exceedance probabilities (the numerator of (31)), each weighted according to how many other points are correlated with it. A large bundle size indicates strong correlation of the weight in question to neighboring weights, and its contribution to the exceedance probability is accordingly decreased. Smaller bundle sizes indicate a more jagged process and give a relatively larger contribution to the exceedance probability. In effect, the space $\mathcal{W}$ is partitioned into regions that tend to "have exceedances together," with a large bundle size indicating a large region. The overall probability can be viewed as a sum over all these regions of the corresponding point exceedance probability. This has a similarity to the Vapnik argument which lumps networks together according to their $n^v/v!$ possible actions on n items in the training set. In this sense the mean bundle size is a fundamental quantity expressing the ability of an architecture to generalize.

7 Empirical Estimates of Bundle Size

The bundle size estimate of lemma 3 is useful in its own right if one has information about the covariance of Z. Other known techniques of finding $EC_b(w)$ exploit special features of the process at hand (e.g. smoothness or similarity to other well-studied processes); the above expression is valid for any covariance structure. In this section we show how one may *estimate* the clump size using the training set, and thus obtain probability approximations in the absence of analytical information about the unknown P and the potentially complex network architecture $\mathcal{N}$.

Here is a practical way to approximate the integral giving $E[D_b \,|\, Z(w) > b]$. For $\gamma < 1$ define a set of significant w'

$$\begin{aligned} S_\gamma(w) &= \{w' \in \mathcal{W} : \zeta(w, w') \le \gamma\} && (32) \\ V_\gamma(w) &= \mathrm{vol}(S_\gamma(w)) && (33) \end{aligned}$$

and note that from the monotonicity of $\bar{\Phi}$

$$E[D_b \,|\, Z(w) > b] \ge \int_{S_\gamma} \bar{\Phi}((b/\sigma)\zeta)\, dw' \ge V_\gamma(w)\, \bar{\Phi}((b/\sigma)\gamma) \qquad .$$

This apparently crude lower bound for $\bar{\Phi}$ is accurate enough near the origin to give satisfactory results in the cases we have studied. For example, we can characterize the covariance $R(w, w')$ of the smooth process of lemma 1 and thus find its ζ function. The bound above is then easily calculated and differs by only small constant factors from the clump size in the lemma.

With this bound we have

$$P(\sup_w Z(w) \ge b) \quad \le \quad \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{V_\gamma(w)\, \bar{\Phi}((b/\sigma)\gamma)}\, dw$$

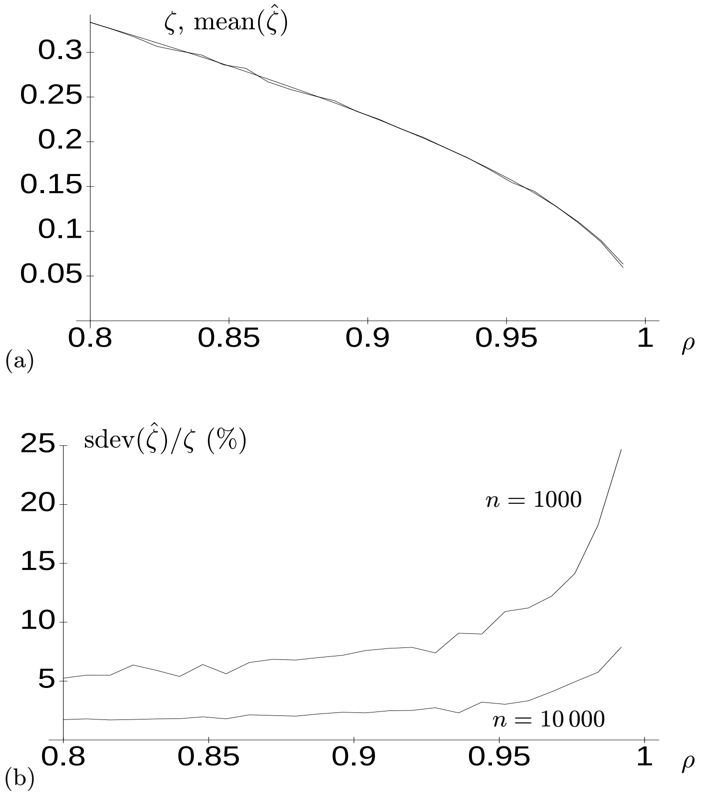

Figure 2: Estimating ζ for binary classification. In the upper plot, the curves nearly coincide. The ratio in the lower plot is shown in percent.

