J. Phys. A: Math. Gen. **30** (1997) L599–L605. Printed in the UK PII: S0305-4470(97)85052-4

LETTER TO THE EDITOR

Functional optimization of online algorithms in multilayer neural networks

R Vicente† and N Caticha‡

Instituto de Física, Universidade de São Paulo, CP66318, CEP 05315-970, São Paulo, SP Brazil

Received 13 June 1997

Abstract. We study the online dynamics of learning in fully connected soft committee machines in the student–teacher scenario. The locally optimal modulation function, which determines the learning algorithm, is obtained from a variational argument in such a manner as to maximize the average generalization error decay per example. Simulations results for the resulting algorithm are presented for a few cases. The symmetric phase *plateaux* are found to be vastly reduced in comparison to those found when online backpropagation algorithms are used. A discussion of the implementation of these ideas as practical algorithms is given.

Learning how learning occurs in artificial systems has caught the attention of the statistical mechanics community in the last decade (see, for example, [1–3]). This interest was ignited for several reasons, among them, the invention of efficient learning-from-examples methods such as backpropagation, that permit learning in computationally complex machines, to the realization that ideas from disordered systems, in particular spin glasses, could be applied to the study of attractor as well as feedforward neural networks and to the generalized interest in complex systems with rugged energy landscapes. The Statistical Mechanics approach has almost invariantly dealt with the thermodynamic limit and has benefitted from the powerful techniques used to calculate the averages over the disorder introduced by the random nature of the examples.

Among several possible approaches to machine learning, online learning [4] has been the subject of an intense research effort due to several factors. In this scheme, examples are used only once, thereby avoiding the need for expensive memory resources, typical of offline methods. This, however, does not translate necessarily into poor performance since efficient methods can be devised that have performance comparable to the memory based ones. Furthermore, learning sequentially from single examples has a greater biological flavour than offline processing. While efficiency, computational economy and biological relevance may be the most relevant factors, the theoretical possibility of rather complete analytical studies has also played an important role. If each one of these factors is, by itself, sufficiently important to make online learning an attractive scheme, together they combine to give a most compelling argument for its thorough study.

In this letter we present results of the optimization of online supervised learning in a model consisting of a fully connected multilayer feedforward neural network, in what has

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

0305-4470/97/170599+07$19.50

become known as the student–teacher scenario. The type of result we present here brings together two separate lines of research that have been recently pursued by several groups.

The study of online backpropagation as put forward by Biehl and Schwarze [5] and later developed in [6,7] has permitted the analytical understanding of several properties of the dynamics of the learning process. The most striking feature being the existence of learning *plateaux* or symmetric phases which signal learning stages where the information available to the student and the form in which it is used do not permit breaking the permutation symmetry among the hidden nodes. Further learning eventually permits the escape from the neighbourhood of these repulsive symmetric fixed points into the broken symmetry, specialized phase. The onset of specialization and different methods to hasten it have been dealt with by several authors [8–11].

The second line of research from which we draw is the variational study of locally optimal online learning. This program deals with the determination of lower bounds for the generalization errors in different models in controlled learning scenarios. The constructive nature of the variational approach has permitted finding update rules that lead to student networks with the optimal generalization performance. The relation of this approach to Bayesian methods has been discussed in [11] and in [12].

The variational method has been previously applied to machines with no internal units [13–16] or with hidden units but non-overlapping receptive fields (RF) [17,18] and also in the case of unsupervised learning [18]. We will introduce the variational method for feedforward machines with overlapping RF. The differences stem from the fact that while in the former case the generalization error is a monotonic decreasing function of the order parameters (student–teacher overlaps), in the latter, the monotonicity is lost, due to the appearance of crossed overlaps.

The main results here presented are the analysis of the locally optimized online learning dynamics of a soft committee. We present results for over-realizable and realizable cases. The striking reduction or complete elimination of the *plateaux* in the learning curves witnesses the great improvement achievable by concentrating in extracting the largest possible amount of information from each example. Rapid escape from the plateaux can be attributed to a fluctuation enhancing mechanism that stimulates permutational symmetry breaking.

