Generalization errors of the simple perceptron

# Generalization errors of the simple perceptron

Jianfeng Feng

Biomathematics Laboratory, The Babraham Institute, Cambridge CB2 4AT, UK

Received 1 August 1997

**Abstract.** To find an exact form for the generalization error of a learning machine is an open problem, even in the simplest case: simple perceptron learning. We introduce a new approach to tackle the problem. The generalization error of the simple perceptron is expressed as a linear combination of extreme values of inputs. With the help of extreme value theory in statistics we then obtain an exact form of the generalization error of the simple perceptron in the case of the worst learning. Generalization errors of the higher-order perceptron taking the form of an inverse power law in the number of examples are also considered.

## 1. Introduction

Understanding a neural network's ability to infer an unknown rule from a set of examples has become a fascinating topic in neural networks. Obtaining an exact form for the generalization errors together with learning errors is of vital importance when assessing how fast a network improves its behaviour. Currently there are three approaches to estimate generalization errors of a learning machine.

• According to the Vapnik–Chervonenkis (VC) theory of learning curves, minimizing empirical error within a function class $\mathcal{F}$ on a random sample of $t$ examples leads to generalization error bounded by $\mathrm{O}(d/t)$ in the case that the target function is contained in $\mathcal{F}$. The bound is universal; it holds for any class of hypothesis function $\mathcal{F}$, for any input distribution and for any target function. The only problem being specific quantity remaining in the bound is the VC dimension $d$, a measure of the complexity of the function class $\mathcal{F}$. There is much research activity on this topic, see for example [4–6, 22, 32, 33].

• Recently, using techniques developed in the physics of disordered systems [17], exact learning curves have been calculated for a variety of rules and network models. Broader reviews can be found in [30, 34], cases of concrete learning problems are discussed in [10, 21, 27, 29, 31, 35]. When the number of examples grows large, and the network parameters assume continuous values, the results obtained for many models suggest that learning curves may have universal asymptotic features. For the important case in which the rule can be implemented exactly by the network, the decay of the so-called generalization error follows an inverse power law in the number of examples, with a constant (often called an 'effective dimension'), that is proportional to the number of adjustable parameters.

• A similar result for the scaling of the so-called entropic error was given using asymptotic methods of statistics [1–3, 26]. It was proved again that the generalization error is of the form $1/t$ with an exactly given coefficient depending on the dimension $m$ of input signals.

The theory of generalization errors is already well developed, however, little is known about the exact form of generalization errors of some concrete learning rules [28]. Even

0305-4470/98/174037+12$19.50

in the simplest case—the simple perceptron—the problem of finding the coefficient of the generalization error is still open except for some very special cases [2]. In this paper, based upon the extreme value theory of statistics, we propose a novel approach aiming to be a complement of the approaches above—to obtain the *exact* form of the generalization errors of some concrete learning algorithms. The idea underlying our approach is straightforward. The generalization error for a given machine is universal, as confirmed by all previous studies, in the sense that it does not depend on the input distribution at all. This fact suggests that to calculate the generalization errors we should build up a model which is as simple as possible. By choosing a specific input we show that the generalization error of the simple perceptron is basically a linear combination of extreme values of input signals. Fortunately, for extreme values of an i.i.d. random sequence we fully understand their properties, which enables us to complete our calculation.

Extreme value theory was first introduced to tackle disordered systems in [18, 13]. Recently good work which compares the extreme value theory approach and Parisi's 'replica symmetry breaking' scheme was presented in [9]. Although in this paper we confine ourselves to the case of the perceptron learning, both linear and higher order, we expect that our approach opens up new possibilities to rigorously consider the generalization errors of a class of learning machines.

## 2. The set-up

### 2.1. The model

We briefly outline the simple perceptron here and refer the reader to [23, p 98] for a more detailed discussion.

Consider the simple perceptron fed with $m$-dimensional independent inputs $\boldsymbol{\xi}(\tau) = (\xi_i(\tau), i = 1, \ldots, m) \in \Omega^+ \cup \Omega^- = \Omega \subset \mathbb{R}^m$, $\tau = 1, 2, \ldots$. Suppose that the two nonintersecting sets $\Omega^+$ and $\Omega^-$ are linear separable which implies that there is a vector of weights $\boldsymbol{w} = \{w_1, w_2, \ldots, w_m\}$ satisfying $\mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{\xi}) \geqslant 0$ if and only if $\xi \in \Omega^+$ (the threshold can always be thought of as a weight subjected to a constant input taking the value 1).

Suppose that at time $\tau$ the weights of the simple perceptron are $\boldsymbol{w}(\tau)$. Then at time $\tau + 1$ with the incoming signal $\xi(\tau+1)$ we update $\boldsymbol{w}(\tau)$ according to the perceptron learning rule

$$\boldsymbol{w}(\tau + 1) = \boldsymbol{w}(\tau) - \gamma \cdot \Theta(\xi(\tau + 1) \cdot \boldsymbol{w})\boldsymbol{\xi}(\tau + 1) \tag{1}$$

where $\gamma > 0$ is the learning rate and $\Theta(x) = 1$ if $x \geqslant 0$ and $\Theta(x) = 0$ otherwise. After repeatedly presenting the examples $\xi(\tau)$, $\tau = 1, \ldots, t$ we find an output function $h(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^m$ which separates examples $\{\boldsymbol{\xi}(\tau), \xi(\tau) \in \Omega^+, \tau = 1, \ldots, t\}$ from $\{\boldsymbol{\xi}(\tau), \xi(\tau) \in \Omega^-, \tau = 1, \ldots, t\}$.

