Lower bounds on generalization errors for drifting rules

1993 J. Phys. A: Math. Gen. 26 6161

(http://iopscience.iop.org/0305-4470/26/22/017)

J. Phys. A: Math. Gen. 26 (1993) 6161–6171. Printed in the UK

Lower bounds on generalization errors for drifting rules

Osame Kinouchi and Nestor Caticha
Instituto de Física, Universidade de São Paulo, CP 20516, 01498 São Paulo, SP, Brazil

Received 5 January 1993

Abstract. The problem of generalization by single-layer perceptrons is studied in the case of time-dependent rules. Lower bounds for the generalization errors within the 'single presentation of examples' case are obtained for randomly drifting rules. These bounds suggest a learning algorithm which uses knowledge of the error itself. Since this error is not readily available it has to be estimated through a mechanism of self-evaluation. The capacity of incorporating recency information into the error estimate is highly desirable. The mechanism proposed has the advantage, beyond good performance, of being self-adaptive, in the sense that it adjusts to changes in the unknown drift rate of the rule. The performance of the rule is also studied for sudden changes in an attempt to mimic the so-called Wisconsin test.

1. Introduction

In an ever-changing environment the ability of organisms to present adaptive behaviour might be an important factor in determining their chances of survival. Thriving under a randomly varying set of rules will depend on the capacity of incorporating useful information from the outside world. Such adaptive behaviour can be modelled within supervised learning in neural nets. Perceptron learning of a time-dependent rule has been recently studied by Biehl and Schwarze [1] in the case where there is a single presentation of the examples. We discuss again this type of problem with the aim of determining what is the optimal way in which the information from the changing environment ('professor' or 'rule') can be used by a single-layer perceptron in order to maximize its generalization ability.

The natural approach to this problem is 'on-line' learning. This means that the examples are used only once and not repeatedly ('iterated learning') until some error measure over all examples is minimized. Iterated learning might not be efficient if old examples are presented again at a later time when they are not any longer representative of the state of the rule. Furthermore, single presentation might also be quite efficient, since in the optimal case (expected stability algorithm) for a static rule it leads [2] to only exactly twice the error of the Bayes limit of Opper and Haussler [3], with just a small fraction of the computational effort.

In section 2 the model [1] and notation are introduced. For a randomly drifting rule we obtain in section 3 the result that, analogously to the time-independent case, suprises carry a high information content, and should be attributed a high weight, while very small importance must be given to obvious facts. More specifically, we obtain an expression for the weight a Hebbian term must have in order to maximize generalization. Examples 'easy' to classify by the student net are those with a high local field at the perceptron's output unit. When an 'easy' example is misclassified it receives a very high weight and a very small one when it is correctly classified. Also in section 3 we present lower bounds for the

0305-4470/93/226161+11$07.50

generalization error as a function of T, where T is a measure of the amplitude of the white noise driving the rule. Calling, as usual, α the ratio of the number of patterns to the number of input units, we obtain that, for $\alpha \to \infty$ and small T, the saturation generalization error for optimal learning grows as $T^{1/3}$, while being proportional to $T^{1/4}$ in the optimal pure Hebbian case [1]. The strategy of learning by 'queries' ('selection of examples') [4, 10] is also analysed.

These bounds can be reached by a learning algorithm which uses the value of the generalization error itself. Since this is not a very realistic feature, we propose in section 4 a new self-adaptive algorithm which relies on the estimate of the generalization error, as judged by recent success. Several estimators can be considered and we discuss the effects on learning of a parameter which measures the persistence of old examples in the actual estimator. Results of simulations are presented for the case of slowly drifting rule as well as for the so-called Wisconsin test [5], where the rule is piecewise constant in time. Section 5 comprises some concluding remarks.

