Globally Optimal Learning Rates for Multilayer Neural Networks

David Saad and Magnus Rattray

Dept. of Comp. Sci. & Applied Math.

Aston University, Birmingham B4 7ET, UK.

Abstract

A method for calculating the globally optimal learning rate in on-line gradient-descent training of multilayer neural networks is presented. The method is based on a variational approach which maximizes the decrease in generalization error over a given time frame. We demonstrate the method by computing optimal learning rates in typical learning scenarios. The method can also be employed when different learning rates are allowed for different parameter vectors as well as to determine the relevance of related training algorithms based on modifications to the basic gradient descent rule.

1 Introduction

Layered feed-forward neural networks are of interest to both theoreticians and practitioners alike for their capability to emulate continuous mappings to any degree of accuracy, if a sufficient number of sigmoidal hidden nodes is provided. In particular, it has been shown that a two layer system with sigmoidal hidden units and a linear output unit is a universal approximator[Cybenko 1989].

In a typical training scenario, a given desired map, of the form $\tilde{f} : X \to Y$, where X and Y represent the input and output space respectively, is instanced by a set of examples. These examples are then used to optimize the network's internal parameters $\{\mathbf{J}\}$ with respect to some measure of the discrepancy between the function implemented by the network $f_{\mathbf{J}}$ and $\tilde{f}$.

The optimization process, which is termed training, may be carried out by a variety of methods. One of the leading techniques in neural networks training, especially for large systems, is *on-line learning*, whereby the network's parameters are modified after each presentation of a training example. On-line learning has been successfully applied to many real-world problems and is arguably the most commonly used neural networks training technique. Many variations of the basic algorithm have been suggested over the years (for a review, see [Bishop 1995]). We will concentrate in this work on basic gradient descent back-propagation, where the dynamics is in the direction of the gradient and the step size is controlled by a single parameter - the learning rate. We will also briefly consider the case of site dependent learning rates, which is a simple extension of the single learning rate case.

Recent studies [Biehl and Schwarze 1995, Saad and Solla 1995a, Saad and Solla 1995b], offer a framework for analysing on-line learning scenarios. We will employ this framework for calculating, within given time windows, globally optimal learning rates. It is worthwhile to point out that the same method can be generalized to accommodate other parameters and learning rules for both smooth and discrete architectures. It can also be employed to assess the usefulness of various modifications to the basic gradient descent rule, or to compare the efficiency of different training techniques, by examining the optimal values assigned to the related coefficients.

2 The general framework

Consider a map from an N-dimensional input space $\boldsymbol{\xi} \in \Re^N$ onto a scalar $\zeta \in \Re$, realized through a map $\sigma(\mathbf{J},\boldsymbol{\xi}) = \sum_{i=1}^{K} g\ (\mathbf{J}_i \cdot \boldsymbol{\xi})$, which can be viewed as a soft committee machine [Biehl and Schwarze 1995], where $g \equiv \mathrm{erf}(x/\sqrt{2})$, is the activation function of the hidden units, $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \leq i \leq K}$ is the set of input-to-hidden adaptive weights for the K hidden nodes and the hidden-to-output weights are set to 1. The activation of hidden node i under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is denoted $x_i^\mu = \mathbf{J}_i \cdot \boldsymbol{\xi}^\mu$. This general configuration represents most of the properties of general multilayer networks and can easily be extended to accommodate adaptive hidden-to-output weights [Riegler and Biehl 1995].