### 2.2. Generalization errors

Without loss of generality we assume that the task for the machine to accomplish is the classification problem—to separate data set $\Omega^+ = \{\boldsymbol{\xi}(\tau), \xi_1(\tau) < 0\}$ from $\Omega^- = \{\boldsymbol{\xi}(\tau), \xi_1(\tau) > 0\}$ and so $\mathrm{sign}(\xi_1(\tau))$ is the so-called target function (see remark 2 below). Suppose that after training with $t$ examples using the simple perceptron learning rule, the output of the learned machine is $h(\boldsymbol{\xi}(t + 1)) \in \{-1, 1\}$ when a new signal $\boldsymbol{\xi}(t + 1)$ arrives. One key assumption (see figure 1) of our approach is that we take into account the case of worst learning (when $\gamma$ is small the following assumption is approximately true).

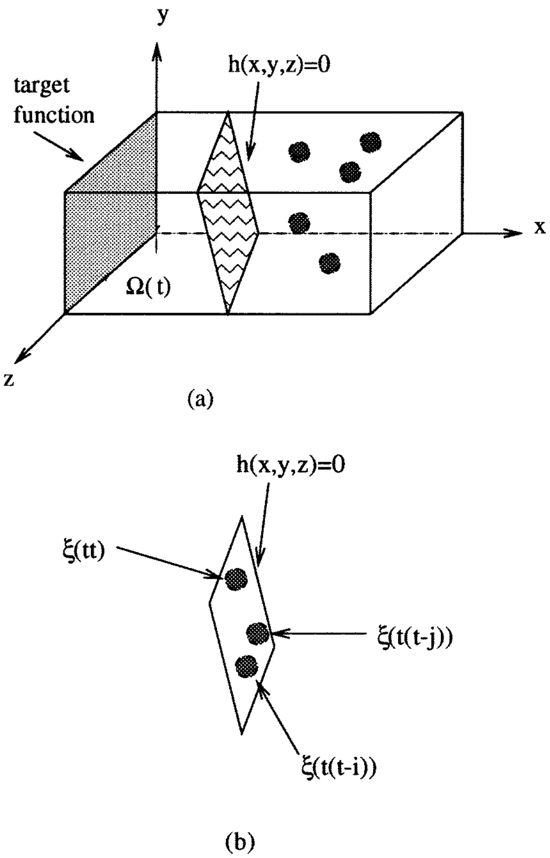

**Figure 1.** Full circle: examples of input signals. ($a$) The target function is sign($x$) (full plane). After learning $t$ examples, a perceptron is capable of separating data on the two sides of the patterned plane $h(x, y, z) = 0$ (redrawn in ($b$)). $\Omega(t)$: the region between the full plane and the patterned plane. ($b$) The plane $h(x, y, z) = 0$ is determined by extreme values $\boldsymbol{\xi}(tt), \boldsymbol{\xi}(t(t-i))$ and $\boldsymbol{\xi}(t(t-j)), i \neq j, i \geqslant 1, j \geqslant 1$.

*Assumption 1.* The plane $h(\boldsymbol{x}) = 0$ passes through $m$ examples so that all examples learnt $\{\boldsymbol{\xi}(\tau) \in \Omega^+, \tau = 1, \ldots, t\}$ are on the one side of it.

Suppose that the distribution of $\boldsymbol{\xi}(\tau)$ is symmetric with respect to $x_1 = 0$, the generalization error can then be defined by

$$\begin{aligned}\epsilon(t, m) &= \langle|h(\boldsymbol{\xi}(t+1)) - \mathrm{sign}(\xi_1(t+1))|\rangle \\ &= \langle P(\boldsymbol{\xi}(t+1) \in \Omega(t)|\mathcal{F}_t)\rangle \end{aligned} \qquad (2)$$

where $\Omega(t)$ is the region (the region between the filled and patterned planes shown in figure 1($a$)) between the target function and output function $h$ and $\Omega(t) \in \mathcal{F}_t$, $\mathcal{F}_t$ is the sigma-algebra generated by $\{\boldsymbol{\xi}(\tau), \tau \leqslant t\}$.

### *2.3. Extreme values*

Extreme value theory in statistics, a well-developed and powerful tool, was first introduced into neural network circles in [18] for considering the capacity of the Hopfield model and other models [11, 14]. A brief account of the results employed in this paper can be found in the appendix and section 6 of [14]. Here we apply it in the estimation of generalization errors.

