On-line learning in parity machines

1996 J. Phys. A: Math. Gen. 29 4859

(http://iopscience.iop.org/0305-4470/29/16/012)

J. Phys. A: Math. Gen. **29** (1996) 4859–4867. Printed in the UK

On-line learning in parity machines

Roberta Simonetti† and Nestor Caticha‡

Instituto de Física, Universidade de São Paulo, Caixa Postal 66318, 05389-970 São Paulo, SP, Brazil

Received 16 April 1996

Abstract. The optimal algorithm for on-line learning in the tree K-parity machine is studied. We introduce a set of recursion relations for the relevant probability distributions, which permit study of the general K case. The generalization error curve is determined and shown to decay to zero for large α as $e_g \approx \alpha^{-1}$ even in the presence of noise. There is no critical noise level. The dynamics of on-line learning is studied analytically near the origin. In the absence of previous knowledge, the learning dynamics has a fixed point at $\alpha = 0$. Previous knowledge is needed in at least $K - 1$ branches for the learning to take place.

1. Introduction

A possible approach to general machine learning theory is to concentrate on the statistical properties of specific artificial neural network simple models. This is done in the hope that it will enable us to identify universal underlying behaviours and features that might transcend the limited scope of the models which can be analytically studied. Among the various possible aspects of learning, we will look at the ability of generalization in on-line supervised [1, 2] learning from a teacher network. Given the limitation to studying the generalization ability, the question that can be addressed is not what is the performance of a predetermined learning algorithm, but rather which algorithm will have the best performance under certain predetermined conditions.

The variational determination of the optimal algorithm, in the sense of generalization, has been done for several particular cases which include the boolean [2, 3] and linear [4] simple perceptrons in the conditions of both on- and off-line learning. The optimized on-line learning has also been addressed for those networks in the presence of noise [5, 6] and in the case of drifting rules [7, 8]. The extension to networks with hidden units such as the committee machine was presented in [9]. As might be expected, several qualitative features are shared by the different optimal algorithms. These include the fact that optimal algorithms are not the same throughout the learning process, but depend on the stage of learning. The common role played by the surprise a new example may bring, as well as the confidence on its correct classification, have been previously stressed. What should not be taken for granted, however, is the fact that several quantitative features are found to be exactly the same for different machines. Among these we mention the $0.88\alpha^{-1}$ decay for the on-line generalization error in the absence of noise for the tree K-committee machine [9], independently of K, where α is the ratio of the number of examples (P) to

† E-mail address: rsimonetti@if.usp.br
‡ E-mail address: nestor@if.usp.br

0305-4470/96/164859+09$19.50

the total number of adjustable connections (N). It has also been verified, from variational arguments, that the optimal on-line energy is closely related to Bayesian ideas [10]. In several different architectures, identical scaling properties of the learning process have also been identified. In this paper we will present the results of such an approach, i.e. variational determination of the optimal on-line learning algorithm, to the case of the parity machine with tree architecture and K hidden units.

On-line learning in the presence of noise in the $K = 2$ parity machine has been recently studied by Kabashima [11] for the so-called 'least action' algorithm (LAA) proposed initially by Mitchison and Durbin. While they [12] study memorization of random examples by doing simulations of an iterated version of the LAA, in [11] the generalization ability for the on-line version of the LAA was studied in the presence of output noise.

The three distinctive features that he found were, first, that in the absence of noise, 'an ability to generalize emerges as the rescaled length of the connection vector J reaches a critical value J_c' (non-zero), as long as the initial conditions are such that $\{\rho_k\}$, the overlaps between student and teacher sub-branches, are $\mathcal{O}(1)$. Second, that the generalization error decays to zero as $\alpha^{-1/3}$ for large α and third, that there is a critical noise level beyond which the generalization error does not decay to zero.

We will show that the optimal learning algorithm, which is in the class of what has been called the 'expected stability' algorithms, gives for those three features a different behaviour. First there is no critical length J_c; that is, if the initial conditions are such that ρ is $\mathcal{O}(1)$, generalization improves as soon as learning starts. Second, for large α the generalization error decays as $0.88\alpha^{-1}$ in the absence of noise, but, third, even in the presence of output noise the error decays asymptotically to zero, as the inverse of α for any noise level. These show that some of the features found in [11] are due to the particular choice of the learning algorithm and disappear for the optimal one.

