UNIVERSITÄT WÜRZBURG

INSTITUT FÜR THEORETISCHE PHYSIK

Am Hubland, D–97074 Würzburg, Germany

Transient dynamics of on–line learning in two–layered neural networks

Michael Biehl, Peter Riegler, and Christian Wöhler

Ref.: WUE-ITP-96-003

e-mail : biehl@physik.uni-wuerzburg.de

ftp : ftp.physik.uni-wuerzburg.de

Transient dynamics of on-line learning in two-layered neural networks

Michael Biehl Peter Riegler Christian Wöhler
Institut für Theoretische Physik
Julius–Maximilians–Universität
Am Hubland, D–97074 Würzburg, Germany
E-mail: biehl{pr}{woehler}@physik.uni-wuerzburg.de

Abstract. The dynamics of on-line learning in neural networks with continuous units is dominated by plateaus in the time dependence of the generalization error. Using tools from statistical mechanics, we show for a soft committee machine the existence of several fixed points of the dynamics of learning that gives rise to complicated behaviour, such as cascade-like runs through different plateaus with a decreasing value of the corresponding generalization error. We find learning-rate dependent phenomena, such as splitting and disappearing of fixed points of the equations of motion. The dependence of plateau lengths on the initial conditions is described analytically and simulations confirm the results.

PACS numbers: 87.10,05.40,07.05

1. Introduction

Layered neural networks [1] are used for implementing input-output maps of relevance to classification and regression tasks. Already networks with one hidden layer are sufficient to represent nontrivial scalar functions of N-dimensional variables [2]. However, the convergence of the learning process is typically very slow due to plateaus in the time dependence of the order parameters describing the state of the neural network (e.g. [3]). While there exist exact theories describing the asymptotics of the learning process for twolayer neural networks [4, 5, 6, 7], little is known about the dynamical properties for transient learning times. An understanding of the dominating processes that lead to the slow convergence in the learning process is essential for an eventual construction of algorithms that overcome these deficiencies. In this paper we examine this crucial regime for the case of on-line gradient descent learning, which is a standard algorithm widely used in practice [1, 8].

The generic architecture of networks discussed here consists of N input units, K hidden units fully connected with the input units, and one linear output unit; for simplicity, the hidden-to-output weights are fixed at unit strength ("soft committee machine" [4, 5]). However, the obtained results will be similar for networks with variable hidden-to-output weights [6].

In the theory of on-line learning [1, 9, 10, 11, 12, 13] it is assumed that a sequence of uncorrelated examples $\{\boldsymbol{\xi}^\mu, \tau^\mu\}$ of an unknown rule $\tau(\boldsymbol{\xi})$ is provided by the environment. The example input vectors are denoted by $\boldsymbol{\xi}^\mu$, and τ^μ is the corresponding correct rule output. Throughout this paper we consider input vectors $\boldsymbol{\xi} \in I\!\!R^N$ with independent identically distributed components of zero mean and unit variance.

The rule defined through a teacher network with M hidden units with a nonlinear activation function g is learned by a student network of the same architecture with K hidden units. The weights of the teacher network are denoted by the vectors $\mathbf{B}_n \in I\!\!R^N$, $n = 1, \ldots, M$, those of the student by $\mathbf{J}_i \in I\!\!R^N$, $i = 1, \ldots, K$. Given a specific teacher network, the generalization error is

$$\epsilon_g(\{\mathbf{J}_i\}) = \langle \epsilon(\{\mathbf{J}_i\}, \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}} \qquad \text{with} \qquad \epsilon(\{\mathbf{J}_i\}, \boldsymbol{\xi}) = \frac{1}{2}(\sigma - \tau)^2, \tag{1}$$

where $\sigma = \sum_{i=1}^K g\,(\mathbf{J}_i \cdot \boldsymbol{\xi})$ is the student and $\tau = \sum_{n=1}^M g\,(\mathbf{B}_n \cdot \boldsymbol{\xi})$ the teacher output, $\langle \ldots \rangle$ denoting the average over the input distribution [14, 15]. In the thermodynamic limit ($N \to \infty$) ϵ_g only depends on the *order parameters*

$$R_{in} = \mathbf{J}_i \cdot \mathbf{B}_n, \qquad Q_{ik} = \mathbf{J}_i \cdot \mathbf{J}_k, \qquad n = 1, \ldots, M, \qquad i, k = 1, \ldots, K. \tag{2}$$

Recently, learning by on-line gradient descent was studied in this framework [5, 6]. In this setting, the variation of the student weights under presentation of example $\{\boldsymbol{\xi}^\mu, \tau^\mu\}$ is given by

$$\mathbf{J}_k^{\mu+1} = \mathbf{J}_k^\mu - \frac{\eta}{N}\, \nabla_{\mathbf{J}_k} \epsilon\left(\{\mathbf{J}_i^\mu\}, \boldsymbol{\xi}^\mu\right), \tag{3}$$

which leads to the following differential equations for the order parameters:

$$\frac{dR_{in}}{d\alpha} = \eta\, \langle \delta_i y_n \rangle, \qquad \frac{dQ_{ik}}{d\alpha} = \eta\, \langle \delta_i x_k + \delta_k x_i \rangle + \eta^2 \langle \delta_i \delta_k \rangle, \tag{4}$$

where $\alpha = \mu/N$ is used as a continuous time, and

$$x_i = \mathbf{J}_i \cdot \boldsymbol{\xi}, \qquad y_n = \mathbf{B}_n \cdot \boldsymbol{\xi}, \qquad \delta_i = g'(x_i) \left[\sum_{n=1}^M g(y_n) - \sum_{i=1}^K g(x_i) \right]. \tag{5}$$

