Analysis of on-line training with optimal learning rates

Magnus Rattray
Computer Science Department, University of Manchester, Manchester M13 9PL, United Kingdom

David Saad
Neural Computing Research Group, Aston University, Birmingham B4 7ET, United Kingdom

(Received 16 June 1998)

We describe a theoretical method of determining optimal learning rates for on-line gradient descent training of a multilayer neural network (a soft committee machine). A variational approach is used to determine the time-dependent learning rate which maximizes the total decrease in generalization error over a fixed time window, using a statistical mechanics description of the learning process which is exact in the limit of large input dimension. A linear analysis around transient and asymptotic fixed points of the dynamics provides insight into the optimization process and explains the excellent agreement between our results and independent results for isotropic, realizable tasks. This allows a rather general characterization of the optimal learning rate dynamics within each phase of learning (we discuss scaling laws with respect to task complexity in particular). Our method can also be used to optimize other parameters and learning rules, and we briefly consider a generalized algorithm in which weights associated with different hidden nodes can be assigned different learning rates. The optimal settings in this case suggest that such an algorithm can significantly outperform standard gradient descent. [S1063-651X(98)08511-0]

PACS number(s): 87.10.+e, 02.50.−r, 05.20.−y

I. INTRODUCTION

Neural networks are the subject of much current research regarding their ability to learn nontrivial mappings from examples (see, for example, [1]). Specifically, we will 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) which may be corrupted by noise. This provides a rather general learning scenario since both the student and teacher can represent a very broad class of functions [2]. 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 on-line learning, which is one of the most popular training methods for feed-forward neural networks, and in particular we focus on stochastic gradient descent learning. The training error is defined to be some measure of discrepancy between the teacher and student and at each learning step the student network's weights are adapted in the direction of negative gradient of this error, calculated according to only the latest in a sequence of training examples. This process is inherently stochastic because a new training example is selected at random each time the training error is determined. This is to be contrasted with batch learning, in which all the training examples are used to determine the training error, leading to a deterministic algorithm. On-line learning can be beneficial in terms of both storage and computation time for large systems.

In this paper we describe a theoretical method of determining optimal learning rates for on-line gradient descent (preliminary results from this work have been reported in [3]). On-line algorithms are often sensitive to the choice of learning parameters and for gradient descent in particular the choice of learning rate can be critical. If the learning rate is chosen too large then the learning process may diverge, but if the learning rate is too low then convergence can take an extremely long time; moreover, in either case the algorithm may get trapped at a suboptimal fixed point. The appropriate learning rate will also vary substantially over time and may require annealing towards the end of the learning process. We employ a framework recently developed for analyzing on-line learning using methods from statistical mechanics [4,5] in order to determine the time-dependent learning rate which provides the maximum decrease in generalization error over the entire learning process. In addition, this method can also be generalized to optimize other parameters and learning rules for both smooth and discrete architectures [6–8]. As an example we briefly consider a generalized algorithm in which weights associated with different hidden nodes can have different learning rates.

An important issue addressed here is the differentiation between local and global optimization. A locally optimal, or greedy, learning rate can be chosen which maximizes the decrease in generalization error at each learning step. This will be far from optimal in many cases, especially when the dynamics is characterized by phases of different nature. For example, it has been shown that the learning time in a multilayer network can be dominated by a symmetric phase in which the student is trapped in a subspace characterized by lack of differentiation between student vectors, resulting in a suboptimal generalization error [4,5]. The *locally optimal* procedure is then to anneal the learning rate towards zero, in which case the student may never leave the symmetric subspace and perfect learning cannot be achieved. In contrast to this, *global optimization* leads to a learning rate which provides the fastest possible escape from the symmetric phase. We will also show how local optimization of the

learning rate may even be suboptimal at late times.

The paper is organized as follows. We first briefly describe a theoretical framework for modeling on-line learning in a soft committee machine (a two-layer network with fixed output weights) in the limit of large input dimension, which uses methods from statistical mechanics. The optimal time-dependent learning rate is then derived for this case, using a variational calculation to optimize the total change in generalization error over a fixed time window. We study the dynamics with the optimal learning rate numerically for realizable noiseless learning and for learning from noise-corrupted examples (output noise). The algorithm is analyzed in the neighborhood of fixed points which dominate the dynamical trajectory and links are made with recent numerical and analytical studies of these fixed points [9] which provide a general characterization of the optimal learning rate dynamics within each phase of learning (we discuss scaling laws with respect to task complexity as an example). Finally, we show how our variational approach can be generalized in order to deal with site-dependent learning rates, leading to some interesting observations.

II. THE GENERAL FRAMEWORK

The method presented in this paper may be applied to optimize training parameters and learning rules in general when the on-line learning dynamics can be represented by differential equations for a set of order parameters [6]. However, we restrict our analysis here to gradient descent learning in a soft committee machine [4] and in this section we establish a framework to describe the learning process in this case.

We consider a student mapping from an N-dimensional input space $\boldsymbol{\xi}\in\mathbb{R}^N$ onto a scalar function $\sigma(\mathbf{J},\boldsymbol{\xi})=\Sigma_{i=1}^{K} g(\mathbf{J}_i\cdot\boldsymbol{\xi})$, which represents a soft committee machine, where $g(x)\equiv\mathrm{erf}(x/\sqrt{2})$ is the activation function of the hidden units, $\mathbf{J}\equiv\{\mathbf{J}_i\}_{1\leqslant i\leqslant K}$ 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 general configuration represents most properties of a general multilayer network and can easily be extended to accommodate adaptive hidden-to-output weights [10,11] (we briefly consider this case in Sec. III).

Training examples are of the form $(\boldsymbol{\xi}^\mu,\zeta^\mu)$ where $\mu=1,2,\ldots$ labels each independently drawn example in a sequence and components of the 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 which may be corrupted by output noise and is of a similar configuration to the student except for a possible difference in the number M of hidden units: $\zeta^\mu=\Sigma_{n=1}^{M} g(\mathbf{B}_n\cdot\boldsymbol{\xi}^\mu)+\rho^\mu$, where $\mathbf{B}\equiv\{\mathbf{B}_n\}_{1\leqslant n\leqslant M}$ is the set of input-to-hidden adaptive weights for teacher hidden nodes and ρ^μ is zero mean Gaussian noise of variance σ^2. The activation of hidden node n in the teacher under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is denoted $y_n^\mu=\mathbf{B}_n\cdot\boldsymbol{\xi}^\mu$. We will use indices i,j,k,l to refer to units in the student network and n,m for units in the teacher network.

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

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

This training error is then used to define the learning dynamics via a gradient descent rule for the update of student weights $\mathbf{J}_i^{\mu+1}=\mathbf{J}_i^\mu+(\eta/N)\delta_i^\mu\boldsymbol{\xi}^\mu$, where $\delta_i^\mu\equiv g'(x_i^\mu)[\Sigma_{n=1}^{M} g(y_n^\mu)-\Sigma_{j=1}^{K} g(x_j^\mu)+\rho^\mu]$ and the learning rate η has been scaled with the input size N. Performance on a typical input in the absence of noise defines the generalization error $\epsilon_g(\mathbf{J})\equiv\langle\epsilon(\mathbf{J},\boldsymbol{\xi})\rangle_{\{\xi\}}|_{\sigma=0}$ through an average over the distribution of input vectors.

Expressions for the generalization error and learning dynamics have been obtained [5] 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:

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

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 [5]. The equations of motion, depending on a closed set of parameters, can be integrated and iteratively solved, providing a full description of the order parameter evolution from which the evolution of the generalization error can be derived. Although the dynamical equations considered here are only strictly valid in the large N limit, they have been shown to describe mean behavior accurately for systems of realistic size [12].

III. GLOBALLY OPTIMAL LEARNING RATE

If we consider the fastest rate of decrease in generalization error as a measure of optimality, it is straightforward to find the *locally optimal* learning rate by determining the value of η that minimizes $d\epsilon_g/d\alpha$, using the equations of motion for R and Q and the fact that the change in generalization error over time depends exclusively on these quantities. The expression obtained for the locally optimal learning rate is then

$$\eta=-\frac{\sum_{in}(\partial\epsilon_g/\partial R_{in})\phi_{in}+\sum_{ik}(\partial\epsilon_g/\partial Q_{ik})\psi_{ik}}{2\sum_{ik}(\partial\epsilon_g/\partial Q_{ik})\upsilon_{ik}}. \quad (3)$$

Although the value of η obtained in this manner may be useful for some phases of the learning process it is likely to be useless for others. For example, the lowest generalization

error for the symmetric phase, characterized by a lack of differentiation between the student nodes, is achieved by gradually reducing the learning rate towards zero; however, decaying the learning rate in the symmetric phase will prevent the system from escaping the symmetric fixed point, thus resulting in a suboptimal solution.

