Learning curves from a modified VC-formalism: a case study

A. Kowalczyk & J. Szymański
Telstra Research Laboratories
770 Blackburn Road,
Clayton, Vic. 3168, Australia
{a.kowalczyk,j.szymanski}@trl.oz.au

R.C. Williamson
Department of Engineering
Australian National University
Canberra, ACT 0200, Australia
Bob.Williamson@anu.edu.au

Abstract

In this paper we present a case study of a 1-dimensional higher order neuron using a statistical approach to learning theory which incorporates some information on the distribution on the sample space and can be viewed as a modification of the Vapnik-Chervonenkis formalism (VC-formalism). We concentrate on learning curves defined as averages of the worst generalization error of binary hypothesis consistent with the target on training samples as a function of the training sample size. The true learning curve is derived and compared against estimates from the classical formalism and its modification. It is shown that the modified VC-formalism improves the VC-learning curve by a factor of ≈ 3 and also produces a meaningful result for small training sample sizes where VC-bounds are void.

1 Introduction

It has been observed on a number of occasions that universal bounds on learning curves and other related predictions from VC-formalism are relatively poor, at least for some cases of neural networks. In particular, certain experiments with a backpropagation network for hand-printed character recognition [10] show that very high generalization can be achieved with the training sample size lower than the number of synaptic weights, while the VC-bounds will give estimates of the training sample size the of order of hundreds times the number of synaptic weights. Similarly, the learning theory based on statistical physics (c.f. [4, 16] and references therein) shows the existence of "phase transitions" to perfect generalization at low training samples (1.2-1.4 × number of synaptic weights for the Ising perceptron), a phenomenon which is not captured by the classical VC-formalism. The conclusion is that an adequate theory of learning and generalization is still an open issue. Two main streams of research in that direction can be noted in the recent publications: theory of learning based on concepts from statistical physics [1, 4, 16], and modifications of the VC-formalism via incorporation of some problem specific information. In the latter stream, particular examples are VC-entropy [13], empirical VC-dimensions [14], efficient complexity [15] and a VC-formalism with error shells [7, 8, 9]. The last formalism is based on a refinement of the "fundamental theorem of computational learning" [2] and its main innovation is to split the set of partitions of a training sample into separate "error shells", each composed of error vectors corresponding to the different error values. The evaluation of the potential of this formalism is of central interest to this work. Since the VC-formalism with error shells can be reduced to the theory developed in [4] in the case of "finite hypothesis space", we have at least one example of a neural network with discrete synaptic weights (Ising perceptron) showing that the new formalism offers a significant improvement over the classical VC-theory. The aim of this paper is to provide such an evaluation in the case of a neural network with continuous synaptic weights and continuous sample space. The particular learning problem studied here is the consistent learning of a constant function by 1-dimensional higher order neuron, which is a polynomial in one variable with superimposed ordinary hard threshold.

2 Summary of the VC-formalism with error shells

We present the results in the typical PAC-style; the notation follows [2, 7, 9, 8]. We consider a space X of samples with a probability measure μ, a subspace H of all binary functions (dichotomies) $X \to \{0,1\}$ called the *hypothesis space* and a *target hypothesis* $t \in H$. For each $h \in H$ and each m-sample $\vec{x} = (x_1, ..., x_m) \in X^m$, where $m \in \{1, 2, ...\}$, we denote by $\epsilon_{h,\vec{x}} \overset{def}{=} \frac{1}{m}\sum_{i=1}^{m} |t-h|(x_i)$ *the empirical error* of h on $\vec{x}$, and by $\epsilon_h \overset{def}{=} \int_X |t-h|(x)\mu(dx)$ the expected error of $h \in H$.

For each $m \in \{1, 2, ...\}$ let us consider the random variable

$$\epsilon_H^{max}(\vec{x}) \overset{def}{=} \sup_{h \in H}\{\epsilon_h \; ; \; \epsilon_{h,\vec{x}} = 0\} \qquad (\vec{x} \in X^m) \qquad (1)$$

defined as the maximal expected error of an hypothesis $h \in H$ consistent with t on $\vec{\boldsymbol{x}}$. The *learning curve of* H, defined as the expected value of ϵ_H^{max},

$$\epsilon_H^{av}(m) \overset{def}{=} \boldsymbol{E}_{X^m}[\epsilon_H^{max}] = \int_{X^m} \epsilon_H^{max}(\vec{\boldsymbol{x}})\mu^m(d\vec{\boldsymbol{x}}), \quad (2)$$

is of central interest to us.

