Baltzer Journals

On-Line Learning in Multilayer Neural Networks

David Saad [1] and Sara A. Solla [2] *

[1] *Department of Physics, University of Edinburgh, Edinburgh EH9 3JZ, UK.*

[2] *CONNECT, The Niels Bohr Institute, Blegdamsvej 17, Copenhagen 2100, Denmark.*

We present an analytic solution to the problem of on-line gradient-descent learning for two-layer neural networks with an arbitrary number of hidden units in both teacher and student networks. The technique, demonstrated here for the case of adaptive input-to-hidden weights, becomes exact as the dimensionality of the input space increases.

Layered neural networks are of interest for their ability to implement input-output maps [1]. Classification and regression tasks formulated as a map from an N-dimensional input space $\boldsymbol{\xi}$ onto a scalar ζ are realized through a map $\zeta = f_{\mathbf{J}}(\boldsymbol{\xi})$, which can be modified through changes in the internal parameters $\{\mathbf{J}\}$ specifying the strength of the interneuron couplings. Learning refers to the modification of these couplings so as to bring the map $f_{\mathbf{J}}$ implemented by the network as close as possible to a desired map $\tilde{f}$. Information about the desired map is provided through independent examples $(\boldsymbol{\xi}^\mu, \zeta^\mu)$, with $\zeta^\mu = \tilde{f}(\boldsymbol{\xi}^\mu)$ for all μ.

A recently introduced approach investigates *on-line learning* [2]. In this scenario the couplings are adjusted to minimize the error after the presentation of each example. The resulting changes in $\{\mathbf{J}\}$ are described as a dynamical evolution, with the number of examples playing the role of time. The average that accounts for the disorder introduced by the independent random selection of an example at each time step can be performed directly. The result is expressed in the form of dynamical equations for *order parameters* which describe correlations among the various nodes in the trained network as well as their degree of specialization towards the implementation of the desired task.

Here we obtain *analytic* equations of motion for the order parameters in a general two-layer scenario: a student network composed of N input units, K hidden units, and a single linear output unit is trained to perform a task defined through a teacher network of similar architecture except that its number M of hidden units is not necessarily equal to K. Two-layer networks with an arbitrary number of hidden units

*Work supported by EU grant ERB CHRX-CT92-0063.

have been shown to be universal approximators [1] for N-to-one dimensional maps. Our results thus describe the learning of tasks of arbitrary complexity (general M). The complexity of the student network is also arbitrary (general K, independent of M), providing a tool to investigate realizable ($K = M$), over-realizable ($K > M$), and unrealizable ($K < M$) learning scenarios.

In this paper we limit our discussion to the case of the soft-committee machine [2], in which all the hidden units are connected to the output unit with positive couplings of unit strength, and only the input-to-hidden couplings are adaptive. Consider the student network: hidden unit i receives information from input unit r through the weight J_{ir}, and its activation under presentation of an input pattern $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ is $x_i = \mathbf{J}_i \cdot \boldsymbol{\xi}$, with $\mathbf{J}_i = (J_{i1}, \ldots, J_{iN})$ defined as the vector of incoming weights onto the i-th hidden unit. The output of the student network is $\sigma(\mathbf{J}, \boldsymbol{\xi}) = \sum_{i=1}^{K} \; g \; (\mathbf{J}_i \cdot \boldsymbol{\xi})$, where g is the activation function of the hidden units, taken here to be the error function $g(x) \equiv \mathrm{erf}(x/\sqrt{2})$, and $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \leq i \leq K}$ is the set of input-to-hidden adaptive weights.