Training examples are of the form $(\boldsymbol{\xi}^\mu, \zeta^\mu)$ where $\mu = 1, 2, \ldots, P$. The components of the independently drawn input vectors $\boldsymbol{\xi}^\mu$ are uncorrelated random variables with zero mean and unit variance. The corresponding output ζ^μ is given by a deterministic teacher of a similar configuration to the student except for a possible difference in the number M of hidden units and is of the form $\zeta^\mu = \sum_{n=1}^{M} g\ (\mathbf{B}_n \cdot \boldsymbol{\xi}^\mu)$, where $\mathbf{B} \equiv \{\mathbf{B}_n\}_{1 \leq n \leq M}$ is the set of input-to-hidden adaptive weights for teacher hidden nodes. The activation of hidden node n under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is denoted $y_n^\mu = \mathbf{B}_n \cdot \boldsymbol{\xi}^\mu$. We will use indices $i, j, k, l \ldots$ to refer to units in the student network and $n, m, \ldots$ for units in the teacher network.

The error made by a student with weights $\mathbf{J}$ on a given input $\boldsymbol{\xi}$ is given by the quadratic deviation

$$\epsilon(\mathbf{J},\boldsymbol{\xi}) \equiv \frac{1}{2} \left[\, \sigma(\mathbf{J},\boldsymbol{\xi}) - \zeta \,\right]^2 = \frac{1}{2} \left[\sum_{i=1}^{K} g(x_i) - \sum_{n=1}^{M} g(y_n) \right]^2 \quad . \tag{1}$$

This error is then used to define the training dynamics via a gradient descent rule for the update of student weights $\mathbf{J}_i^{\mu+1} = \mathbf{J}_i^\mu + \frac{\eta}{N}\ \delta_i^\mu\ \boldsymbol{\xi}^\mu$, where $\delta_i^\mu \equiv g'(x_i^\mu) \left[\sum_{n=1}^{M} g(y_n^\mu) - \sum_{j=1}^{K} g(x_j^\mu)\right]$ and the learning rate η has been scaled with the input size N. Performance on a typical input defines the generalization error $\epsilon_g(\mathbf{J}) \equiv < \epsilon(\mathbf{J}, \boldsymbol{\xi}) >_{\{\xi\}}$ through an average over all possible input vectors $\boldsymbol{\xi}$.

Expressions for the generalization error as well as for the learning dynamics have been obtained [Saad and Solla 1995a] in the thermodynamic limit ($N \to \infty$) and can be represented by a set of macroscopic variables of the form: $\mathbf{J}_i \cdot \mathbf{J}_k \equiv Q_{ik}$, $\mathbf{J}_i \cdot \mathbf{B}_n \equiv R_{in}$, and $\mathbf{B}_n \cdot \mathbf{B}_m \equiv T_{nm}$, measuring overlaps between student and teacher vectors. The overlaps R and Q become the dynamical variables of the system while T is defined by the task. The learning dynamics is then defined in terms of differential equations for the macroscopic variables with respect to the normalized number of examples $\alpha = \mu/N$ playing the role of a continuous time variable:

$$\begin{aligned} \frac{dR_{in}}{d\alpha} &= \eta \; \phi_{in} \;, \\ \frac{dQ_{ik}}{d\alpha} &= \eta \; \psi_{ik} + \eta^2 \; \upsilon_{ik} \;, \end{aligned} \tag{2}$$

where $\phi_{in} \equiv < \delta_i \; y_n >_{\{\xi\}}$, $\psi_{ik} \equiv < \delta_i \; x_k + \delta_k \; x_i >_{\{\xi\}}$ and $\upsilon_{ik} \equiv < \delta_i \; \delta_k >_{\{\xi\}}$. The explicit expressions[Saad and Solla 1995a] for ϕ_{in}, ψ_{ik}, υ_{ik} and ϵ_g depend exclusively on the overlaps Q, R and T.

These equations, depending on a closed set of parameters, can be integrated and iteratively solved, providing a full description of the order parameters evolution, from which the evolution of the generalization error can be derived.