2. The model

We now examine the problem of learning a time-dependent linearly separable rule within an exactly solvable model. A single-layer, single-output unit perceptron, the 'student', learns from examples generated by another such perceptron, the 'professor', whose synaptic weights change randomly in time. Let N be the number of inputs, $S \in (-1, 1)^N$ and $B \in \mathbb{R}^N$ or $(-1/\sqrt{N}, 1/\sqrt{N})^N$ the input vector and the professor synaptic coupling vector, respectively. The correct (professor) output is

$$\sigma_B^\mu = \operatorname{sign}\left(\sum_{i=1}^{N} B_i S_i^\mu\right). \tag{1}$$

The student is described by a synaptic vector $J \in \mathbb{R}^N$ and its output is

$$\sigma_J^\mu = \operatorname{sign}\left(\sum_{i=1}^{N} J_i S_i^\mu\right). \tag{2}$$

The generalization ability is defined as the probabilty that at a given time the outputs of both the professor and student nets be equal. The synaptic vector J is to be constructed from the information contained in the examples, that is, a sequence of pairs (S^μ, σ_B^μ), where μ is the time label or the example position in the sequence.

We first consider, as did Biehl and Schwarze [1], a time evolution of the rule

$$B^\mu = B^{\mu-1} + (1/\sqrt{N})\eta^\mu \tag{3}$$

where η^μ is a random vector satisfying

$$\langle \eta_i^\mu \eta_j^\nu \rangle = (2/N)T\delta_{ij}\delta_{\mu\nu} \tag{4}$$

and we keep the professor norm B^μ constant equal to 1 by imposing the condition

$$\sqrt{N}B^{\mu-1} \cdot \eta^\mu = -\tfrac{1}{2}\eta^\mu \cdot \eta^\mu = -T. \tag{5}$$

Thus

$$B^\mu \cdot B^{\mu-1} = 1 - T/N \tag{6}$$

where T is the drift parameter. As usual the overlap between B and J is the relevant quantity [6, 10]

$$\rho^\mu \equiv R^\mu / J^\mu \qquad R^\mu \equiv B^\mu \cdot J^\mu \qquad J^\mu \equiv \sqrt{J^\mu \cdot J^\mu} \tag{7}$$

since the generalization ability G, the generalization error e_G and ρ are related through

$$e_G = 1 - G = (1/\pi)\cos^{-1}\rho \tag{8}$$

where self average has been assumed. We will use the notation

$$b_\mu \equiv \frac{\boldsymbol{B}^{\mu-1}\cdot\boldsymbol{S}^\mu}{B^{\mu-1}} \qquad \sigma_B^\mu = \mathrm{sign}(b_\mu) \tag{9}$$

for the local field and output of the teacher for the μth example. Our notation is such that the professor changes from $\boldsymbol{B}^{\mu-1}$ to $\boldsymbol{B}^\mu$ only after the presentation of the μth example. Also

$$h_\mu \equiv \frac{\boldsymbol{J}^{\mu-1}\cdot\boldsymbol{S}^\mu}{J^{\mu-1}} \qquad \sigma_J^\mu = \mathrm{sign}(h_\mu) \tag{10}$$

are the normalized local field and the output of the student. The local stabilty of a given example, before it is used to modify $\boldsymbol{J}$, is

$$\Delta_\mu \equiv h_\mu \sigma_B^\mu \tag{11}$$

and the example is correctly classified if Δ_μ is positive. The calculations will also involve a Gardner-like parameter related to the stability on the professor

$$\kappa_\mu \equiv \frac{b_\mu \sigma_B^\mu}{\rho^{\mu-1}}\,. \tag{12}$$

3. Lower bounds on the generalization errors

The learning dynamics that we study has the form

$$\boldsymbol{J}^\mu = \left(1 - \frac{\Omega_\mu}{N}\right)\boldsymbol{J}^{\mu-1} + \frac{\boldsymbol{J}^{\mu-1}}{N} W_\mu \sigma_B^\mu \boldsymbol{S}^\mu + \frac{\boldsymbol{J}^{\mu-1}}{\sqrt{N}}\tilde{\eta} \tag{13}$$

where Ω_μ is a decay term and W_μ is the weight of the Hebbian term. These two quantities depend on the particular example, and our objective is to determine the best possible weight in order to optimize generalization. The last term takes into account possible white noise in the synaptic couplings, but its overall effect is just to increase the amplitude of the drift parameter T by $\frac{1}{2}\langle\tilde{\eta}\cdot\tilde{\eta}\rangle$. The $J^{\mu-1}$ factor is included to simplify further developments. With this notation, $W_\mu = 1/J^{\mu-1}$ for the simple Hebb rule, $W_\mu = \kappa - \Delta_\mu$ for the Adaline algorithm and $W_\mu = (\kappa - \Delta_\mu)\Theta(\kappa - \Delta_\mu)$ for the relaxation algorithm, with κ a positive constant [9]. We now follow closely the method of [2]. In the thermodynamic limit $N \to \infty, \alpha = \mu/N$, the learning dynamics can be written as a differential equation, which after being averaged over all the possible histories, leads to

