On-Line Learning in Soft Committee Machines

David Saad[1] and Sara A. Solla[2]
[1] *Department of Physics, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK.*
[2] *CONNECT, The Niels Bohr Institute, Blegdamsdvej 17, Copenhagen 2100, Denmark.*

The problem of on-line learning in two-layer neural networks is studied within the framework of statistical mechanics. A fully connected committee machine with K hidden units is trained by gradient descent to perform a task defined by a teacher committee machine with M hidden units acting on randomly drawn inputs. The approach, based on a direct averaging over the activation of the hidden units, results in a set of first order differential equations which describe the dynamical evolution of the overlaps among the various hidden units and allow for a computation of the generalization error. The equations of motion are obtained analytically for general K and M, and provide a new and powerful tool used here to study a variety of realizable, over-realizable, and unrealizable learning scenarios, and to analyze the role of the learning rate in controlling the evolution and convergence of the learning process.

87.10.+e, 02.50.-r, 05.20.-y

I. INTRODUCTION

Layered neural networks are the focus of an intense research effort for their ability to implement input-output maps of relevance to classification and regression tasks. Two-layer architectures with N input units, a single internal layer with an arbitrary number H of hidden units, and one output unit, suffice to represent nontrivial scalar functions of N-dimensional variables. Exact representation of Boolean functions requires at most $H = 2^N$ units [1]; continuous functions can be approximated with arbitrary accuracy if the number H of hidden units is not constrained [2,3].

A neural network of fixed architecture is characterized by the internal parameters $\{\mathbf{J}\}$ which quantify the strength of the interneuron couplings [1,4,5]. Specific maps $\zeta = f_{\mathbf{J}}(\boldsymbol{\xi})$ from an N-dimensional input space $\boldsymbol{\xi}$ onto a scalar ζ are selected through the choice of parameters $\{\mathbf{J}\}$. Learning refers to the modification of these couplings so as to bring the map $f_{\mathbf{J}}$ implemented by the network as close as possible to a desired map $\widetilde{f}$. The degree of success is monitored through the *generalization error*, a measure of the dissimilarity between $f_{\mathbf{J}}$ and $\widetilde{f}$.

Learning from examples in layered neural networks is usually formulated as an optimization problem [4,5], based on the minimization of a *learning error* defined as the additive error over a *training set* composed of P independent examples $(\boldsymbol{\xi}^\mu, \zeta^\mu)$, with $\zeta^\mu = \widetilde{f}(\boldsymbol{\xi}^\mu)$ for all $1 \le \mu \le P$. Statistical physics has provided useful tools for investigating the properties of such models, based on the use of the replica method to account for the disorder introduced by the different possible ways in which a training set of fixed size can be chosen. The method has been successfully applied to the analysis of single-layer perceptrons [5] and some simplified two-layer structures (e.g., committee machines [6]). Analysis of more complicated multilayer networks is hampered by technical difficulties due to the complex structure of the solutions in a space of order parameters [7], which describe in this case correlations among the various neurons in the trained network as well as their degree of specialization towards the implementation of the desired task.

An alternative approach is to investigate *on-line learning* [8]. In this scenario the couplings $\{\mathbf{J}\}$ are adjusted after the presentation of each example so as to minimize the corresponding error. The resulting changes in the couplings are described as a dynamical evolution, with the number of examples playing the role of time. The average that accounts for the disorder introduced by the independent random selection of an example at each time step can be performed directly, without invoking the replica method. The resulting equations of motion for the relevant order parameters characterize the structure of the space of solutions and allow for a computation of the generalization error.

In spite of the apparent simplicity resulting from the avoidance of the replica method, this program has up to now been carried out only for single-layer perceptrons [9–11] and some severely restricted two-layer architectures [12–14]. We have applied method outlined in [14] to the analysis of a very general learning scenario: a two-layer student network composed of N input units, K hidden units, and a single linear output unit, trained to perform a task defined through a teacher network of similar architecture except that its number M of hidden units is not necessarily equal to K. The result was unexpected: the dynamical equations for the order parameters can be obtained analytically for general K and M in the large N limit. The resulting equations of motion can be integrated accurately even for large networks, and provide a new powerful tool to study learning in multilayer networks.

In this paper we restrict ourselves to *soft committee machines* [14], for which the output unit is linear and the couplings from all hidden units to the output unit are positive and of unit strength. In Section II we describe the student and teacher networks and define the order parameters needed to compute the generalization error. A gradient descent rule for the update of the student couplings results in first order differential equations for the dynamical evolution of the order parameters. These equations of motion are obtained analytically for general K and M, and provide a tool to study realizable ($K = M$), over-realizable ($K > M$), and unrealizable ($K < M$) learning scenarios. Section III is devoted to a heuristic discussion of the role of the learning rate in the convergence of on-line learning. A rigorous analysis of the structure of the solutions for realizable cases is presented in Section IV, where we discuss a suboptimal transient due to dynamical trapping in the symmetric subspace, the onset of specialization associated with breaking the symmetry among the student hidden units, and the subsequent exponential convergence to an optimal solution with perfect generalization. In Section V we consider two examples which demonstrate the power of the approach developed here when applied to the analysis of over-realizable and unrealizable learning scenarios. Section VI contains a summary and discussion of the results presented in this paper, and some comments on the extension of our approach to the analysis of other learning scenarios.

II. DYNAMICAL EQUATIONS FOR THE ORDER PARAMETERS

Our discussion focuses on the soft committee machine [14], in which all the hidden units are connected to the output unit with positive couplings of unit strength, and only the input-to-hidden couplings are adaptive.

Consider a student network consisting of N input units, K hidden units, and one linear output unit. Hidden unit i receives information from input unit r through the weight J_{ir}, and its activation under presentation of an input pattern $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ is $x_i = \mathbf{J}_i \cdot \boldsymbol{\xi}$, with $\mathbf{J}_i = (J_{i1}, \ldots, J_{iN})$ defined as the vector of incoming weights onto the i-th hidden unit. As all the hidden-to-output weights are fixed to be $+1$, the overall output of the student network is

$$\sigma(\mathbf{J}, \boldsymbol{\xi}) = \sum_{i=1}^{K} g\ (\mathbf{J}_i \cdot \boldsymbol{\xi}) \quad , \tag{1}$$

where g is the activation function of the hidden units and $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \le i \le K}$ is the set of input-to-hidden adaptive weights.

Training examples are of the form $(\boldsymbol{\xi}^\mu, \zeta^\mu)$. The components of the independently drawn input vectors $\boldsymbol{\xi}^\mu$ are uncorrelated random variables with zero mean and unit variance. The corresponding output ζ^μ is given by a deterministic teacher whose internal structure is that of a network similar to the student except for a possible difference in the number M of hidden units. Hidden unit n in the teacher network receives input information through the weight vector $\mathbf{B}_n = (B_{n1}, \ldots, B_{nN})$, and its activation under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is $y_n^\mu = \mathbf{B}_n \cdot \boldsymbol{\xi}^\mu$. The corresponding output is

$$\zeta^\mu = \sum_{n=1}^{M} g\ (\mathbf{B}_n \cdot \boldsymbol{\xi}^\mu) \quad . \tag{2}$$

We will use indices $i, j, k, l \ldots$ to refer to units in the student network, and $n, m, \ldots$ for units in the teacher network.

The error made by a student with weights $\mathbf{J}$ on a given input $\boldsymbol{\xi}$ is given by the quadratic deviation

$$\epsilon(\mathbf{J}, \boldsymbol{\xi}) \equiv \frac{1}{2} \left[\ \sigma(\mathbf{J}, \boldsymbol{\xi}) - \zeta\ \right]^2 = \frac{1}{2} \left[\sum_{i=1}^{K} g(x_i) - \sum_{n=1}^{M} g(y_n) \right]^2 \quad . \tag{3}$$

Performance on a typical input defines the generalization error

$$\epsilon_g(\mathbf{J}) \equiv\ < \epsilon(\mathbf{J}, \boldsymbol{\xi}) >_{\{\xi\}} \tag{4}$$

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

$$\mathcal{P}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{(2\pi)^{M+K} |\mathcal{C}|}}\ \exp\left\{ -\frac{1}{2} (\mathbf{x}, \mathbf{y})^T \mathcal{C}^{-1} (\mathbf{x}, \mathbf{y}) \right\} \ , \tag{5}$$

with

$$\mathcal{C} = \begin{bmatrix} Q & R \\ R^T & T \end{bmatrix} . \tag{6}$$

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

$$\epsilon_g(\mathbf{J}) = \frac{1}{\pi}\Bigg\{ \sum_{ik} \arcsin \frac{Q_{ik}}{\sqrt{1+Q_{ii}}\,\sqrt{1+Q_{kk}}} + \sum_{nm} \arcsin \frac{T_{nm}}{\sqrt{1+T_{nn}}\,\sqrt{1+T_{mm}}} \\ -2 \sum_{in} \arcsin \frac{R_{in}}{\sqrt{1+Q_{ii}}\,\sqrt{1+T_{nn}}} \Bigg\} , \tag{7}$$

where $1 \le i, k \le K$ sum over the student hidden units, and $1 \le n, m \le M$ sum over the hidden units of the teacher. The parameters T_{nm} are characteristic of the task to be learned and remain fixed during training, while the overlaps Q_{ik} among student hidden units and R_{in} between a student and a teacher hidden units are determined by the student weights $\mathbf{J}$ and evolve during training.

A gradient descent rule for the update of the student weights,

$$\mathbf{J}^{\mu+1} = \mathbf{J}^{\mu} - \frac{\eta}{N} \nabla_{\mathbf{J}}\ \epsilon\left(\mathbf{J}^{\mu}, \boldsymbol{\xi}^{\mu}\right) , \tag{8}$$

where the learning rate η has been scaled with the input size N, results in

$$\mathbf{J}_i^{\mu+1} = \mathbf{J}_i^{\mu} + \frac{\eta}{N}\ \delta_i^{\mu}\ \boldsymbol{\xi}^{\mu} , \tag{9}$$

with

$$\delta_i^{\mu} \equiv g'(x_i^{\mu}) \left[\sum_{n=1}^{M} g(y_n^{\mu}) - \sum_{j=1}^{K} g(x_j^{\mu}) \right] \tag{10}$$

defined in terms of both the activation function g and its derivative g'.

The time evolution of the overlaps R_{in} and Q_{ik} is then given by:

$$R_{in}^{\mu+1} - R_{in}^{\mu} = \frac{\eta}{N}\ \delta_i^{\mu}\ y_n^{\mu} \tag{11}$$

and

$$Q_{ik}^{\mu+1} - Q_{ik}^{\mu} = \frac{\eta}{N}\ (\delta_i^{\mu}\ x_k^{\mu} + \delta_k^{\mu}\ x_i^{\mu}) + \frac{\eta^2}{N}\ \delta_i^{\mu}\ \delta_k^{\mu} . \tag{12}$$

The dependence on the current input $\boldsymbol{\xi}^{\mu}$ is only through the activations $\mathbf{x}$ and $\mathbf{y}$, and the corresponding averages can be performed using the joint probability distribution (5). In the thermodynamic limit $N \to \infty$ the normalized example number $\alpha = \mu/N$ can be interpreted as a continuous time variable, leading to the equations of motion:

$$\frac{dR_{in}}{d\alpha} = \eta < \delta_i\ y_n > , \\ \frac{dQ_{ik}}{d\alpha} = \eta < \delta_i\ x_k > + \eta < \delta_k\ x_i > + \eta^2 < \delta_i\ \delta_k > . \tag{13}$$

The averages in Eq. (13) require the evaluation of two types of multivariate Gaussian integrals. Terms proportional to η involve the three-dimensional integral

$$I_3 \equiv < g'(u)\ v\ g(w) > ,$$

where the argument u of g' is one of the components of $\mathbf{x}$, while both v and w can be components of either $\mathbf{x}$ or $\mathbf{y}$. The term proportional to η^2 involves the four-dimensional Gaussian integral

$$I_4 \equiv < g'(u)\ g'(v)\ g(w)\ g(z) > ,$$

where u and v are components of $\mathbf{x}$ while w and z can be components of either $\mathbf{x}$ or $\mathbf{y}$.