3 Globally optimal learning rates

Defining the decrease in generalization error as the measure of optimality, it is straightforward to find the locally optimal learning rate by determining the value of η that minimizes $d\epsilon_g/d\alpha$, using the equations of motion for R and Q and the fact that the generalization error depends exclusively on these parameters. The expression obtained for the locally optimal learning rate is of the form:

$$\eta = -\frac{\sum_{in} \frac{\partial \epsilon_g}{\partial R_{in}} \phi_{in} + \sum_{ij} \frac{\partial \epsilon_g}{\partial Q_{ij}} \psi_{ij}}{2 \sum_{ij} \frac{\partial \epsilon_g}{\partial Q_{ij}} \upsilon_{ij}} \;, \tag{3}$$

Although the value of η obtained in this manner may be useful for some phases of the learning process it is likely to be useless for others. For example, the lowest generalization error for the symmetric phase, characterized by lack of differentiation between the student nodes [Saad and Solla 1995b], is achieved by reducing the learning rate to zero; however, decaying the learning rate in the symmetric phase will prevent the system from escaping the symmetric fixed point, thus resulting in a suboptimal solution.

A *globally optimal* learning scenario in a certain time window $[\alpha_0, \alpha_1]$ corresponds to the largest decrease in generalization error between these two times; i.e., we attempt to minimize $\Delta\epsilon_g = \epsilon_g(\alpha_1) - \epsilon_g(\alpha_0)$ which may be written as an integral of the form:

$$\Delta\epsilon_g = \int_{\alpha_0}^{\alpha_1} \frac{d\epsilon_g}{d\alpha}\, d\alpha \tag{4}$$

Since the generalization error depends exclusively on the overlaps Q, R and T, for which the dynamical equations are known, one can rewrite the integrand $\mathcal{L} = \frac{d\epsilon_g}{d\alpha}$ as

$$\mathcal{L} = \sum_{in} \frac{\partial \epsilon_g}{\partial R_{in}} \frac{dR_{in}}{d\alpha} + \sum_{ik} \frac{\partial \epsilon_g}{\partial Q_{ik}} \frac{dQ_{ik}}{d\alpha} + \sum_{in} \lambda_{in} \left(\frac{dR_{in}}{d\alpha} - \eta\ \phi_{in} \right) + \sum_{ik} \nu_{ik} \left(\frac{dQ_{ik}}{d\alpha} - \eta\ \psi_{ik} - \eta^2\ \upsilon_{ik} \right) \tag{5}$$

The last two right hand terms in Eq.(5) force the correct dynamics using sets of Lagrange multipliers λ_{in} and ν_{ik} for the corresponding equations $dR_{in}/d\alpha$ and $dQ_{ik}/d\alpha$.

Using variational techniques it is straightforward to obtain a set of coupled differential equations for the Lagrange multipliers:

$$\begin{aligned} \frac{d\lambda_{km}}{d\alpha} &= -\eta \sum_{in} \lambda_{in} \frac{\partial \phi_{in}}{\partial R_{km}} - \eta \sum_{ij} \nu_{ij} \frac{\partial \left(\psi_{ij} + \eta\ \upsilon_{ij}\right)}{\partial R_{km}} \\ \frac{d\nu_{kl}}{d\alpha} &= -\eta \sum_{in} \lambda_{in} \frac{\partial \phi_{in}}{\partial Q_{kl}} - \eta \sum_{ij} \nu_{ij} \frac{\partial \left(\psi_{ij} + \eta\ \upsilon_{ij}\right)}{\partial Q_{kl}}\ , \end{aligned} \tag{6}$$

a separate equation for η as a function of the Lagrange multipliers

$$\eta = -\frac{\sum_{in} \lambda_{in} \phi_{in} + \sum_{ij} \nu_{ij} \psi_{ij}}{2 \sum_{ij} \nu_{ij} \upsilon_{ij}}\ , \tag{7}$$

and a set of boundary conditions

$$\lambda_{in}\Big|_{\alpha_1} = \frac{\partial \epsilon_g}{\partial R_{in}}\Big|_{\alpha_1} \quad \text{and} \quad \nu_{ik}\Big|_{\alpha_1} = \frac{\partial \epsilon_g}{\partial Q_{ik}}\Big|_{\alpha_1}\ , \tag{8}$$

which correspond to the greedy optimization of the generalization error with respect to η at α_1. Note that the boundary conditions are identical to the locally optimal solution (Eq.3), reflecting the fact that at α_1 only local information is relevant as the choice of η here does not affect the dynamics at other times.