The aim of learning is to obtain a set of student weights J_{ik} where $i(=1,\ldots,N)$ indexes input layer units and $k(=1,\ldots,K)$ hidden nodes, in such a manner that the student implements as closely as possible the map represented by the teacher network defined by a set of weights B_{in}, where $i(=1,\ldots,N)$ labels the input layer unit and $n(=1,\ldots,M)$ the hidden node. We use $n,m,\ldots$ to label teacher branches and $j,k,\ldots$ for the student branches. Call $\boldsymbol{B}_n=(B_{1n},B_{2n},\ldots,B_{Nn})$, $\boldsymbol{J}_k=(J_{1k},J_{2k},\ldots,J_{Nk})$ the weight branch vectors and B_n and J_k their respective lengths. We define as usual the order parameters $R_{kn}=\boldsymbol{J}_k\cdot\boldsymbol{B}_n$, $Q_{ij}=\boldsymbol{J}_i\cdot\boldsymbol{J}_j$ and $M_{nm}=\boldsymbol{B}_n\cdot\boldsymbol{B}_m$ which will be taken $M_{nm}=\delta_{nm}$, for simplicity.

At each time step μ, an example $\boldsymbol{S}^\mu$ is drawn from a known distribution $P(S)$. We call Σ_B^μ and Σ_J^μ the teacher and student outputs respectively. The internal fields are denoted by $y_n^\mu=\boldsymbol{B}_n\cdot\boldsymbol{S}^\mu$ and $x_k^\mu=\boldsymbol{J}_k\cdot\boldsymbol{S}^\mu$. The available information is used in updating the student weights J_{ik},

$$J_{ik}(\mu+1)=J_{ik}(\mu)+\frac{F_k}{N}S_i^\mu. \tag{1}$$

This is not the most general update possible since a decay term, useful in controlling the length of $\boldsymbol{J}_k$ can be used, we however will not pursue this direction here. The central quantity in this theoretical approach is the set of modulation functions $\boldsymbol{F}=(F_1,F_2,\ldots,F_K)$. The

following analysis will be done in the thermodynamic limit. For any transfer function, the evolution of the order parameters is given by a set of $(K^2 + K)/2 + KM$ first-order differential equations. For fully connected architectures we have:

$$\frac{\mathrm{d}R_{in}}{\mathrm{d}\alpha} = \langle y_n F_i \rangle \qquad \frac{\mathrm{d}Q_{ij}}{\mathrm{d}\alpha} = \langle x_i F_j + x_j F_i + F_i F_j \rangle \tag{2}$$

where as usual, $\alpha = \mu/N$ measures the learning time. We now proceed, first to obtain the best $\boldsymbol{F}$, from a generalization point of view, and then to analyse the dynamical consequences that such a choice will have.

A point of technical importance, which in no way restricts the validity of the general properties of the results here discussed, concerns the choice of an error function for the sigmoidal transfer function g of the internal units and a linear transfer function for the output unit, following [5], since it permits better analytical tractability. Thus $\Sigma_B^\mu = \sum_{n=1,\dots,M} \mathrm{erf}(y_n^\mu/\sqrt{2})$ and $\Sigma_J^\mu = \sum_{k=1,\dots,K} \mathrm{erf}(x_k^\mu/\sqrt{2})$.

For a fixed teacher, the student network will have a generalization error $e_g(\boldsymbol{J}_k) = \langle \frac{1}{2}(\Sigma_B^\mu - \Sigma_J^\mu)^2 \rangle_S$. In the thermodynamic limit, for a uniform distribution of examples, the generalization error can be written as a function of the order parameters:

$$e_g = \frac{1}{\pi} \sum_{i,j} \sin^{(-1)} \left(\frac{Q_{ij}}{\sqrt{1 + Q_{ii}}\sqrt{1 + Q_{jj}}} \right) - \frac{2}{\pi} \sum_{i,n} \sin^{(-1)} \left(\frac{R_{in}}{\sqrt{2(1 + Q_{ii})}} \right) + \frac{M}{6}. \tag{3}$$

Local optimization is obtained by maximizing the average generalization error decay for each example in a given state (R_{kn}, Q_{ij}). We thus look, following [13], at the extremes of the functional $\dot{e}_g[\boldsymbol{F}] = \mathrm{d}e_g[\boldsymbol{F}]/\mathrm{d}\alpha$ that is, the modulation function $\boldsymbol{F}$ which satisfies $\delta \dot{e}_g/\delta F_k = 0$. The solution has the general form

$$\boldsymbol{F} = \mathbf{H}^{-1} \boldsymbol{G} \langle \boldsymbol{y} \rangle_{\mathcal{H}|\mathcal{V}} - \boldsymbol{x} \tag{4}$$

