# Empirically Estimating Generalization Ability of Feedforward Neural Networks

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

### Abstract

We estimate the number of training samples required to ensure that the performance of a neural network on its training data matches that obtained when fresh data is applied to the network. Existing estimates are higher by orders of magnitude than practice indicates. We narrow the gap between theory and practice by transforming the problem into determining the distribution of the supremum of a random field in the space of weight vectors, which in turn is attacked by application of a recent technique called the Poisson clumping heuristic.

## 1 Introduction and Prior Work

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 $w \in \mathcal{W} \subseteq R^d$, taking points $x \in R^k$ as input. 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 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.

We adopt the classification setting as we disucss prior work in the remainder of this section. The best-known result is due to Vapnik [1], introduced to a wider audience by Baum and Haussler [2]:

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

This remarkable bound[1] not only involves no unknown constant factors, but holds independent of the data distribution $P$. Analysis shows that sample sizes of about

$$n_c = (4v/\epsilon^2)\log 3/\epsilon \tag{5}$$

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 $n_c = 68\,000$, which disagrees by orders of magnitude with the experience of practitioners who train such low-complexity networks (about 50 connections). More recently, Talagrand [3] has obtained another upper bound to (4) (having a different functional form) which implies a sufficient condition of order $v/\epsilon^2$. However, the bound involves inaccessible constants so the result is of no practical use.

Formulations providing finer resolution near $\mathcal{E}(w) = 0$ have been examined. Vapnik [1] upper bounds $P(\sup_{w\in\mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|/(\mathcal{E}(w)^{1/2}) \geq \epsilon)$; the normalization $\mathcal{E}(w)^{1/2} \approx Var(\nu_{\mathcal{T}}(w))^{1/2}$ when $\mathcal{E}(w) \approx 0$. Anthony and Biggs [4] work with the equivalent of $P(\sup_{w\in\mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|\, 1_{\{0\}}(\nu_{\mathcal{T}}(w)) \geq \epsilon)$, obtaining the sufficient condition

$$n_c = (5.8v/\epsilon)\,\log 12/\epsilon \tag{6}$$

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.

## 2 Applying the Poisson Clumping Heuristic

We adopt a new 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[2] to suppose that

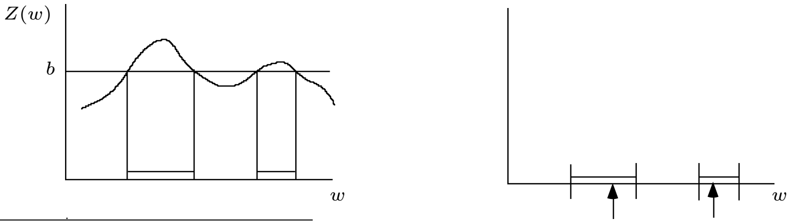
$$P(\sup_{w\in\mathcal{W}} |\,[\nu_{\mathcal{T}}(w) - \mathcal{E}(w)]\,| \geq \epsilon) \simeq P(\sup_{w\in\mathcal{W}} |Z(w)| \geq \epsilon\sqrt{n}) \leq 2P(\sup_{w\in\mathcal{W}} Z(w) \geq \epsilon\sqrt{n})$$

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

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

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 [6] 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 $Z(w)$. At left below, the set $\{w : Z(w) \geq b\}$ is seen 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 are independent. The vertical arrows (below right) illustrate two clump centers (points of the Poisson process); the clumps are the bars centered about the arrows.



[1] The bound above reflects the possible improvement of Vapnik's original exponent by a factor of two.
[2] See chapter 7 of [5] for treatment of some technical details in this limit.

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

$$p_b(w) \simeq \lambda_b(w) E C_b(w). \tag{7}$$

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

$$P(\sup_{w \in \mathcal{W}} Z(w) \geq b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma(w))}{E C_b(w)}\, dw \quad . \tag{8}$$

It remains only to find the mean clump size $EC_b(w)$ in terms of the network architecture and the statistics of $(x, y)$. To give an idea of the results that are possible, suppose that the network activation functions are twice differentiable. Then the $Z$ process can be locally approximated and the clump size determined. This results in estimates of sample size that are of order $d/\epsilon^2$, with a multiplicative factor depending in a simple way on the network and the distribution of the data [7].