The averages are over the $(K + M)$-dimensional Gaussian distribution of the $\{x_i, y_n\}$ which is determined by the correlations

$$\langle x_i x_k \rangle = Q_{ik}, \qquad \langle x_i y_n \rangle = R_{in}, \qquad \langle y_m y_n \rangle = \mathbf{B}_m \cdot \mathbf{B}_n \equiv T_{mn}. \tag{6}$$

The function $g(x) = \mathrm{erf}\left(x/\sqrt{2}\right)$ is used as a sigmoid activation function of the hidden units [4]. With this specific choice, the averaging in the equations of motion (4) can

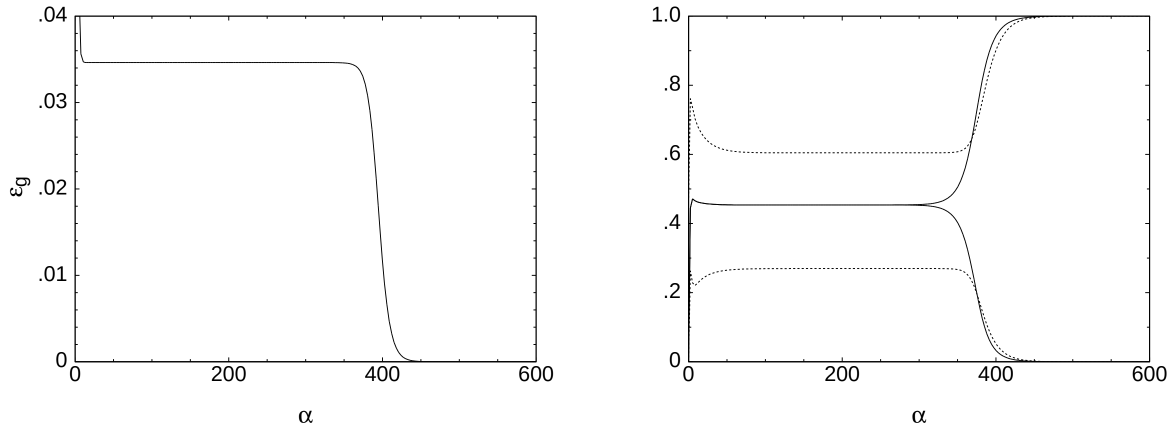

Figure 1. Time evolution of the generalization error (left) and the order parameters R_{in} (right, solid lines) and Q_{ik} (dotted lines) in the $K = M = 2$ learning scenario with an isotropic teacher ($T_{nm} = \delta_{nm}$) and $\eta = 1.5$. The initial conditions are set according to (7).

be performed analytically for general K and M, providing an exact description of the dynamics of the learning process in the thermodynamic limit, see [4, 5] for mathematical details.

In the following we will refer to a rule with $T_{nm} = \delta_{nm}$ as an *isotropic teacher*, and to one with $T_{nm} = n\ \delta_{nm}$ as a *graded teacher* [5]. However, we will generally concentrate on isotropic teachers as the examined plateau phenomena come out most clearly when learning a rule given by an isotropic teacher. We expect, however, that most of the described phenomena also occur when arbitrary sets of teacher vectors are regarded.

2. Plateau states - significant phases of the learning process

2.1. General behaviour during the learning process

A typical learning curve of a soft committee machine with two hidden units learning a rule given by an identical network with an isotropic set of teacher vectors is shown in Figure 1; we have chosen initial values of the order parameters similar to those of randomly drawn teacher and student vectors:

$$R_{in}(0) = U_{R_{in}}[0, 10^{-12}], \qquad Q_{ii}(0) = 0.5, \qquad Q_{ik}(0) = U_{Q_{ik}}[0, 10^{-12}] \quad \forall\ i \neq k, \tag{7}$$

where the expressions $U_{R_{in}}[0, X]$ and $U_{Q_{ik}}[0, X]$ denote arbitrarily chosen (different) numbers of the order $\mathcal{O}(X)$ from the corresponding intervals. For $N \to \infty$, the initial fluctuations would be exactly zero, but a nonzero value has to be taken as otherwise the system would be trapped infinitely long in the suboptimal plateau state. In practice, i. e. for finite N, they are usually much larger when using randomly drawn weight vectors as initial conditions. The general importance of the initial conditions will be regarded in the following section.

In Figure 1, the suboptimal plateau appears in the time dependence of both the generalization error and the order parameters. The "success" of the student, i. e. the time α at which the asymptotic exponential decay of the generalization error begins, is given essentially by the plateau length which will be defined in detail later on.

The plateau states correspond to configurations which are very close to certain fixed points of the set of differential equations (4) for the order parameters. The current explanation for the observed dynamical behaviour of the order parameters is the following: There is a unique fixed point which is symmetric in the sense that at least for $\eta \ll 1$ the relations $R_{in} = R$ and $Q_{ik} = Q$ for all i, k, n are valid. The corresponding generalization error is non-zero, so this symmetric state is called a *suboptimal state.* During the symmetric phase, the student vectors are almost identical and have – apart from small deviations – the same overlap with each teacher vector. This symmetric fixed point is repulsive, so small fluctuations will cause a specialization of the student vectors towards distinct teacher vectors, which then leads to the optimal state, e. g. $R_{in} = \delta_{in}$, $Q_{ik} = \delta_{ik}$, $\epsilon_g = 0$ for the learning scenario in Figure 1. Later on it will be shown that the system can reveal a much richer behaviour with the possible appearance of different plateau states which can be approached in the course of the dynamics (4).