A more useful measure of optimality is the total reduction in generalization error over the entire learning process. 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. We write the change in generalization error as an integral,

$$\Delta\epsilon_g(\eta)=\int_{\alpha_0}^{\alpha_1}\frac{d\epsilon_g}{d\alpha}\,d\alpha=\int_{\alpha_0}^{\alpha_1}\mathcal{L}(\eta,\alpha)d\alpha. \tag{4}$$

This is a functional of the learning rate which we will minimize by a variational calculation. Since the generalization error depends solely on the overlaps Q, R, and T, which are the dynamical variables, we can expand the integrand in terms of these variables,

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

The last two terms in Eq. (5) 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 Eq. (4) with respect to the dynamical variables leads to a set of coupled differential equations for the Lagrange multipliers,

$$\begin{aligned}\frac{d\mu_{jm}}{d\alpha}&=-\eta\sum_{in}\mu_{in}\frac{\partial\phi_{in}}{\partial R_{jm}}-\eta\sum_{ik}\nu_{ik}\frac{\partial(\psi_{ik}+\eta\upsilon_{ik})}{\partial R_{jm}},\\\frac{d\nu_{jl}}{d\alpha}&=-\eta\sum_{in}\mu_{in}\frac{\partial\phi_{in}}{\partial Q_{jl}}-\eta\sum_{ik}\nu_{ik}\frac{\partial(\psi_{ik}+\eta\upsilon_{ik})}{\partial Q_{jl}},\end{aligned} \tag{6}$$

along with a set of boundary conditions,

$$\mu_{in}(\alpha_1)=\left.\frac{\partial\epsilon_g}{\partial R_{in}}\right|_{\alpha_1}\quad\text{and}\quad\nu_{ik}(\alpha_1)=\left.\frac{\partial\epsilon_g}{\partial Q_{ik}}\right|_{\alpha_1}. \tag{7}$$

Then taking variations with respect to η we find a simple expression for the globally optimal learning rate,

$$\eta=-\frac{\sum_{in}\mu_{in}\phi_{in}+\sum_{ik}\nu_{ik}\psi_{ik}}{2\sum_{ik}\nu_{ik}\upsilon_{ik}}. \tag{8}$$

Equations (6), (7), and (8) determine necessary conditions for η to maximize the reduction in generalization error over the interval $[\alpha_0,\alpha_1]$. The boundary conditions correspond to the locally optimal solution in Eq. (3), reflecting the fact that at α_1 the choice of η does not affect the dynamics at other times. To find the learning rate which satisfies this set of conditions we use gradient descent on the functional derivative of $\Delta\epsilon_g$ with respect to η,

$$\begin{aligned}\eta(t+1)&=\eta(t)-\Theta\frac{\delta\Delta\epsilon_g}{\delta\eta},\\\frac{\delta\Delta\epsilon_g}{\delta\eta}&=\sum_{in}\mu_{in}\phi_{in}+\sum_{ik}\nu_{ik}(\psi_{ik}+2\eta\upsilon_{ik}),\end{aligned} \tag{9}$$

where t is the iteration index and Θ is the step size for the iteration process. In order to choose an appropriate value for Θ we employ second order variations,

$$\Theta\propto\left(\frac{\delta^2\Delta\epsilon_g}{\delta\eta^2}\right)^{-1}=\left(2\sum_{ik}\nu_{ik}\upsilon_{ik}\right)^{-1}. \tag{10}$$

Standard heuristics can be used to ensure that the iteration process does not diverge if the second order variations become negative, or close to zero.

All terms required for determining the functional derivatives in Eqs. (9) and (10) can be obtained by integrating the equations for the overlaps forward, using Eqs. (2) and some initial conditions, and then backwards for the Lagrange multipliers, using Eqs. (6) and the boundary conditions expressed in Eq. (7). In our implementation the overlaps are stored during the forward dynamics and reused during the backwards dynamics for the Lagrange multipliers. This process converges within a few iterations and results in an exact function for the optimal learning rate over the given time window.

One limitation of the present model is the assumption of fixed hidden-to-output weights and it is straightforward to include variable hidden-to-output weights [10,11], resulting in an extra set of dynamical equations. However, if the learning rate associated with these weights is chosen to be of the same order as for the input-to-hidden weights then our optimization procedure shows that the learning rate associated with these weights should be set infinitely high, indicating that the chosen scaling is inappropriate and that learning should be on a faster time scale for these weights. This can be incorporated as an adiabatic elimination of the fast variables, as justified in [10] where it is shown that this provides a locally optimal choice for the hidden-to-output weights (a choice which minimizes the generalization error instantaneously). Our analysis therefore indicates that adiabatic elimination is also globally optimal. Since the soft committee machine considered here captures the main features of the dynamics for the remaining input-to-hidden weights, we will not consider hidden-to-output weights further in this work.

IV. NUMERICAL RESULTS

The examples presented in this section will involve students and teachers of equal complexity ($K=M$) and isotropic teachers ($T_{nm}=\delta_{nm}$), although the technique can be used

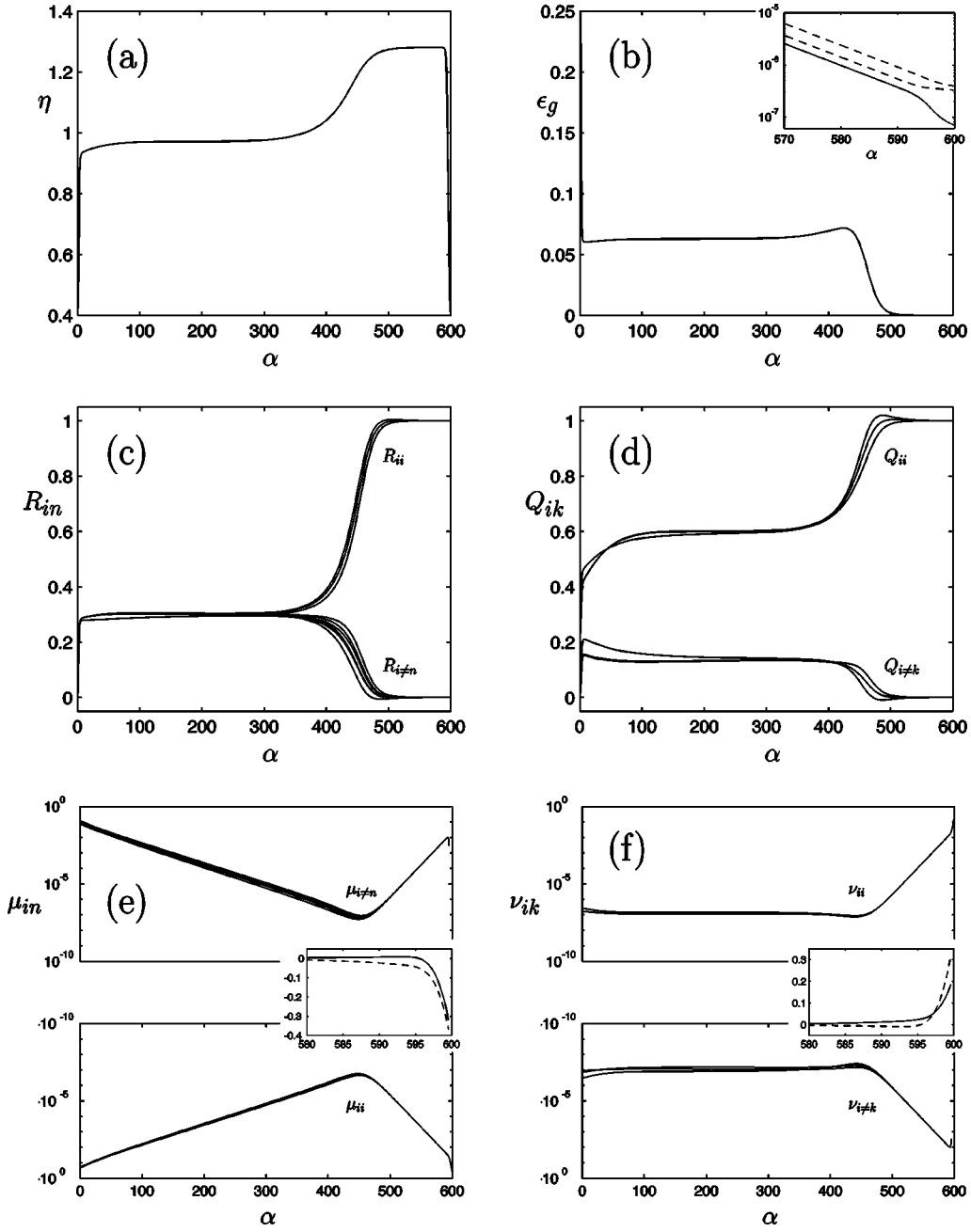

FIG. 1. Results are presented for a three hidden node student trained to emulate an isotropic teacher ($T_{nm}=\delta_{nm}$) of the same configuration. The globally optimal learning rate is shown in (a) along with the corresponding evolution of the generalization error and order parameters in (b), (c), and (d). The inset of (b) shows the generalization error (solid line) and the magnitude of the opposing contributions to the leading term (dashed lines—upper proportional to $2r-q$, lower proportional to $2s-c$). The Lagrange multipliers are shown in (e) and (f) using a log scale, with the later stages magnified in each inset (dashed line for curves associated with the lower figure).