Let $\xi_1(tk)$ be the $(t-k)$th smallest minimum in the set $\{\xi_1(\tau), \tau = 1, \ldots, t\}$ and so

$$\begin{aligned} &\xi_1(tt) = \min\{\xi_1(\tau), \tau = 1, \ldots, t\} \\ &\xi_1(t(t-1)) = \min\{\xi_1(\tau) \geqslant \xi_1(tt), \tau = 1, \ldots, t, \tau \neq tt\} \\ &\vdots \end{aligned} \tag{3}$$

For simplicity of notation we call $\boldsymbol{\xi}(t(t-k))$ the $(t-k)$th smallest minima in the set $\{\boldsymbol{\xi}(\tau), \tau = 1, \ldots, t\}$. Assumption 1 thus indicates that the output function $h = 0$ passes through the global minimum $\boldsymbol{\xi}(tt) = (\xi_1(tt), \xi_2(tt), \ldots, \xi_m(tt))$ and $m-1$ other minima, say $\boldsymbol{\xi}(t(t-k_1)), \boldsymbol{\xi}(t(t-k_2)), \ldots, \boldsymbol{\xi}(t(t-k_{m-1}))$. Note that here $k_i$ depends on the realization of $\{\boldsymbol{\xi}(\tau), \tau = 1, \ldots, t\}$ (figure 1).

When $k$ is fixed there are three types of behaviour for extreme values $\xi_1(tt)$ of a sequence of random variables $\xi_1(1), \xi_1(2), \ldots, \xi_1(t)$. For a full exposition of extreme value theory we refer the reader to [24, 19]. Typically for an extreme $\xi_1(t(t-k))$ of a sequence of random variables, i.e. for the $k$th minimum of a sequence, we have the following property

$$\langle \xi_1(t(t-k)) \rangle = c(k)\mathrm{o}(\gamma(t)) \tag{4}$$

where $c(k)$ is a constant depending on $k$ and $\gamma(t)$ is a vanishing rate of $t$.

When $k_t$ tends to infinity as $t$ tends to infinity, the behaviour of $\boldsymbol{\xi}(t(t-k_t))$ is substantially different from that of $\boldsymbol{\xi}(t(t-k))$ with $k$ independent of $t$, it may take a finite value rather than tending to zero as described in equation (4).

## 3. The simple perceptron

Before proving the main theorem we need a few lemmas which are of interest in themselves. These lemmas provide us with a rudimentary and transparent insight which elucidates the underlying mechanism of the universal property of generalization errors.

*Lemma 1.* Suppose that $\xi_1(\tau) \sim U(0,1)$, the uniform distribution over [0, 1]. When $t \to \infty$ we have

$$P\left(\xi_1(tt) \geqslant \frac{x}{t}\right) = \mathrm{e}^{-x} \tag{5a}$$

$$P\left(\xi_1(t(t-k)) \geqslant \frac{x}{t}\right) = \mathrm{e}^{-x} \sum_{s=0}^{k-1} \frac{x^s}{s!} \tag{5b}$$

$$\langle \xi_1(t(t-k)) \rangle = \frac{k+1}{t} \tag{5c}$$

for $x \geqslant 0$.

*Proof.*

(5$a$) From example 1.7.9 in [24] (see the appendix and section 6 of [14]) we know that $P(\eta(tt) \leqslant 1 - x/t) = \mathrm{e}^{-x}$ for $\eta(tt)$ representing the largest maximum of $\xi_1(\tau)$, $\tau = 1, \ldots t$. Then (5$a$) is a simple consequence of the symmetry between 1 and 0 of the uniform distribution.

(5$b$) This is a simple consequence of theorem 2.2.2 and example 1.7.9 in [24] (see the appendix and section 6 of [14]).

(5$c$) It is a consequence of (5$b$). □

The following lemma tells us that the generalization of the one-dimensional simple perceptron is of the form $1/t$, which is the building-block of generalization errors with $m$-dimensional inputs.

*Lemma 2.* For uniformly distributed inputs $\xi_1(\tau)$, when $t \to \infty$ we have

$$\epsilon(t, 1) := \langle P(\xi_1(t+1) \leqslant \xi_1(tt)|\mathcal{F}_t)\rangle = \frac{1}{t}. \tag{6}$$

*Proof.* By definition of $\epsilon(t, 1)$ (equation (6)) and equation (5) we obtain

$$\begin{aligned}\epsilon(t, 1) &= \langle \xi_1(tt)\rangle \\ &= \int_0^\infty xt\mathrm{e}^{-tx}\,\mathrm{d}x \\ &= \int_0^\infty \mathrm{e}^{-tx}\,\mathrm{d}x \\ &= \frac{1}{t}.\end{aligned} \tag{7}$$

□

We now turn our attention to a more general case: the input signals are continuously distributed random variables. By this we mean that the Radon–Nikodyn derivative of the input distribution is absolutely continuous with respect to the Lebesgue measure. Denote the density

$$f(x) = \mathrm{d}P/\mathrm{d}x.$$

From the definition of $\epsilon(t, 1)$ (equation (6)) we see that

$$\epsilon(t, 1) = \left\langle \int_0^{\xi_1(tt)} f(x)\,\mathrm{d}x\right\rangle. \tag{8}$$

Define a transformation $Y : \mathbb{R}^1 \to \mathbb{R}^1$ by

$$Y(x) = \int_0^x f(u)\,\mathrm{d}u \tag{9}$$

then equation (8) becomes

$$\epsilon(t, 1) = \left\langle \int_{Y(0)}^{Y(\xi_1(tt))} \mathrm{d}Y(x)\right\rangle. \tag{10}$$

Since the function $Y$ is a nondecreasing function we conclude that

$$Y(\xi_1(tt)) \leqslant Y(\xi_1(t(t-1))) \leqslant \cdots \leqslant Y(\xi_1(tk)) \leqslant \cdots \qquad k < t-1$$

which yields the following lemma.