$$\frac{\mathrm{d}\rho}{\mathrm{d}\alpha} = \rho\left(\langle W_\mu(\kappa_\mu - \Delta_\mu) - \tfrac{1}{2}W_\mu^2\rangle - T\right) \tag{14}$$

for the overlap.

The evolution of the length of the synaptic vector $\boldsymbol{J}$ is given by

$$\frac{\mathrm{d}J}{\mathrm{d}\alpha} = J\left(\langle W_\mu\Delta_\mu + \tfrac{1}{2}W_\mu^2 - \Omega_\mu\rangle + T\right). \tag{15}$$

The choice of $\Omega_\mu = W_\mu\Delta_\mu + \frac{1}{2}W_\mu^2 + T$ gives a *local* method to keep the norm of $\boldsymbol{J}$ constant. Nevertheless, this needs *a priori* knowledge about T.

Note that the decay term Ω_μ is absent from the equation for ρ. Then, if W_μ does not depend on $J^{\mu-1}$, the two differential equations are decoupled and the decay term does not

affect the performance (in contrast with the simple Hebb case [1] where $W_\mu = 1/J^{\mu-1}$ and the two equations are coupled).

The choice of the weight function W_μ is made by demanding that the gain per example, e.g. $\mathrm{d}\rho/\mathrm{d}\alpha$, be maximum, with the restriction that it can only depend on b_μ through its sign. Not suprisingly the resulting weight function is the same as in [2]

$$W_\mu^{\mathrm{ES}} = \langle \kappa_\mu \rangle_{\mathrm{ES}} - \Delta_\mu \tag{16}$$

where the triangular brackets $\langle \cdots \rangle_{\mathrm{ES}}$ denote the expected value calculated with the conditional probability distribution of $|b_\mu|$ given σ_B^μ and h_μ, thus the abbreviation ES for expected stabilty. Therefore, the maximum average gain per example can be calculated from

$$\frac{\mathrm{d}\rho}{\mathrm{d}\alpha} = \rho\,(I(\rho) - T) \tag{17}$$

where

$$I(\rho, P(S)) = \sum_{\sigma_B^\mu = \pm 1} \int_{-\infty}^{\infty} \mathrm{d}h_\mu \; P(\sigma_B^\mu, h_\mu) \frac{(W_\mu^{\mathrm{ES}})^2}{2}. \tag{18}$$

Up to now the probabilty distribution of the examples $P(S)$ from which $P(\sigma_B^\mu, h_\mu)$ is obtained has not been chosen. It reflects the strategy in the choice of the learning examples. Two strategies have been considered in the literature [3, 4, 6, 10]: *passive learning*, that is, learning from random examples with a uniform distribuition, which leads to b_μ and h_μ being normal Gaussian variables with correlation ρ; *active learning* (learning by 'selection of examples' or 'queries') where the student demands the answer to the 'hardest to classify' examples. In this case the ES average is performed over $P(|b_\mu| \mid h_\mu = 0)$.

In the case of passive learning the weight function is [2]

$$W_\mu^{\mathrm{ES}} = \frac{\lambda}{\sqrt{2\pi}} \frac{\exp(-\Delta_\mu^2/2\lambda^2)}{H(-\Delta_\mu/\lambda)} \tag{19}$$

where

$$\lambda \equiv \frac{\sqrt{1-\rho^2}}{\rho} = \tan(\pi e_G) \tag{20}$$

and

$$H(x) \equiv \int_x^\infty \mathrm{D}t \qquad \mathrm{D}t \equiv \frac{\mathrm{e}^{-t^2/2}}{\sqrt{2\pi}}\,\mathrm{d}t\,. \tag{21}$$

