The Theory of On-Line Learning - A Statistical Physics Approach

D. Saad

The Neural Computing Research Group
University of Aston, Birmingham, B4 7ET, UK

Abstract: In this paper we review recent theoretical approaches for analysing the dynamics of on-line learning in multilayer neural networks using methods adopted from statistical physics. The analysis is based on monitoring a set of macroscopic variables from which the generalisation error can be calculated. A closed set of dynamical equations for the macroscopic variables is derived analytically and solved numerically. The theoretical framework is then employed for defining optimal learning parameters and for analysing the incorporation of second order information into the learning process using natural gradient descent and matrix-momentum based methods. We will also briefly explain an extension of the original framework for analysing the case where training examples are sampled with repetition.

1 Introduction

Layered neural networks are powerful nonlinear information processing systems, capable of implementing arbitrary continuous and discrete input-output maps to any desired accuracy, given a sufficient number of hidden nodes and a sufficiently large example set. They have been employed successfully in a variety of regression and classification tasks, and have been studied using a wide range of methods (for a review see Bishop (1995)). On-line learning refers to the iterative modification of the network parameters according to a predetermined training rule, following successive presentations of single training examples, each representing a specific input vector and the corresponding output. On-line learning is one of the leading techniques in training large neural networks, especially via gradient descent on a differentiable error measure.

In this review we focus on the use of methods from non-equilibrium statistical mechanics, for analysing on-line learning in multilayer neural network. We concentrate on our contribution to this area and show how these methods can be employed to monitor the learning dynamics, particularly the evolution of the generalisation error, to define optimal learning parameters and to devise and examine improved learning methods. For a general review see Saad (1998) and Mace and Coolen (1998).

The paper is organised as follows: In section 2 we will derive a compact description of the training dynamics using a set of macroscopic variables, setting up the main theoretical framework. This will then be employed to derive optimal training parameters (section 3), to examine analytically the efficacy of natural gradient descent (section 4), and to suggest and examine practical alternatives using matrix-momentum based methods. In section 5 we will explain how the method can be extended to handle scenarios where training examples are sampled with repetition. In section 6 we will point to the main remaining open questions.

2 Learning in multilayer neural networks

For setting up the basic framework, as in Saad and Solla (1995a, 1995b), we consider a learning scenario whereby a feed-forward neural network model, the 'student', emulates an unknown mapping, the 'teacher', given examples of the teacher mapping (in this case another feed-forward neural network); here we restrict the derivation and the examples to the noiseless case although more general scenarios where training examples are corrupted by noise may also be considered. This provides a rather general learning scenario since both student and teacher can represent a very broad class of functions. Student performance is typically measured by the generalization error, which is the student's expected error on an unseen example. The object of training is to minimize the generalization error by adapting the student network's parameters appropriately.

We consider a student mapping from an N-dimensional input space $\boldsymbol{\xi} \in I\!R^N$ onto a scalar function $\sigma(\mathbf{J}, \boldsymbol{\xi}) = \sum_{i=1}^{K} g\,(\mathbf{J}_i \cdot \boldsymbol{\xi})$, which represents a soft Committee machine (SCM - Biehl and Schwarze (1995)), where $g(x) \equiv \mathrm{erf}(x/\sqrt{2})$ is the activation function of the hidden units; $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \le i \le K}$ is the set of input-to-hidden adaptive weights for the K hidden nodes and the hidden-to-output weights are set to one. The activation of hidden node i in the student under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is denoted $x_i^\mu = \mathbf{J}_i \cdot \boldsymbol{\xi}^\mu$. This configuration preserves most properties of a general multi-layer network and can be extended to accommodate adaptive hidden-to-output weights as shown by Riegler and Biehl (1995).