To solve Eq.(7), which is found by setting the functional derivative of $\Delta\epsilon_g$ with respect to η to zero, we use gradient descent $\eta(t+1) = \eta(t) - \theta\ \delta\Delta\epsilon_g/\delta\eta$, where

$$\frac{\delta \Delta\epsilon_g}{\delta \eta} = \sum_{in} \lambda_{in} \phi_{in} + \sum_{ij} \nu_{ij} \left(\psi_{ij} + 2\ \eta\ \upsilon_{ij}\right) \tag{9}$$

Here, t is the iteration index and θ is the learning rate for the optimization process. Second order variations can also be employed to speed up convergence. All terms required for determining this functional derivative can be obtained by integrating the equations forward, using Eq.(2) and some initial conditions for the overlaps, and then backwards for the Lagrange multipliers, using Eq.(6) and the boundary conditions expressed in Eq.(8). This process converges within a few iterations and results in an exact function for the optimal learning rate over the given time window.

4 Examples

In our first example we apply the method to a realizable ($K = M = 2$) noiseless training task in the case of isotropic teacher vectors ($T_{nm} = \delta_{nm}$), to obtain the optimal learning rate throughout the learning process. Initial conditions for the overlaps R_{in} and Q_{ik}, where $i \neq k$, are taken randomly from a uniform distribution between $[0, 10^{-6}]$ while the vector lengths Q_{ii} are taken from a uniform distribution between $[0, 0.5]$. The learning rate was initially fixed to some arbitrary value and the time window taken is $0 \leq \alpha \leq 350$.

Applying the optimization process we obtain the results shown in Fig. 1 for the optimal learning rate and the corresponding evolution of the generalization error. After a rapid initial decay the generalization error stabilizes at an almost fixed value, corresponding to the symmetric phase. At the same time the learning rate grows quickly until stabilizing at an almost fixed value, $\eta \simeq 1.66$, corresponding to the maximal learning rate for which the vectors do not show an uncontrollable growth, thus resulting in the shortest symmetric phase[Saad and Solla 1995b]. This result is in close agreement with values obtained numerically in separate studies[West and Saad 1997]. As the system escapes the symmetric phase, we see an increase in the learning rate towards another fixed value. The new value $\eta = 1.8808$ is identical to the analytical results, obtained independently[Saad and Solla 1995b, Riegler and Biehl 1995, West and Saad 1997] by expanding the dynamical equations (2) around their asymptotic fixed point ($R_{in} = \delta_{in}$ and $Q_{ik} = \delta_{ik}$, once the indices have been reordered).

Towards the end of the time window we see an unexpected drop in the learning rate to a value of about $\eta = 0.59$. Examining the expression for the generalization error in the vicinity of its asymptotic fixed point we see that it is possible to gain an immediate reduction by choosing an appropriate direction for the decay eigenvectors. This is achieved by reducing the learning rate which results in a slower decay of the order parameters. Using the symmetry of the problem we expand the generalization error around the fixed point via $R_{in} = \delta_{in}(1 - r) + (1 - \delta_{in})s$ and $Q_{ik} = \delta_{ik}(1 - q) + (1 - \delta_{ik})c$ to find two contributions to the leading term of opposite sign, proportional to $2r - q$ and $2s - c$ respectively. These quantities are shown in the inset to Fig. 1, for $310 \leq \alpha \leq 350$, which also shows the corresponding generalization error. The constant exponential decay is interrupted by a rapid