Permutational symmetries that arise when some of the arguments in I_3 and I_4 are constrained to be equal result in contributions to the averages in Eq. (13) which require the evaluation of integrals of reduced dimensionality. Terms proportional to η involve not only the three-dimensional integral I_3 for $u \neq v \neq w$, but its three possible two-dimensional reductions: $< g'(u)\ u\ g(v) >$, $< g'(u)\ v\ g(v) >$, and $< g'(u)\ v\ g(u) >$, and the one-dimensional reduction $< g'(u)\ u\ g(u) >$, for a total of five different integrals. A similar counting for I_4 must take into account the initial symmetry under the pairwise exchange $u \leftrightarrow v$ and $w \leftrightarrow z$. The term proportional to η^2 involves not only the four dimensional integral I_4 for $u \neq v \neq w \neq z$ but three distinct three-dimensional reductions: $< g'(u)\ g'(v)\ [g(w)]^2 >$, $< [g'(u)]^2\ g(v)\ g(w) >$, and $< g'(u)\ g'(v)\ g(v)\ g(w) >$; four distinct two-dimensional reductions: $< g'(u)\ g'(v)\ [g(v)]^2 >$, $< [g'(u)]^2\ g(u)\ g(v) >$, $< [g'(u)]^2\ [g(v)]^2 >$, and $< g'(u)\ g'(v)\ g(u)\ g(v) >$; and the one-dimensional reduction $< [g'(u)]^2\ [g(u)]^2 >$, for a total of nine different integrals.

It is a remarkable property of multivariate Gaussian integrals that, as proven in the Appendix, all such integrals as generated from I_3 and I_4 through dimensionality reduction do not need to be evaluated independently: the corresponding results follow from imposing the appropriate constraints on the general expressions for I_3 and I_4. There is no need to evaluate fourteen different integrals, and the equations of motion reduce to a surprisingly compact form in terms of only I_3 and I_4:

$$\begin{aligned}
\frac{dR_{in}}{d\alpha} &= \eta \left\{ \sum_m I_3(i,n,m) - \sum_j I_3(i,n,j) \right\} , \\
\frac{dQ_{ik}}{d\alpha} &= \eta \left\{ \sum_m I_3(i,k,m) - \sum_j I_3(i,k,j) \right\} + \eta \left\{ \sum_m I_3(k,i,m) - \sum_j I_3(k,i,j) \right\} + \\
&\quad \eta^2 \left\{ \sum_{n,m} I_4(i,k,n,m) - 2\sum_{j,n} I_4(i,k,j,n) + \sum_{j,l} I_4(i,k,j,l) \right\} .
\end{aligned} \tag{14}$$

Arguments assigned to I_3 and I_4 are to be interpreted following our convention to distinguish student from teacher activations, i.e, $I_3(i,n,j) \equiv < g'(x_i)\ y_n\ g(x_j) >$, and the average is performed using the three-dimensional covariance matrix C_3 which results from projecting the full covariance matrix $\mathcal{C}$ of Eq. (6) onto the relevant subspace. For $I_3(i,n,j)$ the corresponding matrix is:

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

The equations of motion for the order parameters take the form (14) for any choice of the activation function g, even though it might not always be possible to obtain analytic expressions for the integrals I_3 and I_4.

The two multivariate integrals in Eq. (14) can be performed analytically for $g(x) = \mathrm{erf}(x/\sqrt{2})$. I_3 is given in terms of the components of the C_3 covariance matrix by

$$I_3 = \frac{2}{\pi} \frac{1}{\sqrt{\Lambda_3}} \frac{C_{23}(1+C_{11}) - C_{12}C_{13}}{1+C_{11}} , \tag{15}$$

with

$$\Lambda_3 = (1+C_{11})(1+C_{33}) - C_{13}^2 . \tag{16}$$

The expression for I_4 in terms of the components of the corresponding C_4 covariance matrix is

$$I_4 = \frac{4}{\pi^2} \frac{1}{\sqrt{\Lambda_4}} \arcsin\left(\frac{\Lambda_0}{\sqrt{\Lambda_1}\sqrt{\Lambda_2}} \right) , \tag{17}$$

where

$$\Lambda_4 = (1+C_{11})(1+C_{22}) - C_{12}^2 , \tag{18}$$

and

$$
\begin{aligned}
\Lambda_0 &= \Lambda_4 C_{34} - C_{23}C_{24}(1+C_{11}) - C_{13}C_{14}(1+C_{22}) + C_{12}C_{13}C_{24} + C_{12}C_{14}C_{23} \ , \\
\Lambda_1 &= \Lambda_4(1+C_{33}) - C_{23}^2(1+C_{11}) - C_{13}^2(1+C_{22}) + 2C_{12}C_{13}C_{23} \ , \\
\Lambda_2 &= \Lambda_4(1+C_{44}) - C_{24}^2(1+C_{11}) - C_{14}^2(1+C_{22}) + 2C_{12}C_{14}C_{24} \ .
\end{aligned} \tag{19}
$$

The dynamical equations (14) are the main result of our paper. Together with the analytic expressions for I_3 and I_4, they provide a novel tool for analyzing the learning process for a general soft committee machine with an arbitrary number K of hidden units trained to perform a task defined by a soft committee teacher with M hidden units [15]. Results previously obtained for a soft committee machine with two hidden units trained by a single-layer teacher [14] are recovered for $K = 2$ and $M = 1$. The set of coupled first-order differential equations provided here are exact in the thermodynamic limit; leading corrections are of order $(1/N)$. The equations can be integrated accurately even for large values of K and M to obtain the dynamical evolution of the order parameters, which determine the time evolution of the generalization error (7) and provide valuable insight into the process of learning in multilayer networks.

In what follows we apply the tools developed in this section to the analysis of a variety of learning scenarios. The tasks to be learned are characterized by the number M of teacher hidden units and the matrix $T_{nm} = \mathbf{B}_n \cdot \mathbf{B}_m$. We consider uncorrelated teacher vectors, with $T_{nm} = T_n \ \delta_{nm}$. Two cases are of interest: i) All teacher hidden nodes are equally relevant to the implementation of the target task, as described by an isotropic teacher with $T_n = T$. The actual value of T is of no importance as long as it does not depend on n; we have used $T = 1$ in the analysis and simulations to be presented here. ii) An anisotropic teacher with uncorrelated but graded weight vectors which can be ordered according to their relevance in determining the output: $T_{n_1} \leq T_{n_2}$ for $n_1 < n_2$. As a special case of such a graded teacher we consider $T_n = n$.

The time evolution of the order parameters R_{in} and Q_{ik} follows from integrating the equations of motion (14) from initial conditions determined by a random initialization of the student vectors $\{\mathbf{J}_i\}_{1 \leq i \leq K}$. This initialization results in random norms Q_{ii} for the student weight vectors, represented here through the independent initialization of each Q_{ii} from a uniform distribution in the $[0, 0.5]$ interval, $U[0, 0.5]$. Overlaps Q_{ik} between independently chosen student vectors $\mathbf{J}_i$ and $\mathbf{J}_k$ are of order $1/\sqrt{N}$ and vanishingly small in the regime $N \gg K$. Initial values for Q_{ik}, $i \neq k$, are independently drawn from a uniform distribution $U[0, Q_o]$, with $Q_o \ll 1$. The overlaps R_{in} between a randomly initialized student vector $\mathbf{J}_i$ and an unknown teacher $\mathbf{B}_n$ are also small numbers of order $1/\sqrt{N}$ for $N \gg K$ and $N \gg M$. Initial values for each R_{in} are independently drawn from a uniform distribution $U[0, R_o]$, with $R_o \ll 1$. The numerical results shown in this paper for $Q_o = R_o = 10^{-12}$ are indistinguishable from those obtained with $Q_o = 0$. No differences arise from setting $R_o = 0$ for graded teachers, but it is necessary to keep a nonzero R_o in order to break the symmetry among teacher nodes and achieve specialization in the case of isotropic teachers.

III. THE ROLE OF THE LEARNING RATE

We now examine the role of the learning rate η in the convergence of the training process. The time evolution of the order parameters and the generalization error for different values of η reveals three distinct regimes: a low η regime characterized by a long suboptimal transient due to trapping in a symmetric subspace of solutions; a regime of optimal η values characterized by a rapid escape from the symmetric subspace followed by convergence to the optimal solution; and a high η regime characterized by an uncontrolled growth of the norms of the student vectors.

We illustrate the corresponding evolution of the order parameters through numerical results shown in Fig. 1 for the realizable case $K = M = 3$. The teacher is graded and specified by $T_{nm} = n \ \delta_{mn}$. Three different regimes are clearly observed. Learning at small η, illustrated for $\eta = 0.1$ in Fig. 1a, results in a system trapped for very long times in a symmetric subspace controlled by an unstable suboptimal solution which exhibits no differentiation among student hidden units. The evolution of the overlaps R_{in} indicates that during the transient the student vectors $\mathbf{J}_i$ become identical to each other and model the various teacher units with the same degree of success. The only differentiation is the one among the teacher vectors, due to their different norms $T_{nn} = n$. Trapping in the symmetric subspace prevents the specialization needed to achieve the optimal solution, and the generalization error remains finite, as shown in Fig. 1d. The symmetric solution is unstable, and the perturbation introduced through nonsymmetric initial conditions for the norms Q_{ii} eventually takes over, but the transient can be very long. In the example presented here, the first signs of specialization appear around $\alpha = 750$. Fast specialization is achieved by choosing a larger value of η, as shown in Fig. 1b for $\eta = 0.9$. In this regime the overlaps R_{in} evolve first within the symmetric subspace, but the unstable solution is quickly abandoned as the system evolves towards the optimal solution. The student units become specialized, and the matrix R of student-teacher overlaps becomes identical to the matrix T, except for a permutation symmetry associated with the arbitrary labeling of the hidden units. As shown in Fig. 1d, an early plateau in the generalization error is followed by a monotonic decrease towards zero once the specialization begins. The ability to reach the optimal solution is lost for very large η, as illustrated for $\eta = 3.6$ in Fig. 1c. The large η regime is characterized by an uncontrolled growth of the norm of the student weight vectors with no specialization. The generalization error no longer decays to zero, but approaches a value $\epsilon_\infty \equiv \epsilon_g(\alpha \to \infty) > 0$, as shown in Fig. 1d. The three learning regimes illustrated in Fig. 1 correspond to those found in the detailed analysis of the $K = 2$, $M = 1$ case [14].

The evolution of the generalization error ϵ_g with the normalized number of examples α provides a useful heuristic characterization of the three learning regimes discussed above. Consider the number $\alpha^\star$ of examples needed to achieve a fixed level of performance, chosen here to be $\epsilon_g = 0.01$. We have investigated the dependence of $\alpha^\star$ on η for realizable learning scenarios with $K = M$. Results shown in Fig. 2 for several values of K correspond to learning a teacher specified by $T_{nm} = n\ \delta_{mn}$. The time evolution is followed up to $\alpha = 4000$. The divergence of $\alpha^\star$ at small η signals trapping in the symmetric subspace which prevents the system from achieving the required level of performance within the allotted time. The subsequent monotonic decrease of $\alpha^\star$ with η indicates faster convergence to the optimal solution. The minimum value of $\alpha^\star$ identifies the optimal learning rate η_{opt}; increases in $\alpha^\star$ for $\eta > \eta_{opt}$ culminate in a cutoff at η_{max}. The failure of the system to achieve a low generalization error within the allotted time for $\eta > \eta_{max}$ signals nonconvergent training.

A more rigorous evaluation of η_{opt} and η_{max} follows from the stability analysis of the optimal solution to be found in Section IV. We conclude this section with a discussion of the heuristic estimates of η_{opt} and η_{max} obtained from Fig. 2. The dependence of η_{opt} and η_{max} on the number K of hidden units in both student and teacher networks shown in Fig. 3 exhibits a monotonic decrease suggestive of inverse proportionality. It is of particular interest to examine the minimal number of examples needed to achieve the desired level of performance. The dependence of $\alpha_{opt} = \alpha^\star(\eta_{opt})$ on the number K of hidden units shown in Fig. 4 indicates an exponential increase for $K > 5$.

A final heuristic observation concerns the transient behavior due to trapping in the symmetric subspace. Results shown in Fig. 1 indicate that the time needed to escape from the symmetric subspace increases with decreasing η. This behavior is easily understood by considering the equations of motion (14) in the small η regime: the term proportional to η^2 can be neglected, resulting in a system of equations that becomes independent of η under the rescaling $\alpha\eta \equiv \widetilde{\alpha}$. A universal plateau in the $\epsilon_g(\widetilde{\alpha})$ curve will terminate at a value $\widetilde{\alpha}_T$ which signals the onset of specialization. Trapping times in the small η regime are expected to increase as $\alpha_T = \widetilde{\alpha}_T/\eta$ with decreasing η.