With this weight W the function I appearing in equation (18) is

$$I(\lambda) = \frac{\lambda^3}{2\pi} \int_{-\infty}^{\infty} \mathrm{D}u \; \frac{\mathrm{e}^{-(1-\lambda^2)u^2/2}}{H(u)}. \tag{22}$$

As $\alpha \to \infty$ the error stops decreasing. The saturation error e_G^∞ will increase as T^z, with an exponent z characteristic of the algorithm and learning strategy. The error $e_G^\infty(T)$ is obtained from equation (20) and from

$$I(\lambda^\infty) = T\,. \tag{23}$$

For small T

$$e_G^\infty \simeq \frac{1}{\pi}\lambda^\infty = C T^{1/3} \tag{24}$$

where the constant C is given by

$$C = \left[\frac{\pi^2}{2}\int_{-\infty}^{\infty} \mathrm{D}u\, \frac{\mathrm{e}^{-u^2/2}}{H(u)}\right]^{-1/3} \approx 0.447\,. \tag{25}$$

This $z = \frac{1}{3}$ should be compared to the $z = \frac{1}{4}$ exponent of the optimal Hebb case of reference [1]. For the selection of examples strategy, we have

$$W_\mu^{\mathrm{ES}} = \sqrt{\frac{2}{\pi}}\lambda \tag{26}$$

and

$$I(\lambda) = \frac{\lambda^2}{\pi}\,. \tag{27}$$

The assymptotic value of the error in this case is

$$e_G^\infty = \frac{1}{\pi}\tan^{-1}\sqrt{\pi T} \tag{28}$$

leading to $z = \frac{1}{2}$.

The exponent z is not independent of the generalization exponent x which characterizes the behaviour of e_G for $T = 0$ and large α

$$e_G \propto \alpha^{-x}\,. \tag{29}$$

To show this we proceed as follows. For small T and large α (small λ) we have in general

$$\frac{\mathrm{d}\rho}{\mathrm{d}\alpha} = \rho\left(c_1(T)\lambda^{n_1} + c_2(T)\lambda^{n_2} + \cdots - T\right) \tag{30}$$

where c_1 and c_2 may depend on T. Also $n_1 < n_2$, and if $c_1(T=0) \neq 0$ then $n_1 = 1/z$, giving the relation

$$\frac{1}{x} = \frac{1}{z} - 2 \tag{31}$$

This can be seen by inserting the assymptotic trial solution (for $T = 0$)

$$\rho = 1 - K\alpha^{-2x} \tag{32}$$

where K is some positive constant, into the differential equation leading to

$$\alpha^{-2x-1} \propto e_G^{1/z} \propto \alpha^{-x/z} \tag{33}$$

from which we obtain relation (31). The case $z = \frac{1}{2}$ ($x \to \infty$) indicates an exponential error decay obtained with the learning strategy of 'selection of examples' (see [2]).

If $c_1(T=0) = 0$ then one can only say that

$$\frac{1}{x} > \frac{1}{z} - 2. \tag{34}$$

The bounds on the generalization error $e_G(\alpha)$ obtained by numerically integrating equation (17) with the apropriate weight function are shown in figure 1 for $T = 0.5$. Also the results for the optimal Hebb are presented for comparison. This algorithm requires the knowledge of the value of the drift parameter and uses an optimal weight decay $\Omega(T)$ for the simple Hebb rule. The saturation error is larger in the case of random uniform examples, while in the case of selection of examples the two errors are the same. This result should be compared with that of [1], bearing in mind that we start from *tabula rasa*