Upper bounds on the learning curve can be derived from basic PAC-style estimates as follows. For $\epsilon \geq 0$ we denote by $H_\epsilon \overset{def}{=} \{h \in H \ ; \ \epsilon_h \geq \epsilon\}$ the subset of ϵ-*bad hypothesis* and by

$$\begin{aligned} Q_\epsilon^m \quad &\overset{def}{=} \quad \{\vec{\boldsymbol{x}} \in X^m \ ; \ \exists_{h \in H_\epsilon} \epsilon_{h,\vec{\boldsymbol{x}}} = 0\} \\ &= \quad \{\vec{\boldsymbol{x}} \in X^m \ ; \ \exists_{h \in H} \epsilon_{h,\vec{\boldsymbol{x}}} = 0 \ \& \ \epsilon_h \geq \epsilon\} \end{aligned} (3)$$

the subset of m-samples for which there exists an ϵ-bad hypothesis consistent with t. It can be shown [8] that if

$$\mu^m(Q_\epsilon^m) \leq \psi(\epsilon, m), \quad (4)$$

then

$$\epsilon_H^{av}(m) \leq \epsilon_\psi(m) \overset{def}{=} \int_0^1 \min(1, \psi(\epsilon, m))\mu(d\epsilon), \quad (5)$$

and equality in the assumption implies equality in the conclusion. The central issue from now on is to develop useful upper bounds $\psi(\epsilon, m)$ on $\mu^m(Q_\epsilon^m)$.

Given $\vec{\boldsymbol{x}} = (x_1, ..., x_m) \in X^m$ it is convenient to introduce the transformation (projection) $\pi_{t,\vec{\boldsymbol{x}}} : H \to \{0,1\}^m$ allocating to each $h \in H$ a vector

$$\pi_{t,\vec{\boldsymbol{x}}}(h) \overset{def}{=} (|h(x_1) - t(x_1)|, ..., |h(x_m) - t(x_m)|)$$

from $\{0,1\}^m$ called *the error pattern* of h on $\vec{\boldsymbol{x}}$. The space $\{0,1\}^m$ of all error patterns is the disjoint union of *error shells* $\mathcal{E}_i^m \overset{def}{=} \{(\xi_1, ..., \xi_m) \ ; \ \xi_1 + \cdots + \xi_m = i\}$, where $i = 0, 1, ..., m$. Let $|\pi_{t,\vec{\boldsymbol{x}}}(H_\epsilon) \cap \mathcal{E}_i^m|$ denote the number of different error patterns in the i-th error shell which can be obtained for $h \in H_\epsilon$. We shall employ the special notation for its average called generically *the average error shell size of* H_ϵ:

$$\begin{aligned} |H_\epsilon|_i^m \quad &\overset{def}{=} \quad \boldsymbol{E}_{X^m}[|\pi_{t,\vec{\boldsymbol{x}}}(H_\epsilon) \cap \mathcal{E}_i^m|] \\ &= \quad \int_{X^m} |\pi_{t,\vec{\boldsymbol{x}}}(H_\epsilon) \cap \mathcal{E}_i^m| \mu^m(d\vec{\boldsymbol{x}}). \end{aligned} \quad (6)$$

The central result for this paper, an estimate of the form (4), is obtained by modification of the proof of [7, Theorem 1] which is a refinement of the proof of the "fundamental theorem of computational learning" in [2]. It is a simplified version (to the consistent learning case) of the basic estimate discussed in [9, 6, 8]; the full proof is beyond the scope of this paper and is given in [6].

Theorem 1 *For any integer* $k \geq 0$ *and* $0 \leq \gamma \leq \epsilon \leq 1$

$$\mu^m(Q_\epsilon^m) \leq A_{\epsilon,k,\gamma} \sum_{j=\lceil \gamma k \rceil}^{k} \binom{k}{j} \frac{|H_\epsilon|_j^{m+k}}{\binom{m+k}{j}}, \quad (7)$$

where $A_{\epsilon,k,\gamma} \overset{def}{=} \left(\sum_{\lceil \gamma k \rceil}^{k} \binom{k}{j} \epsilon^j (1-\epsilon)^{k-j} \right)^{-1}$, *for* $k > 0$ *and* $A_{\epsilon,0,\gamma} \overset{def}{=} 1$. □