where $\mathbf{H}$, the functional Hessian matrix and $\boldsymbol{G}$ are defined as $H_{ij} = \delta^2 \dot{e}_g/\delta F_i \delta F_j$ and $G_{kn} = -\partial e_g/\partial R_{kn}$ and the conditional expectation is taken with respect to the assumed examples' probability distribution $P(S)$. The symbols $\mathcal{H}$ and $\mathcal{V}$ stand for the set of hidden or visible information. It is interesting to note that (4) holds for any choice of transfer function or examples' distribution. For the particular case of examples drawn independently from a uniform spherical distribution, we have to solve integrals of the form:

$$\int \prod_n \mathrm{d}y_n \, P_{\mathbf{C}}(\boldsymbol{x}, \boldsymbol{y}) y_m^\epsilon \delta \left(\Sigma_B - \sum_n \mathrm{erf}(y_n^\mu/\sqrt{2}) \right) \tag{5}$$

where $\epsilon = 0, 1$ and $P_{\mathbf{C}}(\boldsymbol{x}, \boldsymbol{y})$ is a $(K + M)$ multivariate Gaussian with correlation matrix

$$\mathbf{C} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{R}^t & \boldsymbol{M} \end{pmatrix}. \tag{6}$$

We now present results obtained by simulating an $N = 5000$ system for the cases $K = 2$, $M = 1$ and $K = M = 2$. Further details will be presented elsewhere [19]. In figures 1 and 2 we show the learning curves for these two cases. Backpropagation results, for the same initial conditions, are included for comparison. Figure 3 shows the evolution of Q_{ik} for the $K = M = 2$ case and suggests that the mechanism used to enhance fluctuations and break the permutation symmetry is to increase synaptic vector norms and stimulate anti-correlated weights.

Whether this solution of the variational problem leads to a maximum generalization or not will be governed by the functional Hessian matrix $\mathbf{H}$. Note that the dependence of the dynamics on the modulation function is only second order, therefore $\mathbf{H}$ is a function of the order parameters and not explicitly of the particular algorithm that led to that state of affairs.

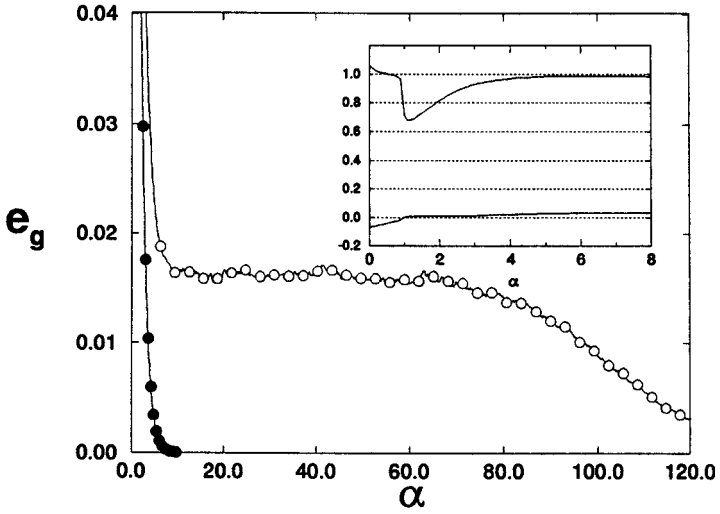

Figure 1. Generalization error learning curves for the $K = 2$, $M = 1$ case obtained by simulating a system of $N = 5000$ with random initial conditions $Q_{11} \in [0, 0.5]$, $Q_{22} \in [0, 1E - 6]$ and $Q_{12} \simeq 0$. Full circles, optimized algorithm; open circles, conventional backpropagation with learning rate $\eta = 1.5$. Inset, eigenvalues of the Hessian **H**. There is a transient where the smallest eigenvalue is negative, it then crosses rapidly into positive values.