The range of validity of the small η scaling can be estimated through the product $\widetilde{\alpha}^\star = \eta\alpha^\star$, shown in Fig. 5 as a function of η for several values of K. Scaling is seen to hold in a range $0 < \eta < \eta^\star$. The decrease of $\eta^\star$ with increasing K signals an earlier onset of quadratic effects in the time evolution of the order parameters, indicating an increase in the magnitude of the I_4 terms relative to the I_3 terms in the equations of motion (14).

IV. STRUCTURE OF THE SOLUTIONS

One of the most important aspects of learning in multilayer networks is the specialization of the hidden units, an essential ingredient to the emergence of generalization ability. Numerical solutions for the time evolution of the order parameters and the generalization error for large networks of the type studied here indicate that the training process takes place in two phases: a symmetric phase which exhibits no differentiation among student hidden units, and a subsequent phase characterized by a specialization of the student nodes leading to optimal network performance.

In this section we present an analysis of the suboptimal solution which controls the symmetric phase, the onset of specialization, and the optimal solution to which the system converges asymptotically. For simplicity we consider a learnable scenario with $K = M$, and focus on an isotropic teacher $T_{nm} = \delta_{mn}$, $1 \leq m, n \leq M$.

A. Suboptimal solution: the symmetric plateau

Curves for the time evolution of the generalization error for different values of η shown in Fig. 6 identify trapping in the symmetric phase as a small η phenomenon. We therefore consider the equations of motion (14) in the small η limit, and neglect terms proportional to η^2. The symmetric phase is characterized by undifferentiated student vectors of similar norms: $Q_{ii} = Q$ for all $1 \leq i \leq K$, similar correlations among themselves: $Q_{ik} = C$ for all $1 \leq i, k \leq K$, $i \neq k$, and similar correlations with the teacher vectors: $R_{in} = R$ for all $1 \leq i, n \leq K$.

The dynamical equations for R, Q, and C follow directly form Eq. (14):

$$
\begin{aligned}
\frac{dR}{d\tilde{\alpha}} &= \frac{2}{\pi}\frac{1}{(1+Q)}\left\{\frac{1+Q-KR^2}{\sqrt{2(1+Q)-R^2}} - \frac{R}{\sqrt{1+2Q}} - \frac{R(K-1)(1+Q-C)}{\sqrt{(1+Q)^2-C^2}}\right\} , \\
\frac{dQ}{d\tilde{\alpha}} &= \frac{4}{\pi}\frac{1}{(1+Q)}\left\{\frac{KR}{\sqrt{2(1+Q)-R^2}} - \frac{Q}{\sqrt{1+2Q}} - \frac{C(K-1)}{\sqrt{(1+Q)^2-C^2}}\right\} , \\
\frac{dC}{d\tilde{\alpha}} &= \frac{4}{\pi}\frac{1}{(1+Q)}\left\{\frac{KR(1+Q-C)}{\sqrt{2(1+Q)-R^2}} - \frac{C}{\sqrt{1+2Q}} - \frac{(1+Q)(Q+C(K-2))-C^2(K-1)}{\sqrt{(1+Q)^2-C^2}}\right\} .
\end{aligned} \tag{20}
$$

The generalization error (7) is given in this regime by

$$
\epsilon_g(\mathbf{J}) = \frac{K}{\pi}\left\{\frac{\pi}{6} + \arcsin\left(\frac{Q}{1+Q}\right) + (K-1)\arcsin\left(\frac{C}{1+Q}\right) - (2K)\arcsin\left(\frac{R}{\sqrt{2(1+Q)}}\right)\right\} . \tag{21}
$$

Fixed point solutions for Eq. (20) follow from setting $dR/d\tilde{\alpha} = dQ/d\tilde{\alpha} = dC/d\tilde{\alpha} = 0$, and require $Q = \pm C$. Since the solutions with $C = -Q$ are unphysical for $K > 2$, we focus on the $Q = C$ subspace to obtain

$$
\begin{aligned}
Q^* = C^* &= \frac{1}{2K-1} , \\
R^* &= \sqrt{\frac{Q^*}{K}} = \frac{1}{\sqrt{K(2K-1)}} .
\end{aligned} \tag{22}
$$

The corresponding generalization error is given by

$$
\epsilon_g^* = \frac{K}{\pi}\left\{\frac{\pi}{6} - K\arcsin\left(\frac{1}{2K}\right)\right\} . \tag{23}
$$

A simple geometrical picture explains the relation $Q^* = C^* = K(R^*)^2$ at the symmetric fixed point. The learning process confines the student vectors $\{\mathbf{J}_i\}_{1\le i\le K}$ to the M-dimensional subspace $\mathcal{S}_B$ spanned by the set of teacher vectors $\{\mathbf{B}_n\}_{1\le n\le M}$. For $T_{nm} = \delta_{nm}$ the teacher vectors form an orthonormal set: $\mathbf{B}_n = \mathbf{e}_n$, with $\mathbf{e}_n \cdot \mathbf{e}_m = \delta_{nm}$, $1 \le n, m \le M$, and provide an expansion for the weight vectors of the trained student: $\mathbf{J}_i^* = \sum_{n=1}^{M} R_{in}\mathbf{e}_n$. The student-teacher overlaps R_{in} are independent of i in the symmetric phase and independent of n for an isotropic teacher: $R_{in} = R^*$ for all $1 \le i \le K$, $1 \le n \le M$. The expansion $\mathbf{J}^* = R^* \sum_{n=1}^{M} \mathbf{e}_n$ results in $Q^* = C^* = M(R^*)^2 = K(R^*)^2$ for the $M = K$ case considered here. This geometrical description identifies $\mathbf{J}^*$ as a vector pointing in the $(1, \ldots, 1)$ direction in the M-dimensional space spanned by the $\{\mathbf{e}_n\}$, but it does not provide information on its norm Q^*.

The symmetric solution discussed here is unstable, and it describes the asymptotic learning behavior only when initial conditions for the order parameters are chosen to satisfy the symmetric constraints. For $M = K$ and $T_{nm} = \delta_{nm}$, the requirements are $R_{in} = R_o$ for all i, n and $Q_{ik} = Q_o$ for all i, k. Results shown in Fig. 7 correspond to a symmetric initialization with $Q_o = 0.5$ and $R_o = 0$ for $M = K = 3$ hidden units. The asymptotic values $Q^* = C^* = 0.2$, $R^* = 0.2582$, and $\epsilon_g^* = 0.0203$ are in agreement with the theoretical predictions of Eqs. (22) and (23).

The specific values assigned to the order parameters as initial conditions are largely irrelevant, as they control the behavior of the system only during the short transient needed to relax onto the symmetric phase described by Eqs. (22) and (23), but nonsymmetric initializations of the student vectors with respect to the teacher vectors, whatever their nature and magnitude, introduce a fundamental perturbation that eventually drives the system away from the symmetric subspace. The length of the symmetric plateau is controlled by the degree of asymmetry in the initial conditions (as observed in [14] for the $K = 2$, $M = 1$ case) and by the learning rate η. The small η analysis developed in this section results in a universal curve for the generalization error as a function of the rescaled variable $\tilde{\alpha} = \eta\alpha$ for any specific choice of initial conditions. As shown in Fig. 8 for $M = K = 3$, trapping in the symmetric subspace is seen to control the generalization error up to $\tilde{\alpha}_T \simeq 600$ for initial conditions $R_{in} = U[0, 10^{-12}]$ and $Q_{ik} = U[0, 0.5]\,\delta_{ik}$, as in Fig. 6. Escape times for identical initial conditions are expected to scale like $\alpha_T \simeq 600/\eta$, in quantitative agreement with the shrinking symmetric plateau shown in Fig. 6.

An additional feature of Fig. 6 remains unexplained by the small η truncation of the equations of motion leading to Fig. 8: the universal curve of ϵ_g as a function of $\tilde{\alpha}$ predicts a unique value of the generalization error at the symmetric plateau. The increase in the height of the plateau with increasing η observed in Fig. 6 is obviously a second order effect; in order to account for it we need to reexamine the structure of the symmetric solution under the full equations of motion (14). We show in Fig. 9 the evolution of the order parameters and generalization error for $M = K = 3$ starting from the same initial conditions used in Figs. 6 and 8. Curves for $\eta = 0.1$ and $\eta = 0.9$ reveal that the symmetric phase is in both cases characterized by student-teacher overlaps $R_{in} = R^* = 1/\sqrt{K(2K-1)}$ for all i, n and student-student overlaps $Q_{ik} = C^* = 1/(2K-1)$ for all $i \neq k$. It is the norms $Q_{ii} = Q$ of the student vectors which deviate from the predictions of the small η analysis: the value of Q does not converge to $Q^* = C^* = 1/(2K-1)$ but remains larger at a value $Q = Q^* + \Delta$. As illustrated in Fig. 9, the gap Δ between diagonal and off-diagonal elements increases with increasing η. This is the mechanism for excess generalization error; a first order expansion of Eq. (21) around $R = R^*$, $C = C^*$, and $Q = Q^* + \Delta$ yields

$$\epsilon_g = \frac{K}{\pi}\left\{\frac{\pi}{6} - K\arcsin\left(\frac{1}{2K}\right) + \sqrt{\frac{2K-1}{2K+1}}\,\Delta\right\} , \tag{24}$$

in agreement with the trend observed both in Fig. 6 and Fig. 9.

The excess norm Δ of the student vectors has a simple interpretation in terms of the geometrical picture developed earlier in this section: learning at finite η results in student weight vectors not completely confined to the subspace $\mathcal{S}_B$. The weight vectors of the trained student can then be written as $\mathbf{J}_i = R^* \sum_{n=1}^{M} \mathbf{e}_n + \mathbf{J}_i^{\perp}$, where $\mathbf{J}_i^{\perp}$ indicates the component of $\mathbf{J}_i$ in the orthogonal subspace: $\mathbf{J}_i^{\perp} \cdot \mathbf{e}_n = 0$ for all $1 \leq n \leq M$. Student weight vectors are not constrained to be identical; they differ through orthogonal components $\mathbf{J}_i^{\perp}$ which are typically uncorrelated: $\mathbf{J}_i^{\perp} \cdot \mathbf{J}_k^{\perp} = 0$ for $i \neq k$. Correlations $Q_{ik} = C$ still satisfy $C = C^* = M(R^*)^2$, but norms $Q_{ii} = Q$ are given by $Q = M(R^*)^2 + \| \mathbf{J}^{\perp} \|^2$. The gap is then identified as $\Delta = \| \mathbf{J}^{\perp} \|^2$. Learning at very small η tends to eliminate $\mathbf{J}^{\perp}$: second order effects become negligible and the student vectors are more effectively confined to $\mathcal{S}_B$.

B. Onset of specialization

Escape from the symmetric subspace signals the onset of hidden unit specialization. As shown in Fig. 9, the process is driven by a breaking of the uniformity of the student-teacher correlations: each student node becomes increasingly specialized to a specific teacher node, while its overlap with the remaining teacher nodes decreases and eventually decays to zero.

The matrix of student-teacher overlaps can no longer be characterized by a unique parameter, as we need to distinguish between a dominant overlap R between a given student node and the teacher node it begins to imitate, and secondary overlaps S between the same student node and the remaining teacher nodes. The student nodes can be relabeled so as to bring the matrix of student-teacher overlaps to the form $R_{in} = R\delta_{in} + S(1-\delta_{in})$. The emerging differentiation among student vectors results in a decrease of the overlaps $Q_{ik} = C$ for $i \neq k$, while their norms $Q_{ii} = Q$ increase. The matrix of student-student overlaps takes the form $Q_{ik} = Q\delta_{ik} + C(1-\delta_{ik})$.