for any soft committee machine. Anisotropic teachers are briefly considered in Sec. VI, when we introduce a site-dependent learning rate. Structurally unrealizable problems ($K<M$) exhibit qualitatively similar behavior to the noise-corrupted teacher which is considered below and are not discussed here. Initial conditions for the overlaps R_{in} and $Q_{i\neq k}$ are taken randomly from a uniform distribution $U[0,10^{-6}]$ while the vector lengths Q_{ii} are taken from $U[0,0.5]$, a choice which corresponds to an input dimension of about $N\simeq 10^{12}$. The choice of initial conditions is not critical, however, and the optimal learning rate in each phase of learning is independent of initial conditions (only the length of the transient fixed point is altered [13]). As already pointed out, the framework used here describes mean behavior accurately in much smaller systems [12].

A. Realizable rules

In our first example we consider a realizable ($K=M=3$) noiseless training task. The time window is $0\leqslant\alpha\leqslant 600$ and the learning rate is initially fixed at some arbitrary value. The update in Eq. (9) is then iterated until convergence and Fig. 1 shows results for the dynamics using the globally optimal learning rate.

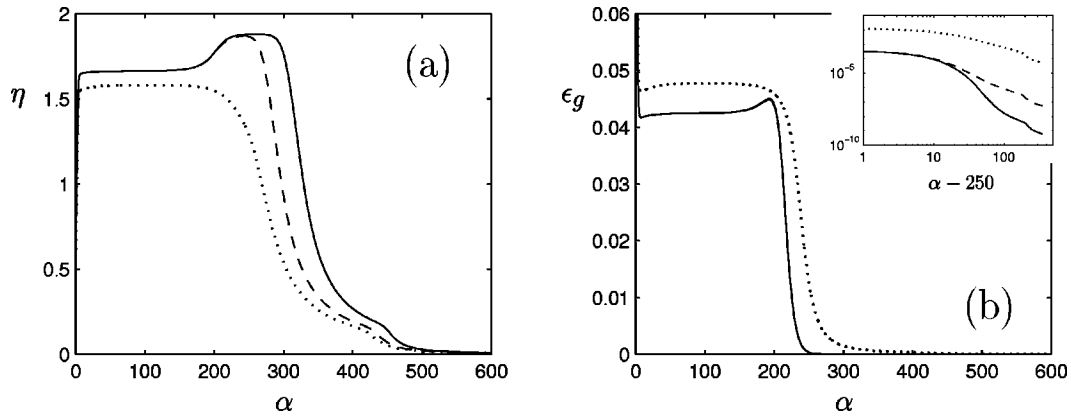

FIG. 2. A two hidden node student is trained on noise-corrupted examples generated by an isotropic teacher ($T_{nm}=\delta_{nm}$) of the same configuration. The optimal learning rate is shown in (a) for three noise levels $\sigma^2=10^{-2}$, 10^{-5}, 10^{-7} (from left to right) over a fixed time window $0\leqslant\alpha\leqslant600$. The corresponding generalization error is shown in (b) with the inset showing a log-log plot for the decay at late times, which indicates that the dashed and solid lines are only split after $\alpha\simeq300$.

Figures 1(a)–1(d) show the optimal learning rate, generalization error, and overlaps, respectively. After a short initial transient both the learning rate and generalization error stabilize at almost constant values, corresponding to a symmetric phase in which the student nodes have not yet specialized to particular teacher nodes, as required to learn perfectly. The overlaps also exhibit a plateau within this phase and Fig. 1(c) shows that the student-teacher overlaps are almost indistinguishable (the indices have been ordered *a posteriori* so that student node i eventually specializes to teacher node i). The learning rate takes a value of about $\eta\simeq0.97$ within the symmetric phase, which is in close agreement to the optimal value obtained numerically in a separate study [9]. Eventually, the student escapes the symmetric phase and the generalization error and overlaps exhibit exponential convergence towards their respective optimal values, as the learning rate increases towards another constant value of $\eta=1.28$, identical to the result obtained independently [9] for the asymptotically optimal learning rate by expanding the dynamical equations for the overlaps around their asymptotic fixed point.

Towards the end of the time window we observe a short transient in which there is an unexpected drop in the learning rate to a value of around $\eta=0.41$ [see Fig. 1(a)]. This can be explained by examining the expression for the generalization error in the vicinity of its asymptotic fixed point. It is possible to gain an immediate reduction in generalization error by choosing an appropriate direction for the decay eigenvectors. 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 are shown in the inset to Fig. 1(b) along with the corresponding generalization error for $570\leqslant\alpha\leqslant600$. By reducing the learning rate it is possible to reduce the difference in magnitude between these opposing contributions, leading to a reduction in generalization error. However, this reduction in learning rate slows down the exponential convergence of the overlaps and is therefore unsustainable in the long term. Thus this greedy drop-off in the learning rate only ever occurs towards the end of the given time window. This example shows how locally optimal learning does not necessarily give good long-term performance, even asymptotically. The long-term goal in this case is to optimize the decay rate of the order parameters, while changes in the decay direction can provide short-term gains but will eventually lead to poorer performance.

The various phases of learning described above are mirrored by the Lagrange multiplier dynamics shown in Figs. 1(e) and 1(f). Figure 1(e) shows how during the symmetric phase μ_{ii} and $\mu_{i\neq n}$ decay exponentially with similar magnitude but opposite sign. At the same time Fig. 1(f) shows that ν_{ii} and $\nu_{i\neq k}$ also have opposite signs but remain almost constant during the symmetric phase. After escaping the symmetric phase all the Lagrange multipliers exhibit an exponential growth with the same constant rate, which is equal in magnitude to the decay rate of the generalization error at this point. The inset to each figure magnifies the short transient at the end of the optimization time window in which the exponential growth is interrupted as each Lagrange multiplier finds the appropriate boundary value.

Notice that the dynamics of the overlaps and Lagrange multipliers forms a small number of bundled similar trajectories, reflecting symmetries in the task. By exploiting these symmetries the dimensionality of the system can be reduced significantly, allowing a compact description for arbitrary K and T. This dimensionality reduction has already been used to study the different phases of learning in [5,9] and in Sec. V we elucidate the relationship between our algorithm and these studies.

B. Noise-corrupted rules

In our second example we consider an unrealizable learning scenario by introducing additive uncorrelated Gaussian noise of zero mean and variance σ^2 to the teacher's output. Qualitatively similar results are obtained for structural unrealizability ($K<M$). The picture that emerges, shown in Fig. 2 for $K=M=2$ and various noise levels ($\sigma^2=10^{-2}$, 10^{-5}, and 10^{-7}), is initially similar to the realizable case but changes dramatically as the system escapes the symmetric phase. As the system begins convergence towards zero generalization error, as shown in Fig. 2(b), the optimal learning rate shown in Fig. 2(a) begins to fall and slowly approaches

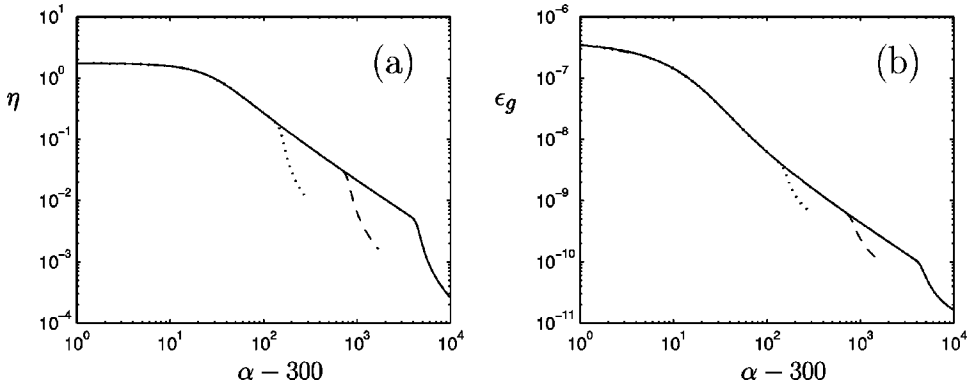

FIG. 3. As in Fig. 2, a two node student is trained on noisy examples from a teacher of the same configuration. Here the noise level is fixed at $\sigma^2 = 10^{-7}$ while the optimization process is carried out over different time windows $0 \leqslant \alpha \leqslant \alpha_1$ with $\alpha_1 = 600$, 2000, and 10^4 (from left to right). The asymptotic decay of the learning rate (a) and generalization error (b) are shown for each case. The initial behavior is similar to that presented in Fig. 2.

a decay inversely proportional to α, proved to be optimal for linear systems (see, for example, [14]), until reaching a greedy phase (after the kink around $\alpha = 440$ in Fig. 2). (Recall that this is the generalization error in the absence of noise. The prediction error for a noisy teacher has an additive constant contribution equal to half the noise variance.) This greedy reduction in generalization error is achieved by changing the decay direction, as for the realizable learning scenario described above.