Training examples are of the form $(\boldsymbol{\xi}^\mu, \zeta^\mu)$. 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 whose internal structure is that of a network similar to the student except for a possible difference in the number M of hidden units. Hidden unit n in the teacher network receives input information through the weight vector $\mathbf{B}_n = (B_{n1}, \ldots, B_{nN})$, and its activation under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is $y_n^\mu = \mathbf{B}_n \cdot \boldsymbol{\xi}^\mu$. The corresponding output is $\zeta^\mu = \sum_{n=1}^{M} \; g \; (\mathbf{B}_n \cdot \boldsymbol{\xi}^\mu)$. 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

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

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 $\boldsymbol{\xi}$, to be performed implicitly through averages over the activations $\mathbf{x} = (x_1, \ldots, x_K)$ and $\mathbf{y} = (y_1, \ldots, y_M)$. Note that both $< x_i > = < y_n > = 0$, while the components of the covariance matrix $\mathcal{C}$ are given by overlaps among the weight vectors associated with the various hidden units: $< x_i \; x_k > \; = \mathbf{J}_i \cdot \mathbf{J}_k \equiv Q_{ik}$, $< x_i \; y_n > \; = \mathbf{J}_i \cdot \mathbf{B}_n \equiv R_{in}$, and $< y_n \; y_m > \; = \mathbf{B}_n \cdot \mathbf{B}_m \equiv T_{nm}$. The averages over $\mathbf{x}$ and $\mathbf{y}$ are performed using a joint probability distribution given by the multivariate Gaussian:

$$\mathcal{P}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{(2\pi)^{K+M} |\mathcal{C}|}} \; \exp\left\{ -\frac{1}{2} (\mathbf{x}, \mathbf{y})^T \mathcal{C}^{-1} (\mathbf{x}, \mathbf{y}) \right\} \; , \text{with } \; \mathcal{C} = \begin{bmatrix} Q & R \\ R^T & T \end{bmatrix} . \tag{2}$$

The averaging yields an expression for the generalization error in terms of the order parameters Q_{ik}, R_{in}, and T_{nm}. For $g(x) \equiv \mathrm{erf}(x/\sqrt{2})$ the result is:

$$\begin{aligned} \epsilon_g(\mathbf{J}) &= \frac{1}{\pi}\left\{ \sum_{ik} \arcsin \frac{Q_{ik}}{\sqrt{1+Q_{ii}}\,\sqrt{1+Q_{kk}}} + \sum_{nm} \arcsin \frac{T_{nm}}{\sqrt{1+T_{nn}}\,\sqrt{1+T_{mm}}} \right. \\ &\left. -2 \sum_{in} \arcsin \frac{R_{in}}{\sqrt{1+Q_{ii}}\,\sqrt{1+T_{nn}}} \right\} . \end{aligned} \tag{3}$$

The parameters T_{nm} are characteristic of the task to be learned and remain fixed, while the overlaps Q_{ik} and R_{in} are determined by the student weights $\mathbf{J}$ and evolve during training.

A gradient descent rule for the update of the student weights results in $\mathbf{J}_i^{\mu+1} = \mathbf{J}_i^{\mu} + \frac{\eta}{N}\,\delta_i^{\mu}\,\boldsymbol{\xi}^{\mu}$, where the learning rate η has been scaled with the input size N, and $\delta_i^{\mu} \equiv g'(x_i^{\mu})\left[\sum_{n=1}^{M} g(y_n^{\mu}) - \sum_{j=1}^{K} g(x_j^{\mu})\right]$ is defined in terms of both the activation function g and its derivative g'.

The time evolution of the overlaps R_{in} and Q_{ik} can be explicitly written in terms of similar difference equations. The dependence on the current input $\boldsymbol{\xi}^{\mu}$ is only through the activations $\mathbf{x}$ and $\mathbf{y}$, and the corresponding averages can be performed using the joint probability distribution (2). In the thermodynamic limit $N \to \infty$ the normalized example number $\alpha = \mu/N$ can be interpreted as a continuous time variable, leading to the equations of motion:

$$\begin{aligned} \frac{dR_{in}}{d\alpha} &= \eta \left\{ \sum_{m} I_3(i,n,m) - \sum_{j} I_3(i,n,j) \right\} , \\ \frac{dQ_{ik}}{d\alpha} &= \eta \left\{ \sum_{m} I_3(i,k,m) - \sum_{j} I_3(i,k,j) \right\} + \eta \left\{ \sum_{m} I_3(k,i,m) - \sum_{j} I_3(k,i,j) \right\} + \\ & \eta^2 \left\{ \sum_{n,m} I_4(i,k,n,m) - 2 \sum_{j,n} I_4(i,k,j,n) + \sum_{j,l} I_4(i,k,j,l) \right\} . \end{aligned} \tag{4}$$

