Vapnik-Chervonenkis generalization bounds for real valued neural networks

by Arne Hole
Department of Mathematics, University of Oslo
Box 1053, Blindern, N-0316 Oslo, Norway
e-mail: arneh@math.uio.no

Abstract

We show how lower bounds on the generalization ability of feedforward neural nets with real ouputs can be derived within a formalism based directly on the concept of VC dimension and Vapnik's theorem on uniform convergence of estimated probabilities. The formalism can be considered as an alternative to the metric dimension based approach used by D. Haussler in connection to his work on generalizing the PAC model.

1. Introduction

Concerning historical background on the PAC (Probably Approximately Correct) learning model and related issues, I refer to Haussler (1992) and the references therein. The results we obtain in this paper are of the following format, roughly described:

> Let a neural net architecture and a learning algorithm be given. Suppose that you choose a training set x consisting of m examples at random, give it as input to the learning algorithm, and observe that the learned function f_l has (mean) error $\leq \gamma\epsilon$ on the training set x. Then the probability that f_l has global mean error larger than ϵ is less than B, where B is a bound.

The bound B will depend on m, and also on some other quantities. Note that the probability we want to bound is a *conditional* probability; it is the probability that f_l has global error larger than ϵ *given* that it has been observed to have error $\leq \gamma\epsilon$ on the training set.

Comparing the results of this paper to the results on feedforward neural networks given in Haussler (1992), the main difference lies in the domain of applicability. The "sharp" learning criterions considered in the first part of this paper is not covered by Haussler's feedforward network results. On the other hand, Haussler's treatment is far more flexible than the formalism presented here, and it covers a vast number of situations where the results of this paper does not apply. However, concerning learning with respect to continuous "loss functions" (which is treated in the second part of this paper), some comparisions of results can be made. The bounds we obtain in section 11 for the special classes of network models considered

there, are stronger than the bounds obtained for such networks in Haussler's paper (and related works). But then again, the domain of applicability for our results are much more limited. Among other things, we rely on special properties of sigmoid-shaped activation functions. Also, we treat only the case of one hidden layer and one output node. In contrast, Haussler's results are valid for almost any kind of activation functions and node types, and for any number of layers.

On notation: The set of real numbers is denoted $\mathbf{R}$. If h is a set, then $\text{card}(h)$ means the cardinality of h, and $\wp(h)$ is the power set of h. The composition of two maps ψ and ϕ is denoted $\psi \circ \phi$, ie. $\psi \circ \phi(\xi) = \psi(\phi(\xi))$. If A and B are sets, the set of functions $f : A \to B$ from A to B is denoted $Map(A, B)$. The notation A^m means the m-fold cartesian product of A with itself, for each integer $m \geq 1$. If $a = (a_1, \ldots, a_m) \in A^m$ and $b = (b_1, \ldots, b_m) \in B^m$, then by $(a; b)$ we mean the element in $(A \times B)^m$ given by $(a; b)_i = (a_i, b_i)$ for $1 \leq i \leq m$. If α and β are events in some probability model with probability measure P, we write the conditional probability of β given α as $P(\beta \mid \alpha)$. Thus $P(\beta \mid \alpha) = P(\alpha \cap \beta)/P(\alpha)$.

Concerning the organization of the article, I have chosen to treat learning with and without noise as two different cases, starting with the noiseless case. The paper is self-contained with respect to definitions and formalism. All the proofs given are "local", ie. they can be skipped without losing the thread of the paper.

2. VC dimension and related concepts

Let h be an arbitrary set, let $m \geq 1$ be an integer, and let $s = (s_1, \ldots, s_m)$ be an arbitrary ordered sequence of m objects. We define

$$s \cap h = \{i \mid 1 \leq i \leq m \text{ and } s_i \in h\}$$

If H is a family of sets, we define $s \cap H = \{s \cap h \mid h \in H\}$, and put $\Delta_H(s) = \text{card}(s \cap H)$. Note that $s \cap H \subseteq \wp(\{1, \ldots, m\})$. If $\Delta_H(s) = 2^m$, then H is said to *shatter* the sequence s. For each integer $m \geq 1$, define

$$\Delta_H(m) = \max\{\Delta_H(s) \mid s \text{ is a sequence of } m \text{ objects}\}.$$

Let $VCdim(H)$ be the greatest integer m such that $\Delta_H(m) = 2^m$, if such an m exists. Otherwise, let $VCdim(H) = +\infty$. It is known (see eg. Vapnik 1982) that if $d = VCdim(H)$ is finite, then

$$\Delta_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d \tag{1}$$

for all $m \geq d$.

Let A, B be sets, and let $F \subseteq Map(A, B)$. We assume $B \subseteq \mathbf{R}$. For each $f \in F$, let

$$f^+ = \{(p, t) \in A \times B \mid t > f(p)\}$$

Put $F^+ = \{f^+ \mid f \in F\}$. Note that $F^+ \subseteq \wp(A \times B)$. The quantities $\Delta_{F^+}(m)$ and $VCdim(F^+)$ will play an important part in the following. It is easily seen that $VCdim(F^+)$ is equal to the so-called pseudo dimension of F, as defined in Haussler (1992). The result below is shown, among other places, in that paper.

Observation. *Let $\psi : \mathbf{R} \to \mathbf{R}$ be increasing (ie. $x > y \Rightarrow \psi(x) \geq \psi(y)$). Let A be a set and $G \subseteq Map(A, \mathbf{R})$. Define $F \subseteq Map(A, \mathbf{R})$ by $F = \{\psi \circ g \mid g \in G\}$. Then for all m we have $\Delta_{F^+}(m) \leq \Delta_{G^+}(m)$.*

3. Sharp, noiseless learning

By a *noiseless learning situation* (abbreviated "QL situation", Q for "quiet") we will mean a 9-tuple

$$\Lambda = (X, P, Y, F, f_0, S, m, \lambda, \nu)$$

where

X is a set (called the *input space*)

P is a probability measure on X

Y is a set (called the *output space*)

$F \subseteq Map(X, Y)$ is a function class

$f_0 : X \to Y$ is a function (called the *target function*)

S is a set

$m \geq 1$ is an integer

$\lambda : X^m \times S \to F$ is a map (called the *learning algorithm*)

ν is a probability measure on $X^m \times S$ such that the marginal of ν on X^m is P^m

We will usually write $\lambda(x, s)$ as λ_x^σ, for each $x \in X^m$ and $\sigma \in S$. By a *criterion map* (or simply a *criterion*) for the QL situation $(X, P, Y, F, f_0, S, m, \lambda, \nu)$ will be meant a map

$$\theta : Map(X, Y) \to \wp(X)$$

For each $f \in Map(X, Y)$, the set $\theta(f) \subseteq X$ will be interpreted as the region of the input space where f "behaves well" relative to the target f_0. We assume in the following that all combinations of criterion maps with QL situations considered are such that the standard measurability condition assumed in connection with Vapnik's theorem (theorem 1 below) is satified. This is a mild condition that one need not worry about in practice. Consult Pollard (1984).

Given a QL situation $\Lambda = (X, P, Y, F, f_0, S, m, \lambda, \nu)$ and a criterion map θ for it, for each $f \in Map(X, Y)$ and $x \in X^m$ we define the error $E(f, \theta, x)$ of f with

respect to θ on x by

$$E(f,\theta,x) = \frac{1}{m} \cdot \text{card}\{i \mid 1 \leq i \leq m \text{ and } x_i \notin \theta(f)\}.$$

For each $f \in Map(X,Y)$ we define the (global) error $E(f,\theta)$ of f with respect to θ by

$$E(f,\theta) = 1 - P(\theta(f))$$

Finally, for each $t \in [0,1]$ let

$$\Omega_\Lambda(t,\theta) = \nu\{E(\lambda_x^\sigma,\theta,x) \leq t\}$$

In the context of the formalism we will develop in section 10, it is natural to refer to the error measures defined above as "sharp". Hence we may refer to learning with respect to criterions θ as defined in this section as "sharp" learning. We will use the following version of Vapnik's theorem.