Figures 3(a) and 3(b) show a log-log plot of the learning rate and generalization error, respectively, as a function of α for optimization over time windows of varying length. One observes that both the learning rate and generalization error approach a decay proportional to $1/\alpha$ and that the curves lie on top of one another until the greedy phase which occurs towards the end of each time window. However, unlike the realizable case where the drop-off in learning rate occurs over a relatively short time, here the final greedy phase increases in length as the total learning time window increases and this phase always takes a significant proportion of the learning time (this is not immediately apparent in Fig. 3 until one considers that the x axis represents a log scale). This is simply a reflection of the slower decay time scale for this problem.

Our results suggest that as symmetry breaks one should gradually modify the decay rate from a constant until it is proportional to $1/\alpha$ (in terms of a rescaled time, which is set to zero close to the point where symmetry is broken). However, it may take a prohibitively long time until the $1/\alpha$ decay becomes optimal, making it irrelevant in many instances. Moreover, if one decays the learning rate at a fixed rate it may take an extremely long time before losses, incurred due to the use of suboptimal learning rates in earlier stages of the dynamics, can be recovered. Annealing the learning rate during the symmetric phase could even lead to trapping, since the length of the symmetric phase scales inversely to η for small η [5].

V. ANALYSIS OF THE OPTIMAL LEARNING DYNAMICS

In general, the dynamical equations (2) and (6) are rather hard to analyze as they are high dimensional and strongly nonlinear. However, as we saw in the preceding section the overlap dynamics is often dominated by fixed points (the symmetric and convergence phases) around which we can make a linear expansion. We therefore carry out a simple analysis of our variational algorithm in the neighborhood of such a fixed point, leading to some valuable insight into how the algorithm optimizes performance.

It would still be rather difficult to solve the linear model for such a high-dimensional system; however, for realizable ($K = M$) learning of an isotropic task ($T_{nm} = T\delta_{nm}$) the analysis can be simplified by exploiting symmetries between the dynamical variables, thereby reducing the dimension to a manageable level and avoiding degeneracies. In this way one can determine generic behavior in terms of the variables T and K. This simplification has recently been used to determine optimal parameters for both the symmetric and convergence phases by an eigenvalue analysis around each fixed point [9] (previous results for the convergence phase [5] made use of an inaccurate assumption). Instead of rederiving many of these results, we focus on showing the close relationship between this work and the variational method and on understanding how our algorithm finds optimal parameters in the simplest scenario, in order to inform our use of the algorithm for more general problems. Details of the fixed points and linearized dynamics considered here are given in [9].

The following analysis requires that the learning rate is fixed in the phases of interest and is therefore only applicable to learning noiseless examples, at least for the convergence phase. The noise-corrupted rules considered in Sec. IV B will require a different approach, perhaps using recent results for optimal annealing schedules in the presence of noise [15]. We leave this analysis for future study.

A. Behavior near a fixed point

Let $\mathbf{y}$ be a vector of dynamical variables, which can be thought of as deviations from some fixed point. In the neighborhood of such a fixed point we have a linearized system of differential equations,

$$\frac{d\mathbf{y}}{d\alpha} = \mathbf{M}\mathbf{y}, \tag{11}$$

which corresponds to decay in the neighborhood of a stable fixed point (the convergence phase) or divergence in the

neighborhood of an unstable fixed point (the symmetric phase). Here, the matrix $\mathbf{M}$ depends on η which is taken to be constant within the region considered. Let $\mathbf{z}$ denote the associated vector of Lagrange multipliers. The linearized equivalent of Eq. (6) is then

$$\frac{d\mathbf{z}}{d\alpha} = -\mathbf{M}^T\mathbf{z}. \tag{12}$$

Let $\mathbf{U}$ denote the matrix whose columns are eigenvectors of $\mathbf{M}$ and let $\boldsymbol{\lambda}$ denote the corresponding vector of eigenvalues. Then $(\mathbf{U}^{-1})^T$ is the matrix of eigenvectors of $-\mathbf{M}^T$ with eigenvalues $-\boldsymbol{\lambda}$. We write the general solutions for $\mathbf{y}$ and $\mathbf{z}$ in component form,

$$y_i = \sum_j \mathbf{U}_{ij}\exp(\alpha\lambda_j), \quad z_i = z_{0i} + \sum_j \beta_j \mathbf{U}^{-1}_{ji}\exp(-\alpha\lambda_j), \tag{13}$$

where the $\{z_{0i}\}$ are components of a fixed point for $\mathbf{z}$ and are independent of α, while β_i weights the ith mode of $\mathbf{z}$ and will depend on the boundary conditions of the fixed point neighborhood.

The functional derivative of $\Delta\epsilon_g$ with respect to η is given by

$$\begin{aligned}\frac{\delta\Delta\epsilon_g}{\delta\eta} = \mathbf{z}^T\frac{\partial\mathbf{M}}{\partial\eta}\mathbf{y} = \mathbf{z}^T\left(\frac{\partial(\mathbf{M}\mathbf{y})}{\partial\eta} - \mathbf{M}\frac{\partial\mathbf{y}}{\partial\eta}\right) &= \sum_i \beta_i\frac{\partial\lambda_i}{\partial\eta} \\ &+ \sum_{ij}\beta_i(\lambda_j-\lambda_i)e^{\alpha(\lambda_j-\lambda_i)}\sum_k \mathbf{U}^{-1}_{ik}\frac{\partial\mathbf{U}_{kj}}{\partial\eta} \\ &+ \sum_{ij} z_{0i}\left[(1+\lambda_j)\mathbf{U}_{ij}\frac{\partial\lambda_j}{\partial\eta} + \lambda_j\frac{\partial\mathbf{U}_{ij}}{\partial\eta}\right]e^{\alpha\lambda_j},\end{aligned} \tag{14}$$

where we have used Eqs. (11), (12), and (13). Equation (14) identifies the various contributions to changes in η under gradient descent on the functional $\Delta\epsilon_g(\eta)$ in the neighborhood of a fixed point. The first term contributes changes in the gradient direction of the eigenvalues while the second term involves derivatives of the eigenvectors with respect to η. The final term involves the fixed point for $\mathbf{z}$. Notice that the first term is independent of α while any contributions from the second term will necessarily depend on α. The final term can only contribute a quantity independent of α if an eigenvalue becomes zero and we do not find zero eigenvalues for either of the fixed points considered here in the neighborhood of the optimal learning rate.

The functional derivative in Eq. (14) will only disappear at constant η if α-dependent terms are negligible (of much lower order than the first term, which is independent of α). This condition is satisfied by ensuring that any term whose exponent is positive and proportional to α has a sufficiently small prefactor. We therefore obtain conditions sufficient, and most likely necessary [for example, these conditions are necessary if each contribution to the second term of Eq. (15) has a different exponent proportional to α, which is also different to any exponents in the final term; this is true so long as each $\lambda_i-\lambda_j$ takes a unique value which also differs from every λ_i], for the existence of a constant η fixed point in an optimal linear system: (1) We require that each component of the fixed point for $\mathbf{z}$ be sufficiently small to ensure the final term in Eq. (14) is negligible, and (2) we further require that $|\beta_i| \ll |\beta_j|$ for at least one j for which $\lambda_j > \lambda_i$ and $\Sigma_k \mathbf{U}^{-1}_{ik}\partial\mathbf{U}_{kj}/\partial\eta$ is nonzero (which implies dependence of the jth eigenvector on η).