$$\simeq \quad \gamma \int_{\mathcal{W}} V_\gamma(w)^{-1} \exp(-(1-\gamma^2)b^2/2\sigma^2)\, dw \qquad , \quad (34)$$

because as long as γ is not too small, both arguments of $\bar{\Phi}$ will be large, justifying use of the asymptotic expansion $\bar{\Phi}(z) \simeq (z\sqrt{2\pi})^{-1} \exp(-z^2/2)$. We now need to find $V_\gamma(w)$, which we term the *correlation volume*, as it represents those weight vectors w' whose errors $Z(w')$ are highly correlated with $Z(w)$.

One simple way to estimate the correlation volume is as follows. Select a weight w' and using the training set compute

$$\begin{array}{l} (y_1 - \eta(x_1; w))^2, \quad \ldots \quad , (y_n - \eta(x_n; w))^2 \quad \text{and} \\ (y_1 - \eta(x_1; w'))^2, \quad \ldots \quad , (y_n - \eta(x_n; w'))^2 \quad . \end{array}$$

It is then easy to estimate σ^2, ${\sigma'}^2$, and ρ, and finally $\zeta(w, w')$, which is compared to the chosen γ to decide if $w' \in S_\gamma(w)$ or not.

It is possible to reliably estimate ζ in this way, even when $\rho \approx 1$. Figure 2 illustrates this for the problem of binary classification. The error $(y - \eta(w; x))^2$ is a Bernoulli random variable, and $\hat{\zeta}$ is formed from n i.i.d. pairs (for w and w') of such variables having a given variance and correlation. Choosing $\sigma = \sigma' = 1/2$, the correlation is then varied from 0.8 to nearly unity, resulting in ζ dropping from about 1/3 to quite small values. (The estimator $\hat{\zeta}$ is forced to find the variances even though they are the same in this example.) The upper panel shows ζ and the sample mean of 100 independent $\hat{\zeta}$ estimates, each of which is computed on the basis of $n = 1000$ pieces of data. This plot shows the scale of ζ and demonstrates that $\hat{\zeta}$ is essentially unbiased, at least for n moderately large. The lower panel shows the ratio of standard deviation of $\hat{\zeta}$ to ζ, expressed as a percentage, for $n = 1000$ (upper curve) and $n = 10\,000$ (lower curve). Only for quite low values of ζ does the variance become significant. However, this variability at very low ζ does not influence estimates of $V_\gamma(w)$ as long as the threshold γ is moderate, which in this simulation would mean greater than 1/20 or so.

The difficulty is that for large d the correlation volume is much smaller than any approximately-enclosing set. Ordinary uniform Monte Carlo sampling and even importance sampling methods fail to estimate the volume of such high-dimensional convex bodies because so few hits can be scored in probing the space [10]. It is necessary to concentrate the search.

The simplest technique is to let $w' = w$ except in one coordinate and sample along each coordinate axis. The correlation volume is then approximated as the product of these one-dimensional measurements.

8 A Simulation

We are now in a position to perform simulation studies to test our ability to estimate the correlation volume and hence the exceedance probability. For simplicity, normalize the process $Z(w)$ by its standard deviation $\sigma(w)$ as indicated earlier. The variance of the scaled process is unity and (34) becomes

$$P(\sup_w \frac{Z(w)}{\sigma(w)} \geq b) \leq \frac{\bar{\Phi}(b)}{\bar{\Phi}(b\gamma)} \int_{\mathcal{W}} V_\gamma(w)^{-1}\, dw \tag{35}$$

which we will estimate by a Monte Carlo integral, using the above method for finding the integrand $V_\gamma(w)$. The only difficulty is the choice of γ, which in turn depends on b. Recomputing the integral for many different γ or b values must be avoided.

This can be done if we make the reasonable assumption that

$$V_\gamma(w) = K(w)\gamma^{2d/\alpha}$$

with $\alpha = 2$ (smooth process) or 1 (rough process). This amounts to supposing the correlation $\rho(w, w')$ falls off quadratically or linearly for w' in a neighborhood of w.[3] The coefficients may change as w varies but the basic form of the correlation does not.