Training examples are of the form $(\boldsymbol{\xi}^\mu, \zeta^\mu)$ where $\mu = 1, 2, ..$ labels each independently drawn example in a sequence. 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 teacher of a similar configuration to the student except for a possible difference in the number M of hidden units: $\zeta^\mu = \sum_{n=1}^{M} g\,(\mathbf{B}_n \cdot \boldsymbol{\xi}^\mu)$, 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 in the teacher under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is denoted $y_n^\mu = \mathbf{B}_n \cdot \boldsymbol{\xi}^\mu$. Indices i, j, k and n, m refer to student and teacher units respectively.

The error made by the student is given by the quadratic deviation,

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

This training error is then used to define the learning 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 $\delta_i^\mu \equiv g'(x_i^\mu)[\sum_{n=1}^{M} g(y_n^\mu) - \sum_{j=1}^{K} g(x_j^\mu)]$ and the learning rate η has been scaled with the input size N. Performance on a typical input defines the generalization error $\epsilon_g(\mathbf{J}) \equiv \langle \epsilon(\mathbf{J}, \boldsymbol{\xi}) \rangle_{\{\xi\}}$ through an average over all possible input vectors $\boldsymbol{\xi}$.

Expressions for the generalization error and learning dynamics have been obtained in the thermodynamic limit ($N \to \infty$), and can be represented by a set of macroscopic variables (order parameters) 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:

$$\frac{dR_{in}}{d\alpha} = \eta\ \phi_{in}\ , \qquad \frac{dQ_{ik}}{d\alpha} = \eta\ \psi_{ik} + \eta^2\ \upsilon_{ik}\ , \quad (2)$$

where $\phi_{in} \equiv \langle \delta_i y_n \rangle_{\{\xi\}}$, $\psi_{ik} \equiv \langle \delta_i x_k + \delta_k x_i \rangle_{\{\xi\}}$ and $\upsilon_{ik} \equiv \langle \delta_i \delta_k \rangle_{\{\xi\}}$. The explicit expressions for ϕ_{in}, ψ_{ik}, υ_{ik} and ϵ_g depend exclusively on the overlaps Q, R and T (Saad and Solla (1995a,1995b)). Equations (2), depend on a closed set of parameters and 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.

Typical plots of the learning dynamics are presented in Fig.1. In this example the learning process prunes unnecessary nodes when the student network has excessive resources. A teacher with $M = 2$ hidden units characterized by $T_{nm} = n\ \delta_{nm}$ is to be learned by a student with $K = 3$ hidden units. The initial values of the order parameters are $R_{in} = 0$ for all i, n, $Q_{ik} = 0$ for all $i \neq k$, while the norms Q_{ii} of the student vectors are initialized independently from a uniform distribution in the $[0, 0.5]$ interval. The time evolution of the various order parameters is shown in Fig. 1a-c for $\eta = 1$. The picture that emerges is one of specialization with increasing α; asymptotically the first student node imitates the first teacher node ($Q_{11} = R_{11} = T_{11}$) while ignoring the second one ($R_{12} = 0$), the second student node imitates the second teacher node while ignoring the first one, and the third student node gets eliminated ($Q_{33} = 0$). The off-diagonal components Q_{ik} shown in Fig.1b indicate that the two surviving student vectors become increasingly uncorrelated. The overlap