While the optimized algorithm for the parity machine (OA) resembles the LAA of Mitchison and Durbin in some of its features, the differences are responsible for the vastly improved behaviour. As for other optimized algorithms, it relies, for improved performance, on elements that can be dubbed surprise and confidence. The use of such elements is shared by the LAA, but the OA also uses a measure of the performance. This means that the algorithm is not the same throughout the learning activity but changes with 'time' or in the case of time-dependent drifting rules it adapts to changes.

This paper is organized as follows. In section 2 we present the differential equations that govern on-line learning in the parity machine for an arbitrary algorithm in the presence of noise. A simple variational argument leads to the OA. In section 3 we analyse its performance near the origin, while the asymptotic behaviour is studied in section 4. In section 5 the modulation function is analysed and some final comments included.

2. Learning dynamics

The parity machine with non-overlapping receptive fields has been previously studied in [12, 11]. We describe it here once again just to establish our notation. It consists of K branches, each one a single layer perceptron of N/K inputs. Every branch perceptron is characterized by an N/K dimensional synaptic weight vector $\boldsymbol{J}_k$, and for a given input ($\boldsymbol{S} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_K)$ where $\boldsymbol{S}_j$ is N/K dimensional) the output is $\Sigma = \mathrm{sgn}(\prod_{j=1}^{K} \boldsymbol{J}_j \cdot \boldsymbol{S}_j)$. We look at possibly the simplest supervised learning scenario, where the task to be learned is defined by a teacher network of the same architecture (characterized by synaptic weights $\boldsymbol{B}_k$, $\boldsymbol{B}_k^2 = 1$) and the examples are drawn from a known distribution. For simplicity we will perform the calculations for a uniform distribution, with $\langle S_{ki} \rangle = 0$ and $\langle S_{ki}^2 \rangle = 1$. Certainly

one could look at more structured distributions and also study the case of unsupervised learning as has been done in [13, 14] for the perceptron.

The aim of learning is to determine a set of $\boldsymbol{J}$ vectors that define the student net such that it approximates the teacher network, based solely on information contained in the learning set $\mathcal{L}\{\boldsymbol{S}^\mu, \xi^\mu\}_{\mu=1\dots P}$. We consider the possibility of noise corrupted outputs ξ^μ so that instead of the true output for a given input, we only have access to the ξ^μ variable and to the conditional probability $P(\xi^\mu|\Sigma^\mu)$. Several noise sources can be considered. We look here only at output noise, so that

$$P(\xi^\mu|\Sigma^\mu) = (1-\chi)\delta_{\xi,\Sigma} + \chi\delta_{\xi,-\Sigma} \tag{1}$$

where χ is the probability that a given output Σ^μ of the teacher is flipped.

Several different performance measures can be defined. The training error e_T measures how well the student network can operate over previously seen examples. The generalization error e_G measures how well the student approximates the noiseless teacher and the prediction error e_P measures the probability that for a new independent input vector, the noisy output is not correctly predicted. It is easy to show that

$$e_P = \chi + (1-2\chi)e_G. \tag{2}$$

As is well known, for this feedforward type of architecture, the generalization error can be expressed in terms of order parameters which describe the student–teacher distance. These so-called overlaps ρ_k, can be shown to be self-averaging quantities in the thermodynamic limit, and are given by $\rho_k = \boldsymbol{B}_k \cdot \boldsymbol{J}_k / J_k$, where $J_k = \|\boldsymbol{J}_k\|$. For the uniform distribution of examples, the internal fields $h_k = \boldsymbol{J}_k \cdot \boldsymbol{S}_k / J_k$ and $b_k = \boldsymbol{B}_k \cdot \boldsymbol{S}_k$ in the student and the teacher nets, respectively, are random Gaussian correlated variables with zero mean and correlation ρ_k, distributed according to

$$P(b_k, h_k) = \frac{1}{2\pi\sqrt{1-\rho_k^2}} \exp\left\{\frac{-b_k^2 - h_k^2 + 2h_k b_k \rho_k}{2(1-\rho_k^2)}\right\}. \tag{3}$$

