Globally Optimal Parameters for On-Line Learning in Multilayer Neural Networks

David Saad and Magnus Rattray
Department of Computer Science and Applied Mathematics, Aston University, Birmingham B4 7ET, UK.

We present a framework for calculating globally optimal parameters, within a given time frame, for on-line learning in multilayer neural networks. We demonstrate the capability of this method by computing optimal learning rates in typical learning scenarios. A similar treatment allows one to determine the relevance of related training algorithms based on modifications to the basic gradient descent rule as well as to compare different training methods.

Feed-forward neural networks have attracted interest in recent years for their ability to implement arbitrary continuous and discrete input-output maps [1], corresponding to general regression and classification tasks. For a given desired map of the form $\tilde{f} : X \to Y$, where X and Y represent the input and output space respectively, one can construct a two layer feed-forward neural network, implementing a function $f_{\mathbf{J}}$ that emulates $\tilde{f}$ to any desired accuracy. The process whereby the network's internal parameters $\{\mathbf{J}\}$ are optimized with respect to a set of examples instancing the underlying rule, and some measure of the discrepancy between $f_{\mathbf{J}}$ and $\tilde{f}$, is termed training and 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* of continuous functions via gradient descent on a differentiable error measure. This technique 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 instance, adding weight decay and momentum terms (for a review, see [2]). These modifications inevitably introduce new parameters which, in addition to the inherent stochasticity of the learning process, makes it very hard to assess their usefulness.

A recent study [3–5], offers a framework for analytically examining different aspects of on-line learning scenarios. We will employ the same framework to suggest a method for calculating, within given time windows, globally optimal parameters. The method will be demonstrated on one of the natural parameters in gradient descent on-line learning, the learning rate, although it can easily be generalized to accommodate other parameters and learning rules, as well as discrete architectures. This method can also be employed to assess the usefulness of various modifications to the basic gradient descent rule, or even to compare the efficiency of different training techniques, by examining the optimal values assigned to the related coefficients. For instance, low optimal values, possibly in certain phases of the learning process, will indicate that these modifications are redundant.

In this letter, we concentrate on maps 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 two layer neural network, where g is the activation function of the hidden units, taken here to be the error function $g(x) \equiv \mathrm{erf}(x/\sqrt{2})$, $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \le i \le K}$ 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, usually referred to as the 'soft committee machine' [3], represents most of the properties of general multilayer networks and can easily be extended to accommodate adaptive hidden-to-output weights [6].

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 $\mathbf{B} \equiv \{\mathbf{B}_n\}_{1 \le n \le M}$ 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

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

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 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 [4] 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 [4] 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.

An 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:

$$\Delta\epsilon_g = \int_{\alpha_0}^{\alpha_1} \frac{d\epsilon_g}{d\alpha}\ d\alpha \tag{3}$$

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

$$\mathcal{L} = \sum_{in} \frac{\partial\epsilon_g}{\partial R_{in}} \frac{dR_{in}}{d\alpha} + \sum_{ik} \frac{\partial\epsilon_g}{\partial Q_{ik}} \frac{dQ_{ik}}{d\alpha} + \sum_{in} \lambda_{in} \left(\frac{dR_{in}}{d\alpha} - \eta\ \phi_{in} \right) + \sum_{ik} \nu_{ik} \left(\frac{dQ_{ik}}{d\alpha} - \eta\ \psi_{ik} - \eta^2\ \upsilon_{ik} \right) \tag{4}$$

The last two right hand terms in Eq.(4) 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:

$$\begin{aligned} \frac{d\lambda_{km}}{d\alpha} &= -\eta \sum_{in} \lambda_{in} \frac{\partial\phi_{in}}{\partial R_{km}} - \eta \sum_{ij} \nu_{ij} \frac{\partial\left(\psi_{ij} + \eta\ \upsilon_{ij}\right)}{\partial R_{km}} \\ \frac{d\nu_{kl}}{d\alpha} &= -\eta \sum_{in} \lambda_{in} \frac{\partial\phi_{in}}{\partial Q_{kl}} - \eta \sum_{ij} \nu_{ij} \frac{\partial\left(\psi_{ij} + \eta\ \upsilon_{ij}\right)}{\partial Q_{kl}}\ , \end{aligned} \tag{5}$$