The two multivariate Gaussian integrals: $I_3 \equiv < g'(u)\; v\; g(w) >$ and $I_4 \equiv < g'(u)\; g'(v)\; g(w)\; g(z) >$ represent averages over the probability distribution (2). The averages can be performed analytically for the choice $g(x) = \mathrm{erf}(x/\sqrt{2})$. Arguments assigned to I_3 and I_4 are to be interpreted following our convention to distinguish student from teacher activations. For example, $I_3(i,n,j) \equiv < g'(x_i)\; y_n\; g(x_j) >$, and the average is performed using the three-dimensional covariance matrix C_3 which results from projecting the full covariance matrix $\mathcal{C}$ of Eq. (2) onto the relevant subspace. For $I_3(i,n,j)$ the corresponding matrix is:

$$C_3 = \begin{pmatrix} Q_{ii} & R_{in} & Q_{ij} \\ R_{in} & T_{nn} & R_{jn} \\ Q_{ij} & R_{jn} & Q_{jj} \end{pmatrix} .$$

I_3 is given in terms of the components of the C_3 covariance matrix by

$$I_3 = \frac{2}{\pi} \frac{1}{\sqrt{\Lambda_3}} \frac{C_{23}(1+C_{11}) - C_{12}C_{13}}{1+C_{11}} \,, \tag{5}$$

with $\Lambda_3 = (1+C_{11})(1+C_{33}) - C_{13}^2$. The expression for I_4 in terms of the components of the corresponding C_4 covariance matrix is

$$I_4 = \frac{4}{\pi^2} \frac{1}{\sqrt{\Lambda_4}} \arcsin\left(\frac{\Lambda_0}{\sqrt{\Lambda_1}\sqrt{\Lambda_2}}\right), \tag{6}$$

where $\Lambda_4 = (1+C_{11})(1+C_{22}) - C_{12}^2$, and

$$\begin{aligned}
\Lambda_0 &= \Lambda_4 C_{34} - C_{23}C_{24}(1+C_{11}) - C_{13}C_{14}(1+C_{22}) + C_{12}C_{13}C_{24} + C_{12}C_{14}C_{23} \,, \\
\Lambda_1 &= \Lambda_4(1+C_{33}) - C_{23}^2(1+C_{11}) - C_{13}^2(1+C_{22}) + 2C_{12}C_{13}C_{23} \,, \\
\Lambda_2 &= \Lambda_4(1+C_{44}) - C_{24}^2(1+C_{11}) - C_{14}^2(1+C_{22}) + 2C_{12}C_{14}C_{24} \,.
\end{aligned}$$

These dynamical equations provide a novel tool for analyzing the learning process for a general soft-committee machine with an arbitrary number K of hidden units, trained to perform a task defined by a soft-committee teacher with M hidden units. This set of coupled first-order differential equations can be easily solved numerically, even for large values of K and M, providing valuable insight into the process of learning in multilayer networks, and allowing for the calculation of the time evolution of the generalization error [3].

In what follows we focus on learning a realizable task ($K = M$) defined through uncorrelated teacher vectors of unit length ($T_{nm} = \delta_{nm}$). The time evolution of the overlaps R_{in} and Q_{ik} follows from integrating the equations of motion (4) from initial conditions determined by a random initialization of the student vectors $\{\mathbf{J}_i\}_{1 \leq i \leq K}$. Random initial norms Q_{ii} for the student vectors are taken here from a uniform distribution in the $[0, 0.5]$ interval. Overlaps Q_{ik} between independently chosen student vectors $\mathbf{J}_i$ and $\mathbf{J}_k$, or R_{in} between $\mathbf{J}_i$ and an unknown teacher vector $\mathbf{B}_n$ are small numbers, of order $1/\sqrt{N}$ for $N \gg K$, and taken here from a uniform distribution in the $[0, 10^{-12}]$ interval. We show in Fig. 1a-c the resulting evolution of the overlaps and generalization error for $K = 3$ and $\eta = 0.1$.