*Lemma 3.* If $\xi_1$ is a continuously distributed random variable we have

$$\epsilon(t, 1) = 1/t$$

or more generally

$$\langle \xi_1(t(t-k))\rangle = \frac{k+1}{t}.$$

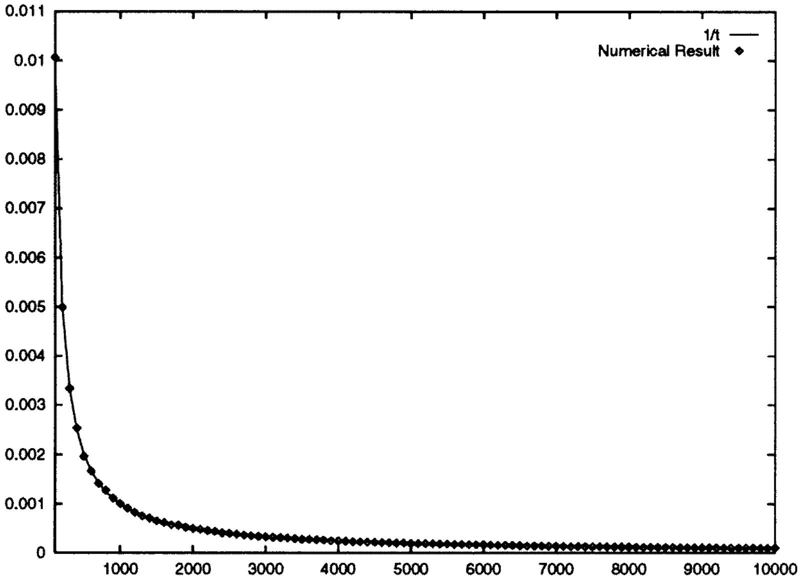

**Figure 2.** Generalization error versus number of examples. The inputs $\xi_1$ are subjected to the Weibull distribution with parameter $A = 10$ and $B = 1$.

Lemma 3 gives rise to a transparent and elementary proof of the universal property of the generalization errors of the simple perceptron in one-dimensional case; $\epsilon(t, 1)$ is independent of the distribution of inputs; $\epsilon(t, 1) = 1/t$ for any continuously distributed inputs.

*Example 1.* $\xi_1(\tau)$ is distributed according to a Weibull distribution with shape parameter $A$ and scale parameter $B$. We note that the density function is then

$$f(x) = \begin{cases} \dfrac{A}{B} x^{A-1} \exp\left(-\dfrac{x^A}{B}\right) & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Figure 2 shows our numerical results with $A = 10$ and $B = 1$.

From lemmas 1–3 we obtain $\epsilon(t, 1)$, a generalization of the one-dimensional case without any geometric structure. What is the simplest inputs embodying the geometric structure of $m$-dimensional inputs? Our answer is as follows.

*Assumption 2.* We suppose that $\xi_2(\tau) \sim 1/m(\delta_{(x_2=0,x_3=0,\ldots,x_m=0)} + \delta_{(x_2=1,x_3=0,\ldots,x_m=0)} + \cdots + \delta_{(x_2=0,x_3=0,\ldots,x_m=1)})$, i.e. input signals are drawn from $m$ lines $(x_2 = 0, x_3 = 0, \ldots, x_m = 0)$, $(x_2 = 1, x_3 = 0, \ldots, x_m = 0), \ldots$ and $(x_2 = 0, x_3 = 0, \ldots, x_m = 1)$ of $\mathbb{R}^m$.

With the help of the above lemmas and assumption 2 we consider the generalization errors of the simple perceptron with $m$-dimensional inputs.

*Theorem 1.* We have the following conclusion

$$\epsilon(t, m) = \begin{cases} \dfrac{1}{t} & \text{if } m = 1 \\ \dfrac{(m-1)!}{(m-1)^{(m-1)}} \left(\dfrac{m}{2} + 1\right) \dfrac{1}{t} & \text{otherwise.} \end{cases} \tag{11}$$

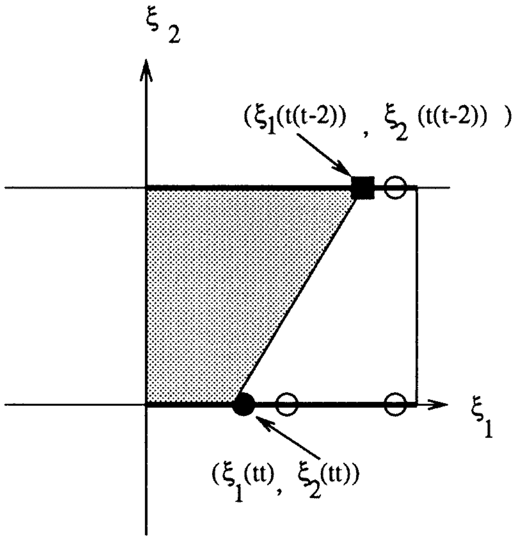

**Figure 3.** When $\xi_2(tt) = \xi_2(t(t-1)) \neq \xi_2(t(t-2))$, $\Omega(t)$ is the region on the left-hand side of the line passing through $(\xi_1(tt), \xi_2(tt))$ and $(\xi_1(t(t-2)), \xi_2(t(t-2)))$.