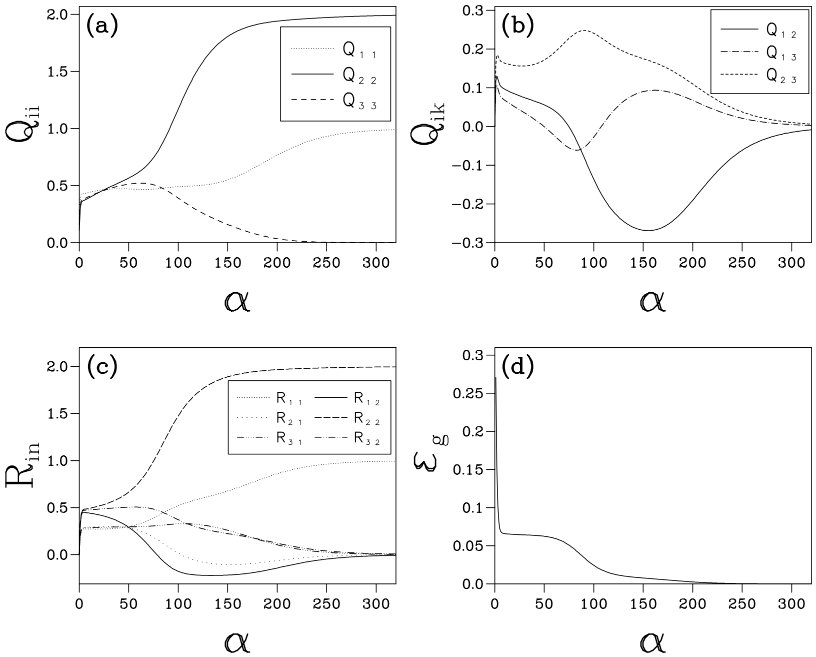

Figure 1: Dependence of the overlaps and ϵ_g on the normalized number of examples α, for $K=3$ and $M=2$: (a) the lengths of student vectors, (b) the correlation between student vectors, (c) the overlap between various student and teacher vectors, and (d) the generalization error.

between student and teacher vectors (Fig.1c) displays a small α behavior dominated by an undifferentiated symmetric solution, followed by a transition onto the specialization required to obtain perfect generalization. The evolution of the generalization error is shown in Fig.1d.

3 Optimal learning parameters

On-line methods are often sensitive to the choice of learning parameters and in particular the choice of learning rate; if chosen too large the learning process may diverge, but if η is too small then convergence can take an extremely long time. The optimal learning rate will also vary substantially over time and may require annealing asymptotically. Most existing analytical results for defining optimal learning rates concentrate on the asymptotic regime where the system may be linearized.

The naive approach to learning rate optimization is to consider the fastest rate of decrease in generalization error as a measure of optimality. To find the locally optimal learning rate one minimizes $d\epsilon_g/d\alpha$, using Eqs.(2), exploiting the fact that the change in ϵ_g over time depends exclusively on the overlaps. The expression obtained for the locally optimal learning rate may be useful for some phases of the learning process but is useless for others (Rattray and Saad (1998)).

A more appropriate measure of optimality is the total reduction in generalization error over the entire learning process as in Saad and Rattray (1997). With this measure one can then define the *globally optimal* learning rate in a given time-window $[\alpha_0, \alpha_1]$ to be that which provides the largest decrease in generalization error between these two times:

$$\Delta\epsilon_g(\eta) \;=\; \int_{\alpha_0}^{\alpha_1} \frac{d\epsilon_g}{d\alpha}\, d\alpha \;=\; \int_{\alpha_0}^{\alpha_1} \mathcal{L}(\eta,\alpha)\; d\alpha \;. \tag{3}$$

Since the generalization error depends solely on the overlaps Q, R and T, which are the dynamical variables (T remains fixed here), we can expand the integrand in terms of these variables,

$$\begin{aligned} \mathcal{L}(\eta,\alpha) \;&=\; \sum_{in} \frac{\partial\epsilon_g}{\partial R_{in}} \frac{dR_{in}}{d\alpha} + \sum_{ik} \frac{\partial\epsilon_g}{\partial Q_{ik}} \frac{dQ_{ik}}{d\alpha} \\ &-\; \sum_{in} \mu_{in} \left(\frac{dR_{in}}{d\alpha} - \eta\;\phi_{in} \right) - \sum_{ik} \nu_{ik} \left(\frac{dQ_{ik}}{d\alpha} - \eta\;\psi_{ik} - \eta^2\;\upsilon_{ik} \right) \;. \end{aligned} \tag{4}$$

The last two terms in equation (4) force the correct dynamics using sets of Lagrange multipliers μ_{in} and ν_{ik} corresponding to the equations of motion for R_{in} and Q_{ik} respectively. Variational minimization of the integral in equation (3) with respect to the dynamical variables leads to a set of coupled differential equations for the Lagrange multipliers along with a set of boundary conditions. Solving these equations over the interval $[\alpha_0, \alpha_1]$ determines necessary conditions for η to maximize $\Delta\epsilon_g(\eta)$. The theory is completely general and may be employed for different learning parameters (e.g., regularization parameters as in Saad and Rattray (1998), site dependent learning rates), various learning scenarios (structurally unrealisable or where examples are corrupted by noise) and for obtaining optimal learning rules (Rattray and Saad (1997).