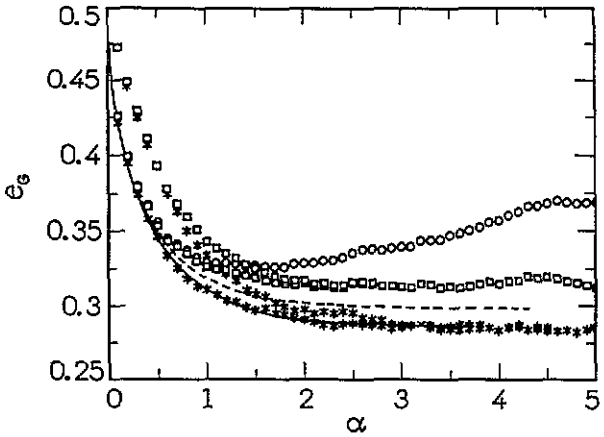

Figure 1. Lower bounds for the generalization error $e_G(\alpha)$ obtained from the integration of equation (17) with W of the expected stability algorithm: passive learning (dashed curve) and active learning (solid curve). Simulation results for the simple Hebb (circles). Optimal Hebb for passive learning (squares), and optimal Hebb with selection of examples (stars) both for *tabula rasa* and starting from $J = 1$ [1]. Simulations starting from *tabula rasa* (lower curves) reach the limit faster than from $J = 1$. Average over 50 runs, $N = 149$.

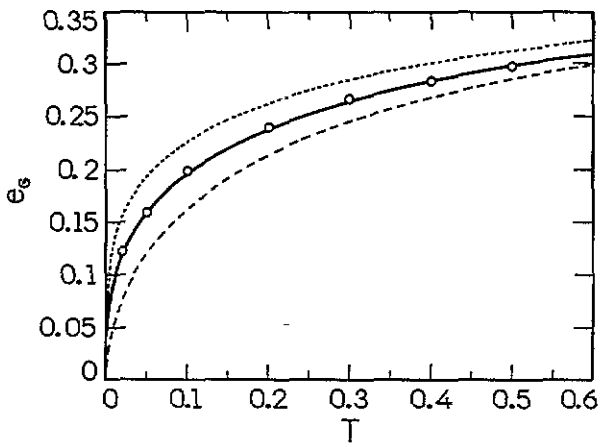

Figure 2. Asymptotic generalization error $e_G(T)$ from equation (23) as a function of the noise parameter T. Passive learning optimal Hebb [1] (short-dashed curve), expected stability (solid curve) and simulation results, $N = 149$, average over 50 runs (circles). Active learning (long-dashed curve).

while they use $J = 1$ as the initial condition and so their approach to the limit will be slower.

In figure 2 our results for the assymptotic errors $e_G^{\infty}(T)$ are shown, as well as the ones for the algorithm of [1]. Also the selection of examples strategy (lower curves) errors are shown, being the same for both algorithms.

Since the optimal weight function W_{μ}^{ES} depends on ρ, the algorithm cannot be implemented without knowledge of the drift parameter T, which is needed to integrate equation (17). As it stands it only provides lower bounds on the average generalization errors. Of course, we can use in the weight function a value ρ_{μ} measured during the simulation to obtain an optimal 'benchmark' curve, against which other algorithms will be

compared. We will call this procedure the 'benchmark' algorithm. In the next section we will study a self-adaptive algorithm that uses an estimate of the overlap ρ and performs very well without knowledge of the noise level.

4. Self-adaptive algorithm

We now turn to a practical implementation of these ideas. First notice that, by equation (20), the optimal weight function can be written as

$$W(e_G, \Delta_\mu) = \frac{1}{\sqrt{2\pi}} \tan(\pi e_G) \exp\left(-\frac{\Delta_\mu^2}{2\tan^2(\pi e_G)}\right)\left[H\left(\frac{-\Delta_\mu}{\tan(\pi e_G)}\right)\right]^{-1}. \tag{35}$$

This suggests that we consider the neural net as endowed with a self-evaluation system which is used to estimate the generalization error, or more specifically $\tan(\pi e_G)$, by the results, failures or successes on the previous examples. The self-evaluator mechanism integrates over a range of time and two competing factors have to be considered. If the integration time is too large, then it will remember results which are no longer meaningful and its estimation will lag the actual value of e_G. If it is too short, it will have no time to accumulate a significant statistics. We will consider an exponential loss of information of old examples and look at the following evaluator:

$$\hat{e}_G^{(\mu)} = \left(1 - \frac{\omega}{N}\right)\hat{e}_G^{(\mu-1)} + \frac{\omega}{N}\varepsilon_\mu . \tag{36}$$