The generalization error of the parity machine

$$e_G^K(\rho_1 \dots \rho_K) = \int \prod_{k=1}^{K} \mathrm{d}b_k \, \mathrm{d}h_k \; P(b_k, h_k)\theta\left(-\prod_{i=1}^{K} b_i h_i\right) \tag{4}$$

satisfies the following recursion relation for tree architectures

$$e_G^{(K)}(\rho_1 \dots \rho_K) = e_G^{(K-1)}(\rho_1 \dots \rho_{K-1}) + e_G^{(1)}(\rho_K) - 2e_G^{(1)}(\rho_K) e_G^{(K-1)}(\rho_1 \dots \rho_{K-1}). \tag{5}$$

This can be also written as

$$e_G^{(K)}(\rho) = \tfrac{1}{2}\left[1 - \prod_{k=1}^{K}(1 - 2e_G^{(1)}(\rho_k))\right] \tag{6}$$

where $e_G^{(1)}(\rho_k) = \pi^{-1}\arccos\rho_k$ is the single branch perceptron generalization error which reduces to the expression given by Opper in [15] in the symmetric case. Under symmetric conditions, where $\rho_k = \rho$ for all k, this shows that the sequence of functions $e_G^{(K)}(\rho)$ converges uniformly, in the limit $K \to \infty$, to $\frac{1}{2}$ for any ρ except at $\rho = 1$. Non-uniform convergence at $\rho = 1$ also occurs in the tree K-committee machine error, but there it signals crossover to a different asymptotic regime, whereas in this case it shows that the parity machine does not learn at all in the infinite K case. At this point we just add that, for this distribution of examples, the generalization error is monotonically decreasing with ρ.

Supervised on-line learning without iterations, that is, by single presentation of each example, is a discrete dynamic stochastic process, where at each time step a random example is used to modify the network's couplings:

$$J_{ki}^{\mu+1} = J_{ki}^{\mu} + \frac{1}{N} F_k^{\mu} S_{ki}^{\mu}. \tag{7}$$

The function F_k^{μ} defines the algorithm and is a measure of the importance that should be given to the example, that is, of the amount of relevant information it carries. We refer to it as the modulation function. Averaging over the possible choices of the example and taking the thermodynamic limit, the infinite set of discrete equations is simplified into $2K$ coupled differential equations for the lengths of the coupling vectors and the order parameters

$$\frac{\mathrm{d}J_k}{\mathrm{d}\alpha} = J_k \left\langle \left(\frac{(F_k^{\mu})^2}{2K J_k^2} + \frac{F_k^{\mu} h_k^{\mu}}{J_k} \right) \right\rangle \tag{8}$$

and

$$\frac{\mathrm{d}\rho_k}{\mathrm{d}\alpha} = \frac{\rho_k}{J_k} \left\langle F_k^{\mu} \left(\frac{b_k^{\mu}}{\rho_k} - h_k^{\mu} - \frac{F_k^{\mu}}{2K J_k} \right) \right\rangle \tag{9}$$

where, dropping the index μ for simplicity,

$$\langle (\cdots) \rangle = \sum_{\xi=\pm 1} \int P^K(b_k, \xi, \{h_i\})\, \mathrm{d}b_k \left[\prod_{i=1}^{K} \mathrm{d}h_i \right] (\cdots). \tag{10}$$

Optimization in the sense of maximal generalization ability can be obtained by maximizing the rate of growth of $\mathrm{d}\rho_k/\mathrm{d}\alpha$. It leads to the following choice of modulation function:

$$F_k = K J_k \left\langle \left(\frac{b_k}{\rho_k} - h_k \right) \right\rangle_{b_k|\xi,\{h_i\}}. \tag{11}$$