Obviously any upper bound (approximation) on coefficients $|H_\epsilon|_j^{m+k}$ will lead to a new upper bound (approximation) on the learning curve via relation (5). It can be shown [9] that under some such crude bounds on $|H_\epsilon|_i^m$ the inequality (7) reduces either to the basic estimate of the VC-formalism (c.f. [2])

$$\mu^m(Q_\epsilon^m) \leq 2^{2 - m\epsilon/2}(2em/d_{VC})^{d_{VC}}, \quad (8)$$

where $d_{VC} < \infty$ is the VC-dimension of H, or to the union bound ,

$$\mu^m(Q_\epsilon^m) \leq \sum_{h \in H_\epsilon} (1 - \epsilon_h)^m,$$

if H has finite cardinality. The union bound is the starting point for the statistical physics based formalism developed in [4]. In this sense both these theories are unified in this upper bound (7), all their conclusions can be derived from it and possibly improved with the use of tighter bounds on $|H_\epsilon|_i^m$'s. On the other hand, sensible approximations to $|H_\epsilon|_i^m$ (rather than strict bounds which are difficult to derive in general) could lead to a reasonable approximate theory of learning curves which could capture at least certain qualitative properties of learning curves.

3 Case of a 1-dimensional higher order neuron

We consider $X \overset{def}{=} (0,1) \subset \mathbf{R}$ (the open segment) with μ defined as the uniform probability distribution. The hypothesis space $H \subset \{0,1\}^X$ is defined as the set of all functions of the form $\theta \circ p(x)$, where p is a polynomial of degree $\leq d$ on $\mathbf{R}$, and θ is the hard threshold function. We consider the constant target $t \equiv 1$ only. It is easy to see that H restricted to any finite subset of $(0,1)$ is exactly the restriction of the family of all piecewise constant functions $\tilde{H} \subset \{0,1\}^{(0,1)}$ with up to d "jumps" between 0 and 1. Thus it is easily seen that the VC-dimension of H is $d_{VC} = d+1$. The VC-theory learning curve (5) given by the upper bound (8) is plotted in Fig. 1 (top solid line).

3.1 True learning curve

Given a generic $\vec{\boldsymbol{x}} \in (0,1)^m$, let $0 < y_1 \leq y_2 \leq ... \leq y_m < 1$ be entries of $\vec{\boldsymbol{x}}$ in ascending order. Let $\lambda_k(\vec{\boldsymbol{x}})$ denote the sum of the k largest among $m+1$ segments $(0, y_1), (y_1, y_2), ..., (y_m, 1)$, respectively. Similarly, let $\tilde{\lambda}_k(\vec{\boldsymbol{x}})$ denote the sum of k largest among $m-1$ segments $(y_1, y_2), ..., (y_{m-1}, y_m)$. Then

$$\epsilon_H^{max}(\vec{\boldsymbol{x}}) = \max\left\{\tilde{\lambda}_{d/2}(\vec{\boldsymbol{x}}), y_1 + 1 - y_m + \tilde{\lambda}_{d/2-1}(\vec{\boldsymbol{x}})\right\}$$

if d is even and

$$\epsilon_H^{max}(\vec{\boldsymbol{x}}) = \max\{y_1 + \tilde{\lambda}_{\lfloor d/2 \rfloor}(\vec{\boldsymbol{x}}), 1 - y_m + \tilde{\lambda}_{\lfloor d/2 \rfloor}(\vec{\boldsymbol{x}}), \\ y_1 + 1 - y_m + \tilde{\lambda}_{\lfloor d/2 \rfloor - 1}(\vec{\boldsymbol{x}})\}$$

if d is odd. Consequently

$$\lambda_{\lfloor d/2 \rfloor}(\vec{\boldsymbol{x}}) \leq \epsilon_H^{max}(\vec{\boldsymbol{x}}) \leq \lambda_{\lfloor d/2 \rfloor + 1}(\vec{\boldsymbol{x}}). \qquad (9)$$

An explicit expression for the expected value of λ_k is known [11]:

$$E(\lambda_k) = \frac{k}{m+1}\left(1 + \sum_{j=k+1}^{m+1} \frac{1}{j}\right),$$

thus a very tight two sided bound on the true learning curve $\epsilon_H^{av}(m)$ defined by (2) can be obtained (tight within $< 200/d$ %):

$$\frac{\lfloor d/2 \rfloor}{m+1}\left(1 + \sum_{j=\lfloor d/2 \rfloor+1}^{m+1} \frac{1}{j}\right) \leq \epsilon_H^{av}(m) \leq \\ \leq \frac{\lfloor d/2 \rfloor + 1}{m+1}\left(1 + \sum_{j=\lfloor d/2 \rfloor+2}^{m+1} \frac{1}{j}\right). \qquad (10)$$

The corresponding upper bound on the true learning curve is shown in Fig. 1 (bottom solid line).