Theorem 1 (Vapnik). *Let $\Lambda = (X,P,Y,F,f_0,S,m,\lambda,\nu)$ be a QL situation, and let θ be a criterion for it. Let $\gamma \in [0,1)$, $\epsilon \in (0,1)$ and $m \geq 4/((1-\gamma)^2\epsilon)$. Then*

$$\nu\Big\{E(\lambda_x^\sigma,\theta) > \epsilon \;\Big|\; E(\lambda_x^\sigma,\theta,x) \leq \gamma\epsilon\Big\} < \frac{2\Delta_{\theta(F)}(2m)}{\Omega_\Lambda(\gamma\epsilon,\theta)e^{\epsilon m(1-\gamma)^2/4}}$$

The above version of the theorem is shown in Hole (1995). The proof follows the original one given in Vapnik (1982) closely. Translating theorem 1 into the usual form, it gives an improvement on the bound given in Anthony and Shawe-Taylor (1993) by a factor of two, and on the bound given in Vapnik (1982) by a factor of four. It may be remarked that if the additional assumption is made that ϵm is an integer, then the bound of theorem 1 can (Hole 1995) be improved by an additional factor of two.

4. Interpretation

In this section I will discuss how the formalism of the preceeding section can be interpreted in terms of neural networks. Let $\Lambda = (X,P,Y,F,f_0,S,m,\lambda,\nu)$ be a QL situation. Then X and Y can be taken as the input space and output space of a network architecture, respectively. The class $F \subseteq Map(X,Y)$ can be viewed as the set of functions defined by the architecture (by varying weights and tresholds). The target $f_0 : X \to Y$ is the (possibly unknown) function we want the network to learn. It is not nescessary that $f_0 \in F$. The elements of $x \in X^m$ are training sequences of length m. The learning algorithm λ associates a function in F to each element (x,σ), where $x \in X^m$ is a training sequence and $\sigma \in S$. The set S is

included to model cases where the learning process used is not deterministic. In the deterministic case, we can take $S = \{0\}$. Then the probability measure ν on $X^m \times S$ reduces to the product measure P^m of P on X^m.

Now let us consider criterion maps θ. As hinted in the previous section, for each $f \in Map(X,Y)$ the set $\theta(f)$ will be interpreted as the set of $p \in X$ such that $f(p)$ is "acceptable" when compared to $f_0(p)$. If $Y = \{-1,1\}$ (the boolean case) the obvious choice for θ is the map θ_b given by

$$\theta_b(f) = \{p \in X \mid f(p) = f_0(p)\}$$

for all $f \in Map(X,Y)$. However, in the general case $Y \subseteq \mathbf{R}$ corresponding to networks with real outputs, the criterion θ_b is too restrictive. Let $\kappa > 0$ be fixed. A natural criterion θ to consider in this context is the map θ_κ defined by

$$\theta_\kappa(f) = \{p \in X \mid \ |f(p) - f_0(p)| \leq \kappa\}$$

for all functions $f \in Map(X,Y)$.

Given θ, the quantity $E(f,\theta,x)$ naturally plays the role as the (mean) error of f on the sequence x of m points in X, and $E(f,\theta)$ represents the global (mean) error of f. The quantity $\Omega_\Lambda(t,\theta)$ is the probability that the learned function $\lambda(x,\sigma)$ has error less than or equal to t on the training sequence x when (x,σ) is drawn at random according to ν. Since the marginal of ν on X^m is assumed to be P^m, taking a random draw according to ν can be interpreted as taking a random draw of $x \in X^m$ according to P^m and giving x as input to the (possibly stochastic) learning process. So whether or not the learning process is stochastic, we may conclude that

> The quantity $\Omega_\Lambda(t,\theta)$ is the probability that the function resulting from the learning process has error $\leq t$ on the training set, when the training set $x \in X^m$ is drawn at random according to X^m.

Note that we are considering noiseless learning here; we assume that we have access to the function values $f_0(x_i)$ for all elements $x_1,\ldots,x_m$ in the training sequence. On the other hand, function values of f_0 on training sequences is the only information about f_0 we need. Theorem 1 now says the following:

> Suppose that you choose a training sequence $x \in X^m$ at random according to P^m, give it as input to the learning process, and observe that the resulting learned function f_l has error less than or equal to $\gamma\epsilon$ on the training sequence x, ie. $E(f_l,\theta,x) \leq \gamma\epsilon$. Then the probability (with respect to choice of x) that $E(f_l,\theta) > \epsilon$ is less than
>
> $$\frac{2\Delta_{\theta(F)}(2m)}{\Omega_\Lambda(\gamma\epsilon,\theta)e^{\epsilon m(1-\gamma)^2/4}}$$

If θ is the boolean criterion θ_b defined above and F is the function class implemented by a feedforward neural network architecture with linear treshold units, the quantity $\Delta_{\theta(F)}(2m)$ appearing in theorem 1 can be estimated as in Baum and Haussler (1989). We will see in the following sections how bounds on $\Delta_{\theta_\kappa(F)}(m)$ can be

obtained. However, in order to apply theorem 1 we also need an estimate of the probability $\Omega_\Lambda(\gamma\epsilon, \theta)$ of "success" on the training set. In some practical cases, it will be possible to estimate this in advance by trying out a number of training sets x and observing for how many of them we get training error $\leq \gamma\epsilon$. In other cases, one may be able to prove (or feel reasonably sure) that the probability is close to one, or at least not smaller than $1/2$.

In the following sections we will derive several results having essentially the same form as theorem 1. The above remarks on interpretation are relevant for these results as well.

5. Reduction to the VC dimension of F^+

To obtain generalization bounds valid for the θ_κ criterions defined in the previous section, we need the following lemma.

Lemma 1. *Let $\kappa > 0$, and let $F \subseteq Map(X, Y)$ be a function class, where $Y \subseteq \mathbf{R}$. Then $\Delta_{\theta_\kappa(F)}(m) \leq [\Delta_{F^+}(m)]^2$.*

Proof. Define the maps $\theta_1, \theta_2 : F \to \wp(X)$ by

$$\theta_1(f) = \{p \in X \mid f(p) \leq f_0(p) + \kappa\}$$
$$\theta_2(f) = \{p \in X \mid f(p) \geq f_0(p) - \kappa\}$$

Then $\theta_\kappa(f) = \theta_1(f) \cap \theta_2(f)$ for each $f \in F$, and therefore for each $x \in X^m$

$$\begin{aligned}\Delta_{\theta_\kappa(F)}(x) &= \text{card}\{x \cap \theta_1(f) \cap \theta_2(f) \mid f \in F\} \\ &\leq \text{card}\{x \cap \theta_1(f) \mid f \in F\} \cdot \text{card}\{x \cap \theta_2(f) \mid f \in F\} \\ &= \Delta_{\theta_1(F)}(x) \cdot \Delta_{\theta_2(F)}(x)\end{aligned}$$

To complete the proof, it is now sufficient to show that $\Delta_{\theta_j(F)}(m) \leq \Delta_{F^+}(m)$ for $j = 1, 2$. We wil first show that $\Delta_{\theta_1(F)}(m) \leq \Delta_{F^+}(m)$.