2.2. *Definition of the plateau length, relevance of the initial conditions*

We now address the question of the plateau length. Near a fixed point, the set of differential equations (4) can be linearized in terms of the deviations of the order parameters from the corresponding values at the fixed point:

$$\frac{d}{d\alpha} \begin{pmatrix} Q_{11} \\ Q_{12} \\ \vdots \\ R_{11} \\ R_{12} \\ \vdots \end{pmatrix} = F \begin{pmatrix} Q_{11} - Q_{11}^{\mathrm{fix}} \\ Q_{12} - Q_{12}^{\mathrm{fix}} \\ \vdots \\ R_{11} - R_{11}^{\mathrm{fix}} \\ R_{12} - R_{12}^{\mathrm{fix}} \\ \vdots \end{pmatrix}, \tag{8}$$

where F is a square matrix of dimension $[KM + K(K+1)/2]$ as Q is a symmetric $K \times K$ matrix with $K(K+1)/2$ independent elements and R an unsymmetric $K \times M$ rectangular matrix.

2.2.1. Results in the thermodynamic limit The eigenvalues of F rule the behaviour of the order parameters for small deviations from their values at the fixed point. The behaviour of an arbitrary order parameter Z is therefore characterized by

$$\frac{dZ}{d\alpha} = \lambda_{\text{esc}} \left[Z(\alpha) - Z^{\text{fix}} \right] \Leftrightarrow Z(\alpha) - Z^{\text{fix}} = \tilde{X}(\alpha_0) e^{\lambda_{\text{esc}}(\alpha - \alpha_0)}, \tag{9}$$

where λ_{esc} is the eigenvalue that rules the repulsion of the order parameters away from the fixed point; this is normally the largest positive eigenvalue of the matrix F. Note that, if the corresponding eigenvector has zero components, different time constants may apply for different order parameters. $\tilde{X}$ is the deviation of Z from its value Z^{fix} at the fixed point at some arbitrary reference point α_0 in the plateau after inset of repulsion. The time α_P at which $Z(\alpha) - Z^{\text{fix}}$ exceeds a given value $B > \tilde{X}$ (which can be chosen arbitrarily) marks the end of the plateau; we thus have

$$Z(\alpha_P) - Z^{\text{fix}} = \tilde{X} e^{\lambda_{\text{esc}}(\alpha_P - \alpha_0)} = B, \tag{10}$$

which leads to the following general expression for the plateau length $\alpha_P - \alpha_0$:

$$\alpha_P - \alpha_0 = \frac{1}{\lambda_{\text{esc}}} \ln \frac{B}{\tilde{X}} = \tau_{\text{esc}} \ln \frac{B}{\tilde{X}}, \tag{11}$$

where τ_{esc} is called the escape time of the fixed point. The difference in length of the plateau obtained with different deviations $\tilde{X}_1$ and $\tilde{X}_2$ is then independent of B:

$$\alpha_P(\tilde{X}_1) - \alpha_P(\tilde{X}_2) = \tau_{\text{esc}} \ln \frac{\tilde{X}_2}{\tilde{X}_1}. \tag{12}$$

Basically, the initial deviation from symmetry is preserved while approaching the plateau, hence $\tilde{X}$ at the reference point α_0 in (9) is proportional to the *initial* deviation X of the configuration of order parameters from symmetry according to (7) or (16). In the following section, however, it will be shown that the actual behaviour in a real learning situation is still a bit more complicated.

2.2.2. Comparison to simulation data, realistic initial conditions For a comparison between simulations and the analytic results for $N \to \infty$ it is not sufficient to fix the initial configuration of the order parameters only by choosing the student and teacher weight vectors on average. In a realistic learning problem the initial values of the mutual student overlaps Q_{ik} can be fixed to arbitrary precision by choice of appropriate weight vectors. The initial student/teacher overlaps R_{in}, however, are unknown and cannot be controlled in a situation with no a priori knowledge about the teacher vectors. In order to demonstrate the logarithmic dependence of the plateau length on the initial deviations from symmetry (12), we now assume that the initial weights of a student with, say, $K = 2$ hidden units are generated randomly subject to the constraints

$$Q_{ii}(0) = \tilde{Q} \quad \text{and} \quad Q_{12}(0) = \tilde{Q} - X, \tag{13}$$

which corresponds to almost identical student vectors for $X \ll \tilde{Q}$. In the previous section it was shown how such a system gets trapped in a perfectly symmetric plateau