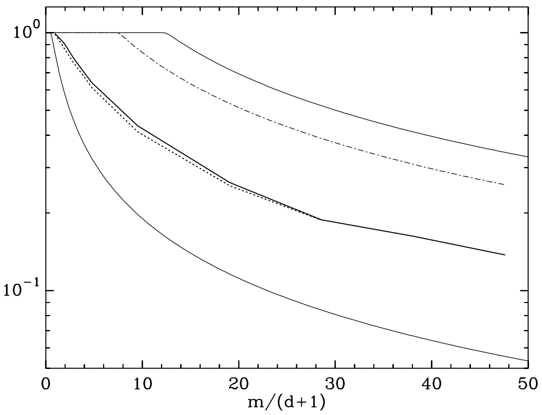

FIG. 1. *Exact learning curves for 1 dimensional higher order neuron. In the top-down order: (i) VC-curve from the upper bound (8), (ii) $|H|_i^m$ learning curve for $k = m$ and $\gamma = \epsilon/2$ (chain line), (iii) $|H|_i^m$ learning curve for $k = m$ and $\gamma = \epsilon$, (iv) $|H|_i^m$ learning curve for $k = 2m$ and $\gamma = \epsilon$ (dotted line), and (v) the upper bound on the true learning curve given by (10).*

3.2 Learning curves based on exact values of $|H|_i^m$

In order to simplify our language, from now on if we have an upper bound $|H_\epsilon|_i^m \leq D_{\epsilon,m,i}$ (or approximation $|H_\epsilon|_i^m \approx D_{\epsilon,m,i}$) then the learning curve $m \mapsto \epsilon_\psi(m)$ defined by RHS of inequality (5) and the bound (7) with $D_{\epsilon,m,i}$ substituted for $|H_\epsilon|_i^m$ will be called shortly *the learning curve for $D_{\epsilon,m,i}$*. In particular we have an obvious upper bound

$$|H_\epsilon|_i^m \leq |H|_i^m \overset{def}{=} \int_{X^m} |\pi_{t,\vec{\boldsymbol{x}}}(H) \cap \mathcal{E}_i^m| \mu^m(d\vec{\boldsymbol{x}}).$$

The coefficients $|H|_i^m$ and thus the $|H|_i^m$-learning curves can be calculated exactly. We now outline these calculations and then derive tight bounds on the $|H|_i^m$-learning curve.

With probability 1 an m-sample $\vec{\boldsymbol{x}} = (x_1, ..., x_m)$ from X^m is such that $x_i \neq x_j$ for $i \neq j$. For such *a generic $\vec{\boldsymbol{x}}$*

$$|\pi_{t,\vec{\boldsymbol{x}}}(H) \cap \mathcal{E}_i^m| = const = |H|_i^m.$$

This observation was used to derive the following relations for the computation of $|H|_i^m$:

$$|H|_i^m = \sum_{\delta=0}^{\min(d,m-1)} |\tilde{H}^{(\delta)}|_i^m + |\tilde{H}^{(\delta)}|_{m-i}^m, \qquad (11)$$