This formal expression holds for any choice of the distribution of examples, although it is only the optimal modulation function in the case where the distribution of examples leads to a generalization error which decreases monotonically with ρ. In this work, as it is not unusual in the literature of the statistical mechanics of neural networks, it will be assumed known. The effects of not knowing exactly the distribution will not be considered here. The average is taken over $P^K(b_k|\xi, \{h_i\})$, that is, over the possible internal fields in the teacher network conditioned on the available noise-corrupted-output ξ and the student internal fields. This shows immediately that the optimal algorithm is non-local in the sense that, in order to train one branch, it needs information about the internal state of the other branches. Actually it is simple to show that non-locality is essential since local algorithms in the parity machine never leave the ground. This should be contrasted with tree committee machines, where local algorithms may still achieve an α^{-1} decay [6]. Equation (11) has the same formal structure of the analogous function in the tree committee machine. Only the conditional probability P^K depends on the network's architecture. To calculate $P^K(b_k|\xi, \{h_i\})$ we only need $P^K(\xi|\{h_i\})$, which is simple to obtain in terms of noiseless conditional probabilities $P_0^K(\Sigma|\{h_i\})$, since it is given by

$$P^K(\xi|\{h_i\}) = (1-\chi) P_0^K(\Sigma = \xi|\{h_i\}) + \chi P_0^K(\Sigma = -\xi|\{h_i\}) \tag{12}$$

and for $K > 1$, $P_0^K(\Sigma|\{h_i\})$ satisfies the recursion relation

$$P_0^K(\Sigma|\{h_i\}) = P_0^1(1|h_k) P_0^{K-1}(\Sigma|\{h_i\}_{i\neq k}) + P_0^1(-1|h_k) P_0^{K-1}(-\Sigma|\{h_i\}_{i\neq k}). \tag{13}$$

Only at this point do we need to make explicit the distribution of examples. As we have already mentioned they are to be chosen independently, from a uniform distribution with $\langle S_{ki}\rangle = 0$ and $\langle S_{ki}^2\rangle = 1$, then

$$P_0^1(\Sigma|h_k) = H\left(-\frac{\Sigma h_k}{\lambda_k}\right) \tag{14}$$

where $H(x) = \int_x^\infty \mathrm{D}t$, $\mathrm{D}t = (2\pi)^{-1/2}\exp(-t^2/2)\,\mathrm{d}t$ and $\lambda = \rho^{-1}\sqrt{1-\rho^2}$.

The optimal modulation function can either be obtained from equation (11) or alternatively by using the on-line optimal energy term $E_{opt} = -\lambda^2 \ln P^K(\xi|\{h_i\})$ through the equation $F_k^i = -S_i J(\partial E_{opt}/\partial \hat{J}_k)$. For a discussion of this point in the committee machine see [6]. The result is

$$F_k = \frac{K}{\sqrt{2\pi}} J_k \lambda_k \mathrm{e}^{-h_k^2/2\lambda_k^2}(1-2\chi)\frac{P_0^{K-1}(\xi|\{h_i\}_{i\neq k}) - P_0^{K-1}(-\xi|\{h_i\}_{i\neq k})}{P^K(\xi|\{h_i\})}. \tag{15}$$

Note that

$$P^K(\xi|\{h_i\}) = \chi\sum_{\{-\}}\left[\prod_i^K H\left(-\frac{\epsilon_i h_i}{\lambda_i}\right)\right] + (1-\chi)\sum_{\{+\}}\left[\prod_i^K H\left(-\frac{\epsilon_i h_i}{\lambda_i}\right)\right] \tag{16}$$

where the notation $\{-\}$ and $\{+\}$ means that the sums are taken over all possible configurations of the set $\{\epsilon_i = \pm 1\}$ subject to the constraints $\prod_i \epsilon_i = -\xi$ and $\prod_i \epsilon_i = +\xi$, respectively. The numerator in equation (15) can be written as

$$P_0^{K-1}(\xi|\{h_i\}_{i\neq k}) - P_0^{K-1}(-\xi|\{h_i\}_{i\neq k}) = \xi\prod_{i\neq k}\left[H\left(-\frac{\epsilon_i h_i}{\lambda_i}\right) - H\left(\frac{\epsilon_i h_i}{\lambda_i}\right)\right]. \tag{17}$$