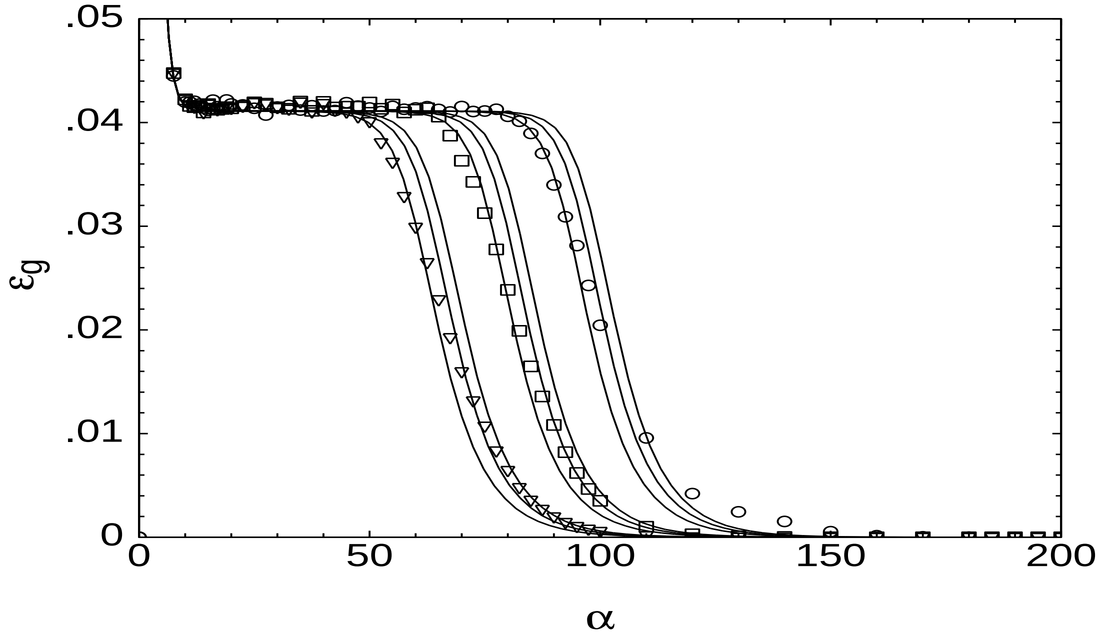

Figure 2. Logarithmic dependence of the plateau length on the initial deviations X from perfect symmetry. As in Figure 1, the learning scenario is $K = M = 2$, $T_{nm} = \delta_{nm}$ at a learning rate $\eta = 1.5$, but the initial conditions are set according to (16). The values of X used in the simulations (dots, the system size is $N = 1000$, each curve is an average over 25 runs) are (from the left) 10^{-6}, 10^{-8}, and 10^{-10}. They are compared to solutions of the equations of motion (4) (solid lines) corresponding to the same values of X. Each triple of curves shows the time dependence of ϵ_g for $N = 1000$, $N = 500$, and $N = 200$, respectively, for constant X according to (16), displaying the logarithmic dependence of the plateau length on N. The logarithmic dependence of the plateau length on X is apparent both in the simulations and in the solutions of the equations of motion (4).

the length of which is determined by $\ln X$ according to equation (11). Equation (13) then implies

$$\mid \mathbf{J}_1 - \mathbf{J}_2 \mid = \sqrt{2X} \tag{14}$$

and thus (for example)

$$R_{11} - R_{21} = (\mathbf{J}_1 - \mathbf{J}_2) \cdot \mathbf{B}_1 = A\sqrt{2X}. \tag{15}$$

with the constant $A = 0$ for random vectors of infinite dimension. However, in any finite system and for uncorrelated random teacher vectors, the R_{in} themselves as well as the factor A in (15) will be fluctuating quantities of order $\mathcal{O}(1/\sqrt{N})$, leading to the following initial conditions which force the system towards the perfectly symmetric fixed point:

$$R_{21}(0) = R_{11}(0) + U_{R_{21}}[0, X_R],$$

$$
\begin{aligned}
R_{22}(0) &= R_{12}(0) + U_{R_{22}}[0, X_R], \\
Q_{11}(0) &= Q_{22}(0) = \tilde{Q}, \\
Q_{12}(0) &= Q_{21}(0) = \tilde{Q} - X,
\end{aligned} \tag{16}
$$

where $X_R = \sqrt{2X/N}$. For simulations and numerical calculations the initial value $\tilde{Q}$ was chosen to be $\tilde{Q} = 0.5$. Thus, the deviations from symmetry in terms of the student/teacher overlaps are inevitably determined by the choice of X in equation (13) and (16). According to (15) they should dominate as $\sqrt{X} \gg X$ and govern the length of the observed symmetric plateau; in analogy to equation (11) one therefore expects the plateau length

$$
\alpha_P - \alpha_0 = \tau_{\rm esc} \ln\left(B\sqrt{\frac{N}{2\tilde{X}}}\right) = \tau_{\rm esc}\left(D - \frac{1}{2}\ln(2X) + \frac{1}{2}\ln N\right), \tag{17}
$$

where the constant D is of the order $\mathcal{O}(1)$ and contains the arbitrarily chosen value of B denoting the end of the plateau as well as the proportionality constant between X and $\tilde{X}$ at the reference point α_0. Hence, the explicit N-dependence is very weak, yet the existence of fluctuations in the finite system drastically affects the dynamics. Note that the naive thermodynamic result (11) is not recovered in the limit $N \to \infty$. Indeed, the observed length of plateaus in Figure 2 (and even more accurately, the difference in length for different X) is half the value predicted by the naive thermodynamic limit (cf. equation (12)).

Figure 2 was obtained by setting $R_{11}(0) = R_{12}(0) = 0$; the actual plateau length, however, does not depend on the particular choice of $R_{11}(0)$ and $R_{12}(0)$, but only on the small initial deviations of the order $\mathcal{O}(\sqrt{X/N})$ from symmetry. Therefore, not only the escape time of the fixed point, but also the properties of the initial configuration of the order parameters play a crucial role for the plateau length and thus for the duration of the learning process. This is in contrast to the conjecture of [5], that "the specific values assigned to the order parameters as initial conditions are largely irrelevant".