Here ω is the inverse integration time or recency span, and $\varepsilon_\mu = \left(1 - \sigma_B^\mu \sigma_J^\mu\right)/2$, which is (one) zero if the last example was (in)correctly classified. Since the tangent diverges at $\pi/2$, it is not very precise to estimate the tangent factor of equation (35) by just calculating the tangent of $\pi\hat{e}_G^{(\mu)}$, and we instead use as estimators the truncated power series expansions

$$\hat{\lambda}_k = \hat{\tan}(\pi e_G)_k = \pi\hat{e}_G + \tfrac{1}{3}(\pi\hat{e}_G)^3 + \tfrac{2}{15}(\pi\hat{e}_G)^5 + \cdots + c_k(\pi\hat{e}_G)^k . \tag{37}$$

In figure 3 we judge the different estimators by comparing them to the benchmark value in a simulation. The third power estimator $\hat{\lambda}_3$ is seen to be the best. In figure 4 the rule is fixed for $\alpha < 5$ and $\alpha > 15$, while it drifts with $T = 0.2$ in between. The performance of

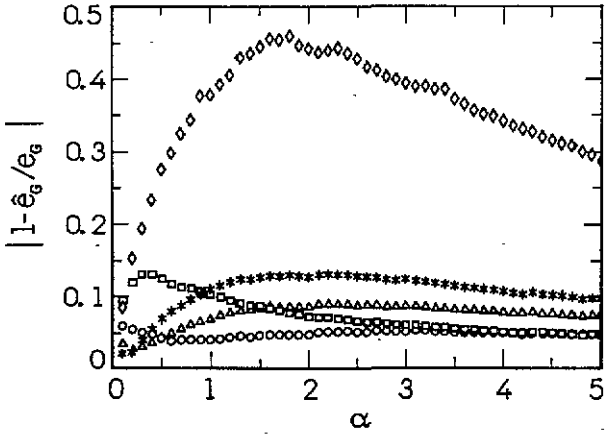

Figure 3. Performance of the different evaluators ($\omega = 2$), $|1 - \hat{e}_G/e_G|$, where $\hat{e}_G$ is the generalization error for the following evaluators: tangent (diamond), $\hat{\lambda}_1$ (squares), $\hat{\lambda}_3$ (circles), $\hat{\lambda}_5$ (triangles), $\hat{\lambda}_7$ (star), and e_G is the benchmark algorithm generalization error.

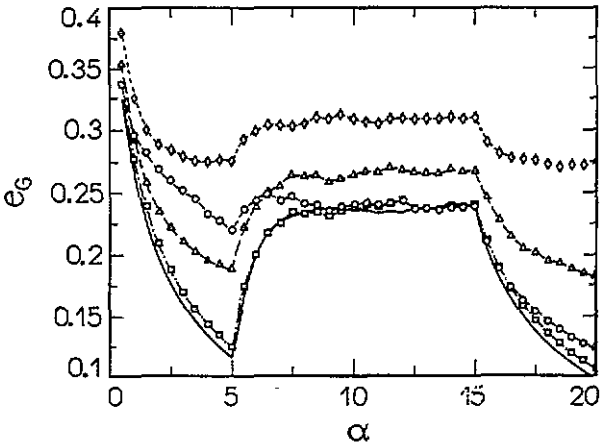

Figure 4. Performance of the $\hat{\lambda}_3$ algorithm for different values of ω. The rule is fixed until $\alpha = 5$, then it drifts until $\alpha = 15$, with a noise parameter $T = 0.2$. $\omega = 0.2$ (circles), $\omega = 2$ (squares), $\omega = 20$ (triangles), $\omega = N$ (diamond), benchmark (solid curve). $N = 99$, average over 50 runs.

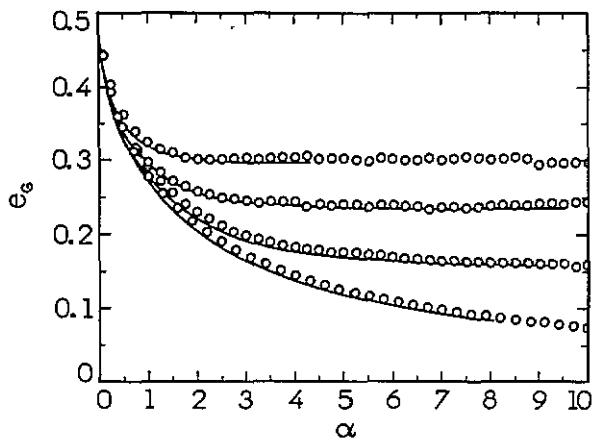

Figure 5. Generalization error for different levels of noise. Self-adaptive algorithm with the $\lambda_3(\omega = 2)$ evaluator (circles) from a simulation with $N = 149$, averaged over 50 runs and for the lower bounds (solid curve). From the lower curve to the top for $T = 0, 0.05, 0.2, 0.5$, respectively.

the $\hat{\lambda}_3$ estimator is measured for different values of ω and a good choice is seen to lie near the value $\omega = 2$.