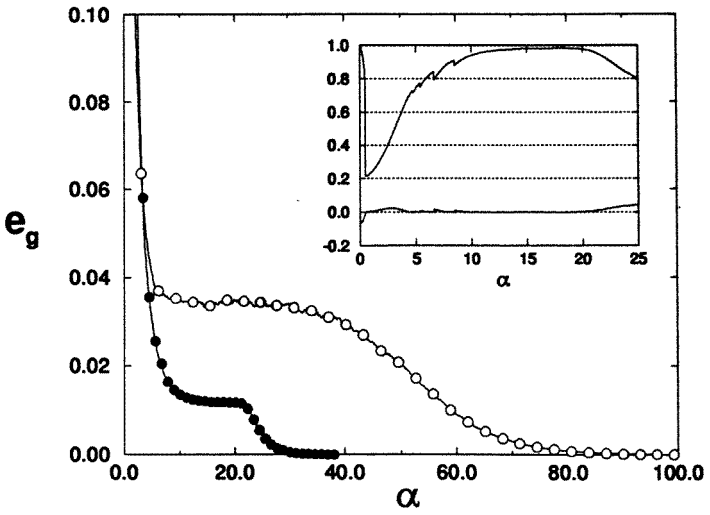

Figure 2. Same as figure 1 but for the $K = M = 2$. Inset, eigenvalues of the Hessian **H**. Note that the smallest eigenvalue stays very close to zero in the *plateau*.

A negative eigenvalue of **H** at a given point in the space of order parameters implies that at that point an optimal algorithm cannot be analytically found.

The evolution of the eigenvalues for both cases is shown as insets in figures 1 and 2. In the space of algorithms, for both cases, at the beginning of the learning process these modulation functions represent saddle points rather than maxima. For the case $K = 2$, $M = 1$ this can be explained as follows. The best generalization would be obtained by using a *correct* architecture, $K = M = 1$, thus the optimal strategy is to trim the student into

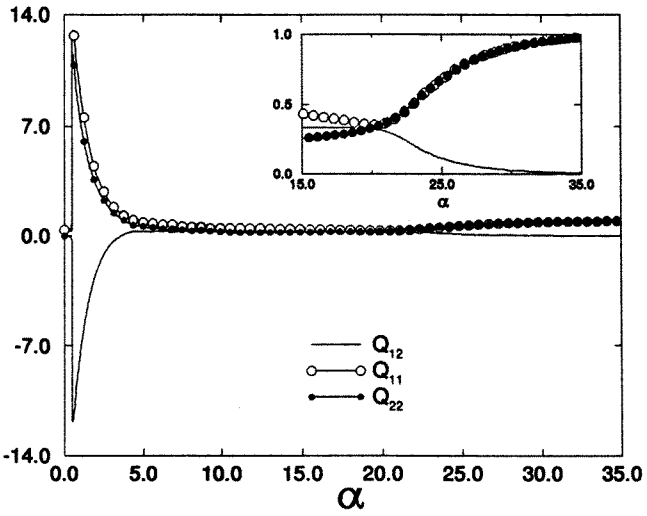

Figure 3. Evolution of the overlaps. Note the anticorrelation that builds up during the transient. Inset, details of the escape from the *plateau*.

the correct architecture and then proceed with the optimized nonlinear perceptron algorithm which could then be obtained by the above variational method. This kind of modulation function cannot be obtained analytically by searching for zero derivatives in the space of algorithms of the $K = 2$ student. The solution found by our method does cut out one of the branches around $\alpha \approx 1$ and turns itself into an effectively $K = 1$ machine quite rapidly, avoiding the long *plateau* of the backpropagation algorithm.

The explanation for the initially negative eigenvalue of $\mathbf{H}$ in the $K = M = 2$ case is not different. The optimal strategy is within the space of students with a $K = 2$ architecture and asymmetric initial conditions, and thus it will not be found by the variational approach. Before there is any information to hint that the permutation symmetry should be broken, it is more efficient locally to learn with a $K = 1$ machine (with an output multiplied by 2). This is however not true after a while, since thus it will never escape the *plateau*. Since the escape is achieved by amplification of symmetry breaking fluctuations, learning initially with a nonlinear perceptron cannot be globally efficient, for it totally suppresses the desired effect of fluctuations.

A null eigenvalue of $\mathbf{H}$ indicates the existence of a class of algorithms with identical performance to that of (4), this can be interpreted as a kind of *functional robustness*. An example of this appears for the case $K = M = 2$, where the smallest eigenvalue stays very close to zero in the *plateau* state (see figure 3). The significance of this is that due to functional robustness, the exact determination of the modulation function is not very critical for learning and eventually escaping the *plateau*.

Although our method permits the locally optimal extraction of information from an example, it does not assure that the system will follow the best global trajectory in the space of order parameters. The global functional optimization has been recently addressed in [20]. They have shown the equivalence between local and global optimizations for the boolean perceptron and the better performance of the global approach in $K = 3$, $M = 1$ case. A thorough investigation on how global and local optimizations are related is an important issue and remains to be done.