The order parameter differential equation is then

$$\frac{\mathrm{d}\rho_k}{\mathrm{d}\alpha} = \frac{K}{2\pi}\rho_k\lambda_k^2(1-2\chi)^2\int \mathrm{D}h_k \mathrm{e}^{-h_k^2/\lambda_k^2}\frac{\prod_{l\neq k}\mathrm{D}h_l[H(-\frac{h_l}{\lambda_l}) - H(\frac{h_l}{\lambda_l})]^2}{P^K(1|\{h_i\})}. \tag{18}$$

As is the case for other optimized dynamics, it is simple to see that the evolution of the lengths is described by a set of identical equations

$$\frac{1}{J_k}\frac{\mathrm{d}J_k}{\mathrm{d}\alpha} = \frac{1}{\rho_k}\frac{\mathrm{d}\rho_k}{\mathrm{d}\alpha}. \tag{19}$$

We now turn to the analysis of the dynamics in both the small- and large-α limit.

3. Escape from the fixed point

Note the factors $[H(-\frac{h_l}{\lambda_l}) - H(\frac{h_l}{\lambda_l})]$ inside the integrals of equation (18) for $K > 1$, which vanish at $\rho_l = 0$. If the learning process starts from *tabula rasa*, on-line learning will not work. To analyse the behaviour at the early stages of learning we have to go beyond the fully symmetric approximation. The differential equations in the small-α limit are

$$\frac{\mathrm{d}\rho_k}{\mathrm{d}\alpha} = \frac{K}{2}\left(\frac{2}{\pi}\right)^K(1-2\chi)^2\frac{1}{\rho_k}\prod_{l\neq k}\rho_l^2. \tag{20}$$

This shows that $\rho_k = 0$, for any two or more branches, is a fixed point of the dynamics. If at most only one branch has zero initial overlap, but the rest are $\mathcal{O}(1)$, the system will manage to learn on-line. Of course, for $K = 1$, $\rho = 0$ is not a fixed point. For $K > 1$

the system needs previous knowledge in order to learn on-line. For $K = 2$, let the initial conditions be $\rho_1^2(0) = \epsilon_1$ and $\rho_2^2(0) = \epsilon_2$. The solution is

$$\rho_{1,2} = \left[\left(\frac{\epsilon_1 + \epsilon_2}{2} \right) \mathrm{e}^{\Lambda_2 \alpha} \pm \left(\frac{\epsilon_1 - \epsilon_2}{2} \right) \mathrm{e}^{-\Lambda_2 \alpha} \right]^{\frac{1}{2}}$$

with $\Lambda_K = K(2/\pi)^K (1 - 2\chi)^2$. For general K, suppose there is previous knowledge for all branches. Call $\rho_k^2(0) = \epsilon_k \neq 0$. Defining $E_k = \prod_{i \neq k} \epsilon_i$, the evolution of the overlaps is given by

$$\rho_k = \left[\epsilon_k + E_k \Lambda_K \alpha + \frac{(\Lambda_K \alpha)^2}{2} \epsilon_k E_k^2 \sum_{i \neq k} \frac{1}{\epsilon_i^2} + \frac{E_k^3}{6} (\Lambda_K \alpha)^3 \left(2\epsilon_k^2 \sum_{i \neq j \neq k} \frac{1}{\epsilon_i^2 \epsilon_j^2} + \sum_i \frac{1}{\epsilon_i^2} \right) + \mathcal{O}[(\Lambda_K \alpha)^4] \right]^{\frac{1}{2}}. \tag{21}$$

Physically, the fixed point of the on-line dynamics is related to the retarded learning phenomenon which appears even in the optimal off-line procedure [16]. Off-line learning only works if a threshold of $\alpha_c N$ examples is surpassed. This means that for on-line learning to work, some previous knowledge is needed in the form of a non-zero initial condition. For finite N, a random choice of initial couplings would suffice to furnish the 'previous knowledge' since it will give $\rho(0) \approx \mathcal{O}(N^{-1/2})$. However, the differential equations are correct to $\mathcal{O}(1)$ only. In order to test the approximations that lead to (18) the initial conditions have to be such that $\mathcal{O}(N^{-1/2}) \ll \rho(0) \ll 1$.