The numerical integration of the differential equations (4) with initial values of type (16) yields learning curves which are in excellent agreement with the simulations. We have chosen the initial conditions (16) rather than those corresponding to purely randomly drawn initial student weight vectors (7) with $Q_{ii}(0) = 0.5$ and $Q_{12}(0) = X$ as in this case a calculation of an asymmetry parameter $X_R^{\rm random}$ in analogy to (16) gives

$$
X_R^{\rm random} = \sqrt{\frac{1-2X}{N}} \approx \frac{1}{\sqrt{N}} \tag{18}
$$

if $X \ll 1$. Thus, simulations with initial conditions of type (7) will show no X dependence of the plateau length, but only a logarithmic N dependence which has already been observed in [4] (cf. Figure 3 therein).

3. The existence of several fixed points of the dynamics

3.1. *General properties, physical validity*

For all student and teacher network sizes $K \geq 2$ and $M \geq 1$ the set of differential equations (4) reveals several roots, each of which is a fixed point of the dynamical behaviour of the order parameters. These fixed points are marked by different values of the corresponding generalization error and different degrees of symmetry. The fixed points were found by using the multi-dimensional Newton-Raphson method with randomly chosen initial values of the order parameters; the obtained results are thus the numerical values of the configuration $\{R_{in}, Q_{ik}\}$ corresponding to the different fixed points.

The completely symmetric fixed point with $R_{in} = R$, $Q_{ik} = Q$ always appears, but also other ones with more irregular values of the order parameters. It must be stressed, however, that not every fixed point corresponds to a physically valid solution, e. g. for geometrical reasons the relation $Q_{12}^2 \leq Q_{11}Q_{22}$ must always be satisfied. Many more similar geometrical constraints of the order parameters can be derived, which are all summed up in the condition that the correlation matrix

$$C = \begin{pmatrix} Q & R \\ R^T & T \end{pmatrix} \tag{19}$$

be positive semidefinite, i. e. all its eigenvalues have to be nonnegative. If this is not the case for a certain fixed point, the network cannot be found in the corresponding configuration of order parameters; the dynamics according to (4) will never approach the state, provided the initial conditions satisfy the constraint stated above. The number of physically valid fixed points is shown in Table 1 for different learning scenarios in the case of an isotropic teacher. The number was determined at intermediate values of η, and only fixed points with different values of ϵ_g are distinguished.

The choice of initial conditions then determines which fixed points are actually being observed as plateau states during the learning process.

3.2. *Learning-rate dependent phenomena*

The values of the order parameters corresponding to a certain fixed point vary with the learning rate η. We restrict ourselves to learning rates $\eta < \eta_c$, where η_c is the learning rate above which perfect learning cannot be achieved for realizable rules [4, 6]. In order to investigate the validity of the $\eta \ll 1$ approximation in [5], we first examine the behaviour of the fixed points at small learning rates.

3.2.1. Small learning rates In the small η regime, the description in [5] is valid for small student and teacher networks ($K = 2$, $M \leq 2$): The two existing suboptimal

Table 1. Number of physically valid fixed points with different values of ϵ_g in various learning scenarios with an isotropic teacher ($T_{nm} = \delta_{nm}$) for intermediate values of η, as found by the numerical procedure described in 3.1. The trivial fixed point with $Q_{ik} = 0$ for all i, k is not counted.

K	M	classification	η	# f. p.
2	1	over-realizable	1.5	3
2	2	realizable	1.5	3
2	3	unrealizable	1.0	4
3	2	over-realizable	1.0	14
3	3	realizable	1.0	10

fixed points merge to the same symmetric configuration of order parameters, i. e. $R_{in} = R$, $Q_{ik} = Q$ (Figure 3) in the limit $\eta \to 0$. From the equations of motion (4) for arbitrary K and M the existence of a completely symmetric fixed point can be shown analytically in the small η regime; a numerical evaluation reveals that it exists as well at intermediate learning rates at least for $K,\ M \leq 20$. This is a consequence of the fact that an identity of all student weight vectors is preserved under application of training algorithm (3) independent of the particular learning scenario. However, in other learning scenarios with larger networks ($K,\ M \geq 3$), there are additional unsymmetric fixed points even in the $\eta \to 0$ limit (Figure 3). These are not taken into account in Reference [5] because there the condition $R_{in} = R,\ Q_{ik} = Q$ is used as an *ansatz* which then reveals the analytical values of R and Q. Our approach is more general, but only gives numerical results.

3.2.2. Intermediate values of the learning rate The configuration of order parameters corresponding to a certain fixed point changes with the value of η. In addition to such smooth variations, the following striking discontinuous effects are observed:

- Splitting of a single fixed point into two distinct fixed points at a certain value η_s (Figure 4). This bifurkation is non-smooth, but with an infinite slope of $R_{in}(\eta)$ at $\eta = \eta_s$.
- A fixed point disappears for learning rates smaller than a certain value.
- Two fixed points merge with decreasing η. At a value η_d, they both disappear (Figure 4).