*Proof.* Since $m = 1$ is proved in lemma 3 we only need to consider the case of $m \geqslant 2$. The following identity is a basic one which indicates that when $\xi_2(tt) \neq \xi_2(t(t-1)) \neq \ldots \neq \xi_2(t(t-m+1))$, $\Omega(t)$ is simply the region on the left-hand side of the plane passing through $(\xi_1(tt), \xi_2(tt))$, $(\xi_1(t(t-1)), \xi_2(t(t-1)))$, ... and $(\xi_1(t(t-m+1)), \xi_2(t(t-m+1)))$, when $\xi_2(tt) = \xi_2(t(t-1))$ but $\xi_2(tt) \neq \xi_2(t(t-m)) \neq \ldots \neq \xi_2(t(t-2)) \neq \xi_2(t(t-1))$ then $\Omega(t)$ is the region on the left-hand side of the plane passing through $(\xi_1(tt), \xi_2(tt))$, $(\xi_1(t(t-2)), \xi_2(t(t-2)))\ldots$ and $(\xi_1(t(t-m)), \xi_2(t(t-m)))\ldots$ (see figure 3). In the following proof we use the convention that for $k_1, k_2, \ldots, k_{m-1} \geqslant 1$

$$\begin{aligned}\{\xi_2(tt) &\neq \xi_2(t(t-k_1)) \neq \cdots \neq \xi_2(t(t-k_1-\cdots-k_{m-1}))\} \\ &= \{\xi_2(tt) = \xi_2(t(t-1)) = \cdots = \xi_2(t(t-k_1+1)) \neq \xi_2(t(t-k_1) \\ &\neq \cdots \neq \xi_2(t(t-k_1-\cdots-k_{m-1})\}\end{aligned}$$

namely $k_i$ is the lowest value of $k$ with the property that $\xi_2(t(t-k)) \neq \xi_2(t(t-k_1-\cdots-k_i))$ where $k < k_1 + \cdots + k_i$, $i = 1, \ldots, m-1$.

The definition of the generalization error $\epsilon(t, m)$ implies that

$$\begin{aligned}\epsilon(t,m) = \langle [&P(\boldsymbol{\xi}(t+1) \in \Omega(t) | \xi_2(tt) \neq \xi_2(t(t-1)) \neq \cdots \neq \xi_2(t(t-m+1))) \\ &\cdot I_{\{\xi_2(tt)\neq\xi_2(t(t-1))\neq\cdots\neq\xi_2(t(t-m+1)))\}} \\ &+ P(\boldsymbol{\xi}(t+1) \in \Omega(t) | \xi_2(tt) \neq \xi_2(t(t-2)) \neq \cdots \neq \xi_2(t(t-m))) \\ &\cdot I_{\{\xi_2(tt)\neq\xi_2(t(t-2))\neq\cdots\neq\xi_2(t(t-m)))\}} \\ &+ P(\boldsymbol{\xi}(t+1) \in \Omega(t) | \xi_2(tt) \neq \xi_2(t(t-3)) \neq \cdots \neq \xi_2(t(t-m-1))) \\ &\cdot I_{\{\xi_2(tt)\neq\xi_2(t(t-3))\neq\cdots\neq\xi_2(t(t-m-1)))\}} \\ &+ \cdots ]\rangle \end{aligned} \tag{12}$$

where $I$ is the indicator function. Therefore to obtain an exact expression of $\epsilon(t, m)$ it suffices for us to consider each term in equation (12). In fact we see that

$$P(\boldsymbol{\xi}(t+1) \in \Omega(t) | \xi_2(tt) \neq \xi_2(t(t-k_1)) \neq \cdots \neq \xi_2(t(t-k_1-\cdots-k_{m-1})))$$

$$= \frac{1}{m}\left[\int_0^{\xi_1(tt)} \mathrm{d}x + \int_0^{\xi_1(t(t-k_1))} \mathrm{d}x + \cdots + \int_0^{\xi_1(t(t-k_1-\cdots-k_{m-1}))} \mathrm{d}x\right]$$

$$= \frac{1}{m}[\xi_1(tt) + \xi_1(t(t-k_1)) + \cdots + \xi_1(t(t-k_1-\cdots-k_{m-1}))]. \tag{13}$$