This example illustrates the successive regimes of the learning process. The system quickly evolves into a symmetric subspace controlled by an unstable suboptimal solution which exhibits no differentiation among the various student hidden units. Trapping in the symmetric subspace prevents the specialization needed to achieve the optimal solution, and the generalization error remains finite, as shown by the plateau in Fig. 1c. The symmetric solution is unstable, and the perturbation introduced through the random initialization of the overlaps R_{in} eventually takes over: the student units become specialized and the matrix R of student-teacher overlaps tends towards the matrix T, except for a permutational symmetry associated with

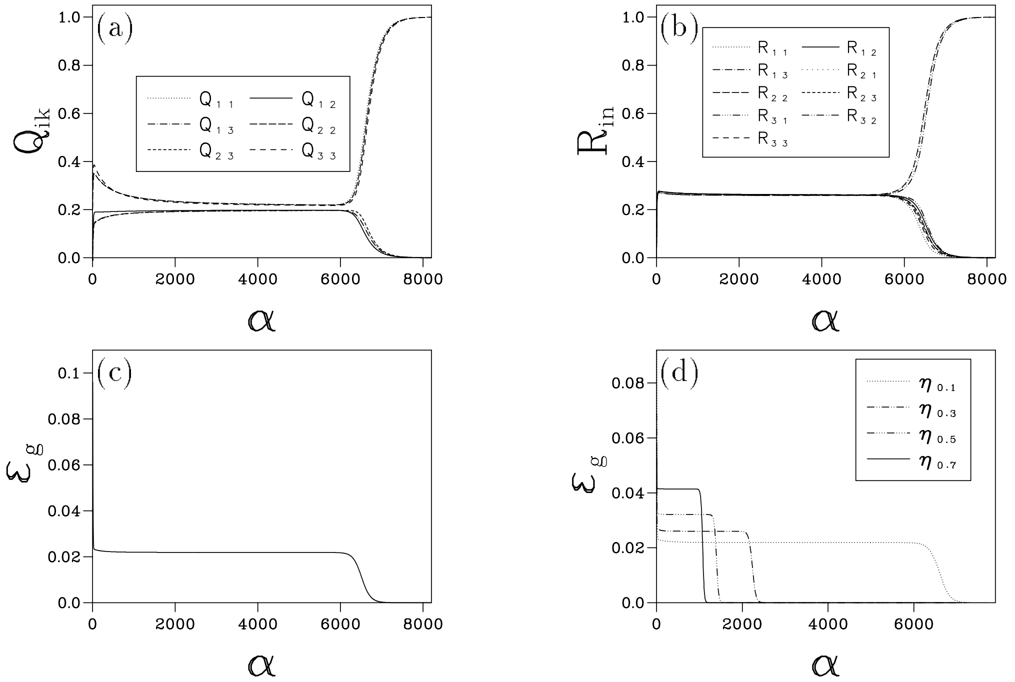

Figure 1: The overlaps and the generalization error as a function of α for a three-node student learning an isotropic teacher ($T_{nm} = \delta_{nm}$). Results for $\eta = 0.1$ are shown for (a) student-student overlaps Q_{ik}, (b) student-teacher overlaps R_{in}, and (c) the generalization error. The generalization error for different values of the learning rate η is shown in (d).

the arbitrary labeling of the student hidden units. The generalization error plateau is followed by a monotonic decrease towards zero once the specialization begins and the system evolves towards the optimal solution.