These effects show that not even the number of fixed points is a constant for a certain learning scenario. It should be emphasized that all the effects mentioned above can be observed within the range of "useful" learning rates, e. g. for $\eta < \eta_c$ in the case of realizable rules. The instability of the optimal fixed point and the occurrence of

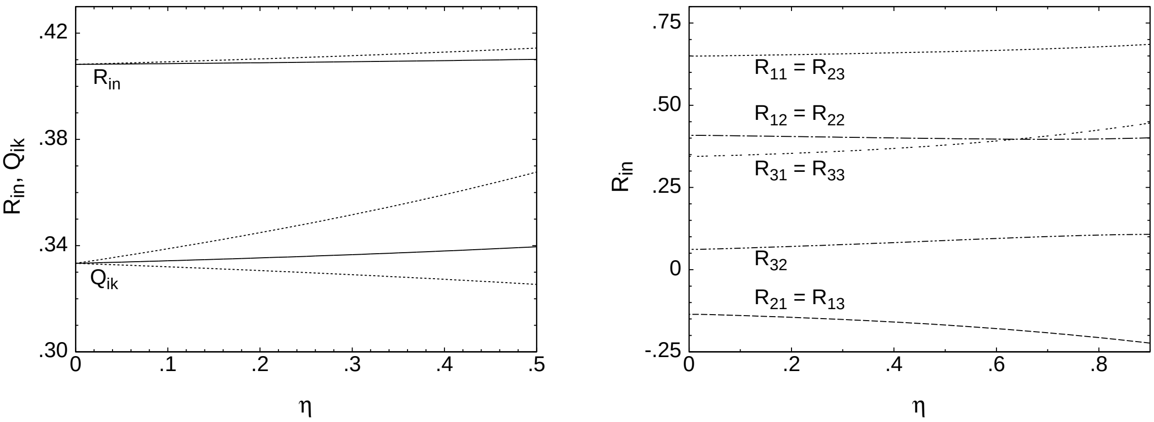

Figure 3. Left: Merging of the two suboptimal fixed points of the $K = M = 2$, $T_{nm} = \delta_{nm}$ scenario. The upper two lines represent the corresponding values of the R_{in}, the lower three lines the Q_{ik} values. Solid lines belong to the perfectly symmetric fixed point, dotted lines to the less symmetric one with $Q_{ii} \neq Q_{12}$. Right: R_{in} for one of the unsymmetric fixed points of the $K = M = 3$, $T_{nm} = \delta_{nm}$ scenario that do not converge towards the perfectly symmetric configuration for $\eta \to 0$.

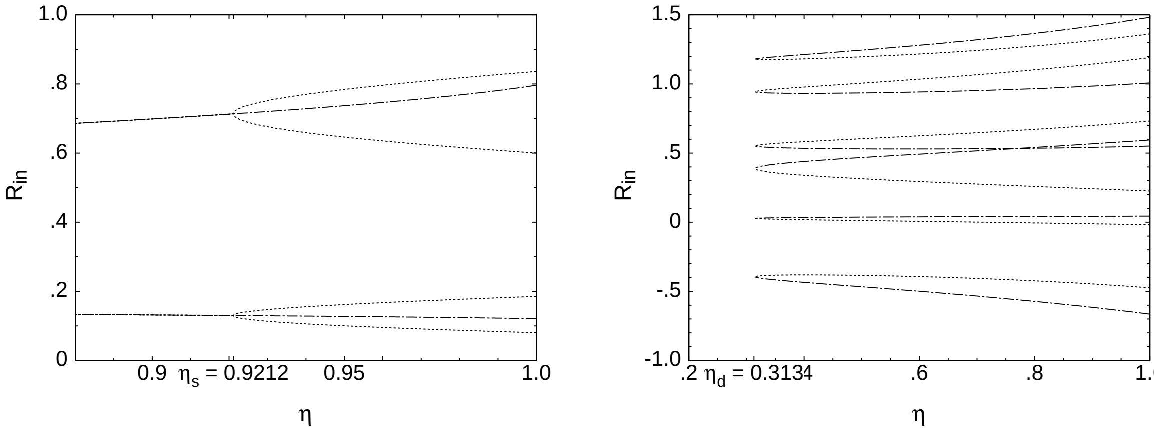

Figure 4. Left: Splitting of a single fixed point into two distinct fixed points at $\eta = \eta_s$ in the over-realizable $K = 3$, $M = 2$, $T_{nm} = \delta_{nm}$ scenario. The R_{in} of the second fixed point reveal an infinite slope at $\eta = \eta_s$. Right: Two fixed points in the realizable $K = M = 3$ scenario with a graded teacher ($T_{nm} = n\ \delta_{nm}$), approaching each other with decreasing η, then merging and disappearing at $\eta = \eta_d$. Single fixed points have as well been observed to disappear suddenly at a certain η value.

attractive suboptimal states for $\eta > \eta_c$ is not analyzed here, see [4] for such a discussion for small networks.

These observations give rise to the speculation that it should be very difficult, if not impossible, to derive a schedule for an optimal, time-dependent learning rate $\eta(\alpha)$ by means of a simple variational principle.

3.3. Dynamical effects resulting from the existence of several fixed points

As expected, the existence of several repulsive fixed points results in various dynamical effects. The first one discussed here is reminiscent of an observation made in [16] in the context of learning with a committee machine with binary threshold units. In certain settings (initial conditions, learning rate) the configuration of order parameters can vary strongly with increasing α while the generalization error remains approximately constant. This type of behaviour occurs when the initial conditions of the system are such that the network is trapped subsequently close to several fixed points with decreasing (yet similar) values of ϵ_g, but completely different sets of order parameters. Such a wave-like evolution of the generalization error was also observed in [5] for learning from a graded teacher ($T_{nm} = \delta_{nm}$), choosing the initial conditions (7).
But even in the isotropic setting with $K = M = 2$ and $T_{nm} = \delta_{nm}$ a cascade-like learning curve is always found when specific initial conditions are prepared. In Figure 5, this behaviour is shown for different learning rates and the initial conditions (16) which correspond to almost identical student weight vectors.