for $0 \leq i \leq m$, where $|\tilde{H}^{(\delta)}|_i^m$, for $\delta = 0, 1, ..., d$, is defined as follows. We initialize $|\tilde{H}^{(0)}|_0^m = |\tilde{H}^{(1)}|_i^m \overset{def}{=} 1$ for $i = 1, ..., m-1$, $|\tilde{H}^{(1)}|_0^m = |\tilde{H}^{(1)}|_m^m \overset{def}{=} 0$ and $|\tilde{H}^{(\delta)}|_i^m \overset{def}{=} 0$ for $i = 0, 1, ..., m$, $\delta = 2, 3, ..., d$, and then, recurrently, for $\delta \geq 2$ we set $|\tilde{H}^{(\delta)}|_i^m \overset{def}{=} \sum_{k=\max(\delta,m-i)}^{m-1} |\tilde{H}^{(\delta-1)}|_{i-m+k}^k$ if δ is odd and $|\tilde{H}^{(\delta)}|_i^m \overset{def}{=} \sum_{k=\delta}^{m-1} |\tilde{H}^{(\delta-1)}|_i^k$ if δ is even. (Here $|\tilde{H}^{(\delta)}|_i^m$ is defined by the relation (6) with the target $t \equiv 1$ for the hypothesis space $H^{(\delta)} \subset \tilde{H}$ composed of functions having the value 1 near 0 and exactly δ jumps in $(0,1)$, exactly at entries of $\vec{\boldsymbol{x}}$; similarly as for H, $|H^{(\delta)}|_i^m = |\pi_{1,\vec{\boldsymbol{x}}} H^{(\delta)} \cap \mathcal{E}_i^m|$ for a generic m-sample $\vec{\boldsymbol{x}} \in (0,1)^m$.)

The $|H|_i^m$ learning curves are shown in Fig. 1 for $d = 20$; evaluations the bound (7) were restricted to $k \leq 2m$. The bound depends only weakly on k but very strongly on γ and the (empirically) optimal γ turns out to be equal to ϵ (c.f. the solid center line with the chain line in Fig. 1).

3.3 Learning curves based on approximations of $|H|_i^m$

Analyzing the embedding of **R** into **R**^{d} and using an argument based on the Vandermonde determinant as in [5, 12], it can be proved that the partition function of H [2], $\Pi_H(m) \overset{def}{=} \max_{\vec{\boldsymbol{x}} \in X^m} |\pi_{0,\vec{\boldsymbol{x}}}(H)|$, is equal to $|\pi_{t,\vec{\boldsymbol{x}}}(H)|$ with probability 1, and is given by Cover's counting function [3]

$$\Pi_H(m) = \sum_{i=0}^{m} |H|_i^m = 2\sum_{i=0}^{d} \binom{m-1}{i}. \qquad (12)$$

This equality gives an independent cross-check of numerical calculations of the coefficients $|H|_i^m$. It also allows us to calculate "the average density of error patterns":

$$\bar{P}_H(m) \overset{def}{=} \int_{X^m} |\pi_{t,\vec{x}}(H)|/2^m \mu^m(d\vec{x}) = \Pi_H(m)/2^m.$$

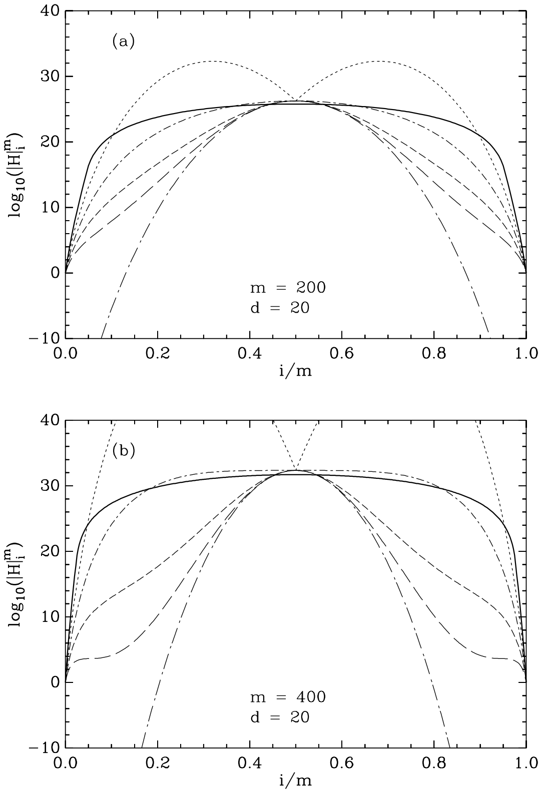

FIG. 2. *Plots of exact values of coefficients $|H|_i^m$ (solid line) and some of its approximations given by (13) for $C = 1$ and $\omega = 1, 2, 3, 4, \infty$ (top to bottom order). Note that for $\omega = \infty$ (long chain line) we have here the approximation $|H_\epsilon|_i^m \approx \binom{m}{i}\bar{P}_H(m)$.*