In order to describe the incipient specialization as the student network escapes from the symmetric subspace we extend the small η analysis of the preceding section to allow for $S \neq R$. The dynamical equations for R, S, Q, and C follow from the truncated form of Eq. (14):

$$\begin{aligned}
\frac{dR}{d\widetilde{\alpha}} &= \frac{2}{\pi}\frac{1}{(1+Q)}\left\{\frac{1+Q-R^2}{\sqrt{2(1+Q)-R^2}} - \frac{RS(K-1)}{\sqrt{2(1+Q)-S^2}} - \frac{R}{\sqrt{1+2Q}} \right.\\
&\qquad\left. - \frac{(1+Q)S(K-1) - RC(K-1)}{\sqrt{(1+Q)^2-C^2}}\right\} ,\\
\frac{dS}{d\widetilde{\alpha}} &= \frac{2}{\pi}\frac{1}{(1+Q)}\left\{\frac{1+Q-S^2(K-1)}{\sqrt{2(1+Q)-S^2}} - \frac{RS}{\sqrt{2(1+Q)-R^2}} - \frac{S}{\sqrt{1+2Q}} \right.\\
&\qquad\left. - \frac{(1+Q)(R+S(K-2)) - SC(K-1)}{\sqrt{(1+Q)^2-C^2}}\right\} ,\\
\frac{dQ}{d\widetilde{\alpha}} &= \frac{4}{\pi}\frac{1}{(1+Q)}\left\{\frac{R}{\sqrt{2(1+Q)-R^2}} + \frac{S(K-1)}{\sqrt{2(1+Q)-S^2}} - \frac{Q}{\sqrt{1+2Q}} - \frac{C(K-1)}{\sqrt{(1+Q)^2-C^2}}\right\} ,\\
\frac{dC}{d\widetilde{\alpha}} &= \frac{4}{\pi}\frac{1}{(1+Q)}\left\{\frac{(1+Q)S - RC}{\sqrt{2(1+Q)-R^2}} + \frac{(1+Q)(R+S(K-2)) - SC(K-1)}{\sqrt{2(1+Q)-S^2}} - \frac{C}{\sqrt{1+2Q}} \right.\\
&\qquad\left. - \frac{(1+Q)(Q+C(K-2)) - C^2(K-1)}{\sqrt{(1+Q)^2-C^2}}\right\} .
\end{aligned} \tag{25}$$

for an isotropic teacher $T_{nm} = \delta_{nm}$. The generalization error (7) is given in this regime by:

$$\epsilon_g(\mathbf{J}) = \frac{K}{\pi}\left\{\frac{\pi}{6} + \arcsin\left(\frac{Q}{1+Q}\right) + (K-1)\arcsin\left(\frac{C}{1+Q}\right) - 2\arcsin\left(\frac{R}{\sqrt{2(1+Q)}}\right)\right.$$
$$\left. - \, 2(K-1)\arcsin\left(\frac{S}{\sqrt{2(1+Q)}}\right)\right\} \ . \tag{26}$$

We now investigate the equations of motion (25) in the vicinity of the symmetric fixed point (22) through deviations $r = R - R^*$, $s = S - S^*$, $q = Q - Q^*$, and $c = C - C^*$, from $R^* = S^* = 1/\sqrt{K(2K-1)}$ and $Q^* = C^* = 1/(2K-1)$. The geometrical interpretation of the symmetric fixed point developed in the preceding section provides an expansion for the student weight vectors $\mathbf{J}_i^* = \sum_{n=1}^{K} R_{in}\mathbf{B}_n$. The orthogonal components $\mathbf{J}_i^{\perp}$ can be neglected in the small η regime, resulting in norms $Q_{ii} = Q$ and overlaps $Q_{ik} = C$ fully determined by the student-teacher overlaps; for $R_{in} = R\delta_{in} + S(1-\delta_{in})$ we obtain $Q = R^2 + S^2(K-1)$ and $C = 2RS + S^2(K-2)$. A first order expansion in the deviations r and s yields $Q = Q^* + 2R^*(r + s(K-1))$ and $C = C^* + 2R^*(r + s(K-1))$. Therefore $q = c = 2R^*(r + s(K-1))$, and the fixed point equality $Q^* = C^*$ is preserved to first order. This observation is consistent with numerical results as illustrated in Fig. 9: it is the differentiation between R and S which signals the escape from the symmetric subspace; the differentiation between Q and C occurs for larger values of α.

The constraint $Q = C$ is consistent with the equations of motion (25): $dQ/d\widetilde{\alpha} = dC/d\widetilde{\alpha}$ to first order in r, s, and q, at $Q = C$. Under this constraint the equations of motion (25) reduce to:

$$\frac{dR}{d\widetilde{\alpha}} = \frac{2}{\pi}\frac{1}{(1+Q)}\left\{\frac{1+Q-R^2}{\sqrt{2(1+Q)-R^2}} - \frac{RS(K-1)}{\sqrt{2(1+Q)-S^2}} - \frac{R+S(K-1)+Q(S-R)(K-1)}{\sqrt{1+2Q}}\right\} ,$$
$$\frac{dS}{d\widetilde{\alpha}} = \frac{2}{\pi}\frac{1}{(1+Q)}\left\{\frac{1+Q-S^2(K-1)}{\sqrt{2(1+Q)-S^2}} - \frac{RS}{\sqrt{2(1+Q)-R^2}} - \frac{R+S(K-1)-Q(S-R)}{\sqrt{1+2Q}}\right\} ,$$
$$\frac{dQ}{d\widetilde{\alpha}} = \frac{4}{\pi}\frac{1}{(1+Q)}\left\{\frac{R}{\sqrt{2(1+Q)-R^2}} + \frac{S(K-1)}{\sqrt{2(1+Q)-S^2}} - \frac{QK}{\sqrt{1+2Q}}\right\} , \tag{27}$$

which are expanded to first order in r, s, and q to obtain:

$$\frac{dr}{d\widetilde{\alpha}} = \frac{1}{\pi}\frac{(2K-1)}{K\,(2K+1)^{3/2}}\left\{-\frac{2K(4K^2+K-1)}{(2K-1)^{3/2}}\,r - \frac{2K(4K^3-2K^2-3K+1)}{(2K-1)^{3/2}}\,s + 3K^{3/2}\,q\right\} ,$$
$$\frac{ds}{d\widetilde{\alpha}} = \frac{1}{\pi}\frac{(2K-1)}{K\,(2K+1)^{3/2}}\left\{-\frac{2K(4K^2+2K-1)}{(2K-1)^{3/2}}\,r - \frac{2K(4K^3-2K^2-4K+1)}{(2K-1)^{3/2}}\,s + 3K^{3/2}\,q\right\} ,$$
$$\frac{dq}{d\widetilde{\alpha}} = \frac{2}{\pi}\frac{(2K-1)}{K\,(2K+1)^{3/2}}\left\{\frac{4K^{5/2}}{(2K-1)}\,r + \frac{4K^{5/2}(K-1)}{(2K-1)}\,s - \frac{K^2(4K-1)}{(2K-1)^{1/2}}\,q\right\} . \tag{28}$$

The corresponding generalization error is given by:

$$\epsilon_g(\mathbf{J}) = \epsilon_g^* - \frac{2}{\pi}\frac{K^{3/2}}{(2K+1)^{1/2}}\,r - \frac{2}{\pi}\frac{K^{3/2}(K-1)}{(2K+1)^{1/2}}\,s + \frac{K^2}{\pi}\frac{(2K-1)^{1/2}}{(2K+1)^{1/2}}\,q , \tag{29}$$

with $\epsilon_g^* = (K/6) - (K^2/\pi)\arcsin(1/2K)$, as in Eq. (23).

The geometry of student vectors confined to $\mathcal{S}_B$ imposes the additional constraint $q = 2R^*(r + s(K-1))$. The first order change in the generalization error (29),

$$\epsilon_g(\mathbf{J}) - \epsilon_g^* = \frac{K^2}{\pi}\frac{(2K-1)^{1/2}}{(2K+1)^{1/2}}\,q - \frac{2}{\pi}\frac{K^{3/2}}{(2K+1)^{1/2}}\,(r + s\,(K-1)) , \tag{30}$$

vanishes under the condition $\sqrt{K(2K-1)}\;q = 2\;(r + s(K-1))$. As long as this geometric constraint is satisfied, the order parameters R, S, and $Q = C$ can experience fluctuations around $R^* = S^*$ and $Q^* = C^*$ without affecting the value ϵ_g^* of the generalization error. The equations of motion for the fluctuations of R and S are:

$$\begin{pmatrix} \dot{r} \\ \dot{s} \end{pmatrix} = -\frac{2}{\pi} \frac{1}{(2K-1)^{1/2}(2K+1)^{3/2}} \begin{pmatrix} (4K^2-5K+2) & (4K^3-8K^2+6K-2) \\ (4K^2-4K+2) & (4K^3-8K^2+5K-2) \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix} . \tag{31}$$

The dynamical evolution described by the linearized equations of motion (31) is characterized by eigenvalues $\lambda_1 = -\upsilon K(2K-1)^2$ and $\lambda_2 = \upsilon K$, with $\upsilon = (2/\pi)(2K-1)^{-1/2}(2K+1)^{-3/2}$. The first mode corresponds to an irrelevant perturbation ($\lambda_1 < 0$ for all K); its associated eigenvector $\mathbf{V}_1 = (1,1)$ describes a perturbation with $r = s$ which does not break the $R = S$ symmetry. The onset of specialization requires an enhancement of the overlap $R = R^* + r$ between a given student node and the teacher node it is learning to imitate, while the overlap $S = S^* + s$ between the same student node and the remaining teacher nodes is weakened. It is the second mode which provides the mechanism: this relevant perturbation ($\lambda_2 > 0$ for all K) breaks the $R = S$ symmetry in the required way. The corresponding eigenvector $\mathbf{V}_2$ is characterized through its direction $s = -\tilde{\gamma} r$ to obtain $\tilde{\gamma} = (4K^2-4K+2)/(4K^3-8K^2+6K-2) = 1/(K-1)$. Note that $\tilde{\gamma} > 0$ for all K, to guarantee $s < 0$ for $r > 0$. The dependence of $\tilde{\lambda} = \lambda_2$ on the number $K = M$ of hidden units is shown in Fig. 10a; as $K \to \infty$, $\tilde{\lambda} \sim (2\pi K)^{-1}$: the time constant associated with the escape from the symmetric subspace increases linearly with the size K of the network.

We now investigate the general conditions under which positive deviations from R^* and negative deviations from S^* will be sustained and enhanced by the dynamical evolution (31). We propose $s = -\gamma r$, with $\gamma > 0$ to guarantee $s < 0$ for $r > 0$. The growth of a positive fluctuation r requires $\dot{r} > 0$, a condition satisfied for $\gamma > \gamma_R = (4K^2-5K+2)/(4K^3-8K^2+6K-2)$. At $\gamma = \gamma_R$, $\dot{r} = 0$, and the deviation r remains stationary. The growth in absolute value of a negative fluctuation s requires $\dot{s} < 0$, a condition satisfied for $\gamma < \gamma_S = (4K^2-4K+2)/(4K^3-8K^2+5K-2)$. At $\gamma = \gamma_S$, $\dot{s} = 0$, and the deviation s remains stationary.

The dependence of both γ_S and γ_R on the number $K = M$ of hidden nodes is shown in Fig. 10b. In order to satisfy both $\dot{r} > 0$ and $\dot{s} < 0$ as required for further specialization, the ratio γ between $|s|$ and r must be confined between the two curves: $\gamma_R < \gamma < \gamma_S$. Note that $\gamma_S > \gamma_R$ for all K, and their difference $\Delta\gamma = \gamma_S - \gamma_R$ vanishes as $1/K^2$ as $K \to \infty$. The eigenmode value $\tilde{\gamma}$ satisfies the condition $\gamma_R < \tilde{\gamma} < \gamma_S$ for all K. As $K \to \infty$, both γ_S and γ_R, as well as $\tilde{\gamma}$, go to zero as $1/K$.

We have thus identified the mechanism for the onset of specialization: positive fluctuations r which enhance the overlap R accompanied by negative fluctuations s which weaken S. The ratio γ between the decrease in S and the increase in R must be in the range (γ_R, γ_S) for the fluctuations to be dynamically amplified, leading to further specialization. In the large K limit both γ_S and γ_R become vanishingly small, and so does the required value of s; the onset of specialization in large networks is primarily controlled by an enhancement in R.