Close to the optimal learning rate only the first term in Eq. (14) will be significant, since any other remaining terms would have a strong α dependence (here we assume that one cannot choose η to make the first term zero while simultaneously setting the prefactor of any remaining α-dependent term to zero). In practice, for a nondegenerate system we often find that a single mode in the Lagrange multiplier dynamics is dominant ($|\beta_j| \gg |\beta_{i\neq j}|$) and in this case the effect of our algorithm is to carry out gradient descent (ascent) on this dominant mode,

$$\frac{\delta\Delta\epsilon_g}{\delta\eta} \propto \frac{\partial\lambda_{\mathrm{dom}}}{\partial\eta}. \tag{15}$$

The second condition above suggests that the dominant mode will have a relatively large eigenvalue, although not necessarily the largest eigenvalue. For example, if the largest eigenvalue is associated with an eigenvector which is independent of η then we can say nothing about its weight relative to modes with smaller eigenvalues. In this case it is necessary to consider the boundary conditions of the fixed point neighborhood in order to determine which mode is dominant. In both the symmetric and convergence phases we find exactly this situation and in the latter phase we find the mode with the second largest eigenvalue to be dominant (each phase is considered in greater detail below). The sign of the proportionality constant in Eq. (15) also depends on the boundary conditions of the fixed point neighborhood and we typically find that the eigenvalue is maximized within an unstable fixed point (maximizing the speed of escape from the symmetric phase), or minimized when converging to a stable fixed point. The case where two or more (nondegenerate) modes contribute is discussed at the end of this section, when we consider the effect of second order contributions to the generalization error.

Note that our discussion is not strictly valid if the fixed point changes with η, as is the case for the symmetric phase considered below. The picture developed here holds as long as these changes are relatively slow and we will therefore neglect any such η dependence.

B. The symmetric phase

As demonstrated in Sec. IV, the learning time can be dominated by a symmetric phase in which student nodes fail to differentiate between teacher nodes, resulting in poor generalization performance. This phase represents an attractive fixed point of the dynamics which becomes unstable as small perturbations due to nonsymmetric initial conditions eventually lead to the symmetry breaking required for the student to improve.

Unfortunately, it seems impossible to study the symmetric phase analytically for finite η and a numerical study of this fixed point was therefore carried out in [9], reducing the dimensionality of the system by exploiting symmetries be-

tween the overlaps in order to determine generic behavior. We employ the same dimensionality reduction in order to analyze our linearized system in the neighborhood of the symmetric fixed point. The overlaps are then represented by $Q_{ik}=Q\delta_{ik}+C(1-\delta_{ik})$ and $R_{in}=R\delta_{in}+S(1-\delta_{in})$, where student node indices have been chosen to correspond with the teacher node with which they will eventually specialize. Following [9], we can make some analytical progress by considering a fixed point characterized by $Q=C$ and $R=S$. This fixed point only exists for significant times when η is small, since any difference between Q and C is quickly increased by a positive eigenvalue of order η^2. The resulting fixed point with $Q>C$ is the one actually observed in our simulations [see Fig. 1(d)] but unfortunately this fixed point cannot be studied analytically. Many features of this fixed point are also observed for the $Q=C$ fixed point, however, and it is therefore instructive to consider this case first.

To aid clarity we choose $T=1$, although qualitatively similar results are found for arbitrary T. The vector of deviations from the order parameter fixed point is $\mathbf{y}=(R-R^*,S-S^*,Q-Q^*,C-C^*)^T$ where $(R^*,S^*,Q^*,C^*)^T$ is the fixed point (we assume that the η dependence of this fixed point is negligible). The conjugate vector of Lagrange multipliers is $\mathbf{z}=(\mu_R,\mu_S,\nu_Q,\nu_C)^T$. The $\mathbf{y}$ dynamics is associated with the following matrix:

$$\mathbf{U}=\begin{pmatrix} K-1 & 1 & U_{(1/2)3} & U_{(1/2)4} \\ -1 & 1 & U_{(1/2)3} & U_{(1/2)4} \\ 0 & U_{32} & 1 & 1 \\ 0 & U_{42} & 1 & 1 \end{pmatrix}, \tag{16}$$

where the ith column of $\mathbf{U}$ is the eigenvector associated with eigenvalue λ_i and the components U_{ij} depend on η and K [9]. The columns of $\mathbf{U}$ have been arranged so that they are associated with eigenvalues which are decreasing from left to right ($\lambda_{i>j}<\lambda_j$). The first eigenvalue is the only positive one (unless η is significantly larger than its optimal setting) and results in the divergence of R and S which eventually leads to escape from the symmetric phase.

Recall that the eigenvectors for the Lagrange multiplier dynamics are given by the columns of $(\mathbf{U}^{-1})^T$ and that the ith column is associated with eigenvalue $-\lambda_i$,

$$(\mathbf{U}^{-1})^T=\begin{pmatrix} 1/K & 0 & \cdots & \cdots \\ -1/K & 0 & \cdots & \\ 0 & (U_{32}-U_{42})^{-1} & & \\ 0 & -(U_{32}-U_{42})^{-1} & & \end{pmatrix}. \tag{17}$$

The entries in the final two columns of the second matrix involve various combinations of the U_{ij} and K which are not shown here as their exact form is not important. From the discussion in Sec. V A we expect that the third and fourth modes for the Lagrange multiplier dynamics will have a very small weight ($\beta_{3/4}\ll\beta_{1/2}$). However, since the first eigenvector is independent of η we cannot determine whether it will be dominant without knowing something about the boundary conditions of the symmetric phase. These boundary conditions are determined according to the dynamics away from the fixed point considered here and therefore we can only proceed by observing what happens in practice.

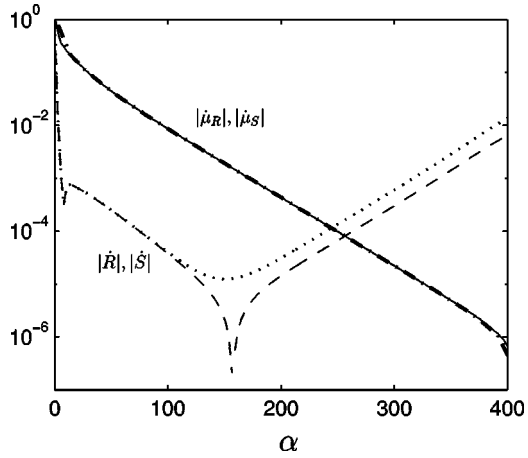

FIG. 4. Magnitudes of the rates of change for R, S, and their conjugate Lagrange multipliers are shown during the symmetric phase for a simulation of the reduced dimensionality equations of motion with $T=1$ and $K=3$. The curves are $|\dot{R}|$ (dotted), $|\dot{S}|$ (dashed), $|\dot{\mu}_R|$ (full), and $|\dot{\mu}_S|$ (dot-dashed). The learning rate is fixed at the optimal value for the symmetric phase and the initial conditions were $R=10^{-6}$, $Q=0.25$, and $S=C=0$.

We find that the $Q=C$ fixed point considered above seems to be a rather good model for the $Q>C$ fixed point observed in simulations. In Fig. 4 we plot the magnitude of $\dot{R}=dR/d\alpha$, $\dot{S}=dS/d\alpha$, $\dot{\mu}_R=d\mu_R/d\alpha$, and $\dot{\mu}_S=d\mu_S/d\alpha$ during the symmetric phase, for simulations of the reduced dimensionality system considered here (with $K=3$ and η fixed at its optimal value for the symmetric phase). Initially $\dot{R}$ and $\dot{S}$ are indistinguishable until differences due to asymmetric initial conditions are amplified and they diverge according to the dominant mode described above (we plot the magnitudes here—the signs of $\dot{R}$ and $\dot{S}$ are different after about $\alpha=160$). Meanwhile, μ_R and μ_S decay with the same rate as the growth of R and S and their rates of change have exactly the same magnitude but opposite sign. This is in agreement with the behavior expected for the first mode, whose eigenvector for the Lagrange multiplier dynamics is shown in the first column of $(\mathbf{U}^{-1})^T$ above. This mode does not contribute to changes in ν_Q and ν_C and because $\dot{\nu}_Q$ and $\dot{\nu}_C$ are observed to have much smaller magnitudes than $\dot{\mu}_R$ and $\dot{\mu}_S$ we conclude that the second mode is associated with a much smaller weight in the Lagrange multiplier dynamics ($\beta_2\ll\beta_1$). The first mode is therefore dominant and the variational algorithm finds the maximum of λ_1 [see Eq. (15)]. This is exactly the mode considered in [9] where λ_1 was maximized numerically and our results agree well with results from that work (more evidence for this is provided in Sec. V E). However, in that study it was very difficult to find the relevant fixed point, requiring much tedious numerical work, and it was not known whether that fixed point really determined the optimal time-dependent learning rate.