Note that

$$\frac{m!}{mm^{k_1}\ldots m^{k_{m-1}}} = P(\xi_2(tt) \neq \xi_2(t(t-k_1)) \neq \cdots \neq \xi_2(t(t-k_1-k_2\cdots-k_{m-1})))$$

together with equation (13) we derive that

$$\begin{aligned}&\langle[P(\boldsymbol{\xi}(t+1) \in \Omega(t)|\xi_2(tt) \neq \xi_2(t(t-k_1)) \neq \cdots \neq \xi_2(t(t-k_1-\cdots-k_{m-1}))\\ &\qquad \cdot I_{\{\xi_2(tt)\neq\xi_2(t(t-k_1))\neq\cdots\xi_2(t(t-k_1-\cdots-k_{m-1}))\}}]\rangle\\ &\qquad = \frac{m!}{m^2}\left[\frac{1}{m^{k_1}}\cdots\frac{1}{m^{k_{m-1}}}(\langle\xi_1(tt)\rangle + \langle\xi_1(t(t-k_1))\rangle + \cdots\right.\\ &\qquad \left. + \langle\xi_1(t(t-k_1-k_2-\cdots-k_m))\rangle)\right].\end{aligned} \tag{14}$$

Substituting equation (14) into equation (12), in terms of lemma 3 we obtain:

$$\begin{aligned}\epsilon(t,m) &= \sum_{k_1,k_2,\ldots,k_{m-1}=1}^{\infty} \frac{1}{m^2}\frac{m!}{m^{k_1}\ldots m^{k_{m-1}}}[\langle\xi_1(tt)\rangle + \langle\xi_1(t(t-1))\rangle + \cdots\\ &\qquad + \langle\xi_1(t(t-k_1-k_2-\cdots k_{m-1}))\rangle]\\ &= \sum_{k_1,k_2,\ldots,k_{m-1}} \frac{1}{m^2}\frac{m!}{m^{k_1}\ldots m^{k_{m-1}}}\left[\frac{1}{t} + \frac{1+k_1}{t} + \cdots + \frac{1+k_1+\cdots+k_{m-1}}{t}\right]\\ &= \sum_{k_1,k_2,\ldots,k_{m-1}} \frac{1}{m^2}\frac{m!}{m^{k_1}\ldots m^{k_{m-1}}}\left[\frac{m}{t} + \frac{(m-1)k_1}{t} + \cdots + \frac{k_{m-1}}{t}\right].\end{aligned} \tag{15}$$

By the identity

$$\sum_{k=1}^{\infty} \frac{k}{m^k} = \frac{m}{(m-1)^2}. \tag{16}$$

Equation (15) becomes

$$\begin{aligned}\epsilon(t,m) &= \frac{m!}{m^2}\left[\frac{m}{t}\frac{1}{(m-1)^{(m-1)}} + \frac{(m-1)}{(m-1)^{m-2}}\frac{m}{(m-1)^2} + \cdots + \frac{1}{(m-1)^{m-2}}\frac{m}{(m-1)^2}\right]\\ &= \frac{m!}{m^2}\left[\frac{m}{t}\frac{1}{(m-1)^{(m-1)}} + \frac{m^2(m-1)}{2t}\frac{1}{(m-1)^m}\right]\\ &= \frac{(m-1)!}{(m-1)^{(m-1)}}\left(\frac{m}{2}+1\right)\frac{1}{t}\end{aligned} \tag{17}$$

which is the desired conclusion. □

Equation (12) is the key identity of our approach which indicates that $\epsilon(t,m)$ is a linear combination of extremes under assumptions 1 and 2. The identity enables us to obtain, in conjunction with extreme value theory, an exact expression for generalization errors of the simple perceptron. It can readily be seen that all conclusions in theorem 1 are valid when $\xi_2(\tau) \sim p_1\delta_{(x_2=0,x_3=0,\ldots,x_m=0)} + p_2\delta_{(x_2=1,x_3=0,\ldots,x_m=0)} + \cdots + p_m\delta_{(x_2=0,x_3=0,\ldots,x_m=1)})$ with $p_i > 0, i = 1,\ldots,m, \sum_{i=1}^m p_i = 1$.

To confirm our theoretical approach above; the coefficient of the generalization error of the simple perceptron is independent of input distributions, here we include a numerical

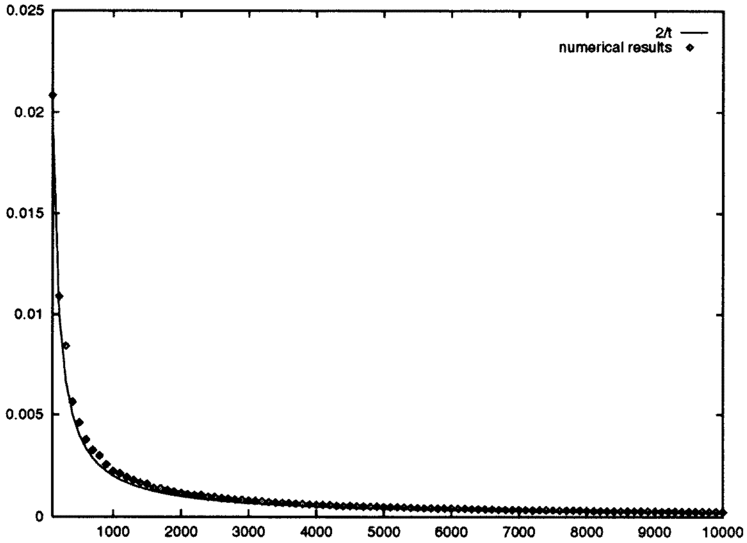

**Figure 4.** Generalization error versus number of examples. Numerical simulations of $\epsilon(t, 2)$ when inputs $(\xi_1(\tau), \xi_2(\tau))$ are i.i.d. uniformly distributed random variables. $\epsilon(t, 2)$ for $t = 100, 200, 300, \ldots, 10\,000$ are numerically calculated.

simulation to estimate the generalization error $\epsilon(t, 2)$. Let *both* $\xi_1(\tau), \xi_2(\tau)$ be i.i.d. and uniformly distributed over [0, 1]. Figure 4 shows the numerical results with 10 000 times simulations for each $t = 100, 200, \ldots, 10\,000$. Numerical results underpin our theoretical approach; the exact form of the generalization error of the simple perceptron can be obtained under assumptions 1 and 2.