Let $x \in X^m$ be fixed, and choose a finite set $\xi \subseteq F$ such that $\Delta_{\theta_1(\xi)}(x) = \Delta_{\theta_1(F)}(x)$. Let

$$d_0 = \min\{f(x_i) - f_0(x_i) - \kappa \mid f \in \xi,\ 1 \leq i \leq m \text{ and } f(x_i) - f_0(x_i) - \kappa > 0\}$$

Define the injection $\phi : X^m \to (X \times \mathbf{R})^m$ by $\phi(x)_i = (x_i, f_0(x_i) + \kappa + d_0)$. For each $f \in \xi$ and $1 \leq i \leq m$, we then have

$$\begin{aligned}x_i \in \theta_1(f) &\iff f(x_i) \leq f_0(x_i) + \kappa \\ &\iff f(x_i) < f_0(x_i) + \kappa + d_0 \iff \phi(x)_i \in f^+\end{aligned}$$

It follows that $\text{card}\{\phi(x) \cap f^+ \mid f \in \xi\} = \Delta_{\theta_1(\xi)}(x)$. So $\Delta_{F^+}(\phi(x)) = \Delta_{\theta_1(F)}(x)$. Since $x \in X^m$ was arbitrary, it follows immediately that $\Delta_{\theta_1(F)}(m) \leq \Delta_{F^+}(m)$.

The proof that $\Delta_{\theta_2(F)}(m) \leq \Delta_{F+}(m)$ is similar. Let $x \in X^m$ be fixed, and choose a finite set $\xi \subseteq F$ such that $\Delta_{\theta_2(\xi)}(x) = \Delta_{\theta_2(F)}(x)$. This time, define $\phi : X^m \to (X \times \mathbf{R})^m$ by $\phi(x)_i = (x_i, f_0(x_i) - \kappa)$. Then for each $f \in \xi$ and $x \in X^m$

$$x_i \in \theta_2(f) \iff f(x_i) \geq f_0(x_i) - \kappa \iff \phi(x)_i \notin f^+$$

Thus $\phi(x) \cap f^+$ is the *complement* of $\{i \mid x_i \in \theta_2(f)\}$ in $\{1, \ldots, m\}$. Again it follows that $\text{card}\{\phi(x) \cap f^+ \mid f \in F\} = \Delta_{\theta_2(F)}(x)$. The rest is similar to the case of $\Delta_{\theta_1(F)}(m)$. ■

To use lemma 2, we need bounds of $\Delta_{F+}(m)$. The simplest case is when F is a vector space of dimension d. Then $VCdim(F^+) \leq d + 1$, as is essentially shown in Cover (1965). A proof is also given in Haussler (1992).

6. First example

In this section, I will derive an upper bound of $\Delta_{F+}(m)$ in the case where F represents the function class defined by a network architecture with the following properties:

(i) The architecture has a single input node, one hidden layer with n nodes and a single output node.

(ii) The activation function in each computation node is $h(t) = \text{erf}(t)$, where erf denotes the error function (ie. the integral of the normal distribution $N(0, 1)$, with $h(0) = 0$). The hidden nodes have no tresholds.

To be precise, we let $F \subseteq Map(\mathbf{R}, \mathbf{R})$ be the class of all functions f on the form

$$f(p) = \text{erf}\Big(a + \sum_{i=1}^{n} b_i \ \text{erf}(c_i p)\Big),$$

where $a, b_1, c_1, \ldots, b_n, c_n \in \mathbf{R}$. Thus the elements in F are analytic functions from $X = \mathbf{R}$ to $Y = \mathbf{R}$.

Lemma 2. *The function class F defined in this section satisfies*

$$\Delta_{F+}(m) \leq \Big(\frac{em}{4n}\Big)^{4n}$$

This lemma is proved in Appendix 1. Note that the total number of parameters in the architecture defining F is $W = 2n + 1$.

7. Second example

In this section, I will estimate $\Delta_{F^+}(m)$ in the case where F is the function class defined by another special kind of network with one hidden layer.

Let $n, k \geq 1$ be integers, let $\phi : \mathbf{R} \to \mathbf{R}$ be increasing, and let $h_1, \ldots, h_n : \mathbf{R} \to \mathbf{R}$ be piecewise linear funcions with s knots each. For fixed n, k, ϕ and $h_1, \ldots, h_n$, let $F \subseteq Map(\mathbf{R}^k, \mathbf{R})$ be the class of all functions $f : \mathbf{R}^k \to \mathbf{R}$ on the form

$$f(p) = \phi\Big(w_{20} + \sum_{i=1}^{n} w_{2i} h_i(w_{1i} + \sum_{j=1}^{k} w_{ij} p^j)\Big)$$

where $w_{\mu\nu} \in \mathbf{R}$ for all $\mu\nu$. The class F can be interpreted as the function class defined by a layered network architecture with the following characteristics:

(i) The architecture has k input nodes, one hidden layer with n nodes, and a single output node.

(ii) The activation function in hidden node number i is h_i, for $1 \leq i \leq n$.

(iii) The activation function in the output node is ϕ.

Lemma 3. *The function class F described in this section satisfies*

$$\Delta_{F^+}(m) \leq \Big(\frac{m}{k}\Big)^{snk} \Big(\frac{m}{W-1}\Big)^{W-1} \leq \Big(\frac{m}{k}\Big)^{(s+1)W}$$

where $W = nk + 2n + 1$ is the number of parameters in F.

This lemma is proved in Appendix 2. It may be remarked that the proof of lemma 3 quite easily can be generalized to the case where the activation functions h_i of the hidden nodes are piecewise polynomial functions of degree $\leq d$, where $d \geq 1$. The details are omitted.

8. VC_δ-dimension

Note that the method used to prove lemma 2 in Appendix 1 depends strongly on the properties of the particular "sigmoid" activation function $h(t) = \mathrm{erf}(t)$ considered. There exist other sigmoid-looking functions for which the bound of the lemma is utterly false. In Sontag (1992), there is even constructed an analytic, sigmoid-shaped, strictly increasing function $h : \mathbf{R} \to \mathbf{R}$ such that the class $F \subseteq Map(\mathbf{R}, \mathbf{R})$ of functions f on the form

$$f(t) = h(wt) + h(-wt)$$

where $w \in \mathbf{R}$ is the only parameter, satifies $VCdim(F^+) = \infty$. Examples of this type indicate that in order to obtain VC generalization bounds valid for real valued

networks using (say) general "sigmoid-shaped" activation functions, we must change our setup somewhat. To this end, we will now define a more "rigid" version of the VC dimension concept for function classes.

Let $F, H \subseteq Map(X, Y)$ where $Y \subseteq \mathbf{R}$, and let $\delta \geq 0$. The class H is said to be *δ-dense in F* if for every $f \in F$ there is a $f' \in H$ such that

$$\sup_{p \in X} |f(p) - f'(p)| \leq \delta$$

Define

$$\Delta^{\delta}_{F+}(m) = \inf\{\Delta_{H+}(m) \mid H \text{ is } \delta\text{-dense in } F\}$$

Note that $\Delta^{0}_{F+}(m) = \Delta_{F+}(m)$. We define $VC_{\delta}dim(F^{+})$ to be the largest integer m such that $\Delta^{\delta}_{F+}(m) = 2^{m}$. If no such m exists, $VC_{\delta}dim(F^{+}) = \infty$.