It is widely known (e.g., [8]) that probabilities like (8) are determined by behavior near the maximum-variance point, where the dominant numerator term takes its largest value. In classification, for example, the maximum-variance point is at $\mathcal{E}(w) = 1/2$. 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

$$P\Bigl(\sup_{w \in \mathcal{W}} \frac{\nu_{\mathcal{T}}(w) - \mathcal{E}(w)}{\sigma(w)} \geq \epsilon\Bigr) \simeq P\Bigl(\sup_{w \in \mathcal{W}} \frac{Z(w)}{\sigma(w)} \geq \epsilon\sqrt{n}\Bigr) \qquad , \tag{9}$$

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

$$\sup_{w \in \mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sigma(w)} = \sup_{w \in \mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sqrt{\mathcal{E}(w)(1 - \mathcal{E}(w))}} < \epsilon \quad , \tag{10}$$

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 (10) is much more powerful than the corresponding unnormalized one. Sample size estimates using this formulation give results having a functional form similar to (6).

# 3 Empirical Estimates of Clump Size

Conditional on there being a clump center at $w$, the upper bound

$$C_b(w) \leq D_b(w) \equiv \int_{\mathcal{W}} 1_{[0,\infty)}(Z(w') - b)\, dw' \tag{11}$$

is evidently valid: the volume of the clump at $w$ is no larger than the total volume of all clumps. (The right hand side is indeed a function of $w$ because we condition on occurrence of a clump center at $w$. See [9] for more on the tightness of this bound.) To compute its mean, we approximate

$$\begin{aligned} ED_b(w) &= \int_{\mathcal{W}} P(Z(w') \geq b | w \text{ a clump center})\, dw' \\ &\simeq \int_{\mathcal{W}} P(Z(w') \geq b | Z(w) \geq b)\, dw' \qquad . \end{aligned} \tag{12}$$

The point is that occurrence of a clump center at $w_0$ is a smaller class of events than merely $Z(w_0) \geq b$: the latter can arise from a clump center at a nearby $w \in \mathcal{W}$ capturing $w_0$. 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

$$\begin{aligned} \zeta = \zeta(w, w') &= (\sigma/\sigma')\frac{1 - \rho\sigma'/\sigma}{\sqrt{1-\rho^2}} && (13) \\ &= \big((1-\rho)/(1+\rho)\big)^{1/2} \quad \text{(constant variance case)} \qquad . && (14) \end{aligned}$$

Evaluating the conditional probabilities of (12) presents no problem, and we obtain the estimate

$$EC_b(w) \simeq ED_b(w) \simeq \int_{\mathcal{W}} \bar{\Phi}\left((b/\sigma)\zeta\right)\, dw' \qquad . \tag{15}$$

This clump size estimate 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.

This approximation will be used with (8) to find

$$P(\sup_w Z(w) > b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{\int_{\mathcal{W}} \bar{\Phi}\left((b/\sigma)\zeta\right)\, dw'} dw \qquad . \tag{16}$$

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 $\zeta$ is nonnegative. By comparison with results for the differentiable process [7], we expect the estimate (15) 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 clump sizes resulting if $(\forall w')\zeta(w, w') \geq M \gg 0$. To achieve such polynomial sizes, $\zeta$ 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 clump size.* We also remark that this approximation to clump size can serve in calculating a lower bound to the true exceedance probability (not the PCH approximation); see [9].

Here is a practical way to approximate the integral giving $ED_b(w)$ 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 $\mathcal{N}$. For $\gamma < 1$ define a set of significant $w'$

$$\begin{aligned} S_\gamma(w) &= \{w' \in \mathcal{W} : \zeta(w, w') \leq \gamma\} && (17) \\ V_\gamma(w) &= \text{vol}(S_\gamma(w)) && (18) \end{aligned}$$

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

$$ED_b(w) \geq \int_{S_\gamma} \bar{\Phi}((b/\sigma)\zeta)\, dw' \geq V_\gamma(w)\, \bar{\Phi}((b/\sigma)\gamma) \qquad .$$

This apparently crude lower bound for $\bar{\Phi}$ is accurate enough near the origin—noting the above comments this is the region that counts—to give satisfactory results in the cases we have studied.

With this bound we have

$$P(\sup_w Z(w) \geq b) \leq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{V_\gamma(w)\, \bar{\Phi}((b/\sigma)\gamma)}\, dw \simeq \gamma \int_{\mathcal{W}} V_\gamma(w)^{-1} \exp(-(1-\gamma^2)b^2/2\sigma^2)\, dw \quad , \tag{19}$$

because as long as $\gamma$ 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

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

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

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.

# 4 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. We normalize the process $Z(w)$ by its standard deviation $\sigma(w)$ as indicated earlier. The variance of the scaled process is unity and (19) becomes

$$P(\sup_w \frac{Z(w)}{\sigma(w)} \geq b) \leq \frac{\bar{\Phi}(b)}{\bar{\Phi}(b\gamma)} \int_{\mathcal{W}} V_\gamma(w)^{-1}\, dw \tag{20}$$

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 $\gamma$, which in turn depends on $b$. Recomputing the integral for many different $\gamma$ or $b$ values must be avoided.