4 Natural Gradient Descent

The same theoretical framework may be used for examining novel training methods. Natural gradient descent (NGD) was recently proposed by Amari (1998) as a principled alternative to standard on-line gradient descent (GD). When learning to emulate a stochastic rule with some probabilistic model, e.g. a feed-forward neural network, NGD has the desirable properties of asymptotic optimality, given a sufficiently rich model which is differentiable with respect to its parameters, and invariance to re-parameterization of our model distribution. These properties are achieved by viewing the parameter space of the model as a Riemannian space in which local distance is defined by the Kullback-Leibler divergence. The Fisher information matrix provides the appropriate metric in this space. If the training error is defined as the negative log-likelihood of the data under our probabilistic model, then the direction of steepest descent in this Riemannian space is found by premultiplying the error gradient with the inverse of the Fisher information matrix; this defines the NGD learning direction.

Studying the learning performance of NGD in the case of isotropic tasks and structurally matched student and teacher ($K = M$ and $T = T\delta_{nm}$) we determined generic behaviour in terms of task complexity K and non-linearity T (Rattray et al (1998)). An analysis of the transient, using globally optimal learning parameters reveals that trapping time in the symmetric phase for the NGD optimized system scales as K^2, compared to a scaling of $K^{8/3}$ for optimal GD. Asymptotically, NGD saturates the universal bounds on generalization performance and provides a significant improvement over optimized GD, especially for small T.

However, in practical applications there will be an increased cost required in estimating and inverting the Fisher information matrix as it requires an average over the input distribution and a matrix inversion. An on-line matrix momentum algorithm (Orr and Leen (1994)) was introduced in order to invert an estimate of the Hessian efficiently on-line. We propose to use this method to compute the inverse of the Fisher information matrix as required for NGD. This method is particularly efficient since the inversion is replaced by a matrix-vector multiplication which can be carried out by a back-propagation step. Since the true Fisher information matrix will not be known in general we use a single step approximation of it, which can be determined on-line. We compared the efficiency of the proposed matrix momentum NGD with that of standard GD and true NGD in training two-layer networks. It turns out to provide a significant improvement over gradient descent learning but with some sensitivity to parameter choice, due to noise in the Fisher information estimate (Scarpetta et al (1999)).

5 Restricted Training Sets

In a realistic scenario the number of training examples scales with the number of free parameters, and examples are therefore sampled with repetition. This gives rise to correlations between the network parameters and the training examples, which clearly affect the learning process. One of the most significant aspects of having a fixed example set is the distinction between the two key performance measures: the *training error*, measuring the network performance with respect to the restricted training set, and the *test (generalisation) error*, calculated for all possible inputs sampled from the true distribution. The former may be monitored in practical training scenarios, while the latter can only be assessed. Another important aspect of learning from restricted training sets which have been corrupted by noise is the emergence of *overfitting* and the need to employ regularization techniques (e.g., weight decay, early stopping - see Bishop (1995)).

The fundamental difference between the infinite and restricted training set scenarios is that the joint probability distribution $P(\mathbf{x}, \mathbf{y})$ for the student and teacher node activations, which is Gaussian in the former case, takes here a more general form, which depends on the training patterns and changes dynamically during the learning process. In fact, we define $P(\mathbf{x}, \mathbf{y})$ as one of the macroscopic variables to be monitored continuously, together with the overlaps R and Q (Coolen and Saad (2000)). To follow the dynamics, one derives a set of coupled differential equations describing the evolution of the macroscopic variables in the limit $N \to \infty$. This set of equations cannot be closed in general; closure is obtained by invoking the dynamical replica theory. The resulting equations can be solved numerically with some simplifications.