In figure 5 the results of a simulation are shown for an algorithm which uses the best estimator ($\hat{\lambda}_3$) with $\omega = 2$. This is a realistic algorithm, since neither the drift parameter nor the value of the overlap ρ are needed. It is seen to work quite well since it approximates the lower bounds very efficiently for different drift rates. This in fact means that the algorithm is self-adaptive, a fact which can be better seen in figure 6 where the rule is fixed until $\alpha = 5$, performs a random walk with $T = 0.2$ until $\alpha = 15$, after which it remains fixed again.

We also have performed the so-called Wisconsin test [5]. In this situation the rule suddenly changes and the adaptation of the student to the new environment is analysed. In

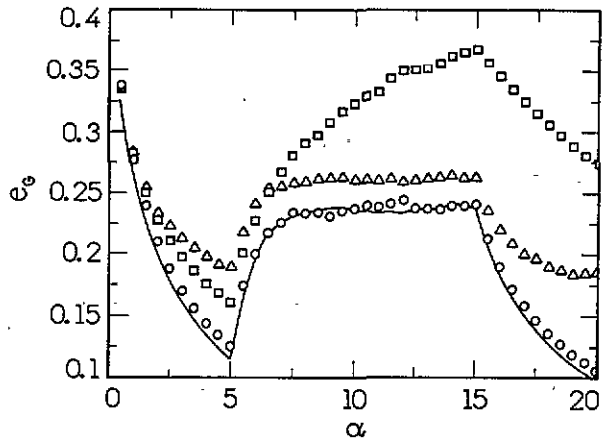

Figure 6. Performance of different algorithms for changing drifts. For $\alpha < 5$ and $\alpha > 15$, $T = 0$ and in between $T = 0.2$. Simple Hebb (squares), optimal Hebb (triangles), self-adaptive algorithm with the $\hat{\lambda}_3(\omega = 2)$ (circles) and the benchmark (solid curve). $N = 99$, average over 50 runs.

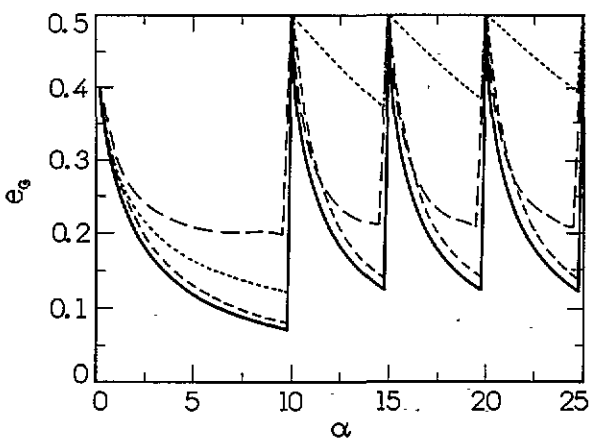

Figure 7. The Wisconsin test: the rule is piecewise constant in time. It suddenly changes at $\alpha = 10, 15, 20$. The dotted curve is the generalization ability of the simple Hebb algorithm. It is quite efficient at the begining, but is very slow in adapting to a new rule. The dashed curve is obtained for the self-adaptive algorithm with the $\hat{\lambda}_3(\omega = 2)$ evaluator. The solid curve is obtained using the measured value of ρ (benchmark algorithm).

figure 7 the performance of our algorithm in the Wisconsin test is compared with the simple Hebb rule, and can be seen to be significantly better.