C. The convergence phase

Once the student nodes specialize to specific teacher nodes the dynamics quickly leaves the symmetric phase and

approaches a convergence phase in which the overlaps and generalization error exhibit an exponential convergence towards their optimal values (in the absence of noise). In this case it is possible to study the fixed point analytically [9]. We first discuss a completely linear system, in which higher order contributions to the generalization error are negligible. Inclusion of second order terms is required for a more complete description, as discussed in Sec. V D.

As in the symmetric phase there is a mode whose eigenvector is independent of η (recall our second condition for a fixed point in an optimal linear system in Sec. V A). This turns out to be the slowest (least negative) mode for the overlap dynamics and is orthogonal to the leading term of the linearized generalization error, so not contributing to its decay. In this case the mode associated with the next largest eigenvalue dominates the Lagrange multiplier dynamics [and therefore the first term in Eq. (14)] at late times. It is therefore this eigenvalue which is minimized by our variational algorithm [recall Eq. (15)].

In general it is difficult to study analytically how the boundary conditions affect the Lagrange multiplier dynamics because of the greedy drop-off in η at the end of the optimization time window which was described in Sec. IV (see Fig. 1). This greedy phase is not described by our linear system (which requires a constant η) and therefore the final boundary conditions occur outside the region in which our linear model provides a good approximation. However, for the perceptron ($M=K=1$) this greedy phase does not occur and the boundary conditions are well defined for our linear system. It is therefore instructive to consider the perceptron as a special case.

For the perceptron (with $T=1$) we expand around the convergence fixed point via $\mathbf{y}=(r,q)^T$ where $r=1-R$ and $q=1-Q$ with associated Lagrange multipliers $\mathbf{z}=(z_r,z_q)^T$. The $\mathbf{y}$ and $\mathbf{z}$ dynamics are associated with the following matrices, respectively (recall that the columns are eigenvectors):

$$\mathbf{U}=\begin{pmatrix} 1 & U_{12} \\ 2 & U_{22} \end{pmatrix}, \quad (\mathbf{U}^{-1})^T \propto \begin{pmatrix} U_{22} & -2 \\ -U_{12} & 1 \end{pmatrix}, \tag{18}$$

where U_{12} and U_{22} are functions of η (see [5] for details). The linearized generalization error is proportional to $2r-q$, so that the boundary conditions for the Lagrange multipliers are (ignoring a multiplicative constant) $\mathbf{z}(\alpha_1)=(2,-1)^T$ [see Eq. (7)]. This boundary condition corresponds exactly to the second column in $(\mathbf{U}^{-1})^T$ (except for a change in sign) so that the mode associated with this column must be completely dominant for the $\mathbf{z}$ dynamics at late times. Therefore the mode whose eigenvector is orthogonal to the generalization error (the first column in $\mathbf{U}$) in the $\mathbf{y}$ dynamics is irrelevant to determining the optimal learning rate in a completely linear system (however, as discussed below this mode may still contribute to second order terms in the generalization error).

Although we cannot extend this argument to a general multilayer system, because the boundary conditions occur outside the region in which our linear model is reliable, a similar effect is observed in general. Recall the reduced dimensionality equations of motion which were used to analyze the symmetric fixed point ($Q_{ik}=Q\delta_{ik}+C(1-\delta_{ik})$, $R_{in}=R\delta_{in}+S(1-\delta_{in})$). Figure 5 shows the evolution of the overlap deviations ($r=1-R$, $q=1-Q$, $s=S$, $c=C$) and their conjugate Lagrange multipliers during the convergence phase for a simulation of the reduced dimensionality system. The generalization error is also shown and exhibits a faster decay than the overlaps, because the slow mode which determines the decay of the overlaps is orthogonal to the linearized generalization error (higher order contributions to the generalization error are negligible in this case). It is therefore the second slowest mode which determines the decay rate of the generalization error. As for the simpler perceptron case described above, it is this second mode which is mirrored in the Lagrange multiplier dynamics and Fig. 5 shows how the Lagrange multipliers grow with the same rate as the generalization error decays. The second mode is therefore dominant and the variational algorithm minimizes the associated eigenvalue [see Eq. (15)]. This eigenvalue was minimized explicitly in [9] and again we find excellent agreement with results from our variational algorithm (see Sec. V E).

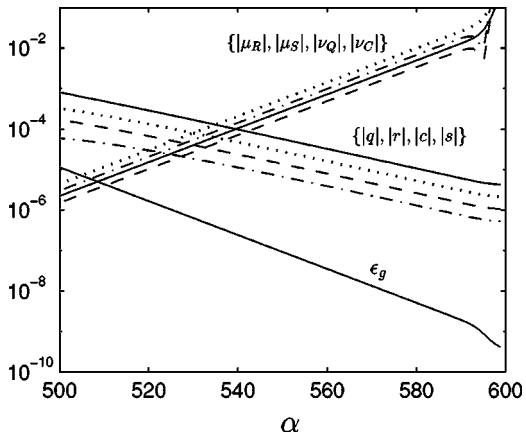

FIG. 5. The magnitudes of the overlap deviations ($r=1-R$, $q=1-Q$, $c=C$, $s=S$) and Lagrange multipliers are shown during the convergence phase for a simulation of the reduced dimensionality equations of motion with $T=1$, $K=3$, and optimal learning rate. The curves for the overlap deviations are $|q|$ (full), $|r|$ (dotted), $|c|$ (dashed), and $|s|$ (dot-dashed) and their conjugate Lagrange multipliers are given the same line type. The lower solid curve shows the generalization error. The initial conditions were $R=10^{-6}$, $Q=0.25$, and $S=C=0$.

D. The effect of quadratic terms in the generalization error

The above discussion does not address the consideration that although we can assume an essentially linear dynamical system in the neighborhood of fixed points, we may still have to consider higher order contributions to the generalization error. Optimizing the convergence rate of the linearized dynamics is only optimal in general if we are minimizing a linear function of the dynamical variables. Indeed, it was noted in [9,11] that quadratic contributions to the generalization error determine the slow mode for convergence for realizable learning ($M=K$) with an isotropic teacher ($T_{nm}=T\delta_{nm}$) when T is above some critical value T_{crit}. This is because there is an eigenvector orthogonal to the linearized generalization error which therefore does not contribute to its decay, but which does contribute to the decay of quadratic terms in the generalization error. For $T>T_{\mathrm{crit}}$ the optimal asymptotic learning rate is the value which allows the

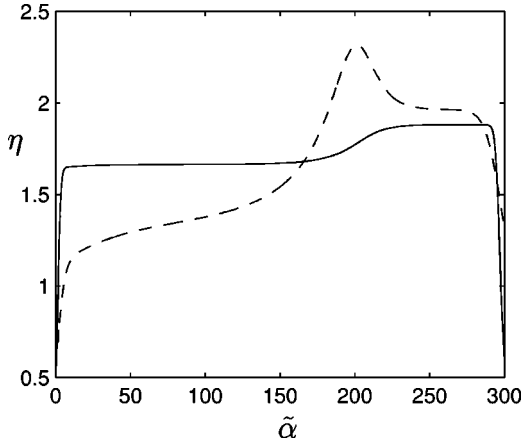

FIG. 6. The optimal learning rate is shown for a two hidden node student learning an isotropic teacher ($T_{nm}=T\delta_{nm}$) also with two hidden nodes. The two curves are for teacher lengths $T=1$ (solid curve) and $T=5$ (dashed curve). The larger teacher length is above the critical value T_{crit} where second order contributions to the generalization error become significant at late times. The learning time has been rescaled so that the curves fit on the same graph, with $\tilde{\alpha}=\alpha$ for $T=1$ and $\tilde{\alpha}=1.875\alpha$ for $T=5$.

linear and quadratic terms to decay at the same rate. One might therefore expect our algorithm to approach this asymptotically optimal value at late times. However, in view of the fact that the end-point boundary conditions for the Lagrange multipliers [see Eq. (7)] are first derivatives of the generalization error, it seems unlikely that second order terms in the generalization error will cause significant changes to the Lagrange multiplier dynamics at late times. In fact, we find that the algorithm deals with quadratic effects *before* converging to the optimal learning rate for a linear error.