In Fig. 2 we plot values of $|H|_i^m$ as a function of i/m along with some approximations to it. We observe that for "large" error shells (in terms of their cardinality $|\mathcal{E}_i^m| = \binom{m}{i}$), i.e. for $i/m \approx 1/2$, we have $|H|_i^m \approx \bar{P}_H(m)|\mathcal{E}_i^m| = \bar{P}_H(m)\binom{m}{i}$. In this sense $\bar{P}_H(m)$ determines the maximal value of $|H|_i^m$. However for error shells of smaller cardinality, i.e. for $i \not\approx m/2$, we observe a gradual departure form this relation (compare the solid curve and the long chain curve in Fig. 2). In particular, for $i/m = 0, 1$ we have $|H|_i^m$ equal to 1 (since $t \in H$) rather than $1 \times \bar{P}_H(m)$ (which is $\approx 10^{-33}$ in the case of Fig. 2.a). One may heuristically interpret this fact as an apparent increase of the density of error patterns for small error shells, from $\bar{P}_H(m)$ to 1 as $|i/m - 1/2|$ increases from 0 to 1/2. It is natural to try some simple heuristic "corrections" to the density which display this type of behavior. The ultimate aim of such heuristic approximations is to possibly capture a general dependence which can be later used to model more complex neural systems as well. It could also be useful for qualitative explanation of different properties of learning curves. In our considerations we have used $\bar{P}_H(m)^{1-|1-2i/m|^\omega}$ where $\omega \geq 1$ is a real parameter referred to as *an order of uniformity* [6, 8]. This has lead to the family of approximations

$$|H|_i^m \approx C\binom{m}{i}\bar{P}_H(m)^{1-|1-2i/m|^\omega}, \qquad (13)$$

for $0 \leq i \leq m$ and different values of ω and the constant C. In Fig. 2 we plot examples of such approximations for the selected values of $\omega = 1, 2, 3, 4$ and $C = 1$. By inspection we find that the approximation for $\omega = 2$ is the best. The learning curve for $\omega = 2$ is shown in Fig. 3.a (the lower solid curve). For comparison we plot also the learning curve for the approximation (13) with $C = 10^5$ (the upper solid curve). Both learning curves approximate quite well the exact $|H|_i^m$-learning curve for relatively small m ($0 \leq m \leq 12d_{VC} = 12(d+1)$), then overestimate it. Note that in the whole range of plotted values these estimates are better than the VC-learning curve, especially for small m ($m < 12d_{VC}$) where VC-learning curve is trivial, i.e. $\equiv 1$.

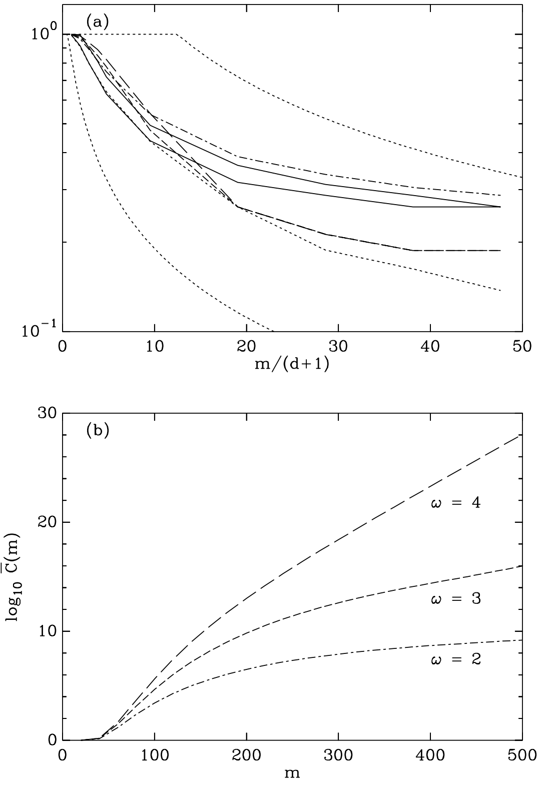

FIG. 3. *(a) Approximation and upper bound on the $|H|_i^m$ learning curve (center dotted line). Top and bottom dotted lines show the VC-curve and the true learning curve, respectively. Solid lines are for the approximation (13) with $\omega = 2$ and $C = 10^5, 1$ (the top-bottom order). Upper bounds implied by the inequality (14) for $\omega = 2, 3, 4$ with $\bar{C}(m)$ set to the smallest possible constant are given by the chain, broken and long broken lines, respectively. The values of such constants are plotted in Figure (b). We have used here consistently $k = m$ and $\gamma = \epsilon$ to derive all these plots.*