Specialization as described here and illustrated in Fig. 9 is a simultaneous process in which each student node acquires a strong correlation with a specific teacher node while becoming decorrelated from the remaining teacher nodes. Such synchronous escape from the symmetric phase is characteristic of learning scenarios where the target task is defined through an isotropic teacher. In the case of a graded teacher we find that specialization occurs through a sequence of escapes from the symmetric subspace, ordered according to the relevance of the corresponding teacher nodes. The process is illustrated for $K = M = 4$ in Fig 11. The evolution of the student norms shown in Fig. 11a for $\eta = 0.03$ demonstrates the asymptotic specialization in which each student node imitates a specific teacher node: $Q_{11} \to T_{44} = 4$, $Q_{33} \to T_{33} = 3$, $Q_{44} \to T_{22} = 2$, and $Q_{22} \to T_{11} = 1$. Sequential escape of the student weight vectors from the symmetric subspace follows the order imposed by the relevance of the corresponding teacher weight vectors. The evolution of the generalization error shown in Fig. 11b reflects these successive transitions: a plateau characteristic of trapping in the symmetric subspace is followed by a monotonic decrease where three observable inflection points correspond to the specialization of $i = 1$ onto $n = 4$, followed by that of $i = 3$ onto $n = 3$, and that of $i = 4$ onto $n = 2$. There is no visible signature of the subsequent specialization of $i = 2$ onto $n = 1$. Such structure in the $\epsilon_g(\alpha)$ curve, not uncommon in realistic learning scenarios, signals the existence of graded teacher nodes.

C. Optimal solution: convergence to perfect generalization

The onset of specialization has been described in the preceding section as a breaking of the $R = S$ symmetry: each student node becomes specialized to a specific teacher node (R increases), while its correlation with the remaining teacher nodes weakens (S decreases); the secondary overlap S decays to zero as the process continues. Further specialization involves the decay to zero of the student-student correlations C and the growth of the norms Q of the student vectors. The subsequent evolution of the system converges to an optimal solution with perfect generalization.

Numerical experiments as illustrated in Fig. 9 indicate that the decay of the off-diagonal elements S and C to zero precedes the convergence of R and Q to their asymptotic values. This observation is confirmed by a linear analysis of the truncated equations of motion (25) around the asymptotic fixed point at $S^* = C^* = 0$, $R^* = Q^* = 1$. We therefore describe convergence to the optimal solution by applying the full equations of motion (14) to a phase characterized by $R_{in} = R\delta_{in}$ and $Q_{ik} = Q\delta_{ik}$. The resulting dynamical equations for R and Q are:

$$
\begin{aligned}
\frac{dR}{d\alpha} &= \frac{2}{\pi}\frac{\eta}{(1+Q)}\left\{\frac{1+Q-R^2}{\sqrt{2(1+Q)-R^2}} - \frac{R}{\sqrt{1+2Q}}\right\} , \\
\frac{dQ}{d\alpha} &= \frac{4}{\pi}\frac{\eta}{(1+Q)}\left\{\frac{R}{\sqrt{2(1+Q)-R^2}} - \frac{Q}{\sqrt{1+2Q}}\right\} + \frac{4}{\pi^2}\frac{\eta^2}{\sqrt{1+2Q}} \\
&\times \left\{(K-1)\left[\frac{\pi}{6} + \arcsin\left(\frac{Q}{1+Q}\right) - 2\arcsin\left(\frac{R}{\sqrt{2(1+Q)}}\right)\right]\right. \\
&+ \left.\left[\arcsin\left(\frac{1+2Q-2R^2}{2(1+2Q-R^2)}\right) + \arcsin\left(\frac{Q}{1+3Q}\right) - 2\arcsin\left(\frac{R}{\sqrt{2(1+3Q)(1+2Q-R^2)}}\right)\right]\right\} .
\end{aligned}
\tag{32}
$$

The generalization error (7) is given in this regime by

$$
\epsilon_g(\mathbf{J}) = \frac{K}{\pi}\left\{\frac{\pi}{6} + \arcsin\left(\frac{Q}{1+Q}\right) - 2\arcsin\left(\frac{R}{\sqrt{2(1+Q)}}\right)\right\} . \tag{33}
$$

The fixed point solution of Eq. (32) follows from setting $dR/d\alpha = dQ/d\alpha = 0$ to obtain

$$
(R^*)^2 = Q^* = 1 . \tag{34}
$$

This fixed point corresponds to the optimal solution, with $\epsilon_g^* = 0$.

The asymptotic behavior follows from linearizing the equations of motion (32) around the fixed point at $R^* = Q^* = 1$. We first consider the small η regime, and neglect terms proportional to η^2 in the equation of motion for Q. The resulting truncated version of Eq. (32) is independent of the size K of the networks. The time evolution of the deviations $r = 1 - R$ and $q = 1 - Q$ is given by:

$$
\begin{pmatrix} \dot r \\ \dot q \end{pmatrix} = \frac{2}{\pi}\frac{\sqrt{3}}{9}\,\eta \begin{pmatrix} -4 & 3/2 \\ 4 & -3 \end{pmatrix}\begin{pmatrix} r \\ q \end{pmatrix} , \tag{35}
$$

in the small η regime. The eigenvalues are $\lambda_1 = -6\nu\eta$ and $\lambda_2 = -\nu\eta$, with $\nu = (2\sqrt{3})/(9\pi)$. The corresponding eigenvectors are $\mathbf{V}_1 = \frac{1}{5}(3,-4)$ for the fast mode and $\mathbf{V}_2 = \frac{1}{\sqrt{5}}(1,2)$ for the slow mode. Note that $\lambda_1 = 6\lambda_2$.

The linearization of the full equations of motion (32) around the $R^* = Q^* = 1$ fixed point leads to

$$
\begin{pmatrix} \dot r \\ \dot q \end{pmatrix} = \frac{2}{\pi}\frac{\sqrt{3}}{9}\,\eta \begin{pmatrix} -4 & 3/2 \\ (4-2\eta\mu) & (-3+\eta\mu) \end{pmatrix}\begin{pmatrix} r \\ q \end{pmatrix} , \tag{36}
$$

with $\mu = \sqrt{3}\,(2/\pi)(K - 1 + 3/\sqrt{5})$. The eigenvalues are $\lambda_1 = -\nu\eta(6 - \eta\mu)$ and $\lambda_2 = -\nu\eta$, with $\nu = (2\sqrt{3})/(9\pi)$ as before. Note that λ_2 still depends only linearly on η, while λ_1 has acquired a quadratic contribution of opposite sign. The eigenvector $\mathbf{V}_2 = \frac{1}{\sqrt{5}}(1,2)$ remains unchanged, while $\mathbf{V}_1$ acquires a dependence on η which is easily obtainable but of no relevance to the analysis that follows. The dependence of both eigenvalues on η is shown in Fig. 12 for $M = K = 3$.

The existence of two negative eigenvalues for a finite range of values of η implies exponential convergence of the order parameters R and Q to their optimal values. In the small η regime convergence is controlled by the slow eigenvalue λ_2, and both r and q decay as $\exp(\lambda_2\alpha)$. The relaxation time $\tau = -1/\lambda_2$ decreases with increasing η until the crossing of the eigenvalues illustrated in Fig. 12. As η increases further, the slow mode is the one associated with λ_1, and $r, q \sim \exp(\lambda_1\alpha)$. The relaxation time $\tau = -1/\lambda_1$ increases with increasing η, and diverges as $\eta \longrightarrow \eta_{max}$ defined by $\lambda_1(\eta = \eta_{max}) = 0$. The fixed point at $R^* = Q^* = 1$ becomes unstable as λ_1 turns positive; the optimal solution with $\epsilon_g^* = 0$ is not accessible for $\eta > \eta_{max}$. Exponential convergence of the order parameters to their optimal value is guaranteed for $0 < \eta < \eta_{max}$, with

$$
\eta_{max} = \frac{6}{\mu} = \frac{\pi\sqrt{3}}{K - 1 + 3/\sqrt{5}} . \tag{37}
$$

As $\eta \longrightarrow \eta_{max}$ the relaxation time diverges as

$$\tau = \frac{A}{(\eta_{max} - \eta)} \ , \tag{38}$$

with $A = (\pi\sqrt{3})/4$, independent of K.

The generalization error decays to $\epsilon_g^* = 0$ for all learning rates in the range $(0, \eta_{max})$. In order to identify the corresponding relaxation time consider the expansion of Eq. (33) to second order in r and q:

$$\epsilon_g \approx \frac{K}{\pi}\frac{\sqrt{3}}{3}\left[(2r-q) - \frac{1}{12}(2r-q)^2 - \frac{1}{3}q(q-r)\right] \ . \tag{39}$$

Since the mode associated with λ_2 cannot contribute to the asymptotic decay of the linear combination $(2r-q)$, the linear term decays as $\exp(\lambda_1\alpha)$. This rate of convergence is to be compared to that of the quadratic terms q^2 and qr, which decay as $\exp(2\lambda_2\alpha)$ in the small η regime. Since in this regime $\lambda_1 = 6\lambda_2$, it is the quadratic terms which control the decay of the generalization error. The corresponding relaxation time $\tau = -1/2\lambda_2$ decreases monotonically with increasing η and reaches its optimal value at the crossing between $2\lambda_2$ and λ_1. As η increases beyond η_{opt} defined by $\lambda_1(\eta = \eta_{opt}) = 2\lambda_2(\eta = \eta_{opt})$, the relaxation time $\tau = -1/\lambda_1$ increases with increasing η, and diverges as $\eta \rightarrow \eta_{max}$, as described in Eq. (38). The learning rate η_{opt} which guarantees the fastest asymptotic decay for the generalization error is given by

$$\eta_{opt} = \frac{4}{\mu} = \frac{2}{3}\frac{\pi\sqrt{3}}{K - 1 + 3/\sqrt{5}} \ . \tag{40}$$

Results (37) and (40) for the maximal and optimal learning rates establish the asymptotic $1/K$ decay of both η_{max} and η_{opt}, and imply a surprisingly general result: $\eta_{opt} = (2/3)\ \eta_{max}$ for all K. The full dependence of both quantities on the number $K = M$ of hidden units, shown in Fig. 13, is in good agreement with the heuristic data of Fig. 3. The analytic results obtained here for an isotropic teacher ($T_{nm} = \delta_{nm}$) provide reliable predictions for more complex learning scenarios ($T_{nm} = n\ \delta_{nm}$). The optimal decay rate $\lambda_{opt} = -2\lambda_2(\eta = \eta_{opt})$ is given by

$$\lambda_{opt} = \frac{8}{9}\frac{1}{K - 1 + 3/\sqrt{5}} \ , \tag{41}$$

and shown as a function of K in Fig. 14. As $K \rightarrow \infty$, $\lambda_{opt} \sim 8/(9K)$: the time constant associated with convergence to the optimal solution increases linearly with the size K of the network. For the $K = 3$ curves shown in Fig. 12 the corresponding values are $\eta_{max} = 1.628$ (for $\lambda_1 = 0$), $\eta_{opt} = 1.086$ (for $\lambda_1 = 2\lambda_2$), and $\lambda_{opt} = 0.266$.

V. UNREALIZABLE AND OVER-REALIZABLE LEARNING SCENARIOS

The discussion of the preceding two sections has focused on a learning scenario in which both student and teacher networks have the same number $K = M$ of hidden units. The equations of motion (14) describe the evolution of the order parameters for arbitrary K and M, and provide a tool for investigating both over-realizable ($K > M$) and unrealizable ($K < M$) scenarios. In this section we consider two examples which demonstrate the power of the approach developed here when applied to the analysis of general learning scenarios. We focus on a graded teacher with $T_{nm} = n\ \delta_{nm}$ for all $1 \leq n, m \leq M$.

In the over-realizable case $K > M$ the learning process is found to prune unnecessary hidden nodes. As an example consider a teacher with $M = 2$ hidden units to be learned by a student with $K = 3$ hidden units. The time evolution of the order parameters is shown in Fig. 15a-c for $\eta = 1$. The picture that emerges is one of specialization with increasing α: asymptotically the first student node imitates the first teacher node ($R_{11} \rightarrow T_{11}$) while ignoring the second one ($R_{12} \rightarrow 0$), the second student node imitates the second teacher node ($R_{22} \rightarrow T_{22}$) while ignoring the first one ($R_{21} \rightarrow 0$), and the third student node gets eliminated. The evolution of the student norms shown in Fig. 15a demonstrates $Q_{11} \rightarrow T_{11} = 1$, $Q_{22} \rightarrow T_{22} = 2$, and $Q_{33} \rightarrow 0$ as $\alpha \rightarrow \infty$. The student-student overlaps Q_{ik} shown in Fig. 15b reveal an intermediate regime in which both surviving student nodes are anticorrelated while correlated with the node to be pruned. Overlaps involving the third student node decay to zero with Q_{33}, and the two surviving student nodes become increasingly uncorrelated. The overlap between student and teacher hidden nodes shown in Fig. 15c clearly displays a small α behavior dominated by the symmetric solution, followed by a transition onto the specialization required to obtain perfect generalization. The corresponding evolution of the generalization error is shown in Fig. 15d.