Curves for the time evolution of the generalization error for different values of η shown in Fig. 1d for $K = 3$ identify trapping in the symmetric subspace as a small η phenomenon. We therefore consider the equations of motion (4) in the small η regime. The term proportional to η^2 is neglected and the resulting truncated equations of motion are used to investigate a phase characterized by students of similar norms: $Q_{ii} = Q$ for all $1 \leq i \leq K$, similar correlations among themselves: $Q_{ik} = C$ for all $i \neq k$, and similar correlations with the teacher vectors: $R_{in} = R$ for all $1 \leq i, n \leq K$. The resulting dynamical equations exhibit a fixed point solution at $Q^* = C^* = 1/(2K-1)$ and $R^* = \sqrt{Q^*/K} = 1/\sqrt{K(2K-1)}$. The corresponding generalization error is given by $\epsilon_g^* = (K/\pi)\left\{\pi/6 - K \arcsin\left((2K)^{-1}\right)\right\}$.

A simple geometrical picture explains the relation $Q^* = C^* = K(R^*)^2$ at the symmetric fixed point. The learning process confines the student vectors $\{\mathbf{J}_i\}$ to the subspace $\mathcal{S}_B$ spanned by the set of teacher vectors $\{\mathbf{B}_n\}$. For $T_{nm} = \delta_{nm}$ the teacher vectors form an orthonormal set: $\mathbf{B}_n = \mathbf{e}_n$, with $\mathbf{e}_n \cdot \mathbf{e}_m = \delta_{nm}$ for $1 \leq n, m \leq K$, and provide an expansion for the weight vectors of the trained student: $\mathbf{J}_i^* = \sum_n R_{in} \mathbf{e}_n$. The student-teacher overlaps R_{in} are independent of i in the symmetric phase and independent of n for an isotropic teacher: $R_{in} = R^*$ for all $1 \leq i, n \leq K$. The expansion $\mathbf{J}_i^* = R^* \sum_n \mathbf{e}_n$ results in $Q^* = C^* = K(R^*)^2$.

The length of the symmetric plateau is controlled by the degree of asymmetry in the initial conditions [2] and by the learning rate η. The small η analysis predicts trapping times inversely proportional to η, in quantitative agreement with the shrinking plateau of Fig. 1d. The increase in the height of the plateau with decreasing η is a second order effect [3], as the truncated equations of motion predict a unique value of $\epsilon_g^* = 0.0203$ at $K = 3$.

Escape from the symmetric subspace signals the onset of hidden unit specialization. As shown in Fig. 1b, the process is driven by a breaking of the uniformity of the student-teacher correlations [3]: each student node becomes increasingly specialized to a specific teacher node, while its overlap with the remaining teacher nodes decreases and eventually decays to zero. We thus distinguish between a growing overlap R between a given student node and the teacher node it begins to imitate, and decaying secondary overlaps S between the same student node and the remaining teacher nodes. Further specialization involves the decay to zero of the student-student correlations C and the growth of the norms Q of the student vectors. The student nodes can be relabeled so as to bring the matrix of student-teacher overlaps to the form $R_{in} = R\delta_{in} + S(1 - \delta_{in})$; the matrix of student-student overlaps is of the form $Q_{ik} = Q\delta_{ik} + C(1 - \delta_{ik})$.

The subsequent evolution of the system converges to an optimal solution with perfect generalization, characterized by a fixed point at $(R^*)^2 = Q^* = 1$ and $S^* = C^* = 0$, with $\epsilon_g^* = 0$. Linearization of the full equations of motion around the asymptotic fixed point results in four eigenvalues, of which only two control convergence. An initially slow mode is characterized by a negative eigenvalue that decreases monotonically with η, while an initially faster mode is characterized by an eigenvalue that eventually increases and becomes positive at $\eta_{max} = (\pi/K)[75 - 42\sqrt{3}]/[25\sqrt{3} - 42]$, to first order in $1/K$. Exponential convergence of R, S, C, and Q to their optimal values is guaranteed for all learning rates in the range $(0, \eta_{max})$; in this regime the generalization error decays exponentially to $\epsilon_g^* = 0$, with a rate controlled by the slowest decay mode.

References

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

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

[3] D. Saad and S. A. Solla, *Phys. Rev. Lett.* **74**, 4337, (1995); *Phys. Rev. E* **52**, 4225, (1995).