Theorem 2 (Sharp, noiseless learning). *Let $\Lambda = (X, P, Y, F, f_0, S, m, \lambda, \nu)$ be a QL situation. Let $\gamma \in [0,1)$, $\epsilon \in (0,1)$, $\kappa > 0$, $\delta \geq 0$ and $m \geq 4/((1-\gamma)^2\epsilon)$. Then*

$$\nu\Big\{E(\lambda^{\sigma}_{x}, \theta_{\kappa}) > \epsilon \;\Big|\; E(\lambda^{\sigma}_{x}, \theta_{\kappa-2\delta}, x) \leq \gamma\epsilon\Big\} < \frac{2[\Delta^{\delta}_{F+}(2m)]^2}{\Omega_{\Lambda}(\gamma\epsilon, \theta_{\kappa-2\delta})e^{\epsilon m(1-\gamma)^2/4}}$$

Proof. Choose H such that H is δ-dense in F and $\Delta_{H+}(2m) = \Delta^{\delta}_{F+}(2m)$. Define $\beta : X^m \times S \to H$ such that $|\beta^{\sigma}_{x}(p) - \lambda^{\sigma}_{x}(p)| \leq \delta$ for all x, σ, p. Then for all x, σ, p we have

$$\begin{aligned} p \notin \theta_{\kappa-\delta}(\beta^{\sigma}_{x}) &\iff |\beta^{\sigma}_{x}(p) - f_0(p)| > \kappa - \delta \\ &\implies |\beta^{\sigma}_{x}(p) - \lambda^{\sigma}_{x}(p)| + |\lambda^{\sigma}_{x}(p) - f_0(p)| > \kappa - \delta \\ &\implies |\lambda^{\sigma}_{x}(p) - f_0(p)| > \kappa - 2\delta \\ &\iff p \notin \theta_{\kappa-2\delta}(\lambda^{\sigma}_{x}) \end{aligned}$$

Hence $E(\beta^{\sigma}_{x}, \theta_{\kappa-\delta}, x) \leq E(\lambda^{\sigma}_{x}, \theta_{\kappa-2\delta}, x) \leq \gamma\epsilon$. In the same manner as above, one can show that $E(\lambda^{\sigma}_{x}, \theta_{\kappa}) \leq E(\beta^{\sigma}_{x}, \theta_{\kappa-\delta})$. So if $E(\lambda^{\sigma}_{x}, \theta_{\kappa-2\delta}, x) \leq \gamma\epsilon$ and $E(\lambda^{\sigma}_{x}, \theta_{\kappa}) > \epsilon$, then $E(\beta^{\sigma}_{x}, \theta_{\kappa-\delta}, x) \leq \gamma\epsilon$ and $E(\beta^{\sigma}_{x}, \theta_{\kappa-\delta}) > \epsilon$. The result now follows by applying theorem 1 to the QL situation Λ' obtained from Λ by replacing F by H and λ by β, under the criterion $\theta_{\kappa-\delta}$. ■

9. Third example

To estimate $\Delta^{\delta}_{F+}(m)$ for a given function class F, a natural strategy is to find a class H such that (i) H is δ-dense in F, and (ii) we are able to bound $\Delta_{H+}(m)$. In this section, we will estimate $\Delta^{\delta}_{F+}(m)$ in the case where F is the function class defined by a quite general network architecture with one hidden layer, using a class covered by lemma 3 as H.

Let F be defined as in section 7, except that now (i) we allow the activation functions h_i in the hidden nodes to be arbitrary functions, (ii) we assume that there is a real constant M such that

$$\sum_{i=1}^{n} |w_{2i}| \leq M$$

for all $f \in F$, and (iii) we assume that the activation function ϕ of the output node satisfies the Lipschitz bound $|\phi(t_1) - \phi(t_2)| \leq |t_1 - t_2|$ for all $t_1, t_2 \in \mathbf{R}$. Note that as in section 7, the total number W of parameters in F is given by $W = nk + 2n + 1$. Combined with theorem 2, the following lemma yields a generalization bound valid for the class F.

Lemma 4. *Let F a function class of the type described above. Suppose that for each $1 \leq i \leq n$ there is a piecewise linear function g_i with s knots such that $|g_i(t) - h_i(t)| \leq \delta/M$ for all $t \in \mathbf{R}$. Then*

$$\Delta^{\delta}_{F+}(m) \leq \left(\frac{m}{k}\right)^{(s+1)W}$$

Proof. Define $\psi : F \to Map(X, \mathbf{R})$ by letting $\psi(f)$ be the function obtained from f by replacing h_i with g_i for all i. Put $H = \psi(F)$. Let $f \in F$ have parameter values w, and put

$$A(p) = w_{1i} + \sum_{j=1}^{k} w_{ij} p^j$$

By utilizing the Lipschitz bound on ϕ, we see that for all $p \in X$

$$\begin{aligned} |\psi(f)(p) - f(p)| &\leq \left| \sum_{i=1}^{n} w_{2i} \Big[g_i(A(p)) - h_i(A(p)) \Big] \right| \\ &\leq \sum_{i=1}^{n} |w_{2i}| \frac{\delta}{M} \leq \delta \end{aligned}$$

Hence H is δ-dense in F. By lemma 3, $\Delta_{H+}(m) \leq (m/k)^{(s+1)W}$. The result follows by the definition of $\Delta^{\delta}_{F+}(m)$. ∎

10. Noise and general loss functions

In this section I will describe how the preceeding results can be adapted to situations where a fixed, noiseless target function f_0 is not given, or where one works with a "non-sharp" learning criterion which cannot be expressed in terms of a map θ of the type we have been considering.

By a *noisy learning situation* (NL situation) we will mean a 8-tuple

$$\Lambda = (X, Y, P, F, S, m, \lambda, \nu)$$

where

X is a set (called the *input space*)

Y is a set (called the *output space*)

P is a probability measure on $X \times Y$

$F \subseteq Map(X, Y)$ is a function class

S is a set

$m \geq 1$ is an integer

$\lambda : (X \times Y)^m \times S \to F$ is a map (called the *learning algorithm*)

ν is a probability measure on $(X \times Y)^m \times S$ such that the marginal of ν on $(X \times Y)^m$ is P^m

We use the letter Z to denote the product $X \times Y$, and we denote the image $\lambda(Z^m \times S)$ by F_λ. Note that $F_\lambda \subseteq F$. As before, we write $\lambda(z, \sigma)$ as λ_z^σ.

A map $L : \mathbf{R} \times \mathbf{R} \to [0, \infty)$ will be called a *loss function* provided there is an increasing map $\mu_L : [0, \infty) \to [0, \infty)$ such that

$$L(a, b) = \mu_L(|a - b|)$$

Typical examples are $L(a, b) = (a - b)^2$ (quadratic loss) and $L(a, b) = |a - b|$ (standard distance loss). We assume in the following that all combinations of loss functions L with NL situations considered are such that for all $f \in F$ the function $(p, t) \mapsto L(t, f(p))$ defined on $X \times Y$ is measurable, and such that the standard measurability condition needed for the use of Vapnik's theorem below is satisfied (cf. the comments in section 3). Again, these are mild conditions that can be ignored in practice.

Given a NL situation $\Lambda = (X, Y, P, F, S, m, \lambda, \nu)$ and a loss function L, for each $f \in Map(X, Y)$ and $z = (x; y) \in Z^m$ we define the error $E(f, L, z)$ of f with respect to L on z by

$$E(f, L, z) = \frac{1}{m} \sum_{i=1}^{m} L(y_i, f(x_i))$$

For each $f \in Map(X, Y)$ we define the (global) error $E(f, L)$ of f with respect to L by

$$E(f, L) = \int_{X \times Y} L(t, f(p)) \; dP(p, t)$$

Finally, for each $t \in [0, 1]$ let

$$\Omega_\Lambda(t, L) = \nu\{E(\lambda_z^\sigma, L, z) \leq t\}$$

The main difference between a QL situation and an NL situation is that in the latter case the probability distribution P is defined on $X \times Y$ instead of on X only. We do not have access to any particular target function f_0, and instead we are trying to learn an input-output *relation* on $X \times Y$. Thus the probability distribution P itself plays the role as "target" in an NL situation. The "sharp" loss function L_κ defined by

$$L_\kappa(a,b) = \begin{cases} 1 & \text{if } |a-b| > \kappa \\ 0 & \text{if } |a-b| \leq \kappa \end{cases}$$

where $\kappa > 0$ is fixed, corresponds to the θ_κ learning criterion considered in the previous sections. The only difference between the previous setup and the present one is that now the model is designed to treat noisy situations. However, our main result goes through exactly as before:

Theorem 3 (Sharp, noisy learning). *Let $\Lambda = (X, Y, P, F, S, m, \lambda, \nu)$ be an NL situation. Let $\gamma \in [0,1)$, $\epsilon \in (0,1)$, $\kappa > 0$ and $\delta \geq 0$. Then*

$$\nu\Big\{ E(\lambda_z^\sigma, L_\kappa) > \epsilon \;\Big|\; E(\lambda_z^\sigma, L_{\kappa-2\delta}, z) \leq \gamma\epsilon \Big\} < \frac{2[\Delta_{F^+}^\delta(2m)]^2}{\Omega_\Lambda(\gamma\epsilon, L_{\kappa-2\delta}) e^{\epsilon m(1-\gamma)^2/4}}$$

Proof. Assume first $\delta = 0$. Define $\theta_\kappa : F \to \wp(Z)$ by $\theta_\kappa(f) = \{(p,t) \mid L_\kappa(p,t) = 0\}$. Then a variant of theorem 1 yields the formula of theorem 3 with $\Delta_{\theta_\kappa(F)}(2m)$ instead of $[\Delta_{F^+}(2m)]^2$ (consult Hole (1995) for details). The proof that $\Delta_{\theta_\kappa(F)}(2m) \leq [\Delta_{F^+}(2m)]^2$ in this situation is analogous to the proof of lemma 1. Define maps $\theta_1, \theta_2 : F \to \wp(Z)$ by $\theta_1(F) = \{(p,t) \in X \times Y \mid t \geq f(p) - \kappa\}$ and $\theta_2(F) = \{(p,t) \in X \times Y \mid t \leq f(p) + \kappa\}$. Then

$$\Delta_{\theta_\kappa(F)}(x;y) \leq \Delta_{\theta_1(F)}(x;y) \cdot \Delta_{\theta_2(F)}(x;y)$$

for all $(x;y) \in (X \times Y)^m$. To show that $\Delta_{\theta_1(F)}(m) \leq \Delta_{F^+}(m)$, let $(x;y) \in (X \times Y)^m$ be fixed, and choose a finite set $\xi \subseteq F$ such that $\Delta_{\theta_1(\xi)}(x;y) = \Delta_{\theta_1(F)}(x;y)$. Let

$$d_0 = \min\{f(x_i) - \kappa - y_i \mid f \in \xi,\, 1 \leq i \leq m \text{ and } f(x_i) - \kappa - y_i > 0\}$$

Define the injection $\phi : (X \times Y)^m \to (X \times Y)^m$ by $\phi(x;y)_i = (x_i, y_i + \kappa + d_0)$. For each $f \in \xi$ and $1 \leq i \leq m$, we then have

$$\begin{aligned} z_i \in \theta_1(f) &\iff y_i \geq f(x_i) - \kappa \\ &\iff y_i > f(x_i) - \kappa - d_0 \iff \phi(x;y)_i \in f^+ \end{aligned}$$

So $\text{card}\{\phi(x;y) \cap f^+ \mid f \in \xi\} = \Delta_{\theta_1(\xi)}(x;y)$. Thus since $(x;y)$ was arbitrary, $\Delta_{\theta_1(F)}(m) \leq \Delta_{F^+}(m)$. To show $\Delta_{\theta_2(F)}(m) \leq \Delta_{F^+}(m)$, let again $(x;y) \in (X \times Y)^m$ be fixed, and choose a finite set $\xi \subseteq F$ such that $\Delta_{\theta_2(\xi)}(x;y) = \Delta_{\theta_2(F)}(x;y)$. Define

the injection $\phi : (X \times Y)^m \to (X \times Y)^m$ by $\phi(x;y)_i = (x_i, y_i - \kappa)$. For each $f \in \xi$ and $1 \leq i \leq m$, we then have

$$z_i \in \theta_2(f) \iff y_i \leq f(x_i) + \kappa \iff \phi(x;y)_i \notin f^+$$

The conclusion $\Delta_{\theta_2(F)}(m) \leq \Delta_{F+}(m)$ follows.

Then consider the case $\delta > 0$. Expressing things in terms of the map θ_κ introduced above, this follows from the case $\delta = 0$ of the theorem by an argument very similar to the proof of theorem 2. The details are omitted. ■

Now let $f : X \to Y$ be a function, where $Y \subseteq \mathbf{R}$. Let L be an arbitrary loss function. For each $\tau \in [0, \infty)$, let $f_\tau^L = \{(p,t) \in X \times Y \mid L(t, f(p)) > \tau\}$. Also, let

$$m_2^L(f) = \sqrt{\int_{X \times Y} [L(t, f(p))]^2 dP(p,t)}$$

if this moment exists. If $F \subseteq Map(X,Y)$ is a function class, let us define $F_L = \{f_\tau^L \mid f \in F \text{ and } \tau \in [0, \infty)\}$.

We say that a loss function L is c-Lipschitz if there is a $c \in \mathbf{R}$ such that $|\mu_L(a) - \mu_L(b)| \leq c|a - b|$ for all $a, b \in [0, \infty)$. If the map μ_L is continuous and strictly increasing, then we call L continuous and strictly increasing (abbreviated CASI) as well.

Lemma 5. *Let $F \subseteq Map(X,Y)$, where $Y \subseteq \mathbf{R}$. Assume that F is closed under addition of constant functions, and that L is CASI. Then*

$$\Delta_{F_L}(m) \leq [\Delta_{F+}(m)]^2$$

Proof. We use our standard trick once again. Let $z = (x;y) \in Z^m$ be arbitrary, and choose a finite set $\xi \subseteq F$ such that $\Delta_{(\xi)_L}(z) = \Delta_{F_L}(z)$. Let τ be a point in the range of μ_L (clearly it is enough to consider such τ). Let $a = \mu_L^{-1}(\tau)$, and put

$$b = \min \left\{ |y_i - f(x_i)| - a \;\middle|\; f \in \xi,\ 1 \leq i \leq m \text{ and } (x_i, y_i) \in f_\tau^L \right\}$$

Let $\gamma_\tau = a + b/2$. For each $f \in \xi$, let

$$f^{up} = \{(p,t) \in Z \mid t > f(p) + \gamma_\tau\} = (f + \gamma_\tau)^+$$
$$f^{down} = \{(p,t) \in Z \mid t \leq f(p) - \gamma_\tau\} = Z \setminus (f - \gamma_\tau)^+$$

where $f + \gamma_\tau$ and $f - \gamma_\tau$ denote the functions obtained from f by adding the constants γ_τ and $-\gamma_\tau$, respectively. Observe that $z \cap f_\tau^L = (z \cap f^{up}) \cup (z \cap f^{down})$, so

$$\begin{aligned} \text{card}(z \cap F_L) &\leq \text{card}\{z \cap f^{up} \mid f \in F\} \cdot \text{card}\{z \cap f^{down} \mid f \in F\} \\ &\leq \Delta_{F+}(z) \cdot \Delta_{F+}(z). \quad \blacksquare \end{aligned}$$

Let $\Lambda = (X, Y, P, F, S, m, \lambda, \nu)$ be an NL situation, let L be a loss function, and let $\delta \geq 0$, $\tau \geq 1$. A map $\psi : F \to Map(X, Y)$ will be called a *(δ, L)-balancer for Λ with bound τ* if (i) $|\psi(f)(p) - f(p)| \leq \delta$ for all $f \in F$ and $p \in X$, and (ii) we have $m_2^L(\psi(\lambda_z^\sigma)) \leq \tau E(\psi(\lambda_z^\sigma))$ for all $(z, \sigma) \in Z^m \times S$.