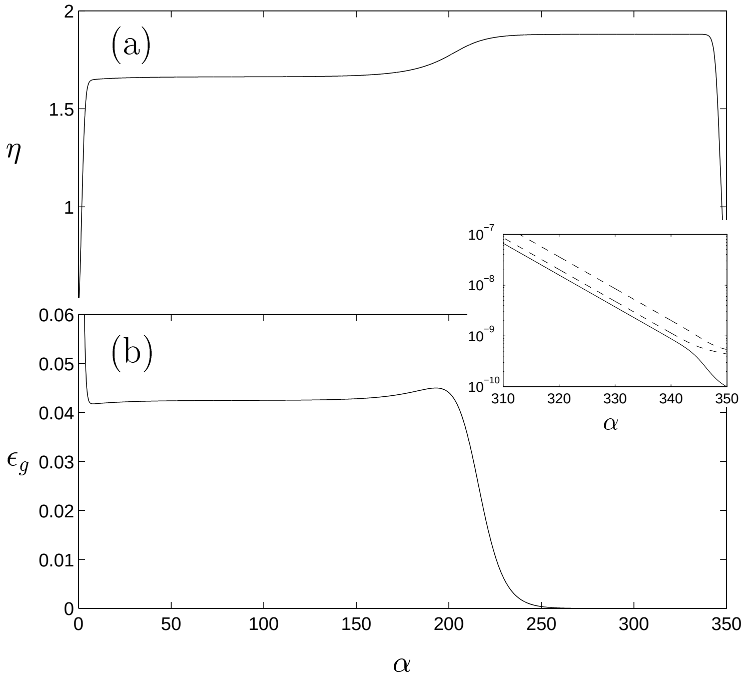

Figure 1: The optimal learning rate (a) and the resulting generalization error (b) as a function of α for the case of a two hidden node student trained to emulate a teacher of a similar configuration. Inset - the evolution of the generalization error (solid line) and the magnitude of the opposing contributions to the leading term (dashed lines - upper line proportional to $2r - q$, lower line proportional to $2s - c$) for $310 \leq \alpha \leq 350$.

reduction in the difference between these two opposing contributions to the generalization error. The greedy procedure slows the asymptotic decay of the order parameters and is therefore unsustainable in the long term. Thus, this drop off in the learning rate only ever occurs towards the end of the given time window.

In the second example we apply our method to an unrealizable learning scenario, by introducing additive uncorrelated Gaussian output noise of zero mean and some variance σ^2 to the examples. Similar results are obtained for structural unrealizability ($K < M$). The picture that emerges, shown in Fig. 2(a) for various noise levels ($\sigma^2 = 10^{-2}, 10^{-5}$ and 10^{-7}), is initially similar to that of the realizable case but changes dramatically as the system escapes the symmetric phase towards the asymptotic regime. In this case the learning rate starts from a fixed value but decays increasingly