This can be done if we make the reasonable assumption that

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

with $\alpha = 1$ or $2$ according as the activation functions are differentiable or not. This amounts to supposing the correlation $\rho(w, w')$ falls off quadratically or linearly for $w'$ in a neighborhood of $w$. 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 $\gamma_0$, it can be scaled to a desired $\gamma \ll 1$ via

$$P\Big(\sup_w \frac{Z(w)}{\sigma(w)} \geq b\Big) \leq \frac{\bar{\Phi}(b)}{\bar{\Phi}(b\gamma)\gamma^{\alpha d}} \left(\gamma_0^{\alpha d} \int_{\mathcal{W}} V_{\gamma_0}(w)^{-1}\, dw\right) \quad . \tag{21}$$

Upon differentiating we find the optimal $\gamma$ equals $\sqrt{\alpha d}/b$, and

$$\begin{aligned} P\Big(\sup_w \frac{Z(w)}{\sigma(w)} \geq b\Big) &\leq \left(\frac{b^2}{d}\right)^{\alpha d/2} \bar{\Phi}(b) \left[\frac{\gamma_0^{\alpha d} \int_{\mathcal{W}} V_{\gamma_0}(w)^{-1}\, dw}{\alpha^{\alpha d/2}\, \bar{\Phi}(\sqrt{\alpha d})}\right] && (22) \\ &= (b^2/d)^{\alpha d/2}\, \bar{\Phi}(b) \exp(dQ) && (23) \end{aligned}$$

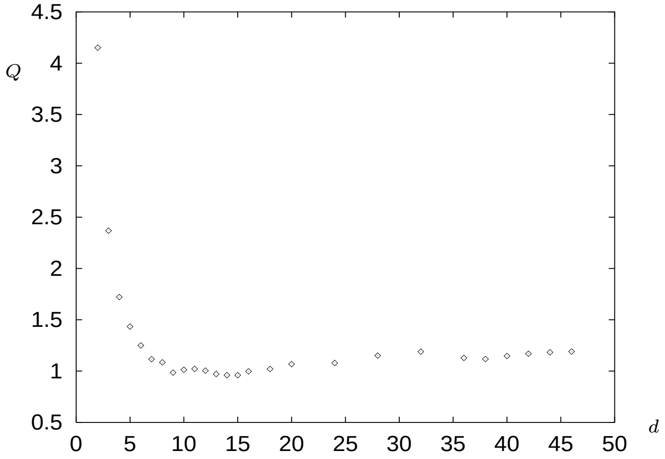
where the final line defines $Q$.

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 $\alpha = 2$.

Below 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

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

This falls below zero at $b^2/d = 5.4$, implying that sample sizes above the critical value $n_c = 5.4d/\epsilon^2$ are enough to ensure (10) with high probability. As in the remarks below that equation, if there is a net having $\nu_{\mathcal{T}}(w) = 0$, we see that sample sizes above $n_c = 5.4d/\epsilon$ will guarantee $\mathcal{E}(w) < \epsilon$ with high probability, which compares favorably with (6).



## 5 Discussion and Conclusions

To find realistic estimates of sample size we transform the original problem into 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.

We obtain a useful estimate for the clump size of a general process in terms of the correlation volume $V_\gamma(w)$. For normalized error, (19) becomes approximately

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

where the expectation is taken with respect to a uniform distribution on $\mathcal{W}$. The probability of reliable generalization is roughly given by an exponentially decreasing factor (the exceedance probability for a single point) times a number representing degrees of freedom. The latter is the mean size of an equivalence class of “similarly-acting” networks. There is an obvious 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. 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.*

To capture this dependence 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 indeed yield informative estimates of sample complexity.

## References

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

[2] E. Baum and D. Haussler. What size net gives valid generalization? In *NIPS 1*, pages 81–90. 1989.

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

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

[5] D. Pollard. *Convergence of Stochastic Processes.* Springer, 1984.

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

[7] M. J. Turmon and T. L. Fine. Sample size requirements for feedforward neural networks. In *NIPS 7.* 1995.

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

[9] M. J. Turmon and T. L. Fine. Generalization in feedforward neural networks. Submitted to *IEEE 1995 Intern. Sympos. Inform. Theory.*

[10] L. Lovász. Geometric algorithms and algorithmic geometry. In *Proc. Internat. Congr. Mathemat.* The Math. Soc. of Japan, 1991.

[11] M. J. Turmon. *Assessing Generalization Ability of Feedforward Neural Networks.* PhD thesis, Cornell, 1995.