*Remark 1.* Surprisingly, our numerical and theoretical results are both different from the results obtained in terms of the replica trick approach in which it is estimated that $\epsilon(t, m) = 0.62m/t$. The deviation can be understood from the following two reasons. First the replica trick approach, as we have already referred to at the beginning of the paper, is only valid when $m$ tends to infinity in proportion to $t$. Secondly, the behaviour of extreme values also changes substantially when $k$ is in proportion to $t$, see for example [7, 8]. However, when $m$ is small this effect will not play a role in our estimation since in equation (12) the term with large $k$ is already quite small. But when $m \to \infty$ is in proportion to $t$ we have to take this effect into account in equation (12).

*Remark 2.* According to lemma 3 we have the same result when the target plane is arbitrary rather than $x = 0$.

## 4. High-order perceptron

The higher-order simple perceptron is a generalization of the simple perceptron considered in the previous section. These higher-order neurons, called sigma–pi units by Rumelhart *et al* [23] can be employed to define the conditions of invariant perceptron. The action of the higher-order synapses can be understood, from a more general point of view, as the evaluation of nonlocal correlations between input patterns. Bialek and Zee [23] have argued, in the context of statistical mechanics, that such nonlocal operations are an essential requisite of invariant perception. There is no doubt that human vision allows for a very large class of invariances, achieving close to optimum performance, but it is not known to what extent the brain relies on nonlocal information processing for that purpose. It is well known that a higher-order simple perceptron, without resorting to multilayer structures and the BP algorithm, is capable of solving any classification problems. In this section we carry out a calculation of generalization errors of the higher-order perceptron.

*Lemma 4.* For uniformly distributed inputs $\xi_1(\tau)$, when $t \to \infty$ we have

$$\epsilon(t,1) := \langle P(\xi_1(t+1) \leqslant \xi_1^p(tt)|\mathcal{F}_t)\rangle = \frac{\Gamma(p+1)}{t^p} \qquad p > 0 \tag{18}$$

where $\Gamma(x)$ is the gamma function.

*Proof.* By the definition of $\epsilon(t,1)$ (equations (6) and (5)) we obtain

$$\begin{aligned}\epsilon(t,1) &= \langle \xi_1^p(tt)\rangle \\ &= \int_0^\infty x^p t \mathrm{e}^{-tx}\,\mathrm{d}x \\ &= \frac{1}{t^p}\int_0^\infty x^p \mathrm{e}^{-x}\,\mathrm{d}x \\ &= \frac{\Gamma(p+1)}{t^p}.\end{aligned} \tag{19}$$

□

The high order of perceptron does not have an influence on $\xi_2(\tau)$ and so, combining conclusions in the previous section, we conclude the following theorem.

*Theorem 2.* Under assumptions 1 and 2 for the $p$-order perceptron, $p > 0$, we have

$$\epsilon(t,m) = \begin{cases} \dfrac{\Gamma(p+1)}{t^p} & \text{if } m = 1 \\ \dfrac{(m-1)!\Gamma(p+1)}{(m-1)^{(m-1)}}\left(\dfrac{m}{2}+1\right)\dfrac{1}{t^p} & \text{otherwise.} \end{cases} \tag{20}$$

## 5. Conclusions

Although the perceptron learning rule is now almost 40 years old it does not seem to have lost much of its attraction [20]. On the contrary, there are several appealing features, on both a practical and theoretical level, that make it appear advantageous—the perceptron rule is easy to implement since the corrections are simple additions or subtractions; the famous perceptron convergence theorem states that any set of examples that has a solution vector will be classified correctly after learning. In this paper we have calculated the generalization error of the $p$-order perceptron of the worst learning where $p > 0$.

There are many questions requiring further investigation. For example, a challenging problem is to generalize our approach to consider algorithms such as the BP algorithm etc [12, 15, 16]. It is promising to replace the line we considered in this paper by a curve reflecting the nonlinearity of the BP and the curve is determined by a few (more than two in the two-dimensional case) extreme values of input signals; taking a similar approach to that which we developed here, we would expect to obtain a learning curve for the BP algorithm.

In summary, our approach reported in this paper opens up new possibilities for rigorous analyses of generalization errors which reflect intricate nonlinear properties underlying most learning algorithms in neural networks.

## Acknowledgments

I thank David Brown and Elinor Stephens for valuable comments on the manuscript.