In this over-realizable scenario in which the student has more resources than necessary for the implementation of the task as defined by the teacher, error minimization results in a pruning of the excessive student nodes. The resulting learning process is a special case of realizable learning of an anisotropic teacher: consider a teacher with $\widetilde{M} = K$ hidden units, characterized by $T_{nm} = T_n\ \delta_{nm}$, with $T_n = n$ for $1 \leq n \leq M$ and $T_n = 0$ for $M < n \leq \widetilde{M} = K$. The task is to be learned by a student network with $K = \widetilde{M}$ hidden units. The specialization required to achieve perfect generalization results in a student network in which M nodes become specialized to the M active teacher nodes, in a one-to-one correspondence, while the remaining $\widetilde{M} - M = K - M$ nodes specialize to the artificially introduced null teacher nodes, becoming themselves disconnected from the input layer to imitate $\mathbf{B}_n = 0$ for $M < n \leq K$.

In the unrealizable case $K < M$ the student does not have enough resources to implement the task, and cannot achieve perfect generalization as $\alpha \rightarrow \infty$. As an example consider a teacher with $M = 4$ hidden units to be learned by a student with $K = 3$ hidden units. The time evolution of the order parameters shown in Fig. 16a-c for $\eta = 0.6$ reveals an initial behavior dominated by a symmetric solution in which all three student nodes have the same overlap with any given teacher node, and the only differentiation is due to the graded norm $T_{nn} = n$ of the teacher weight vectors. Trapping in the symmetric subspace is followed by a process in which each student node specializes to one of the three dominant teacher nodes. The specialization of student node i to teacher node n results in $R_{in}^2 \rightarrow Q_{ii}T_{nn}$, so that $R_{in} \rightarrow T_{nn}$ as $Q_{ii} \rightarrow T_{nn}$. The evolution of the norm of the student vectors shown in Fig. 16a demonstrates $Q_{11} \rightarrow T_{22} = 2$, $Q_{22} \rightarrow T_{44} = 4$, and $Q_{33} \rightarrow T_{33} = 3$ as $\alpha \rightarrow \infty$. The student-teacher overlaps shown in Fig. 16c indicate that as each student node imitates one of the dominant teacher nodes, it ignores the other two dominant nodes ($R_{12} \rightarrow T_{22}$ while R_{13} and $R_{14} \rightarrow 0$; $R_{24} \rightarrow T_{44}$ while R_{22} and $R_{23} \rightarrow 0$, $R_{33} \rightarrow T_{33}$ while R_{32} and $R_{34} \rightarrow 0$), but all three student nodes retain some overlap with the less dominant teacher node $n = 1$ (note the residual asymptotic value of R_{11}, R_{21}, and R_{31}). The nonvanishing component in the direction of $\mathbf{B}_1$ results in persistent correlations among the three student vectors, as shown in Fig. 16b. Note that the specialization of the student nodes does not occur simultaneously, but is ordered according to the relevance of the corresponding teacher nodes, resulting in a cascade of specialization transitions. The evolution of the generalization error shown in Fig. 16d reveals this structure: a plateau characteristic of trapping in the symmetric subspace is followed by a monotonic decrease where two observable inflection points correspond to the specialization of $i = 2$ onto $n = 4$ followed by that of $i = 3$ onto $n = 3$. An asymptotic residual error $\epsilon_\infty \neq 0$ is the signature of unrealizable learning.

VI. CONCLUSIONS AND DISCUSSION

We have investigated on-line learning of continuous functions through gradient descent in a very general learning scenario. The target function is generated by a soft committee machine with M hidden units. The student is a network of the same architecture with K hidden units; its weights are updated after the presentation of each randomly drawn example.

The average over the input distribution is performed analytically in the thermodynamic limit and yields equations of motion for the order parameters which describe the correlations among student nodes and their overlaps with the teacher nodes they are learning to imitate. The dynamical equations are exact and can be integrated accurately, providing a novel and powerful tool to monitor the specialization of hidden units and the emergence of generalization ability in multilayer networks. The solution is valid for arbitrary M and K, allowing for the investigation of realizable ($K = M$), over-realizable ($K > M$), and unrealizable ($K < M$) learning scenarios.

For the realizable learning of a task defined by an isotropic uncorrelated teacher, the student network converges to the globally optimal solution when trained with a fixed and sufficiently small learning rate $\eta < \eta_{max}$. The asymptotic convergence is exponential, and optimal decay of the generalization error to zero is achieved with $\eta_{opt} = (2/3)\ \eta_{max}$.

This fast decay is to be contrasted with the one recently found for on-line learning of realizable dichotomies [11], for which learning at fixed η results in a residual error $\epsilon_\infty \propto \eta$. Asymptotic convergence to the optimal solution requires in this case a monotonically decreasing learning rate; if $\eta \sim \alpha^{-z}$ with $z < 1$, the generalization error decays to zero as $\epsilon_g(\alpha) \sim \alpha^{-z}$. The intrinsic slowness of this process is due to the binary character of the corresponding error signal: even as the student weight vector $\mathbf{J}$ approaches the optimal solution $\mathbf{J}^*$ with $\epsilon_g^* = 0$, the error $\epsilon(\mathbf{J}, \boldsymbol{\xi})$ made on an arbitrary input $\boldsymbol{\xi}$ is either 0 or 1. If an example is misclassified the error signal conveys no information about the closeness between $\mathbf{J}$ and $\mathbf{J}^*$; if η is kept fixed the large weight adjustments $\Delta\mathbf{J}$ made in response to such error signals cause persistent fluctuations which prevent the convergence of $\mathbf{J}$ to $\mathbf{J}^*$.

This observation highlights the advantage of building networks of continuous as opposed to discrete units: as $\mathbf{J}$ approaches $\mathbf{J}^*$ the error signal (3) is intrinsically small for arbitrary inputs $\boldsymbol{\xi}$, allowing for the learning process to converge exponentially fast at fixed η. The use of error functions of the type (3) for multilayer networks with continuous units leads to training by the widely used gradient descent algorithm, as investigated here; networks with discrete units require the use of perceptron-type learning algorithms which have generated much theoretical interest [4,5] but are of limited practical use. To those concerned with the desirability of implementing binary classifications, we remark that linearly separable dichotomies of the type discussed in [9–11] can be well approximated in the model analyzed here by setting $M = K = 1$ and increasing the effective steepness of the nonlinear activation function $g(x)$ through increasing the norm T of the teacher vector.

Learning at a constant η in continuous neural networks has been investigated using stochastic approximation theory to obtain a master equation for the dynamical evolution of the student weight vectors $\mathbf{J}$ [8]. Nonvanishing quadratic deviations $\| \mathbf{J} - \mathbf{J}^* \|^2 \propto \eta$ were found to persist asymptotically, indicating lack of convergence to the desired solution $\mathbf{J}^*$. The statistical-mechanics approach implemented here involves a change of representation from the student weight vector $\mathbf{J} = \{\mathbf{J}_i\}_{1 \leq i \leq K}$ to the order parameters $R_{in} = \mathbf{J}_i \cdot \mathbf{B}_n$ and $Q_{ik} = \mathbf{J}_i \cdot \mathbf{J}_k$ which are self-averaging in the thermodynamic limit. It is the $N \to \infty$ limit that provides a continuous time description in which fluctuations are eliminated and asymptotic convergence to the global solution can be achieved at finite, constant η.

Realizable learning for a soft committee machine trained by an isotropic uncorrelated teacher is the continuous version of the model analyzed by Schwarze in what remains the most complex investigation of off-line learning in multilayer networks [6]. Our analysis of the continuous model within the on-line learning paradigm yields a dynamical description of the learning process that confirms and expands the results in [6]. The early and intermediate stages of the dynamics are controlled by a strongly attractive solution which is symmetric under permutation of the student hidden nodes. The system eventually escapes this suboptimal solution and evolves towards an optimal solution in which each student node is correlated with a particular hidden node of the teacher network. We are able to monitor the dynamics of specialization, the process which characterizes the transition between these two regimes, and find that escape from the symmetric phase is synchronous when the target task is defined through an isotropic teacher, but occurs through a sequence ordered according to the relevance of the corresponding teacher nodes when the teacher is graded.

The ability to follow the dynamical evolution of the student network from arbitrary initial conditions to asymptotic convergence reveals a crucial aspect of training a committee machine: trapping in the symmetric subspace provides a substantial and unwelcome contribution to the total training time (or number of examples) needed to achieve a desired level of generalization ability. Attempts at reducing the total training time by fine tuning the asymptotic decay of the generalization error to zero are misguided and overlook the basic role of the trapping time as a limiting factor. The strategy to reduce trapping times is to use the largest possible training rate η compatible with asymptotic convergence to the optimal solution. This observation cautions against schedules for a monotonic decrease of the learning rate as proposed in [11], which result in inefficiently low values of η at intermediate times controlled by the symmetric solution.

The detailed investigation of realizable learning is complemented in this paper by two examples which illustrate the power of the method developed here when applied to the analysis of over-realizable and unrealizable learning scenarios. In the over-realizable case $K > M$, learning leads to pruning of unnecessary student nodes, a process easily understood when interpreted as a special case of realizable learning of an anisotropic teacher. The unrealizable case $K < M$ leads to novel adaptive behavior to compensate for the lack of resources as would be needed to implement the target task. A detailed analysis of this frequently encountered learning scenario will be reported elsewhere [16].

The theoretical framework developed in Section II has been extended into a tool to investigate learning from noisy data with weight-decay regularization [16]. Other possible extensions allow for nonlinear output units, unrestricted and adaptive hidden-to-output weights, correlated teacher vectors, and correlated input components.

We conclude this discussion with a general comment on the relation between the on-line and the off-line learning paradigms. Off-line learning is formulated as a problem in equilibrium statistical mechanics, in which averages over the distribution of student weight vectors $\mathbf{J}$ and averages over the disorder introduced by the random selection of training examples occur on different time scales. Training examples are held fixed while the exploration of $\mathbf{J}$ space that leads to thermal equilibrium takes place. The ensemble average over different realizations of the training set is assumed to occur over a much longer time scale; the replica method is used to perform the corresponding *quenched* average.

One way to avoid the technical difficulties intrinsic to quenched averaging is to invoke the *annealed* approximation, based on neglecting the separation between time scales: weights and examples are assumed to undergo simultaneous equilibration through a joint exploration of $\mathbf{J}$ and $\boldsymbol{\xi}$ spaces [5]. The resulting Gibbs distribution, controlled by the learning error, favors training examples which are compatible with the current student hypothesis as represented by $\mathbf{J}$. Annealing is not an efficient learning strategy; efficient learning strategies are based on precisely the opposite selection criterion [17]. Examples that contradict the current hypothesis and are associated with large errors $\epsilon(\mathbf{J}, \boldsymbol{\xi})$ are to be preferred, as they convey sizeable information about the target task and result in a reduction of the entropy associated with the effective volume of the current space of hypothesis.

On-line learning can be considered as a mechanism to restore the separation between time scales while reversing the role of $\mathbf{J}$ and $\boldsymbol{\xi}$. A current network configuration $\mathbf{J}$ is held fixed while the average over all possible ways of selecting the next training example is performed. This average avoids the technical complications of the replica method, and generates a dynamical evolution in the space of student weights. The average over $\boldsymbol{\xi}$ at fixed $\mathbf{J}$ is dominated by the examples which give the largest contribution to the gradient $\nabla_{\mathbf{J}}\, \epsilon(\mathbf{J}, \boldsymbol{\xi})$. Such examples reveal a large discrepancy between the current hypothesis and the target, and are most useful to the training process.

Acknowledgement

The work was supported by the EU grant CHRX-CT92-0063. DS would like to thank the Niels Bohr Institute and the CONNECT group for their hospitality.