The value of X used in Figure 5 is $X = 10^{-10}$. Following the discussion of the previous section, we choose $X_R = \sqrt{2X/N}$ with $N = 2000$ and $R_{11}(0) = R_{12}(0) = 0$. In the particular example of $\eta = 2.0$, the system first "visits" a completely symmetric state given by

$$R_{\rm fix}^{(1)} = \begin{pmatrix} 0.487 & 0.487 \\ 0.487 & 0.487 \end{pmatrix}, \qquad Q_{\rm fix}^{(1)} = \begin{pmatrix} 0.606 & 0.606 \\ 0.606 & 0.606 \end{pmatrix}, \qquad \epsilon^{(1)}_{g,\,\rm fix} = 0.124, \tag{20}$$

where it would stay trapped for $X = 0$. Due to the imposed deviations, however, it then approaches the less symmetric fixed point with

$$R_{\rm fix}^{(2)} = \begin{pmatrix} 0.520 & 0.520 \\ 0.520 & 0.520 \end{pmatrix}, \qquad Q_{\rm fix}^{(2)} = \begin{pmatrix} 1.081 & 0.133 \\ 0.133 & 1.081 \end{pmatrix}, \qquad \epsilon^{(2)}_{g,\,\rm fix} = 0.0652, \tag{21}$$

and finally evolves towards the attractive perfect solution with $\epsilon_g = 0$. The difference in "height" $\left(\epsilon^{(1)}_{g,\,\rm fix} - \epsilon^{(2)}_{g,\,\rm fix}\right)$ is η-dependent and can be rather small, yielding a learning curve very similar to those described in [16]. This is compared to simulations (Figure 5) in which the value of X is as well fixed to 10^{-10} by imposing the initial conditions exactly on a set of random weight vectors by means of a generalization of the Gram-Schmidt-orthogonalization. The cascade-like run through two different plateau states only comes out clearly in single-run simulations as shown in Figure 5, because as a consequence of the finite dimension of the system, some single runs show this behaviour while others do

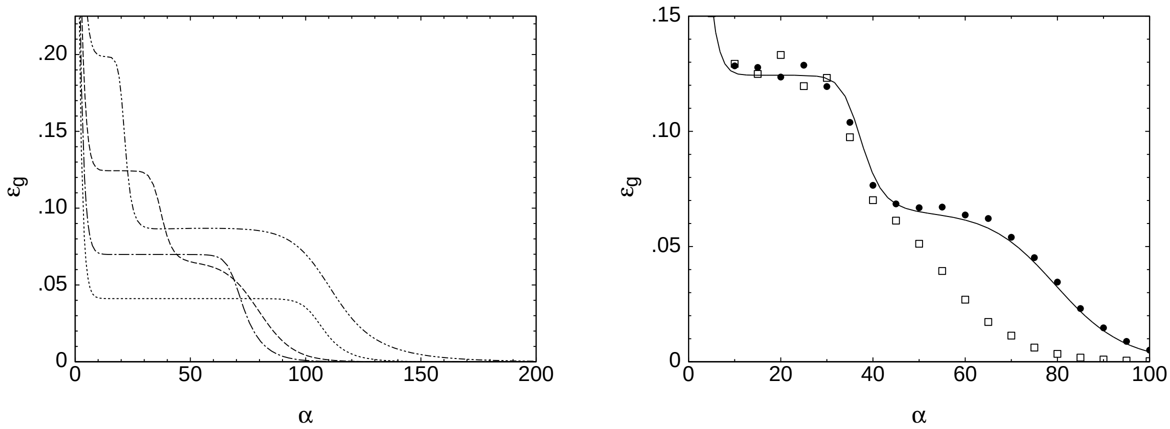

Figure 5. Left: Time evolution of the generalization error in the $K = M = 2$, $T_{nm} = \delta_{nm}$ scenario at (from below) $\eta = 1.5$, 1.75, 2.0, 2.25 (numerical solutions of the equations of motion (4)). Initial conditions are set according to (16) with $X = 10^{-10}$, $R_{11} = R_{12} = 0$, $R_{21} = 0.2 \cdot X_R$ and $R_{22} = X_R$, where $X_R = (2X/N)^{1/2}$ and $N = 2000$. Right: The cascade-like behaviour is also observed in simulations ($\eta = 2.0$). The simulations shown are single runs of a system with $N = 2000$; the corresponding numerical solution of the equations of motion (4) (solid line) is in excellent agreement with the single run displaying both plateaus.

not, so that the effect is partially wiped out when averaging over many simulation runs. The fraction of simulation runs displaying both plateaus is rising with increasing N, which means that the two fixed points are indeed relevant for the dynamical behaviour of networks with a large number of input units under realistic circumstances. As in the previous section, the plateau length is given by (17) and is thus governed by $\ln\sqrt{X/N}$. In this setting, nonzero initial values of R_{11} and R_{12} result in learning curves which are identical to the ones shown in Figure 5; the learning curve is determined only by the choice of X and N.

3.4. The repulsive properties of the fixed points

We have systematically examined the eigenvalues of all the fixed points found by evaluating the matrix F obtained by a linearization of the differential equations (4) around the different fixed points. It appears to be a general property that the most symmetric fixed point always has both the largest positive eigenvalue and the highest number of positive eigenvalues (or complex eigenvalues with a positive real part), leading to a relatively strong repulsive behaviour. Regarding the other suboptimal fixed points, we always find at least one repulsive eigenvalue preventing the student from being caught in a suboptimal state *except in one very special case*: In the learning scenario $K = 3$, $M = 2$ there is one quite unsymmetric suboptimal fixed point with no positive eigenvalue, i. e., once having approached this fixed point, the student will never escape into the optimal state. However, this situation seems to appear rarely, and very carefully chosen initial conditions are necessary to encounter this exceptional fixed point. It is an open question if the existence of purely attractive suboptimal fixed points is typical of over-realizable scenarios for larger K and M.