The solutions describe the dynamics of both training and generalization errors (and the various overlaps, Coolen et al (2000), Xiong and Saad (2000)), provide insight to the link between the number of examples and the breaking of internal symmetries as well as some asymptotic scaling laws. Our ability to provide analytical solutions is limited due to the complexity of the equations; however, such solutions are highly desirable for deriving analytically generic scaling laws in both the symmetric phase and asymptotically, and to make a quantitative link between the noise level and the optimal regularization to be used.

6 Conclusion

We showed how the methods of statistical physics can provide insight into the dynamics of on-line learning as well as play an important role in defining optimal learning parameters and in examining the properties of new learning algorithms. Several open questions remain, for instance, finding principled methods for optimising the generalisation ability in the case of restricted training sets and the dependence of the length of the symmetric phase on the number of training examples.

References

AMARI, S. (1998): Natural Gradient Works Efficiently in Learning. *Neural Computation, Vol. 10, 251–276.*

BIEHL, M. and SCHWARZE, H. (1995): Learning by Online Gradient Descent. *Jour. Phys. A, Vol. 28, 643–656.*

BISHOP, C. M. (1995): Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

COOLEN, A. C. C. and SAAD, D. (2000): Dynamics of Learning with Restricted Training Sets. *Phys. Rev. E., Vol. 62, 5444–5487.*

COOLEN, A. C. C., SAAD, D. and XIONG, Y. (2000): On-line Learning from Restricted Training Sets in Multilayer Neural Networks. *Europhys. Lett., Vol. 51, 691–697.*

MACE, C. W. H. and COOLEN, A. C. C. (1998): Statistical Mechanical Analysis of the Dynamics of Learning in Perceptrons. *Statistics and Computing, Vol. 8 55–88.*

ORR, G. B. and LEEN, T. K. (1994):Using Curvature Information for Fast Stochastic Search. in Cowan, Tesauro and Alspector (Eds.): Advances in Neural Information Processing Systems, NIPS Vol. 6, Morgan Kaufmann, San Mateo CA, 477–484.

RATTRAY, M. and SAAD, D. (1997): Globally Optimal Rules for Online Learning in Multilayer Networks. *Jour. Phys. A, Vol. 30, L771–776.*

RATTRAY, M. and SAAD, D. (1998): An analysis of on-line training with optimal learning rates. *Phys. Rev. E., Vol. 58, 6379–6391.*

RATTRAY, M., SAAD, D. and AMARI, S. (1998): Natural Gradient Descent for On-line Learning. *Phys. Rev. Lett., Vol. 81, 5461–5464.*

RIEGLER, P. and BIEHL, M. (1995): Online Backpropagation in Two Layered Neural Networks. *Jour. Phys. A, Vol. 28, L507–513.*

SAAD, D. (Editor) (1998): On-Line Learning in Neural Networks. Publications of the Newton Institute, Cambridge University Press, Cambridge.

SAAD, D. and RATTRAY, M. (1997): Globally Optimal Parameters for On-line Learning in Multilayer Networks. *Phys. Rev. Lett., Vol. 79, 2578–2581.*

SAAD, D. and RATTRAY, M. (1998): Learning with Regularizers in Multilayer Neural Networks. *Phys. Rev. E., Vol. 57, 2170–2176.*

SAAD, D. and SOLLA, S. A. (1995): Exact Solution for On-Line Learning in Multilayer Neural Networks. *Phys. Rev. Lett., Vol. 74, 4337–4340.*

SAAD, D. and SOLLA, S. A. (1995): On-Line Learning in Soft Committee Machines. *Phys. Rev. E, Vol. 52, 4225–4243.*

SCARPETTA, S., RATTRAY, M. and SAAD, D. (1999): Matrix Momentum for Practical Natural Gradient Learning. *Jour. Phys. A, Vol. 32, 4047–4059.*

XIONG, Y. and SAAD, D. (2001): Noise, Regularizers and Unrealizable Scenarios in On-line Learning From Restricted Training Sets. *submitted.*