Figure 1 shows the results of numerically integrating the differential equations for the special case $K = 2$ and of simulations of the learning process.

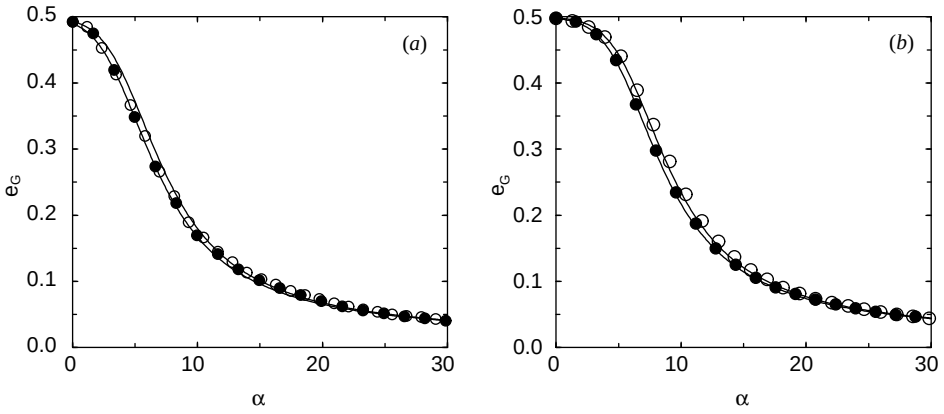

Figure 1. Generalization error obtained from numerical integration of $\mathrm{d}\rho/\mathrm{d}\alpha$ (full curves), simulation using the overlap (●) and J (○) for (*a*) $N = 302$ and $0.18 < \rho < 0.22$, (*b*) $N = 1002$ and $0.09 < \rho < 0.11$.

For each set of data, a window was defined by choosing two slightly different limiting initial conditions. The two lines of each set are the results of numerically integrating the differential equations (18) for those initial conditions. In the simulation, the student networks' initial couplings were chosen randomly, those whose initial conditions were inside the predetermined ρ-window were kept. The lengths of the coupling vectors were normalized according to equation (19). Two sets of simulations were done, one with the

modulation function which uses the value of ρ and another with J in its place. The simulations and the numerical integrations are seen to agree quite well.

The origin of the retardation is due to the symmetry presented by this machine [17]. If in an even number of branches $\boldsymbol{J}$ is transformed into $-\boldsymbol{J}$, the output is unchanged. In the fully connected smooth committee machines the existence of plateaux in the learning dynamics is also due to symmetries between the different branches. In the escape from symmetric plateaux in both cases, finite size effects are most important and the usefulness of this type of description, by a set of differential equations, is questionable. Outside its immediate vicinity, however, it is an excellent tool for understanding the learning dynamics.

The relation between the retarded learning effect in off-line learning and the repulsive fixed point at zero overlap in the dynamics of on-line learning seems to be general. It has also been detected in the unsupervised learning of a structured distribution of examples by a perceptron [18, 19].

4. Asymptotic behaviour

In the limit $\alpha \to \infty$, the overlaps $\rho \to 1$ and the generalization error can be written as $e_G^{(K)}(\rho) = (K/\pi)\arccos\rho$. Under these conditions the ρ-differential equation is

$$\frac{\mathrm{d}\rho}{\mathrm{d}\alpha} = \frac{K}{2\pi}\frac{(1-\rho^2)^{\frac{3}{2}}}{\rho^2} I(\chi)$$

where

$$I(\chi) = (1-2\chi)^2 \int \mathrm{D}x \frac{\mathrm{e}^{-x^2/2}}{\hat{H}(x)}$$

and for convenience we define $\hat{H}(x) = \chi + (1-2\chi)H(x)$.