4. Summary and Conclusion

We have investigated the occurrence of a variety of fixed points of the dynamics (4) of the learning process of a soft committee machine with K hidden units learning a rule defined through a teacher network of an identical architecture, but with M hidden units. Apart from the completely symmetric fixed point already discussed in [5], several further less symmetric fixed points of (4) arise in any learning scenario, even in very simple ones with small values of K and M.

We have analyzed the crucial importance of the initial conditions for the convergence of the learning process towards the optimal state characterized by a minimum of the generalization error. We give an analytical expression for the length of the observed learning plateaus which depends on both the initial deviations of the order parameters from the fixed point's symmetry and the escape time of the fixed points following from a linearization of the equations of motion (4). The obtained results enable us to extract the finite-size effects visible when comparing the analytic results to simulations: The initial configuration of the order parameters must be given exactly not only in the model, but also in the simulations in order to obtain comparable results. Our simulations are in excellent agreement with the behaviour predicted by the thermodynamic model.

In an examination of the variety of fixed points in several learning scenarios we show that in general, the number of fixed points in a given scenario not even remains constant under variation of the learning rate η. The dynamical effects resulting from the existence of several fixed points, such as cascade-like runs through two or more learning plateaus after preparing the initial conditions in a special manner, were examined both in the thermodynamic model ($N \to \infty$) and in simulations.

We then study the repulsive and attractive properties of the observed fixed points of the dynamics. It becomes clear that nearly all suboptimal fixed points have at least one repulsive eigenvalue except one very asymmetric one in the $K = 3,\ M = 2$ scenario which is purely attractive in spite of displaying a non-zero generalization error. For all learning scenarios taken into account, a completely symmetric fixed point with $R_{in} = R,\ Q_{ik} = Q$ (i. e. identical student weight vectors) exists which always possesses both the largest positive eigenvalue and the highest number of positive eigenvalues when comparing it to the other, less symmetric suboptimal fixed points.

A possible strategy for an efficient reduction of the plateau length is thus to prepare the initial conditions in the way given by (16) with, say, $X \approx 10^{-3}$. This is

realizable in practice as nothing has to be known about the teacher vectors. Then the initial fluctuations X are still small enough to force the student to approach the most symmetric and most repulsive fixed point, but also large enough to guarantee a quick escape into the optimal final state. This could be a successful strategy for obtaining a higher efficiency of the on-line gradient descent algorithm under rather general circumstances, as the completely symmetric, highly repulsive fixed point of the dynamics (4) exists for realizable, over-realizable and non-realizable scenarios.
Combinations of this method with other symmetry-breaking mechanisms [17] should be examined in order to provide a tool widely usable in practice to obtain a better effectivity of multilayer network training. Moreover, it will be interesting to study the effects of the existence of many fixed points in large networks ($K,\ M \gg 1$), the dependence of the number of fixed points on the values of K and M, and the relevance of the fixed points for the dynamics of the learning process in such more general learning scenarios.

We thank W. Kinzel, G. Reents and R. Urbanzcik for useful discussions and a critical reading of the manuscript. We are grateful to H. Sompolinsky for bringing to our attention the effect observed in [16]. C. W. was supported by the Studienstiftung des deutschen Volkes, P. R. by the Deutsche Forschungsgemeinschaft. Simulations were performed on the Cray Y-MP-EL of the Rechenzentrum der Universität Würzburg.

[1] J.A. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison–Wesley (Redwood City, CA), 1991.
[2] G. Cybenko, Math. Control Signals Systems **2**, 303 (1989).
[3] Y.K. Kim and J.B. Ra, in *International Joint Conference on Neural Networks*, IEEE (New York), 2396 (1991)
[4] M. Biehl and H. Schwarze, J.Phys. A **28**, 643 (1995).
[5] D. Saad and S. A. Solla, Phys.Rev.Lett. **74**, 4337 (1995).
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] W. Wiegerinck, T. M. Heskes, preprint, 1996.
[8] Y. Chauvin and D.E. Rumelhart (eds.), *Backpropagation: Theory, Architectures, and Applications*, Erlbaum (Hillsdale, NJ), 1995.
[9] S. Amari 1967 IEEE Trans. Elect. Comput. **EC-16**, 299; 1993 Neurocomp. **5**, 185.
[10] O. Kinouchi and N. Caticha, J.Phys. A **25**, 6243 (1992).
[11] M. Copelli and N. Caticha, J.Phys. A **28**, 1615 (1995).
[12] T. M. Heskes and B. Kappen, Phys.Rev. A **26**, 63 (1993).
[13] N. Barkai, H. S. Seung, H. Sompolinsky, Phys.Rev.Lett. **75**, 1415 (1995).
[14] T.L.H. Watkin, A. Rau, and M. Biehl, Rev.Mod.Phys. **65**, 499 (1993).
[15] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys.Rev. A **45**, 6056 (1992).
[16] H. Sompolinsky, N. Barkai, and H. S. Seung, in: *Neural Networks: The Statistical Mechanics Perspective*, eds. J.-H. Oh, C. Kwon, and S. Cho, Word Scientific (Singapore), 1995.
[17] D. Barber, P. Sollich, D. Saad, preprint, 1995.