In Fig. 3.a we also plot the upper bound on the $|H|_i^m$-learning curve implied by the estimates

$$|H|_i^m \leq \bar{C}(m) \binom{m}{i} \bar{P}_H(m)^{1-|1-2i/m|^\omega}, \qquad (14)$$

for $0 \leq i \leq m/2$, where $\bar{C}(m)$ is the smallest constant satisfying this inequality for $\omega = 2, 3, 4$. In Fig. 3.b we plot values of $\bar{C}(m)$. Note that the upper bounds on the $|H|_i^m$ learning curve obtained in this way for $\omega = 3$ and $\omega = 4$ are quite tight and significantly better than for $\omega = 2$.

4 Conclusions

The importance of the 1-dimensional higher order neuron stems from the fact that strict results (e.g. learning curves) can be derived which can then serve as benchmarks for approximate solutions. We have shown on the example of a particular learning problem (consistent learning of a constant) that by incorporating the limited knowledge of the statistical distribution of error patterns in the sample space one dramatically improves results by shifting the ($|H|_i^m$) learning curve from the VC result at least half way (on a log scale) towards true learning curve. This is particularly important for small training sample sizes ($m \leq 12\times$ VC-dimension) where the VC-bounds are void. It is worth noting that for this particular neural network the $|H|_i^m$-learning curve is a distribution independent upper bound on the true learning curve (which is distribution dependent).

We have also shown that approximate approaches consisting of replacing $|H|_i^m$ by a simple estimate (13) can produce learning curves very close to exact $|H|_i^m$-learning curves suggesting that generalization to the systems where $|H|_i^m$ cannot by calculated might be possible. This could lead to a sensible approximate theory capturing at least certain qualitative properties of learning curves.

Acknowledgements. The permission of the Director of Telstra Research Laboratories to publish this paper is gratefully acknowledged. This work was supported by the Australian Research Council.

References

[1] S. Amari, N. Fujita, and S. Shinomoto. Four types of learning curves. *Neural Computation,* **4(4):605–618, 1992.**

[2] **M. Anthony and N. Biggs. *Computational Learning Theory.* Cambridge University Press, 1992.**

[3] **T.M. Cover. Geometrical and statistical properties of linear inequalities with applications to pattern recognition. *IEEE Trans. Elec. Comp.,* EC-14:326–334, 1965.**

[4] **D. Haussler, M. Kearns, H.S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. In *Proc. 7th Annual ACM Conf. on Computational Learning Theory,* pages 76–87, 1994.**

[5] **A. Kowalczyk. Estimates of storage capacity of multi-layer perceptron with threshold logic hidden units. Neural networks, to appear.**

[6] **A. Kowalczyk. VC-formalism with explicit bounds on error shells size distribution. A manuscript, 1994.**

[7] **A. Kowalczyk and H. Ferra. Generalisation in feedforward networks. NIPS'94, to appear, 1994.**

[8] **A. Kowalczyk, J. Szymański, P.L. Bartlett, and R.C. Williamson. Examples of learning curves from a modified VC-formalism. Submitted, 1995.**

[9] **A. Kowalczyk, J. Szymanski, and H. Ferra. Combining statistical physics with VC-bounds on generalisation in learning systems. In *Proceedings of the Sixth Australian Conference on Neural Networks, ACNN'95,* Sydney, 1995. University of Sydney.**

[10] **G.L. Martin and J.A. Pitman. Recognizing handprinted letters and digits using backpropagation learning. *Neural Comput.,* 3:258–267, 1991.**

[11] **J.G. Mauldon. Random division of an interval. *Proc. Cambridge Phil. Soc.,* 47:331–336, 1951.**

[12] **A. Sakurai. n-h-1 networks store no less $n\;h+1$ examples but sometimes no more. In *Proceedings of the 1992 International Conference on Neural Networks,* pages III–936–III–941. IEEE, June 1992.**

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

[14] **V. Vapnik, E. Levin, and Y. Le Cun. Measuring the VC-dimension of a learning machine. *Neural Computation,* 6 (5):851–876, 1994.**

[15] **C. Wang and S.S. Venkantesh. Temporal dynamics of generalisation in neural networks. NIPS'94, 1994.**

[16] **L.H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning rule. *Rev. Mod. Phys.,* 65:499–556, 1993.**