The effects of finite size N have not been systematically investigated and therefore the advantages of these methods, if any, over conventional algorithms remains to be proved. Nevertheless, learning is easier in smaller networks and a straightforward use of the modulation function in regimes where the central limit theorem cannot yet be used leads to a successful learning prescription as can be seen from simulating learning for the rather small network with $N = 15$, $K = M = 2$ [19].

The main difficulties of using this approach to construct practical algorithms concern the assumed knowledge of several unavailable quantities. First of all the examples' probability distribution is needed in order to calculate the integrals in equation (2). Then, the resulting modulation function depends on unknown order parameters, such as R_{in}, and worst, these order parameters are only self-averaging in the thermodynamic limit. We first discuss rapidly the first two points. Optimality is hard to define, several different possible criteria lead to different results. Also, given a definition, such as the one we use here of maximizing generalization, the optimal prescription will depend on the amount of available information and on the environment where learning takes place. Although we do not attempt to solve these problems here, a short digression is in order. A parametric representation of $P(S, \Sigma_B) \approx P_{\boldsymbol{w}}(S, \Sigma_B)$ permits introducing an extra set of p differential equations for the online estimation of the distribution parameters $\boldsymbol{w} = (w_1, w_2, \ldots, w_p)$. Also the order parameters can be analogously estimated online, as has been done in [21], even in the case of time dependent or drifting rules.

How robust these 'optimal' algorithms are in the absence or misestimation of this information, as well as its response to learning in noisy environments remains to be seen. The last issue has been addressed recently in [22] for boolean machines. They found a large robustness to noise-level-misestimation, as well as efficient online noise level estimators which manage to steer the dynamics into an efficient learning phase.

These comments about the need for extra knowledge to implement these methods as algorithms can be seen as drawbacks for the variational program. We rather think of them as calling our attention to the further work that has to be done in order to obtain efficient adaptive practical algorithms, and pointing out directions in which these objectives can be reached. Whatever point of view is chosen, the validity of these results and their relation to improving the generalization ability remains.

The authors thank O Kinouchi and M Copelli for several useful discussions and M Biehl, P Riegler, S Solla and C Van den Broeck for discussions during the early stages of this work. This work received partial financial support from CNPq and Finep (RECOPE).

References

[1] Seung H, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[2] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[3] Opper M and Kinzel W 1996 *Statistical Mechanics of Generalization, Physics of Neural Networks* ed J L van Hemmen, E Domany and K Schulten (Berlin: Springer)
[4] Amari S 1967 *IEEE Trans.* **EC-16** 299
[5] Biehl M and Schwarze H 1995 *J. Phys. A: Math. Gen.* **28** 643
[6] Saad D and Solla S A 1995 *Phys. Rev.* E **52** 4225
[7] Biehl M and Riegler P 1994 *Europhys. Lett.* **28** 525
[8] Biehl M, Riegler P and Wöhler C 1996 *J. Phys. A: Math. Gen.* **29** 4769
[9] West A H L and Saad D 1997 Online learning with adaptive backpropagation in two-layer networks, submitted
[10] Amari S 1997 Natural gradient works efficiently in learning *Preprint*
[11] Kinouchi O and Caticha N 1996 *Phys. Rev.* E **54** R54
[12] Opper M 1996 *Phys. Rev. Lett.* **77** 4671

[13] Kinouchi O and Caticha N 1992 *J. Phys. A: Math. Gen.* **25** 6243
[14] Biehl M and Schwarze H 1993 *Europhys. Lett.* **20** 733
[15] Biehl M and Riegler P 1994 *Europhys. Lett.* **28** 525
[16] Copelli M and Caticha N 1995 *J. Phys. A: Math. Gen.* **28** 1615
[17] Simonetti R and Caticha N 1996 *J. Phys. A: Math. Gen.* **29** 4859
[18] Van den Broeck C and Reimann 1996 *Phys. Rev. Lett.* **76** 2188
[19] Vicente R and Caticha N 1997 in preparation
[20] Saad D and Rattray M 1997 Globally optimal online learning rules in multilayer neural networks *Europhys. Lett.* submitted
[21] Kinouchi O and Caticha N 1993 *J. Phys. A: Math. Gen.* **26** 6161
[22] Copelli M, Eichhorn R, Kinouchi O, Biehl M, Simonetti R, Riegler P and Caticha N 1997 *Europhys. Lett.* **37** 427