Figure 6 shows the optimal learning rate for a two hidden node student learning an isotropic teacher of the same configuration. The two curves are for $T=1$ ($<T_{\mathrm{crit}}$) and $T=5$ ($>T_{\mathrm{crit}}$). For $T<T_{\mathrm{crit}}$ we get a similar picture to Fig. 1(a) with two well defined plateaus determining the symmetric and convergence phases. However, for $T>T_{\mathrm{crit}}$ the optimal learning rate rises after the symmetric phase towards, but not reaching, the asymptotically optimal value identified in [9] ($\eta\simeq 2.79$) for the full generalization error and then falls to the optimal value for a linearized generalization error ($\eta=1.963$). Forcing the learning rate to approach the asymptotically optimal value for the full generalization error at late times leads to poorer performance, suggesting that the algorithm does indeed find the optimal learning rate. For very long time windows we would expect the algorithm to reach the asymptotically optimal learning rate after the symmetric phase and to stay there, before dropping to the optimal learning rate for a linearized generalization error. However, this would seem to require a very low final generalization error for the cases we have considered (at least 10^{-20}, depending on T and K) and since errors of this order are of little physical relevance and require greater numerical accuracy than achieved by our implementation we do not pursue this regime here.

E. Generic behavior

Using the reduced dimensionality equations of motion makes it possible to find the optimal learning rate for arbitrary K, since we avoid the increase in computation time necessary to deal with large systems. (The computation time still seems to grow with K because of increased precision required in the numerical integration.) We can therefore run our algorithm for various values of K and T in order to deduce scaling laws for the optimal learning rate within the two phases of learning. However, as we saw above our algorithm simply performs gradient descent (ascent) on the relevant eigenvalue within each phase and the results should therefore compare closely to results obtained by optimizing with respect to this eigenvalue directly.

In Fig. 7 we compare the optimal learning rate determined by our method to the value found by a direct numerical eigenvalue analysis [9] for various values of K. The results agree well and any discrepancies may be due to the variational algorithm not converging completely, or because there is some variance in the learning rate within the symmetric phase (we used an approximate halfway point to determine η_{sym}). This outcome is rather fortuitous, as the scaling laws previously determined for each phase in [9] also describe the scaling behavior for the globally optimal learning rate. This was not obvious *a priori* since it was not known that the optimal time-dependent learning rate would be dominated by two plateaus, each with a constant learning rate (indeed, the

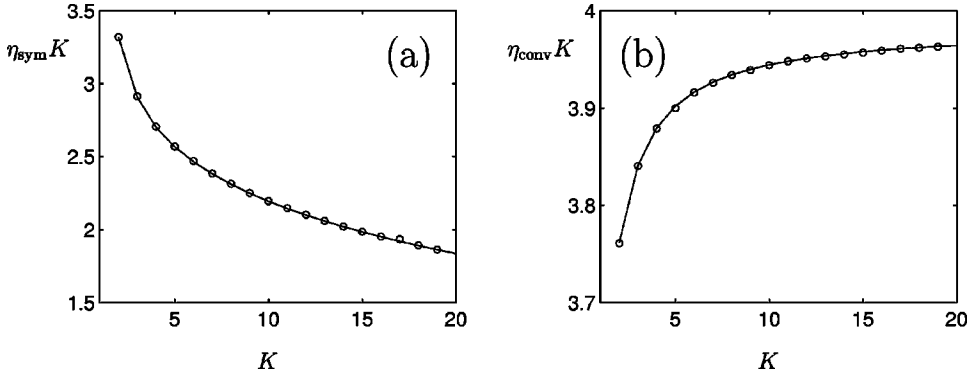

FIG. 7. We compare results for the optimal learning rate found by our variational method (circles) to results from a direct optimization of the relevant eigenvalue (solid line) as a function of task complexity K for (a) the symmetric phase and (b) the convergence phase. A reduced dimensionality system describing an isotropic, realizable task ($T_{nm}=\delta_{nm}$) was used to obtain both sets of results.

discussion in Sec. V D shows that this is not always the case).

It is interesting to summarize some of the scaling laws deduced in [9] (we will only consider fixed T here). During the symmetric phase the optimal learning rate scales as $K^{-5/3}$ for large K and the trapping time within the symmetric phase, which determines the total training time, scales as $K^{8/3}$. During the convergence phase the optimal learning rate scales as $1/K$. The maximal learning rate in each phase, above which perfect learning is impossible, scales in the same way as the optimal learning rate. One therefore finds that using a learning rate which is optimal asymptotically will be very bad during the transient, either leading to trapping within the symmetric phase or divergence of the student weight vector norms. For an account of other scaling phenomena (for example, with respect to task nonlinearity T), see [9]. Of course, in many cases it would be very difficult to carry out an eigenvalue analysis (we have only considered the simplest scenario of a single parameter and an isotropic, noiseless task) and our algorithm therefore provides a powerful alternative tool for determining generic results.

F. Limitations to the variational approach

The analysis in Sec. V A suggests a possible limitation of our variational algorithms if eigenvalues are multimodal, as might be the case if the equations of motion involved higher powers, or other nonlinear functions, of the parameters being optimized. In this case the algorithm may get stuck in functional local minima, since we are effectively carrying out gradient descent (ascent) on the dominant eigenvalue.

Certain initial settings for the learning rate may also result in convergence to functional local minima which are suboptimal. For example, if the learning rate is chosen to be very low initially and the time window is not sufficiently long for the system to leave the symmetric phase then the algorithm may anneal the learning rate in an attempt to optimize performance without leaving the symmetric subspace. As with any differential method we are at the mercy of local minima; our conditions for the globally optimal learning rate are necessary but not sufficient. From careful study of the dynamics we are satisfied that all the solutions presented in this work are globally optimal.

VI. SITE-DEPENDENT LEARNING RATES

It is straightforward to extend our method to more complex learning rules and to different learning parameters. As a simple example we consider a generalized gradient descent algorithm in which different learning rates are associated with different hidden nodes, so that the new update rule is given by $\mathbf{J}_i^{\mu+1}=\mathbf{J}_i^{\mu}+(1/N)\,\eta_i\,\delta_i^{\mu}\boldsymbol{\xi}^{\mu}$. This enables the system to explore more complex routes to breaking symmetry and converging to the optimal solution. The derivation for the optimal site-dependent learning rate follows the discussion in Sec. III closely and is therefore not included here.

A. Isotropic teacher

In our first example we train a three hidden node system on examples generated by a three node isotropic teacher ($M=K=3$), using three different learning rates related to the various hidden nodes. The optimal learning rates and the corresponding generalization error are shown in Figs. 8(a) and 8(c), respectively. Comparison with Fig. 1(b) shows a significant improvement over standard gradient descent, although the dynamics is still dominated by a symmetric plateau (the initial conditions were the same in both cases). A shortened symmetric plateau is achieved by setting one learning rate close to zero while the other two are assigned a high value, much higher than the optimal learning rate for standard gradient descent on the same problem (but lower than the optimal learning rate for standard gradient descent with $M=K=2$). Once the two nodes associated with the high learning rate become associated to specific teacher nodes, the learning rate associated with the third node rises rapidly as it learns the remaining teacher node. Eventually, all learning rates converge to the same constant and asymptotically optimal value, which is the same as for standard gradient descent [see Fig. 1(a)], before the greedy drop-off in η which was explained in Sec. IV. The same pattern of successively active learning rates and hence a phased symmetry breaking is repeated for larger isotropic systems, suggesting that using different learning rates for different nodes may be beneficial for speeding up the learning process in general and escaping the symmetric phase in particular.

B. Graded teacher

In our second example we train the same three hidden node system on examples generated by a graded three node teacher ($T_{nm}=n\,\delta_{nm}$). The optimal learning rates and the corresponding generalization error are shown in Figs. 8(b) and 8(d), respectively. The optimal site-dependent learning rate shows a much richer behavior in this example. Initially, the learning rate associated with the node learning the largest teacher vector (solid line) is highest, followed by phases in which the learning rates associated with nodes learning the intermediate and smallest teacher vectors increase in turn. This corresponds to what one might expect, since the system specializes to teacher nodes in order of decreasing impact to the teacher output. Eventually each learning rate approaches an asymptotically constant value, before dropping off towards the end of the time window due to the greedy effect explained in Sec. IV. Here, we see that the order of learning rate magnitudes has changed, so that the learning rate for the node associated with the largest teacher node is smallest and vice versa. This is a somewhat unintuitive result, since the optimal asymptotic learning rate for standard gradient descent increases with increasing T [9]. Unfortunately, an analytical study of this scenario is hampered by the lack of dimensionality reducing symmetries in this case. It would be most interesting to study the relationship between values assigned to the learning rates in this case and the corresponding effective values suggested by other methods which incorporate information about curvature of the mean error surface and have been proved asymptotically optimal [16].