Define $\alpha_p = K\alpha$, which measures the number of examples divided by the number of couplings in one single perceptron branch. In terms of this scale, every perceptron branch learns at the same rate independently of K. If the single layer perceptron error decays asymptotically as C/α_p , then the generalization error of the K-parity machine will be

$$e_G^{(K)}(\rho) = K\frac{C}{\alpha_p} = \frac{C}{\alpha} \tag{22}$$

independently of K. So the K-parity machine will learn with a much higher error than each individual branch; actually it is the sum of the errors. However, the correct scale to measure the decay is α and not α_p. This leads to cancellation of the K factors leaving the nominal error independent of K. This most interesting feature is shared by the tree K-committee machine and will occur for any optimized algorithm [20] for treelike architectures. It is easy to show that $C = 2/I(\chi)$. This is the same coefficient found in [5, 6] for the simple perceptron. After all, this is a K-independent result and the perceptron is a $K = 1$ parity machine. In the absence of noise this leads to the ubiquitous $0.88\alpha^{-1}$ decay which is exactly twice the Bayes optimal. It is interesting that even in the presence of noise the error decays with the optimal power decay but with a larger coefficient, since $I(\chi) \approx \sqrt{2}(1-2\chi)^2$ for large χ (near $\frac{1}{2}$). This shows that there is no critical noise level. To implement the optimal algorithm we need to know the actual noise level. This is certainly hardly the case. Noise level evaluation by on-line monitoring and the robustness to noise level determination, both in the case of output as well as weight noise will be the topic of a forthcoming work. At this point we can just claim that these results represent mean-field lower bounds for the generalization errors.

5. The modulation function

The modulation functions have three relevant properties that we want to note. First of all, the modulation functions, and thus the learning algorithm, depend explicitly on ρ. From equation (19), it can be seen that the values of the synaptic vectors length can be used instead. We want to stress the fact that this dependence implies that the optimal algorithm is not the same throughout the learning process. The algorithm that should be used by a student in an early stage is not the same as one that a rather experienced one uses. The evolution of the modulation functions scales with ρ in a simple manner and figure 2 shows, for $K = 3$, the rescaled modulation functions $F_k/(\lambda J_k)$ in terms of the rescaled internal field $y_1 = h_1/\lambda$, for a particular choice of $y_{2,3} = h_{2,3}/\lambda$.

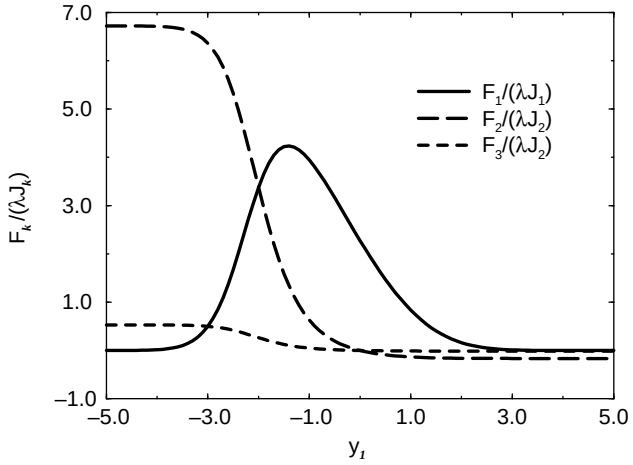

Figure 2. rescaled modulation function in terms of the rescaled internal field in the first branch y_1 for a $K = 3$ parity machine.

The second feature is related to what can be called the value of information or the importance that a new example has. For small α, whether an example is correctly classified or not is not very relevant. In both cases the same order of magnitude synaptic changes occur. However, later on, the fact that an example is wrongly predicted implies that at least one modulation function is appreciably large. Correctly classified inputs lead to small changes.

If, for a given input, the internal fields h_k are large, then the student can be said to be very confident in the probability of correctly predicting its output. A surprising result, meaning a wrong output of an easy example, leads to a high modulation on at least one branch. However, a very confident prediction by one branch that leads to an overall wrong output may be explained by some other factor, leading to an attenuation of the surprise factor. The presence of output noise, which we have studied here (see also [5, 6]), or a less confident branch may be blamed for wrong predictions. This leads to the third characteristic we want to stress, which is blame attribution crossover. Figure 2 shows a noiseless situation in which the teacher output is $\Sigma = 1$, and $y_2 = -2$, $y_3 = -3$. Therefore, when $y_1 > 0$ the student is giving the correct output. When y_1 starts decreasing, F_1 starts increasing, since for $y_1 < 0$ the prediction is wrong. For $y_1 > -2$, we are in a situation where the least confident branch, i.e. number 1, gets the highest modulation. As y_1 goes below -2, blame attribution for the wrong output crosses over from the first to the second branch. The third

branch starts getting a fraction of the blame as y_1 gets larger, in module, than y_3, but since it is always more confident than the second branch, it never gets a high modulation.