## References

[1] Amari S and Murata N 1993 Statistical theory of learning curves under entropic loss criterion *Neural Comput.* **5** 140–53

[2] Amari S, Murata N and Ikeda K 1995a Statistical theory of learning curves *Neural Networks: The Statistical Mechanics Perspective* ed J Oh, Ch Kwon and S Chao, pp 3–17

[3] Amari S, Murata N, Müller K-R, Finke M and Yang H 1995b Asymptotic statistical theory of overtraining and cross-validation *METR 95-06* Department of Mathematical Engineering and Information Physics, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan

[4] Baum E B 1990 The perceptron algorithm is fast for nonmalicious distribution *Neural Comput.* **2** 248

[5] Baum E B and Haussler D 1989 What size net gives valid generalization *Neural Comput.* **4** 151–60

[6] Cohn D and Tesauro G 1992 How tight are the Vapnik–Chervonenkis bounds *Neural Comput.* **4** 249–69

[7] Balkema A A and De Haan L 1978 Limit distributions for order statistics. I *Theory Prob. Appl.* **XXIII** 77–92

[8] Balkema A A and De Haan L 1978 Limit distributions for order statistics. II *Theory Prob. Appl.* **XXIII** 341–58

[9] Bouchaud J-P and Mézard M 1997 Universality classes for extreme value statistics *J. Phys. A: Math. Gen.* **30** 7997–8015

[10] Engel A and den Broeck C V 1993 Statistical mechanics calculation of Vapnik Chervonenkis bounds for perceptron *J. Phys. A: Math. Gen.* **26** 6893–914

[11] Feng J 1997 Behaviours of spike output jitter in the integrate-and-fire model *Phys. Rev. Lett.* **79** 4505–8

[12] Feng J 1997 Lyapunov functions for neural nets with nondifferentiable input-output characteristics *Neural Comput.* **9** 45-51

[13] Feng, J., and Brown, D.(1998), Fixed-point attractor analysis for a class of neurodynamics *Neural Comput.* **10** 189–213

[14] Feng J and Brown D 1998 Spike output jitter, mean firing time and coefficient of variation *J. Phys. A: Math. Gen.* **31** 1239–52

[15] Feng J and Hadeler K P 1996 Qualitative behaviors of some simple neural networks *J. Phys. A: Math. Gen.* **29** 5019–33

[16] Feng J, Pan H and Roychowdhury V P 1996 On neurodynamics with limiter function and Linsker's developmental model *Neural Comput.* **8** 1003–19

[17] Feng J and Tirozzi B 1995 The SLLN for the free-energy of the Hopfield and spin glass model *Helv. Phys. Acta* **68** 365–79

[18] Feng J and Tirozzi B 1997 Capacity of the Hopfield model *J. Phys. A: Math. Gen.* **30** 3383–91

[19] Galambos J 1984 *Introductory Probability Theory* (New York: Marcel Dekker) pp 164–8

[20] Gray M S, Lawrence DT, Golomb B A and Sejnowski T J 1995 A perceptron reveals the face of sex *Neural Comput.* **7** 1160–4

[21] Haussler D, Kearns M and Shapire R 1991 Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension *Proc. 4th Annu. Workshop on computational Learning Theory* (San Mateo, CA: Morgan Kaufmann) pp 61–74

[22] Haussler D, Littlestone N and Warmuth K 1988 Predicting $\{0, 1\}$ functions on randomly drawn points *Proc. COLT'88* (San Mateo, CA: Morgan Kaufmann) pp 280–95

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

[24] Leadbetter M R, Lindgren G and Rootzén H 1983 *Extremes and Related Properties of Random Sequences and Processes* (New York: Springer)

[25] Levin E, Tishby N and Solla S A 1990 A statistical approach to learning and generalization in layered neural networks *Proc. IEEE* **78** 1568–74

[26] Murata N, Yoshizawa S and Amari S 1994 Network information criterion-determinate the number of hidden units for an artificial neural network model *IEEE Trans. NN* **6** 865–72

[27] Opper M and Haussler D 1991 Calculation of the learning curve of Bayes optimal classification algorithm for learning perceptron with noise *Proc. 4th Annu. Workshop on Computer Learning Theory* pp 75–87

[28] Opper M and Haussler D 1995 Bounds for predictive errors in the statistical mechanics of supervised learning *Phys. Rev. Lett.* **75** 3772–5

[29] Romeo A 1993 Generalization transitions in hidden-layer neural networks for 3rd-order feature discrimination *Phys. Rev.* E **47** 2162–71

[30] Seung H S, Sompolinsky H and Tishbby N 1992 Statistical mechanics of learning from examples *Phys. Rev.* A **45** 6056–91

[31] Sompolinsky H and Tishby N 1990 Learning in a 2-layer neural network of edge detectors *Europhys. Lett.* **13** 567–72

[32] Vapnik V N and Chervonenkis A Y 1971 On the uniform convergence of relative frequencies of events to their probabilities *Theory Prob. Appl.* **16** 264–80

[33] Vapnik E, Levin E and LeCun Y 1994 Measuring the VC dimension of a learning machine *Neural Comput.* **5** 851–76

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

[35] Yamanishi K 1991 A loss bound model for on-line stochastic prediction strategies *Proc. 4th Annu. Workshop on Computer Learning Theory* pp 290–302