Theorem 4 (Smooth, noisy learning). *Let $\Lambda = (X, Y, P, F, S, m, \lambda, \nu)$ be an NL situation, let L be a CASI loss function, and let ψ be a (δ, L)-balancer for Λ with bound $\tau \geq 1$, where $\delta \geq 0$. Assume that F is closed under addition of constants. Let $\gamma \in [0, 1)$, $\epsilon \in (0, 1)$ and $m \geq 4\tau^4/(1-\gamma)^4$. If $\delta > 0$, assume that L is c-Lipschitz. Then*

$$\nu\Big\{E(\lambda_z^s, L) > \epsilon \;\Big|\; E(\lambda_z^s, L, z) \leq \gamma\epsilon - 2c\delta\Big\} < \frac{2[\Delta_{\psi(F)+}(2m)]^2}{\Omega_\Lambda(\gamma\epsilon - 2c\delta, L)e^{m(1-\gamma)^4/(4\tau^4)}}$$

This theorem is proved in Appendix 3. The bound of theorem 4 has the advantage over our previous bounds that the expression in the exponent does not depend on ϵ. What gives theorem 4 its extra strength, is the assumption

$$\text{There exists a } (\delta, L)\text{-balancer for } \Lambda \text{ with bound } \tau \tag{2}$$

It is easy to see that if L is a "sharp" loss function L_κ of the type considered in theorem 3, then (2) does not hold under any reasonably general conditions. For CASI loss functions such as $L(a, b) = (a - b)^2$ or $L(a, b) = |a - b|$ however, the assumption is not so unreasonable. For example, if we assume that we have a map $\psi : F \to Map(X, Y)$ such that $|\psi(f)(p) - f(p)| \leq \delta$ for all $f \in F$ and $p \in X$, and

$$\begin{aligned}&\text{The random variable } X_f(p, t) = t - \psi(f)(p) \text{ is normally}\\ &\text{distributed under } P \text{ on } X \times Y \text{ for all } f \in F_\lambda\end{aligned} \tag{3}$$

then easy calculations show that for the loss functions $L(a, b) = (a - b)^2$ and $L(a, b) = |a - b|$, the condition (2) holds with $\tau = \sqrt{3}$ and $\tau = \sqrt{\pi/2}$, respectively. The condition (3) may often be a good approximation in practice. Some additional comments on these matters can be found in Vapnik (1982) and Bottou and Vapnik (1993).

11. Fourth example

Consider an NL situation $\Lambda = (X, Y, P, F, S, m, \lambda, \nu)$ where $Z = X \times Y = \mathbf{R}^k \times \mathbf{R}$, and where $F \subseteq Map(\mathbf{R}^k, \mathbf{R})$ is defined as in section 7, except for the following: (i) the activation functions h_i of the hidden nodes are allowed to be arbitrary functions, (ii) the activation function ϕ of the output node is the identity, and (iii) there is a constant M such that

$$\sum_{i=1}^{n} |w_{2i}| \leq M$$

for all $f \in F$. Let $L(a,b) = |a-b|$ be the standard distance loss function, and let the integer s be such that for each $1 \leq i \leq n$ there is a piecewise linear function $g_i : \mathbf{R} \to \mathbf{R}$ having s knots with $|g_i(t) - h_i(t)| \leq \delta/M$ for all $t \in \mathbf{R}$, where $\delta \geq 0$. Define $\psi : F \to Map(X,Y)$ by letting $\psi(f)$ be the function obtained from f by replacing h_i by its approximation g_i for $1 \leq i \leq n$. Then we have the following result:

Corollary. *Consider the setup described in this section, and assume that the normality condition (3) holds. Let $\epsilon \in (0,1)$, $\delta \geq 0$, $m \geq 4\tau^4/(1-\gamma)^4$, and let $W = nk + 2n + 1$ be the total number of parameters in F. Then*

$$\nu\Big\{E(\lambda_x^\sigma, L) > \epsilon \;\Big|\; E(\lambda_x^\sigma, L, x) \leq (\epsilon/2) - 2\delta\Big\} < \frac{2(m/k)^{2(s+1)W}}{\Omega_\Lambda(\gamma\epsilon - 2\delta, L)e^{m/(16\pi^2)}}$$

Proof. Reasoning as in the proof of lemma 4 and using the normality assumption (3), we see that ψ is a (δ, L)-balancer for Λ with bound $\tau = \sqrt{\pi/2}$. Note that L is CASI and 1-Lipschitz, and that F is closed under addition of constants. Combine theorem 4 with the bound on $\Delta_{\psi(F)^+}(m)$ given by lemma 3, taking $\gamma = 1/2$. ■

Let us consider the following three special cases. In all three cases we take $h_1 = h_2 = \ldots = h_n$, ie. the hidden nodes have a common activation function. This function will be denoted by h.

Case 1: h is piecewise linear. In this case, we can take δ as zero and s as the number of knots in h. An easy calculation shows that if

$$m \geq \Big(64(s+1)\pi^2 W\Big) \ln\Big(64(s+1)\pi^2 W\Big)$$

then the bound of the corollary is less than $2\Omega_\Lambda(\epsilon/2, L)^{-1} e^{-6(s+1)W}$. As a particular example, we may take the popular activation function h defined by $h(t) = -1$ for $t > -1$, $h(t) = t$ for $t \in [-1,1]$ and $h(t) = 1$ for $t > 1$. In this case $s = 2$.

Case 2: h is the truncated sigmoid given by $h(t) = \tanh(-a)$ for $t < a$, $h(t) = \tanh t$ for $t \in [-a,a]$ and $h(t) = \tanh a$ for $t > a$, where $a > 0$ is fixed. To simplify some estimates, assume $\epsilon \leq M/2$. Choose $\delta = \epsilon/8$. Let $t_1, t_2 \in \mathbf{R}$ with $t_2 > t_1$, and let $\zeta : \mathbf{R} \to \mathbf{R}$ be the linear function passing through the two points $p_i = (t_i, \tanh t_i)$ for $i = 1,2$. Assume that the length of the straight line segment of ζ between p_1 and p_2 is less than or equal to $3\sqrt{\delta/M}$. It is easy to check that the graph of $\tanh t$ (considered as a curve in $\mathbf{R}^2$) has curvature less than $1/2$ for all t. Then by comparing to a circular arc of radius 2 (which has constant curvature $1/2$) and remembering that $d/dt(\tanh t) \leq 1$ for all t, it follows that

$$|\zeta(t) - \tanh t| < \delta/M$$

for all $t \in [t_1, t_2]$. Using line segments of this type, we can construct a piecewise linear g such that $|g(t) - h(t)| < \delta/M$ for all t. We take g continuous, place all the s knots of g on the graph of h, and put $g(t) = \tanh a$ for $t \geq a$, $g(t) = \tanh(-a)$ for $t \leq -a$. The arc length along the graph of h from $(-a, \tanh(-a)))$ to $(a, \tanh a)$ is clearly less than $2a + 2$. So

$$s + 1 = 2 + \frac{2a+2}{3\sqrt{\delta/M}} < \frac{(2/3)(a+2)\sqrt{8M}}{\sqrt{\epsilon}}$$

is sufficient. The inequality of the corollary can now be written

$$\nu\Big\{E(\lambda_x^\sigma, L) > \epsilon \;\Big|\; E(\lambda_x^\sigma, L, x) \leq \epsilon/4\Big\} < \frac{2(m/k)^{(4/3)(a+2)W(8M/\epsilon)^{1/2}}}{\Omega_\Lambda(\epsilon/4, L)e^{m/(16\pi^2)}}$$

So in this case, if

$$m \geq \left(\frac{128}{3}\pi^2(a+2)W\sqrt{\frac{8M}{\epsilon}}\right)\ln\left(\frac{128}{3}\pi^2(a+2)W\sqrt{\frac{8M}{\epsilon}}\right)$$

then the bound of the corollary is less than $2\Omega_\Lambda(\epsilon/4, L)^{-1}e^{-6W}$.