This discussion certainly agrees with the rationale behind the least action algorithm of Mitchison and Durbin as implemented by Kabashima for the $K = 2$ parity machine. But it is important to notice, along with their similarities, their main differences. First, the algorithm changes with 'time'. It depends on ρ as discussed above, which is in principle an unavailable quantity. We have just suggested, among the many possible solutions to this problem, substituting for ρ_k the value of J_k. But since ρ_k can be estimated on-line [8], we have a truly adaptive algorithm which can adjust to changes in the environment or drifting teachers. Second, crossings are not sharply defined. There are regions, e.g. $y_1 \approx -2$ in figure 2, where all branches are getting contributions. The fact that no information is being discarded leads to the α^{-1} decay for the generalization error, as opposed to the $\alpha^{-1/3}$. It also makes the error decay to zero, with the same exponent in the presence of any noise level below $\chi = \frac{1}{2}$.

The main criticism that can be raised about this kind of approach is its dependence on some unknown quantities. These are the distribution of examples, the overlaps and the noise level. All of these can be, to a large extent, estimated on-line, as the learning procedure takes place. The characteristics of the resulting algorithms will be the subject of future work.

Acknowledgments

The authors thank M Copelli and O Kinouchi for helpful discussions. RS was partially supported by CAPES and CNPq. NC was partially supported by CNPq.

References

[1] Kinzel W and Ruján P 1990 *Europhys. Lett.* **13** 473
[2] Kinouchi O and Caticha N 1992 *J. Phys. A: Math. Gen.* **25** 6243
[3] Biehl M and Riegler P 1994 *Europhys. Lett.* **28** 525
[4] Kinouchi O and Caticha N 1995 *Phys. Rev.* E **52** 2878
[5] Biehl M, Riegler P and Stechert M 1995 *Phys. Rev.* E **52** R4624
[6] Copelli M, Kinouchi O and Caticha N *Phys. Rev.* E to appear
[7] Biehl M and Schwarze H 1992 *Europhys. Lett.* **20** 733
[8] Kinouchi O and Caticha N 1993 *J. Phys. A: Math. Gen.* **26** 6161
[9] Copelli M and Caticha N 1995 *J. Phys. A: Math. Gen.* **28** 1615
[10] Kinouchi O and Caticha N 1995 A learning algorithm that gives the Bayes generalization limit for perceptrons *Preprint* Universidade de São Paulo *Phys. Rev.* E to appear
[11] Kabashima Y 1994 *J. Phys. A: Math. Gen.* **27** 1917
[12] Mitchison G J and Durbin R M 1989 *Biol. Cybern.* **60** 345
[13] Riegler P, Biehl M, Solla S A and Marangi C 1996 On-line learning from clustered input examples *Proc. 7th Italian Workshop on Neural Networks* (Singapore: World Scientific)
Marangi C, Solla S A, Biehl M and Riegler P 1995 Off-line learning from clustered input examples *Proc. 7th Italian Workshop on Neural Networks* (Singapore: World Scientific)
[14] Van den Broeck C and Reimann P 1996 *Phys. Rev. Lett.* **76** 2188
[15] Opper M 1995 *Phys. Rev.* E **51** 3613
[16] Opper M 1994 *Phys. Rev. Lett.* **72** 2113
[17] Hansel D, Mato G and Meunier C 1992 *Europhys. Lett.* **20** 471
[18] Biehl M and Mietzner 1993 *Europhys. Lett.* **24** 421
[19] Biehl M 1994 *Europhys. Lett.* **25** 391
[20] Copelli M Private communication