Thus, once the integral is computed for a reference γ_0, it can be scaled to a desired $\gamma \ll 1$ via

$$P\Bigl(\sup_w \frac{Z(w)}{\sigma(w)} \geq b\Bigr) \leq \frac{\bar{\Phi}(b)}{\bar{\Phi}(b\gamma)\gamma^{2d/\alpha}} \left(\gamma_0^{2d/\alpha} \int_{\mathcal{W}} V_{\gamma_0}(w)^{-1}\, dw\right) \quad . \tag{36}$$

Upon differentiating we find the optimal γ satisfies $\gamma^2 b^2 = 2d/\alpha$, and

$$\begin{aligned} P\Bigl(\sup_w \frac{Z(w)}{\sigma(w)} \geq b\Bigr) &\leq \left(\frac{b^2}{d}\right)^{d/\alpha} \bar{\Phi}(b) \left[\frac{\gamma_0^{2d/\alpha} \int_{\mathcal{W}} V_{\gamma_0}(w)^{-1}\, dw}{(2/\alpha)^{d/\alpha}\, \bar{\Phi}(\sqrt{2d/\alpha})}\right] &(37)\\ &= \left(\frac{b^2}{d}\right)^{d/\alpha d} \bar{\Phi}(b)\exp(dQ) &(38)\end{aligned}$$

where the final line defines Q.

As a brief demonstration of the potential accuracy of the method outlined above, consider the following example of a perceptron. Nets are $\eta(x;w) = 1_{[0,\infty)}(w^T x)$ for $w \in \mathcal{W} = R^d$, and data x is uniform on $[-1/2, 1/2]^d$. Suppose $y = \eta(x;w^*)$ and $w^* = [1 \cdots 1]$. This is a version of the *threshold function* in R^d. Nets are discontinuous so $Z(w)$ is 'rough' with $\alpha = 1$.

In figure 3 is the empirically determined Q versus d for the threshold function. At each d twenty independent estimates of Q are averaged. Each estimate is found via a Monte Carlo integral, as described above, with correlation volumes determined from a training set of size $100d$.

Over the range, say, $7 \leq d \leq 50$, we see $Q \approx 1$ and

$$\begin{aligned} P\Bigl(\sup_w \frac{Z(w)}{\sigma(w)} \geq b\Bigr) &\leq e^d\,(b^2/d)^d\ \bar{\Phi}(b)\\ (1/d)\log P\Bigl(\sup_w \frac{Z(w)}{\sigma(w)} \geq b\Bigr) &\leq 1 + \log(b^2/d) - (1/2)(b^2/d)\end{aligned}$$

This falls below zero at $b^2/d = 5.4$, implying that sample sizes above the critical value

$$\begin{aligned} \frac{b^2}{d} &= \frac{n_c\epsilon^2}{d} \simeq 5.4\\ n_c &= \frac{5.4d}{\epsilon^2} &(39)\end{aligned}$$

are enough to ensure that with high probability,

$$\sup_{w\in\mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sqrt{\mathcal{E}(w)(1-\mathcal{E}(w))}} < \epsilon \quad .$$

[3] For example, a nondifferentiable process must have a covariance $R(w,w')$ that does not have a derivative in w' at $w' = w$. The most familiar example of such a process is the Brownian bridge on $[0,1]$, for which $R(w,w') = w \wedge w' - ww'$ is 'tent-shaped' at $w = w'$.

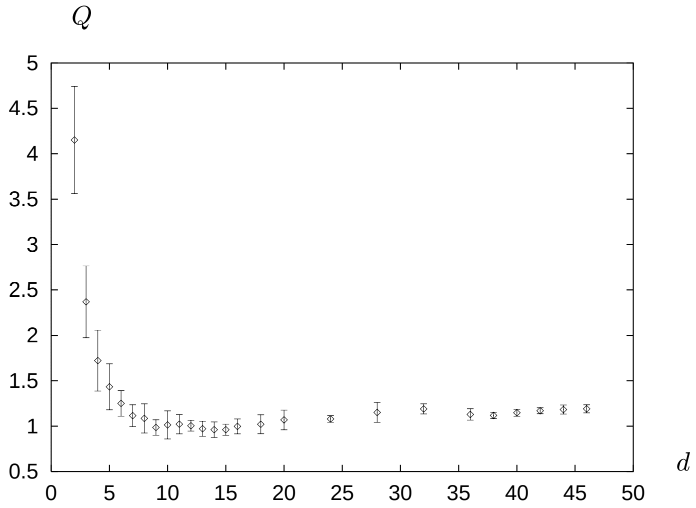

Figure 3: Empirical estimate of the leading constant for a perceptron architecture. Error bars span one sample standard deviation in each direction from the sample mean.