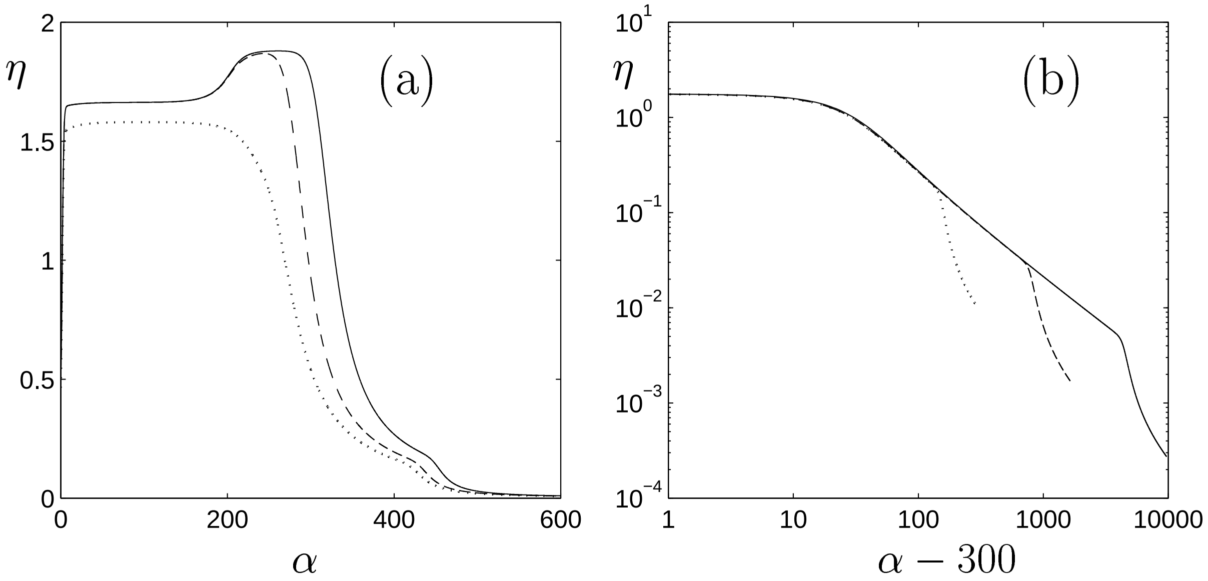

Figure 2: Optimal learning rate for a two hidden node student trained on corrupted examples generated by a teacher of a similar configuration. (a) shows behaviour for three noise levels $\sigma^2 = 10^{-2}$, 10^{-5} and 10^{-7} (from left to right) over a fixed time window $0 \leq \alpha \leq \alpha_1 = 600$. (b) shows the asymptotic decay for $\sigma^2 = 10^{-7}$ over different time windows, with $\alpha_1 = 600$, 2000 and 10^4 (from left to right). The curves lie on top of one another until a drop off towards the end of each curve which corresponds to a greedy minimization of the generalization error. The overall trend before this point is towards a decay inversely proportional to α.

rapidly until it reaches a decay inversely proportional to α, proved to be optimal for linear systems (for a review, see [White 1998]). As in the realizable case one observes a greedy selection of the learning rate for obtaining an instantaneous reduction of the generalization error, in the form of a kink in the curve after $\alpha = 420$. The log-log plot in Fig. 2(b) shows the optimal learning rate as a function of α for various time windows (increasing α_1). The drop off towards the end of each time window is due to the greedy effect discussed above and corresponds to a similar fast reduction in the generalization error. Before this point is reached the decay of the learning rate and generalization error becomes inversely proportional to α asymptotically, which presumably corresponds to the optimal sustainable learning schedule in this regime. As the symmetry breaks one should therefore gradually modify the decay rate from a constant until it is proportional to $1/\alpha$. However, it will often take a prohibitively long time until the $1/\alpha$ decay rate becomes optimal, making it completely irrelevant in many instances. Moreover, if one decays the learning rate at a fixed rate (for example, inversely with α) it may take an extremely long time before losses, incurred due to the use of sub-optimal learning rates in earlier stages of the dynamics, can be recovered.

5 Site dependent learning rate

This framework for choosing the optimal learning rate can easily be extended to accommodate different learning rates for weights associated with different hidden nodes. This enables the system to explore more complex routes of breaking the symmetry and converging to the optimal solution.

In the following example, shown in Fig.3, we train a three hidden node system on examples generated by a three node teacher, using three different learning rates related to the various hidden nodes. The following picture emerges when applying the optimisation process: At first, the method separates the three nodes by assigning a high value to two of the learning rates, higher than the value obtained for a single learning rate in the case of a realizable three node scenario, and a very low value to the third. Then, after the symmetry of the two nodes is broken, the third learning rate increases and the third model vector specializes on the remaining teacher vector. Eventually, all learning rates converge to the same constant and asymptotically optimal learning rate.