Case 3: h is the *standard sigmoid* given by $h(t) = \tanh t$ for all t. Again assume $\epsilon \leq M/2$, and choose $\delta = \epsilon/8$. As before, we choose g constant on each side of an interval of the type $[-a, a]$. The only difference between this case and the previous one, is that now we need $1 - \tanh a = \delta/M$, ie. $a = \frac{1}{2}\ln(2M/\delta - 1)$. Thus

$$s + 1 = 2 + \frac{\ln(2M/\delta - 1) + 2}{3\sqrt{\delta/M}} < \sqrt{\frac{8M}{\epsilon}}\ln\frac{8M}{\epsilon}$$

is sufficient. (Remember $\delta/M \leq 1/16$.) The inequality of the corollary can now be written

$$\nu\Big\{E(\lambda_x^\sigma, L) > \epsilon \;\Big|\; E(\lambda_x^\sigma, L, x) \leq \epsilon/4\Big\} < \frac{2(m/k)^{2W(8M/\epsilon)^{1/2}\ln(8M/\epsilon)}}{\Omega_\Lambda(\epsilon/4, L)e^{m/(16\pi^2)}}$$

So in this case, if

$$m \geq \left(64\pi^2 W\sqrt{\frac{8M}{\epsilon}}\ln\frac{8M}{\epsilon}\right)\ln\left(64\pi^2 W\sqrt{\frac{8M}{\epsilon}}\ln\frac{8M}{\epsilon}\right)$$

then the bound of the corollary is less than $2\Omega_\Lambda(\epsilon/4, L)^{-1}e^{-6W}$.

To sum up, the length m of the training sequence x needed to have "high" probability of global error $E(\lambda_x^\sigma, L) < \epsilon$ given that $E(\lambda_x^\sigma, L, x) \leq \epsilon/4$, scales as least as good as follows in the three cases considered above:

$$\begin{aligned}
&\textit{Case 1:} \quad m \sim (KW)\ln(KW)\\
&\textit{Case 2:} \quad m \sim \left(\frac{KW}{\sqrt{\epsilon}}\right)\ln\left(\frac{KW}{\sqrt{\epsilon}}\right)\\
&\textit{Case 3:} \quad m \sim \left(\frac{KW}{\sqrt{\epsilon}}\ln\frac{K'}{\epsilon}\right)\ln\left(\frac{KW}{\sqrt{\epsilon}}\ln\frac{K'}{\epsilon}\right)
\end{aligned}$$

where K, K' are constants estimated above in each case. (Note that in fact we worked with $\epsilon/2$ instead of $\epsilon/4$ in case 1.) Thus the scaling laws obtained in the three cases above are all better than the corresponding ones obtained in Haussler (1992).

Appendix 1: Proof of lemma 2

By (1) of section 2, it is enough to show that $VCdim(F^+) \leq 4n$. Since erf is strictly increasing, by the observation of section 2 it is enough to consider the class G of all functions $f : \mathbf{R} \to \mathbf{R}$ of the form

$$f(p) = a + \sum_{i=1}^{n} b_i \ \mathrm{erf}(c_i p)$$

where $a, b_1, c_1, \ldots, b_n, c_n \in \mathbf{R}$. Assume that $VCdim(G^+) = N$. Then there exists a sequence $(x; y) \in (\mathbf{R} \times \mathbf{R})^N$ with $x_1 < x_2 < \ldots < x_N$ and functions $f_1, f_2 \in G$ such that for all $1 \leq i \leq N$ we have $f_1(x_i) > y_i$ for i odd, $f_1(x_i) \leq y_i$ for i even, while $f_2(x_i) \leq y_i$ for i odd, $f_2(x_i) > y_i$ for i even. Let $g = f_1 - f_2$. Then $g(x_i) < 0$ for i odd, and $g(x_i) > 0$ for i even. It follows that g has at least $N - 1$ zeros in the interval $(x_1, x_N) \subseteq \mathbf{R}$. But then the derivative g' must have at least $N - 2$ zeros in the same interval. Further, g can be written in the form

$$g(p) = a + \sum_{i=1}^{2n} b_i \ \mathrm{erf}(c_i p)$$

so

$$g'(p) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{2n} b_i c_i e^{-c_i^2 p^2/2}$$

It is known (see Braess (1986), chapter IV, for instance) that exponential sums of the type

$$\sum_{i=1}^{r} \alpha_i e^{\beta_i u}$$

(with $\alpha_i, \beta_i \in \mathbf{R}$ for all i) has at most $r - 1$ zeros for $u \in \mathbf{R}$. Since the map $p \mapsto p^2$ is at most two to one, it follows that $N - 2 \leq 2(2n - 1)$, or $N \leq 4n$. Thus $VCdim(F^+) \leq 4n$. ■

Appendix 2: Proof of lemma 3

By the observation of section 2, we may take ϕ to be the identity. Let $(x; y) \in (X \times \mathbf{R})^m$ be given. Let $\beta_{i1}, \ldots, \beta_{is}$ be the knots of h_i for each i. For each $f \in F$,

consider the sn associated half spaces $H_{f,ir}$ in $\mathbf{R}^k$ consisting of those $p \in \mathbf{R}^k$ satisfying

$$w_{1i} + \sum_{j=1}^{k} w_{ij} p^j \geq \beta_{ir}$$

for $1 \leq i \leq n$ and $1 \leq r \leq s$, where the parameters w correspond to f. Let

$$\begin{aligned} \Theta_{f,ir} &= \{\mu \mid x_\mu \in H_{f,ir}\} \\ \Theta_{ir} &= \{\Theta_{f,ir} \mid f \in F\} \\ \Theta_f &= \{\Theta_{f,ir} \mid 1 \leq i \leq n \text{ and } 1 \leq r \leq s\} \end{aligned}$$

Define an equivalence relation $\sim$ on F by $f \sim g \iff \Theta_f = \Theta_g$. Since for each combination of i and r we have (Cover, 1965)

$$\text{card } \Theta_{ir} \leq 2 \sum_{i=0}^{k} \binom{m-1}{i} \leq \left(\frac{m}{k}\right)^k$$

it follows that the number K of equivalence classes under $\sim$ satisfies

$$K \leq \prod_{i=1}^{n} \prod_{r=1}^{s} \text{card } \Theta_{ir} \leq \left(\frac{m}{k}\right)^{snk} \tag{4}$$

Let F_0 be one of the equivalence classes. Pick $f_0 \in F_0$. Define an equivalence relation $\cong$ on $\{x_1, \ldots, x_m\}$ by letting $x_\mu \cong x_\nu$ iff

$$\{(i,r) \mid x_\mu \in H_{f_0,ir}\} = \{(i,r) \mid x_\nu \in H_{f_0,ir}\}$$

Let $D_1, \ldots, D_N$ be the equivalence classes under $\cong$. Let C_ξ be the convex hull in $\mathbf{R}^k$ of the set of points in D_ξ, for $1 \leq \xi \leq N$. Then for each ξ and i, there exist real numbers a_i^ξ and b_i^ξ such that

$$f_0(p) = w_{20}^0 + \sum_{i=1}^{n} w_{2i}^0 \left[a_i^\xi (w_{1i}^0 + \sum_{j=1}^{k} w_{ij}^0 p^j) + b_i^\xi \right]$$

for all $p \in C_\xi$, where the parameters w^0 correspond to f_0. Moreover, the restriction of an *arbitrary* $f \in F_0$ to C_ξ can be written

$$f(p) = w_{20} + \sum_{i=1}^{n} w_{2i} \left[a_i^\xi (w_{1i} + \sum_{j=1}^{k} w_{ij} p^j) + b_i^\xi \right] \tag{5}$$

for the *same* numbers a_i^ξ and b_i^ξ. The equation (5) can be rewritten in the form

$$f(p) = w_{20} + \sum_{i=1}^{n} w_{2i} b_i^\xi + \sum_{i=1}^{n} w_{2i} w_{1i} a_i^\xi + \sum_{i=1}^{n} \sum_{j=1}^{k} w_{2i} w_{ij} a_i^\xi p^j$$