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

$$\eta = -\frac{\sum_{in} \lambda_{in}\phi_{in} + \sum_{ij} \nu_{ij}\psi_{ij}}{2\sum_{ij} \nu_{ij}\upsilon_{ij}}\ , \tag{6}$$

and a set of boundary conditions

$$\lambda_{in}\Big|_{\alpha_1} = \frac{\partial\epsilon_g}{\partial R_{in}}\Big|_{\alpha_1} \quad \text{and} \quad \nu_{ik}\Big|_{\alpha_1} = \frac{\partial\epsilon_g}{\partial Q_{ik}}\Big|_{\alpha_1}\ , \tag{7}$$

which correspond to the greedy optimization of the generalization error with respect to η at α_1.

To solve Eq.(6), which is found by setting the functional derivative of $\Delta\epsilon_g$ with respect to η to zero, we use gradient descent (second order variations can also be employed to speed up convergence). All terms required for carrying out the optimization of Eq.(6) 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.(5) and the boundary conditions expressed in Eq.(7). This process converges within a few iterations and results in an exact function for the optimal learning rate over the time window.

We demonstrate the results obtained by this method via two simple examples. In the 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; 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 characterized by the lack of differentiation between different teacher vectors [5]. At the same time the learning rate grows quickly until stabilizing at an almost fixed value. This value, $\eta \simeq 1.66$, corresponds to the maximal learning rate for which the vectors do not show an uncontrollable growth, thus resulting in the shortest symmetric phase [5]. This result is in close agreement with values obtained numerically in separate studies [7]. 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 [5–7] 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 reduction in the difference between these two opposing contributions to the generalization error. This 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 rapidly until it reaches a decay inversely proportional to α, proved to be optimal for linear systems (for a review see [8]). 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.

In this letter we have merely demonstrated the capability of the method for a single parameter. However, a similar approach can be applied to incorporate information about the curvature, e.g., in the form of different learning rates applied to the various student vectors, 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 discrete machines [9]. In addition, by constraining the differential equations (5) on the basis of the numerical solutions, 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 such as noise and weight decay. 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.