6 Conclusions

This paper introduces a method for optimizing the learning rate in on-line learning scenarios over a given time frame, using a variational approach. First we consider the case of a single learning rate, optimized in two learning scenarios: realizable and unrealizable. The results, which are consistent with numerical values obtained separately, provide the constant learning rate that should be used in the symmetric phase in both scenarios as well as the final constant learning rate and the asymptotic decay rate required for the realizable and unrealizable cases respectively. We then demonstrate the capability of the method for handling site dependent learning rates in a realizable scenario and explain the results obtained.

A similar approach can be applied to incorporate information about curvature and to examine the relevance of many modifications that have been suggested over the years to the basic gradient descent rule. It is also possible to determine globally optimal learning rules, extending existing results for

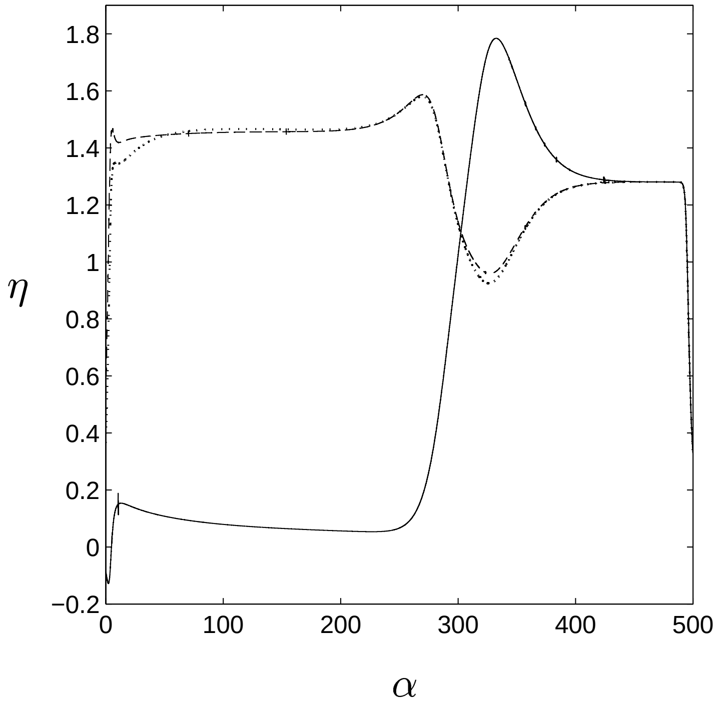

Figure 3: Optimized site dependent learning rates. Both student and teacher are three node soft committee machines; examples are generated by an isotropic teacher.

discrete machines [Kinouchi and Caticha 1992]. In addition, by constraining the differential equations (6) on the basis of the numerical solutions, using the fact that symmetries in the Lagrange multipliers dynamics mirror those of the order parameters, one can analyse the behavior of the differential equations for specific phases in the evolution of η to obtain a more generic description for its behavior as a function of the network size and other relevant parameters. These aspects and others will be discussed in future publications.

Acknowledgement This work was supported by the EPSRC grant GR/L19232. The authors would like to thank Ansgar West and Bernhard Schottky for useful discussions.

References

[Cybenko 1989] G. Cybenko, *Math. Control Signals and Systems* **2**, 303 (1989).

[Bishop 1995] C.M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, Oxford (1995).

[Biehl and Schwarze 1995] M. Biehl and H. Schwarze, *J. Phys. A* **28**, 643 (1995).

[Saad and Solla 1995a] D. Saad and S. A. Solla, *Phys. Rev. Lett.* **74**, 4337 (1995).

[Saad and Solla 1995b] D. Saad and S.A. Solla *Phys. Rev. E* **52** 4225 (1995).

[Riegler and Biehl 1995] P. Riegler and M. Biehl, *J. Phys. A* **28**, L507 (1995).

[West and Saad 1997] A.H.L. West and D. Saad, *Phys. Rev. E*, in press (1997).

[White 1998] H. White, *Neural Computation* **1**, 425 (1989).

[Kinouchi and Caticha 1992] O. Kinouchi and N. Caticha *J. Phys. A* **25**, 6243 (1992).