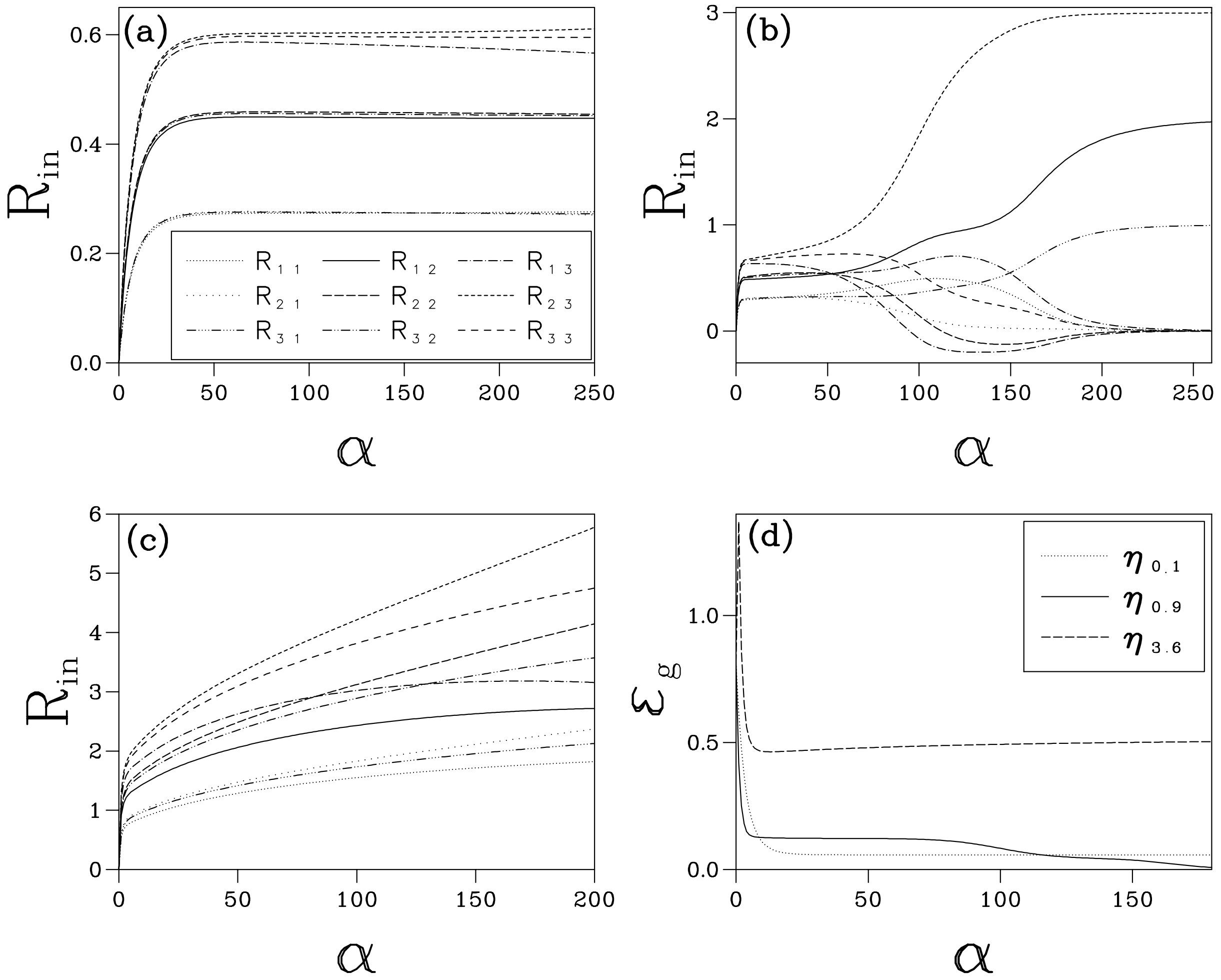

FIG. 1. Dependence of the overlaps between various student and teacher vectors and the generalization error on the normalized number of examples α, for several values of the learning rate η. The overlaps are shown for (a) $\eta = 0.1$, (b) $\eta = 0.9$, and (c) $\eta = 3.6$. The generalization error for these three cases is shown in (d). The teacher is characterized by $T_{nm} = n\ \delta_{nm}$. Initial conditions are $R_{in} = 0$ and $Q_{ik} = U[0, 0.5]\ \delta_{ik}$.

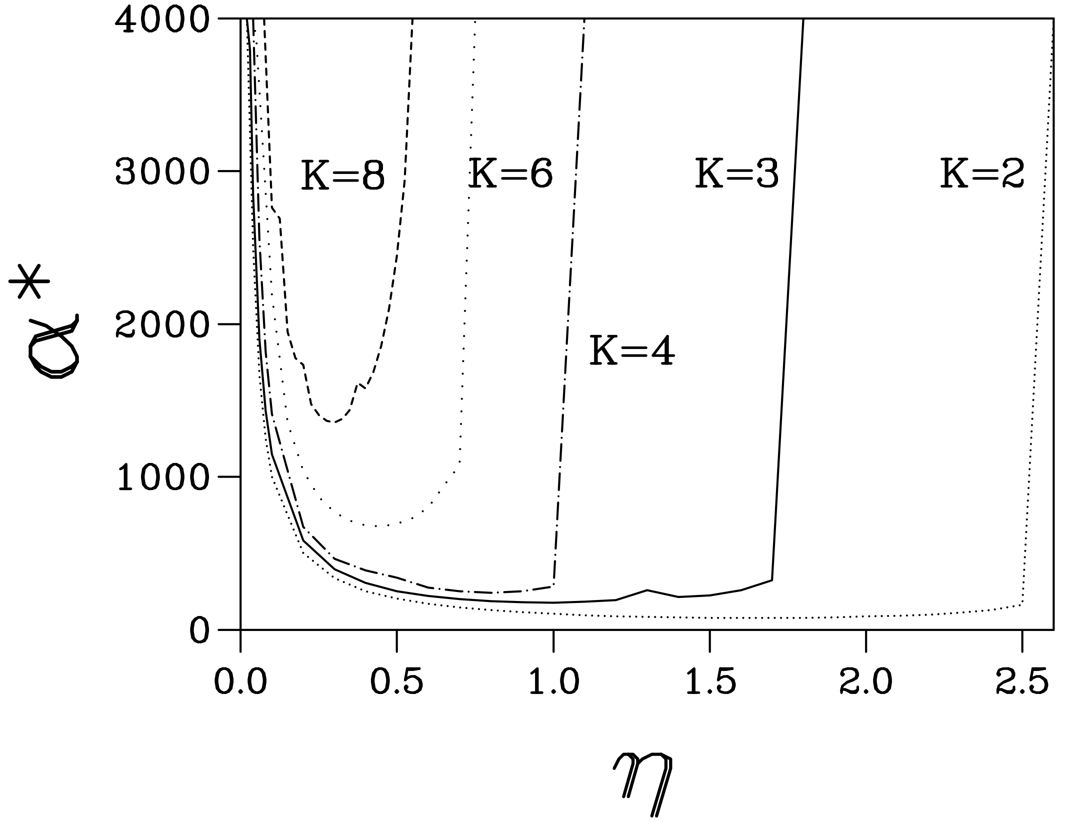

FIG. 2. Dependence of the number of examples α^* needed to achieve $\epsilon_g = 0.01$ on the learning rate η for several values of the number $K = M$ of hidden units.

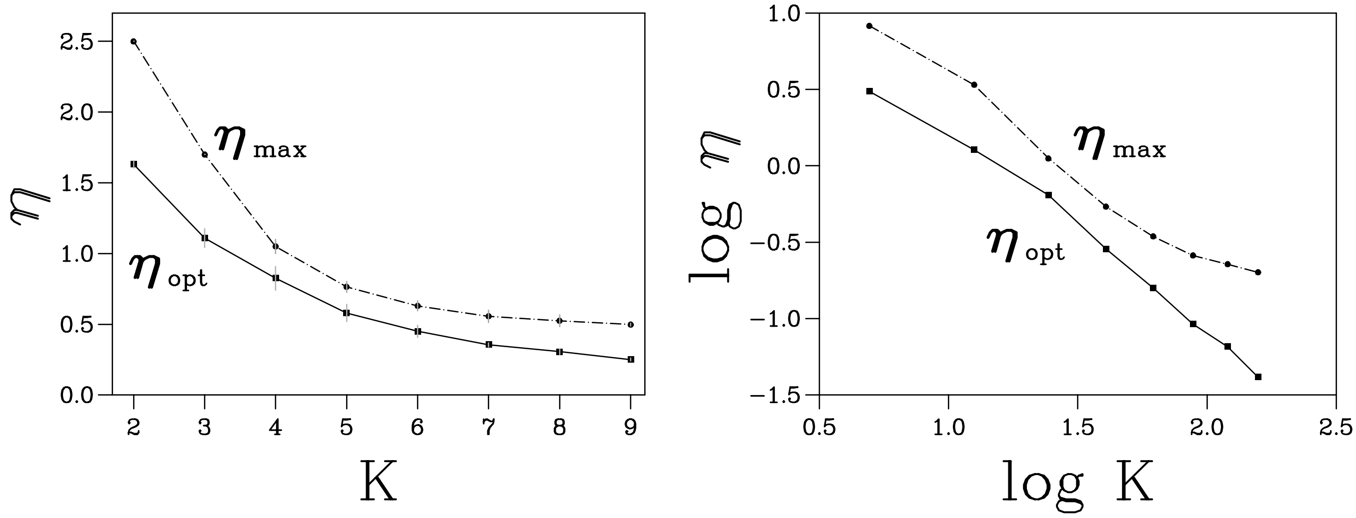

FIG. 3. On the left, dependence of the optimal and maximal learning rates on the number $K = M$ of hidden units. On the right, the same data is shown on logarithmic scales.

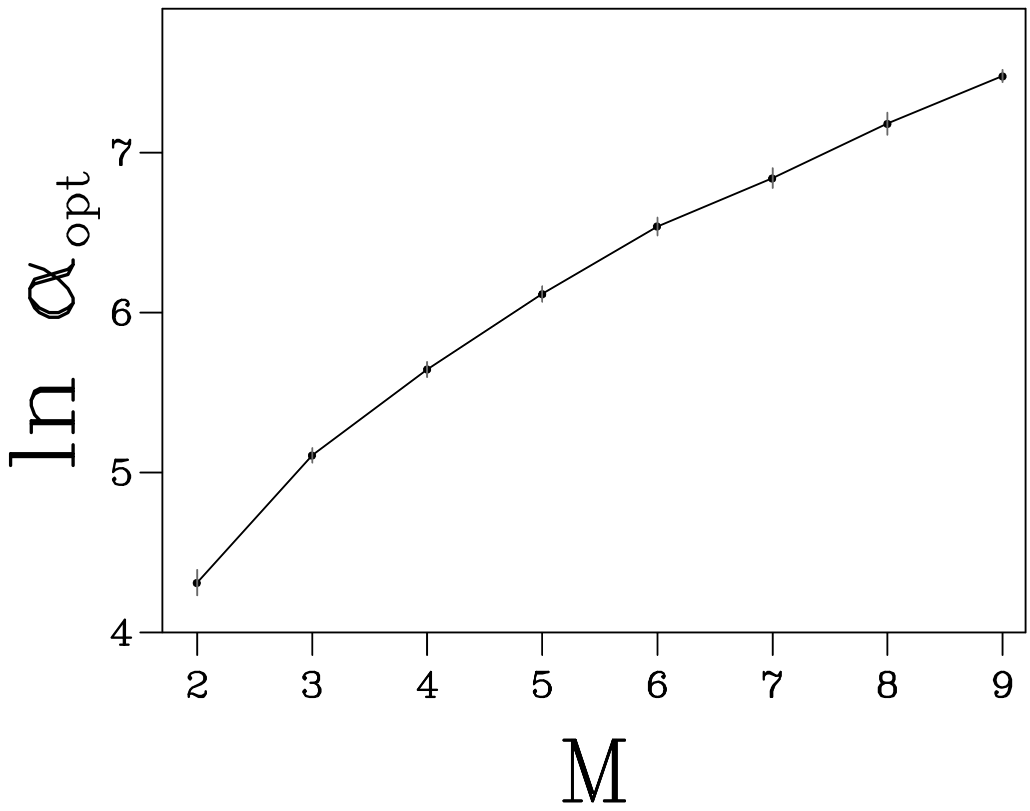

FIG. 4. The logarithm of the minimal number of examples needed to achieve $\epsilon_g = 0.01$ as a function of the number $K = M$ of hidden units.