The failure to use the optimal algorithm might be thougth of as a lesion to the learning mechanism of a neural net. Althougth by using the simple Hebb algorithm a perceptron can learn the first rule, it will have a great difficulty to detect a rule change. A similar behaviour has been noticed in patients with 'pre-frontal syndrome' [5], and has been previously modelled by Danchin and Changeux [7] and by Levine *et al* [8] with neural networks. It is interesting that the same effects appear in a natural way within the much simpler perceptron architecture *which has not been pre-designed to model this phenomenon.*

5. Conclusions

We have analysed the problem of learning a linearly separable rule drifting randomly with time, through the single presentation of examples. Lower bounds to the generalization errors are given. For low rule drift (or low synaptic noise) the saturation error increases as T^z. The exponent for passive learning can be improved from the Hebbian with optimal weight decay result of Biehl and Schwarze, $z = \frac{1}{4}$, to $z = \frac{1}{3}$, while $z = \frac{1}{2}$ is achieved in the active learning strategy. For on-line learning we obtained a relation between the generalization exponent x and the noise exponent z, $1/x = 1/z - 2$, if $c_1(T = 0) \neq 0$. This conditon is true for a large class of algorithms, and a careful study of this will appear elsewhere. It will be interesting to check if this relation is also valid for iterated learning.

These bounds inspire a realistic self-adaptive algorithm which, in order to assess the actual generalization abilty, needs a self-evaluating mechanism. The performance of this algorithm is studied numerically for slow drifts as well as for sudden changes of the rule, the so-called Wisconsin test. The algorithm approximates very well the optimal performance, does not require previous knowledge of the drift parameter and is robust to changing rule drifts.

This algorithm belongs to the class which we have previously called expected stability algorithms [2], since the weight function depends on the expectation value of the stability of an example in the professor net as judged by the student. The learning proceeds by giving higher weight to the suprising examples, that is, those with a high local field in the student net, which have been misclassified. This differentiated treatment of the examples is shared by other algorithms, but with a weight function independent of the learning stage (e_G). An important characteristic is that the optimal weight function depends on the generalization error, and this demands that the net be endowed with an extra feature, that is a self-evaluator mechanism. The failure of the Hebb algorithm in keeping information about temporal ordering leads to poor adaptability under changing rules or 'persistence' and lack of 'recency' information, which is similar to the behaviour of some quite complex neural nets [7, 8] which have been specifically built to model 'pre-frontal' syndrome.

The extension of this kind of results for multilayer nets would be of great interest both for the theoretical aspects as well as for their possible applications. We are presently working on these ideas and this will be the topic of future publications.

Acknowledgments

We thank M Biehl and H Schwarze for communicating their results prior to publication, and H S del Nero for indicating references [5, 8] and for discussions of the 'pre-frontal' syndrome.

OK has been supported by a CNPq fellowship, and NC has received partial support from CNPq and FAPESP.

Note added. After this work was finished we received a preprint from Biehl and Schwarze where essentially the same lower bounds are obtained for the continuous drifting rule case.

References

[1] Biehl M and Schwarze H 1992 Online learning of a time dependent rule *Europhys. Lett.* **20** 733

[2] Kinouchi O and Caticha N 1992 Optimal generalization in perceptrons *J. Phys. A: Math. Gen.* **25** 6243

[3] Opper M and Haussler D 1991 Generalization performance of Bayes optimal classification algorithm for learning a perceptron *Phys. Rev. Lett.* **66** 2677
[4] Kinzel W and Ruján P 1990 Improving a network generalization ability by selecting examples *Europhys. Lett.* **13** 473
[5] Shallice T 1988 *From Neuropsychology to Mental Structure* (Cambridge: Cambridge University Press)
[6] Opper M, Kinzel W, Kleinz J and Nehl R 1990 On the ability of the optimal perceptron to generalise *J. Phys A: Math. Gen* **23** L581
[7] Changeux J P 1992 *La Recherche 244* **23** 705
[8] Levine D S, Leven S and Prueitt P S 1992 Integration, disintegration, and the frontal lobes *Motivation, Emotion and Goal Direction in Neural Nets* ed D S Levine and S J Leven (Hillsdale, NJ: Erlbaum)
[9] Abbot L F 1990 Learning in neural networks memories *Network* **1** 105
[10] Watkin T L H, Rau A and Biehl M 1993 The statistical mechanics of learning a rule *Rev. Mod. Phys.* in press