As in the remarks below (22), if there is a net having $\nu_{\mathcal{T}}(w) = 0$, we see that sample sizes above

$$n_c = \frac{5.4d}{\epsilon} \tag{40}$$

will guarantee $\mathcal{E}(w) < \epsilon$ with high probability, which compares favorably with (9). The condition above also tells us about nets having small but nonzero error.

9 Summary and Conclusions

In an effort to provide guidelines for intelligent choice of a network architecture based on the amount of data available to select a network, we have related ability to generalize correctly (ϵ) to network complexity (v or d) and training set size (n).

To do this we transform the neural network problem to one of finding the distribution of the supremum of a derived Gaussian random field, which is defined over the weight space of the network architecture. The latter problem is amenable to solution via the Poisson clumping heuristic. In terms of the PCH the question becomes one of estimating the mean clump size, that is, the typical volume of an excursion above a given level by the random field. In the "smooth" case we directly find the clump volume and obtain estimates of sample size that are in general of order d/ϵ^2. The leading constants depend on simple properties of the architecture and the data—which has the advantage of being tailored to the given problem but the potential disadvantage of our having to compute them.

We also obtain a useful estimate for the clump size of a general process. When this estimate is put back into the expression for exceedance probability, a simple interpretation of the clump size in terms of the number of "degrees of freedom" of the problem results. Related to this clump size is the correlation volume $V_\gamma(w)$. Paraphrasing (34) in the constant-variance case,

$$P(\sup_{w \in \mathcal{W}} \frac{\nu_{\mathcal{T}}(w) - \mathcal{E}(w)}{\sigma(w)} \geq \epsilon) \approx E \left[\frac{\mathrm{vol}(\mathcal{W})}{V_\gamma(w)} \right] e^{-(1-\gamma^2)n\epsilon^2/2}$$

where the expectation is taken with respect to a uniform distribution over the weight space. The probability of reliable generalization is roughly given by an exponentially decreasing factor (the worst-case exceedance probability for a single point) times a number representing degrees of freedom. The latter is the mean number of networks needed to cover the space. The parallel with the Vapnik approach, in which a worst-case exceedance probability is multiplied by a growth function bounding the number of classes of networks in $\mathcal{N}$ that can act differently on n pieces of data, is striking. In this fashion the correlation volume is an analog of the VC dimension, but one that depends on the interaction of the data and the architecture.

Lastly, we have proposed practical methods of estimating the correlation volume empirically from the training data. Initial simulation studies based on a perceptron with input uniform on a region in R^d show that these approximations can indeed yield informative estimates of sample complexity.

References

[1] D. Aldous. The harmonic mean formula for probabilities of unions: Applications to sparse random graphs. *Discrete Mathematics*, 76:167–176, 1989.

[2] D. Aldous. *Probability Approximations via the Poisson Clumping Heuristic.* Springer, New York, 1989.

[3] M. Anthony and N. Biggs. *Computational Learning Theory.* Cambridge Univ., 1992.

[4] A. R. Barron. Approximation and estimation bounds for artificial neural networks. In *Proc. Fourth Ann. Workshop on Computational Learning Theory (COLT'91)*, pages 243–249. Morgan-Kauffman, 1991. Reprinted in *Machine Learning*, **14**, 1994.

[5] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993.

[6] E. Baum and D. Haussler. What size net gives valid generalization? In D. S. Touretzky, editor, *Neural Information Processing Systems 1*, pages 81–90. Morgan-Kauffman, 1989.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Jour. Assoc. Comp. Mach.*, 36(4):929–965, 1989.

[8] L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Trans. Patt. Anal. and Mach. Intell.*, 10(4):530–543, 1988.

[9] D. Haussler. Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[10] L. Lovász. Geometric algorithms and algorithmic geometry. In *Proceedings of the International Congress of Mathematicians*. The Mathematical Society of Japan, 1991.

[11] D. Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

[12] M. Talagrand. Small tails for the supremum of a Gaussian process. *Ann. Inst. Henri Poincaré*, 24(2):307–315, 1988.

[13] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994.

[14] M. Turmon. *Assessing Generalization of Feedforward Neural Networks*. PhD thesis, Cornell University, Ithaca, N.Y., 1995.

[15] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, New York, 1982.

[16] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. and its Apps.*, 16(2):264–280, 1971.