VII. CONCLUSION

We have described a method for determining optimal learning rates for on-line learning in a soft committee machine, using a variational calculation to maximize the total reduction in generalization error over a fixed time window.

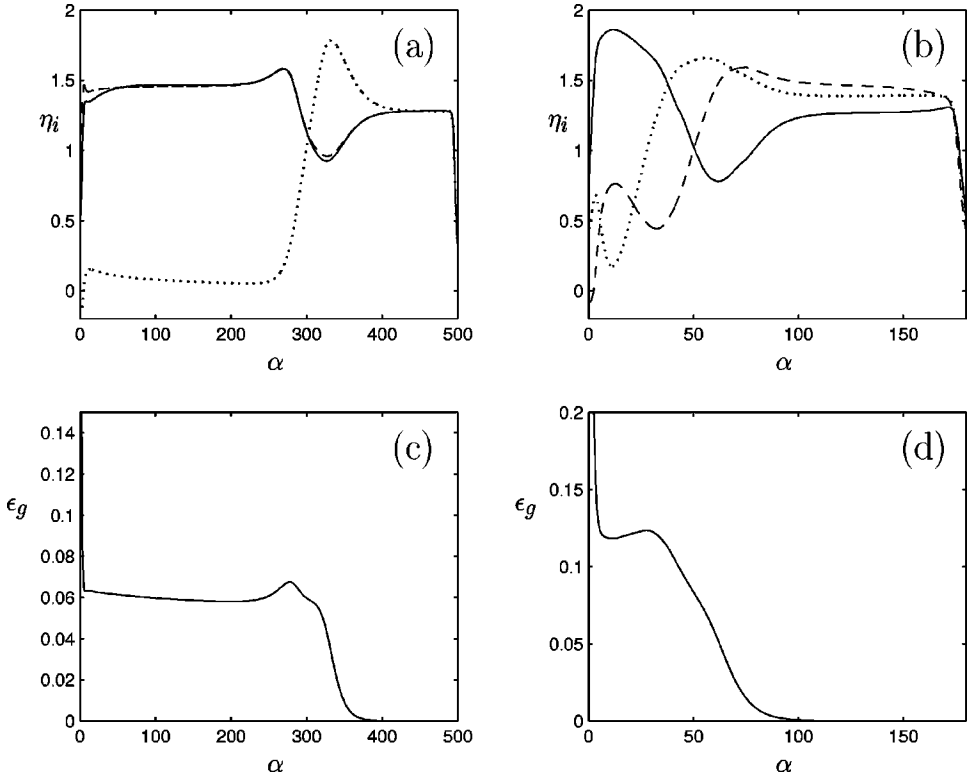

FIG. 8. A three hidden node student with site-dependent learning rates is trained to emulate a teacher with the same number of hidden nodes. The globally optimal learning rates are shown in (a) for an isotropic teacher ($T_{nm}=\delta_{nm}$) and in (b) for a graded teacher ($T_{nm}=n\delta_{nm}$). The corresponding generalization errors are shown in (c) and (d), respectively. In (b) the learning rates are associated with nodes learning the following teacher nodes—dashed line for the node learning the first teacher node ($T_{11}=1$), dotted line for the second ($T_{22}=2$), and solid line for the third ($T_{33}=3$).

The method makes use of a recent statistical mechanics model which allows a compact and exact description of the learning process for large input dimension via differential equations for a small number of macroscopic quantities.

Learning with the optimal learning rate still suffers from trapping in a symmetric phase reported in [4,5], which dominates training time, and the fastest escape time is achieved by maximizing the only positive eigenvalue within this unstable fixed point. An analytical study of our variational algorithm in the neighborhood of fixed points, which uses a linear model, shows this to be the expected behavior of our algorithm and explains the excellent agreement with previous results for isotropic, realizable tasks [9], in which the escape eigenvalue within this phase was maximized explicitly, allowing a rather general characterization of the optimal learning rate dynamics. During the convergence phase the slowest mode, which is orthogonal to the first order term in the generalization error, does not contribute to the optimization of η at late times and the dominant mode was the next slowest. This result is also found to be consistent with our analysis of the variational algorithm near a fixed point. For the perceptron it is possible to show exactly how the boundary conditions lead to the exclusion of the slowest mode, but in general this is not possible because the boundary conditions occur outside the region in which our simple linear model holds due to a greedy minimization of the generalization error at the end of the optimization time window. The main difference between our results for this example and the analysis in [9] is in the consideration of second order contributions to the generalization error, which in some situations determine the fastest asymptotic decay. We find that these second order effects play a role immediately after the symmetric phase, but that the algorithm always approaches the optimal learning rate for a linearized generalization error at late times. In fact, these second order effects are only relevant when the optimization time window is sufficiently large to allow a very low generalization error and they may therefore be irrelevant in practice.

Learning from corrupted examples provides a very different picture. After leaving the symmetric phase the optimal learning rate is gradually annealed towards a decay inversely proportional to α. As for the noiseless case there is a greedy phase towards the end of the optimization time window, reflecting a change in the decay direction, which provides a short-term improvement but is unsustainable for longer times. Our results suggest that there is some danger in annealing the learning rate too early, since losses due to an initially low learning rate might never be recovered and the learning process could even become trapped within the symmetric phase.

The main benefit of the present approach lies in its generality. We can apply our optimization scheme to learning scenarios or phases of learning for which analysis is infeasible, either because the dynamics cannot be captured by any sufficiently simple model or because we have insufficient insight. Indeed, we have recently used our approach to determine globally optimal learning rules [6], extending previous results for locally optimal rules [17]. We have also used this framework to determine the efficacy of a quadratic regularizer [7] and in order to quantify the performance of natural

gradient learning [8], an on-line variable-metric algorithm which was recently introduced by Amari [16]. In the latter study we show how learning time scales better with increasing numbers of hidden nodes for this algorithm when compared with optimized gradient descent. In the present paper we showed how one can apply the optimization procedure to a generalized algorithm in which different nodes are associated with different learning rates. The picture which emerges shows a rich behavior for the optimal learning rates, especially in the case of a graded teacher. There is significant evidence that such a generalized algorithm will provide gains over standard gradient descent. However, a question which remains to be answered is whether one can find practical methods for selecting parameters close to the optimal ones determined here. This has been carried through in some cases for the asymptotic stages of learning (see, for example, [18]) and asymptotically optimal algorithms also exist [16], but as we have seen here the transient stages of learning will often dominate the training time. We hope that the present analysis will aid the search for a principled solution to this problem.

ACKNOWLEDGMENTS

We would like to thank Ansgar West and Bernhard Schottky for many helpful discussions. This work was supported by the EPSRC under Grant No. GR/L19232.

[1] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, London, 1995).
[2] G. Cybenko, Math. Control Signals Systems **2**, 303 (1989).
[3] D. Saad and M. Rattray, Phys. Rev. Lett. **79**, 2578 (1997).
[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); Phys. Rev. E **52**, 4225 (1995).
[6] M. Rattray and D. Saad, J. Phys. A **30**, L771 (1997).
[7] D. Saad and M. Rattray, Phys. Rev. E **57**, 2170 (1998).
[8] M. Rattray and D. Saad, in *Proceedings of the 8th International Conference on Artificial Neural Networks*, edited by L. Niklasson, M. Bodòden, and T. Ziemke (Springer-Verlag, London, 1998), p. 165.
[9] A. H. L. West and D. Saad, Phys. Rev. E **56**, 3426 (1997).
[10] P. Riegler, Ph.D. thesis, University of Würzburg, 1997.
[11] P. Riegler and M. Biehl, J. Phys. A **28**, L507 (1995).
[12] D. Barber, D. Saad, and P. Sollich, Europhys. Lett. **34**, 151 (1996).
[13] M. Biehl, P. Riegler, and C. Wöhler, J. Phys. A **29**, 4769 (1996).
[14] H. White, Neural Comput. **1**, 425 (1989).
[15] T. K. Leen, B. Schottky, and D. Saad, in *Advances in Neural Information Processing Systems*, edited by M. I. Jordan, M. J. Kearns, and S. A. Solla (MIT Press, Cambridge, MA, 1998), Vol. 10, p. 301.
[16] S. Amari, Neural Comput. **10**, 251 (1998).
[17] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992).
[18] Y. LeCun, P. Y. Simard, and B. Pearlmutter, in *Advances in Neural Information Processing Systems*, edited by C. L. Giles, S. J. Hanson, and J. D. Cowan (Morgan Kaufmann, San Mateo, CA, 1993), Vol. 5, p. 156.