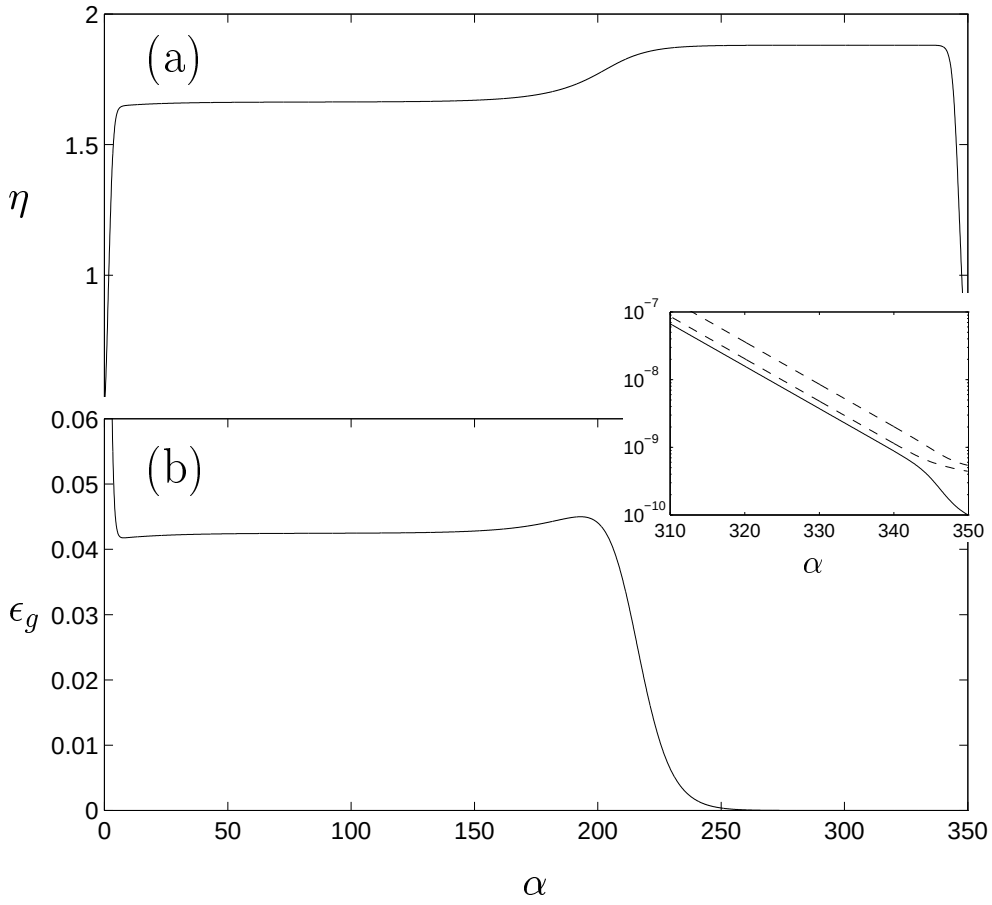

FIG. 1. The optimal learning rate (a) and the resulting generalization error (b) as a function of α for the case of a two hidden node student trained to emulate a teacher of a similar configuration. Inset - the evolution of the generalization error (solid line) and the magnitude of the opposing contributions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

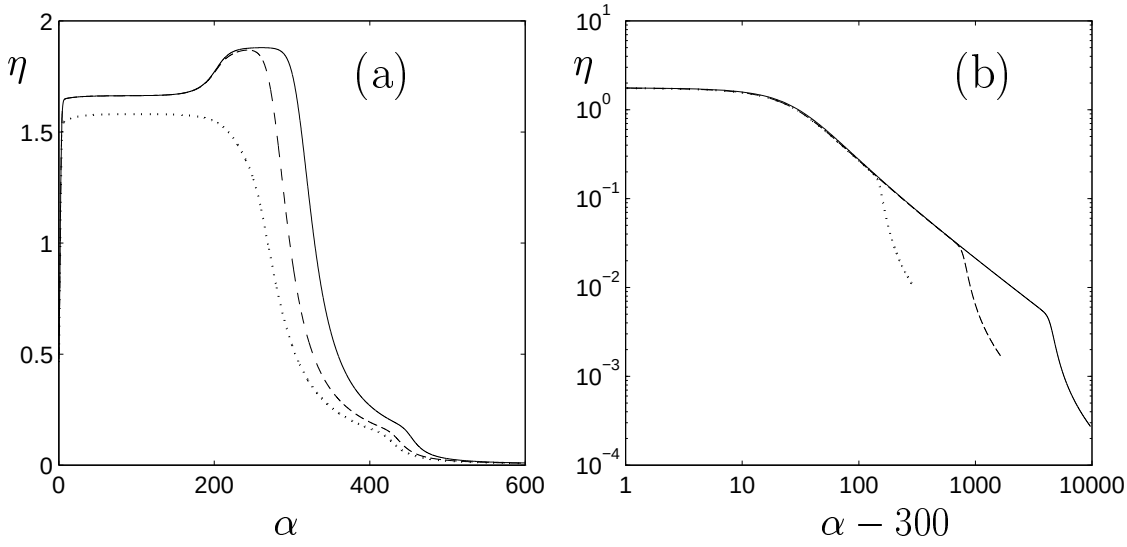

FIG. 2. Optimal learning rate for a two hidden node student trained on corrupted examples generated by a teacher of a similar configuration. (a) shows behaviour for three noise levels $\sigma^2 = 10^{-2}$, 10^{-5} and 10^{-7} (from left to right) over a fixed time window $0 \leq \alpha \leq \alpha_1 = 600$. (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.

---

[1] G. Cybenko, *Math. Control Signals and Systems* **2**, 303 (1989).
[2] C.M. Bishop, *Neural networks for pattern recognition*, (Oxford University Press, Oxford, 1995).
[3] M. Biehl and H. Schwarze, *J. Phys. A* **28**, 643 (1995).
[4] D. Saad and S. A. Solla, *Phys. Rev. Lett.* **74**, 4337 (1995).
[5] D. Saad and S.A. Solla *Phys. Rev. E* **52** 4225 (1995).
[6] P. Riegler and M. Biehl, *J. Phys. A* **28**, L507 (1995).
[7] A.H.L. West and D. Saad, unpublished.
[8] H. White, *Neural Computation* **1**, 425 (1989).
[9] O. Kinouchi and N. Caticha *J. Phys. A* **25**, 6243 (1992).