For each pair of integers $1 \leq \xi \leq N$ and $1 \leq J \leq W$, define the map $\psi_J^\xi : \mathbf{R}^k \to \mathbf{R}$ by

$$\begin{aligned}
&\psi_1^\xi(p) = 1 \\
&\psi_{1+i}^\xi(p) = b_i^\xi \qquad \text{for } 1 \leq i \leq n \\
&\psi_{1+n+i}^\xi(p) = a_i^\xi \qquad \text{for } 1 \leq i \leq n \\
&\psi_{1+2n+(j-1)k+i}^\xi(p) = a_i^\xi p^j \qquad \text{for } 1 \leq i \leq n,\ 1 \leq j \leq k
\end{aligned}$$

Then define $\psi_J : \bigcup_\xi C_\xi \to \mathbf{R}$ for $1 \leq J \leq W$ by $\psi_J(p) = \psi_J^\xi(p)$ for $p \in C_\xi$. Finally, define $\psi : \bigcup_\xi C_\xi \to \mathbf{R}^W$ by putting $\psi(p) = (\psi_1(p), \ldots, \psi_W(p))$. Now assume $x_\mu \in C_{\xi_\nu}$. Let $f \in F_0$ have parameters w. Then

$$\begin{aligned}
f(x_\mu) > t \iff\ & w_{20} + \sum_{i=1}^n w_{2i} b_i^{\xi_\nu} + \sum_{i=1}^n w_{2i} w_{1i} a_i^{\xi_\nu} \\
& + \sum_{i=1}^n \sum_{j=1}^k w_{2i} w_{ij} a_i^{\xi_\nu} x_\mu^j > y_\mu \\
\iff\ & (w_{20} - y_\mu)\psi_1^{\xi_\nu}(x_\mu) + \sum_{i=1}^n w_{2i} \psi_{1+i}^{\xi_\nu}(x_\mu) \\
& + \sum_{i=1}^n w_{2i} w_{1i} \psi_{1+n+i}^{\xi_\nu}(x_\mu) \\
& + \sum_{i=1}^n \sum_{j=1}^k w_{2i} w_{ij} \psi_{1+2n+(j-1)k+i}^{\xi_\nu}(x_\mu) > 0 \\
\iff\ & (w_{20} - y_\mu)\psi_1(x_\mu) + \sum_{i=1}^n w_{2i} \psi_{i+1}(x_\mu) \\
& + \sum_{i=1}^n w_{2i} w_{1i} \psi_{1+n+i}(x_\mu) \\
& + \sum_{i=1}^n \sum_{j=1}^k w_{2i} w_{ij} \psi_{1+2n+(j-1)k+i}(x_\mu) > 0
\end{aligned}$$

for $1 \leq \mu \leq m$. Hence we see that f induces a particular homogenously linearly separable dichotomy on the set $\{\psi(x_1), \ldots, \psi(x_m)\}$ in $\mathbf{R}^W$. Moreover, this dichotomy uniquely determines $(x; y) \cap f^+$. Thus it follows from Cover's formula (Cover 1965) that

$$\Delta_{F_0^+}(x; y) \leq 2 \sum_{i=0}^{W-1} \binom{m-1}{i} \leq \left(\frac{m}{W-1}\right)^{W-1}$$

But since this estimate will be valid for all the $\sim$ equivalence classes, it follows that $\Delta_{F^+}(x; y) \leq K \cdot (m/(W-1))^{W-1}$. Substituting (4), the result follows. ■

Appendix 3: Proof of theorem 4

First assume $\delta = 0$. Then ψ is the identity. Using lemma 5 to replace the VC dimension bounds, its follows from the proof of theorem 7.6 in (Vapnik 1982) that for all $a \leq 1$

$$\nu\left\{\frac{E(\lambda_z^\sigma, L) - E(\lambda_z^\sigma, L, z)}{E(\lambda_z^\sigma, L)} > \tau V_2(a)\right\} \leq \Gamma \tag{6}$$

where $\Gamma = c_0[\Delta_{F^+}(2m)]^2 e^{-a^2 m/4}$, c_0 is a constant and $V_2(a) < \sqrt{a}$. Vapnik uses the value 8 for c_0. However, theorem 1 of section 2 improves the bound of the underlying theorem on uniform convergence by a factor of 4, so we may take $c_0 = 2$. Assume $z \in Z^m$ is such that $E(\lambda_z^\sigma, L) > \epsilon$ and $E(\lambda_z^\sigma, L, z) \leq \gamma\epsilon$. Then $E(\lambda_z^\sigma, L, z) < \gamma E(\lambda_z^\sigma, L)$, so

$$\frac{E(\lambda_z^\sigma, L) - E(\lambda_z^\sigma, L, z)}{E(\lambda_z^\sigma, L)} > 1 - \gamma$$

Now use (6), with $\sqrt{a}$ substituted for $V_2(a)$ and $1 - \gamma = \tau\sqrt{a}$. The condition $\tau \geq 1$ ensures that $a = (1-\gamma)^2/\tau^2 \leq 1$. Finally, apply the formula $\nu(A \mid B) = \nu(A \cap B)/\nu(A)$ for conditional probabilities, with the obvious choices for A and B.

Assume now $\delta > 0$. Write $\beta_z^\sigma = \psi(\lambda_z^\sigma)$. Then for all z, σ

$$\begin{aligned}
E(\beta_z^\sigma, L, z) &= (1/m)\sum_{i=1}^m \mu_L(|y_i - \beta_z^\sigma(x_i)|) \\
&\leq (1/m)\sum_{i=1}^m \mu_L(|y_i - \lambda_z^\sigma(x_i)| + |\lambda_z^s(x_i) - \beta_z^\sigma(x_i)|) \\
&\leq (1/m)\sum_{i=1}^m \mu_L(|y_i - \lambda_z^\sigma(x_i)| + \delta) \\
&\leq (1/m)\sum_{i=1}^m [\mu_L(|y_i - \lambda_z^\sigma(x_i)|) + c\delta] \\
&= E(\lambda_z^\sigma, L, z) + c\delta
\end{aligned}$$

In the same manner, one can show that $E(\lambda_z^\sigma, L) \leq E(\beta_z^\sigma, L) + c\delta$. So if $E(\lambda_x^\sigma, L) > \epsilon$ and $E(\lambda_x^\sigma, L, z) \leq \gamma\epsilon - 2c\delta$, then $E(\beta_z^\sigma, L) > \epsilon - c\delta$ and $E(\beta_z^\sigma, L, z) \leq \epsilon\gamma - c\delta \leq \gamma(\epsilon - c\delta)$. The result follows by applying the case $\delta = 0$ of the theorem to the NL situation Λ' obtained from Λ by replacing F by $\psi(F)$ and λ by β, substituting $\epsilon - c\delta$ for ϵ. ■

REFERENCES

Anthony, M., and Shawe-Taylor, J. 1993. A result of Vapnik with applications. *Discrete Appl. Math.* **47**, 207-217.

Baum, E., and Haussler, D. 1989. What Size Net Gives Valid Generalization? *Neural Comp.* **1**, 151-160.

Braess, D. 1986. *Nonlinear Approximation Theory.* Springer-Verlag, Berlin.

Cover, T. 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* **EC-14**, 326-334.

Haussler, D. 1992. Decision Theoretic Generalizations of the PAC Model for Neural Net and other Learning Applications. *Inform. Comput.* **100**, 78-150.

Hole, A. 1995. Two variants on a theorem by Vapnik. *Preprint Series, Institute of mathematics, University of Oslo.*

Pollard, D. 1984. *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Sontag, E. 1992. Feedforward Nets for Interpolation and Classification. *J. Comp. Syst. Sci.* **45**, 20-48.

Vapnik, V. 1982. *Estimation of Dependencies Based on Empirical Data.* Springer-Verlag, New York.

Vapnik, V., and Bottou, L. 1993. Local algorithms for Pattern Recognition and Dependencies Estimation. *Neural Comp.* **5**(6), 893-909.