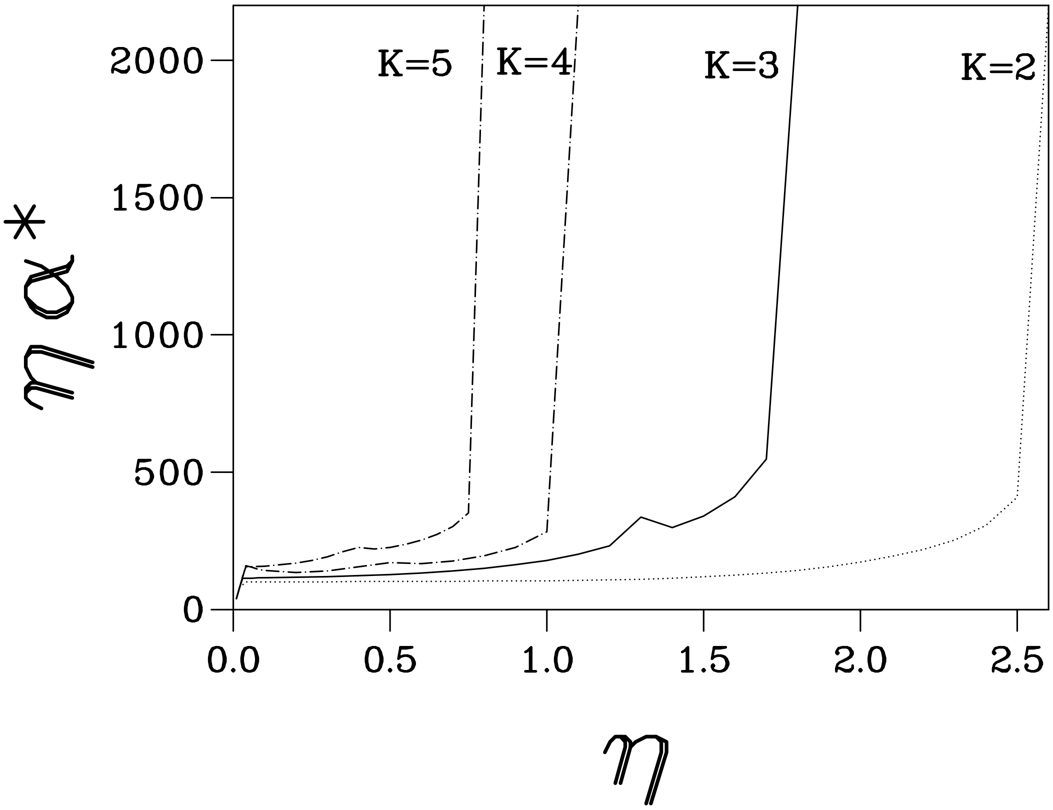

FIG. 5. Small η scaling for several values of the number $K = M$ of hidden units. The product $\widetilde{\alpha}^{\star} = \eta\alpha^{\star}$ remains constant in the small η regime.

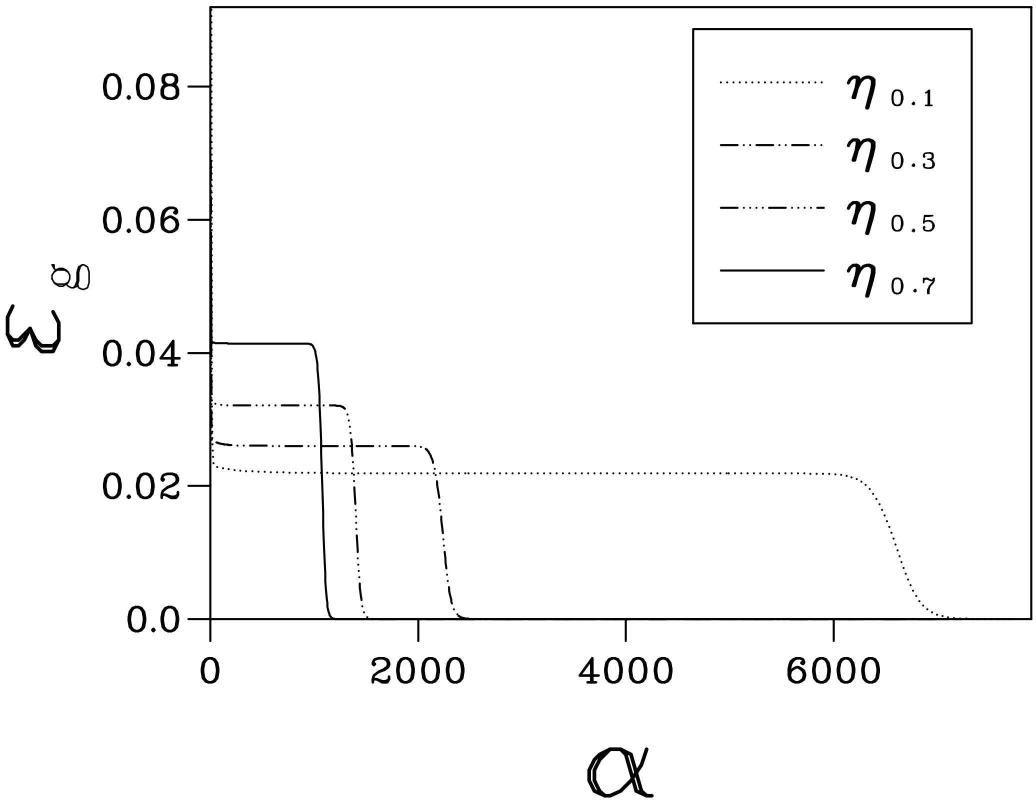

FIG. 6. Dependence of the generalization error on the number of examples α for different values of the learning rate η. Results are shown for $M = K = 3$. The teacher is characterized by $T_{nm} = \delta_{nm}$. Initial conditions are $R_{in} = U[0, 10^{-12}]$ and $Q_{ik} = U[0, 0.5]\ \delta_{ik}$.

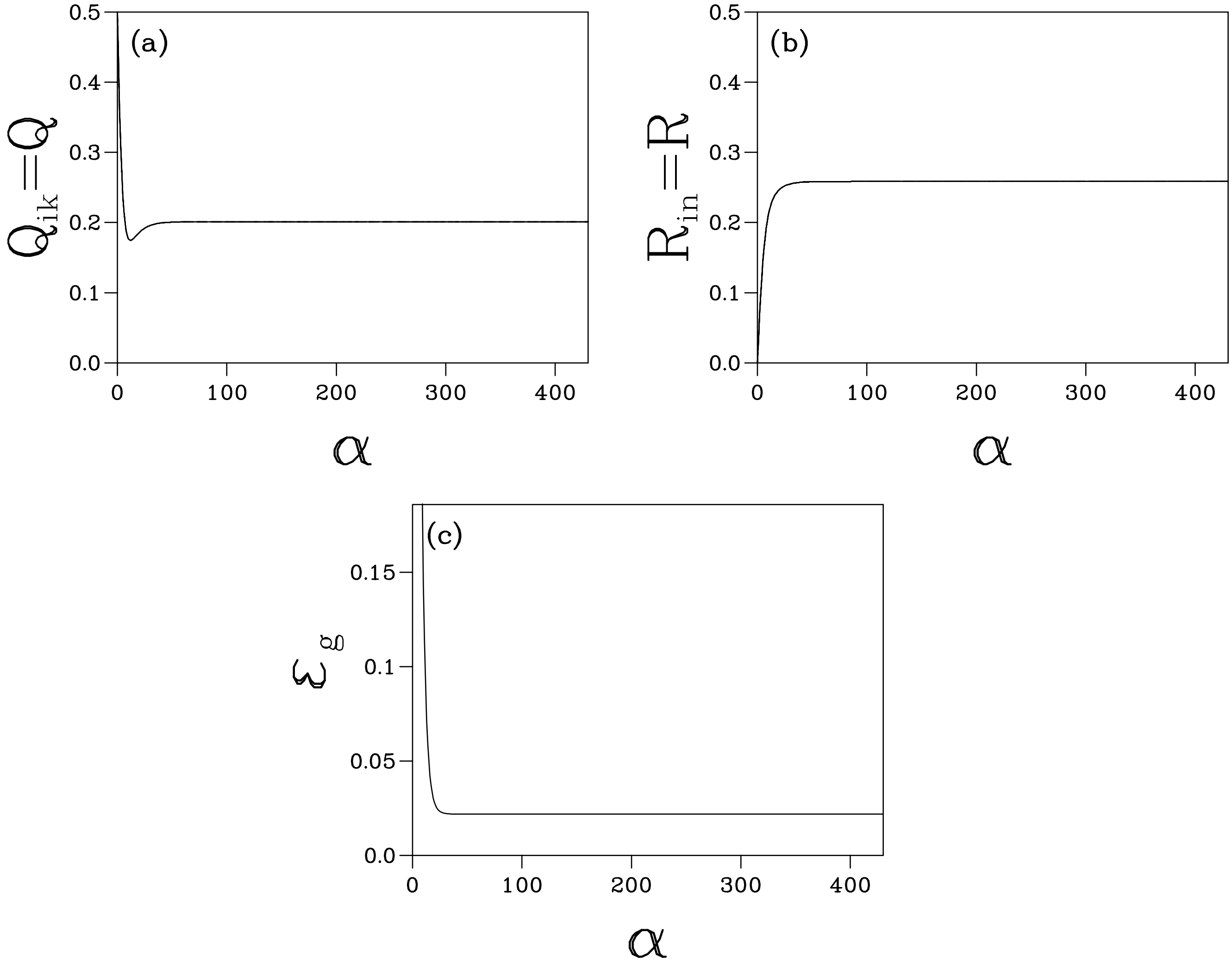

FIG. 7. Evolution of the order parameters and generalization error in the symmetric subspace. Results for $M = K = 3$ are shown here for (a) the student-student overlap Q, (b) the student-teacher overlap R, and (c) the generalization error. The teacher is characterized by $T_{nm} = \delta_{nm}$. Initial conditions are $R_{in} = R_o = 0$ and $Q_{ik} = Q_o = 0.5$.

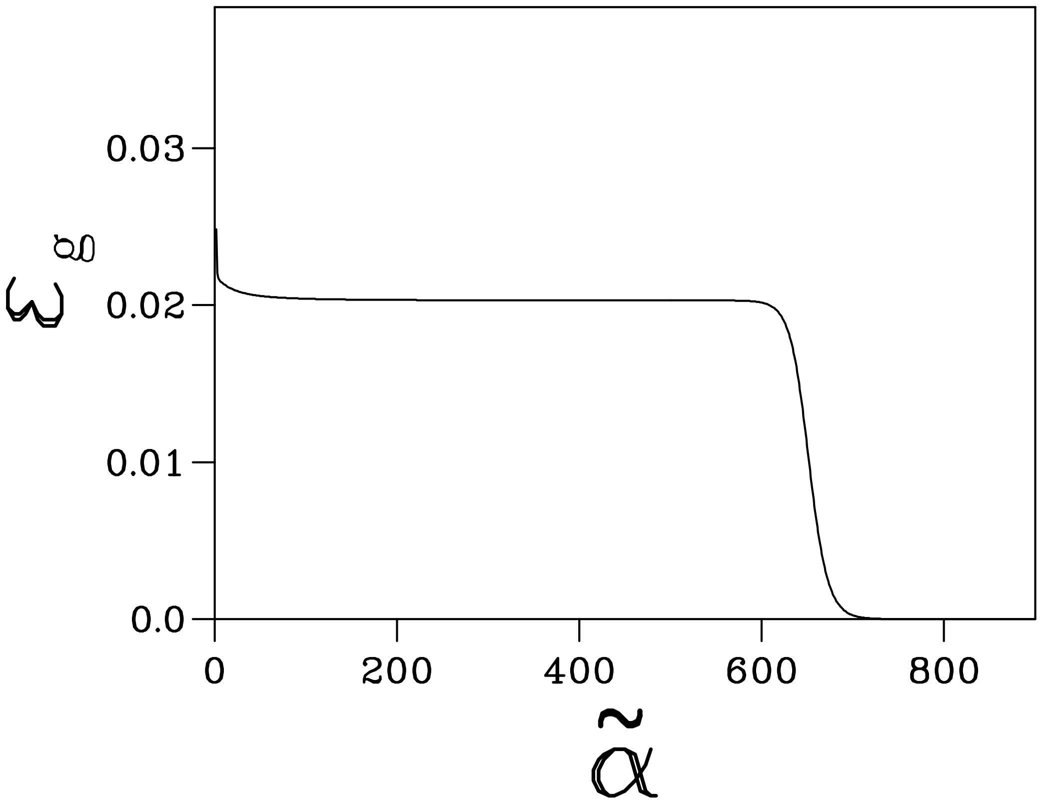

FIG. 8. Universal dependence of the generalization error on $\widetilde{\alpha} = \eta\alpha$ for $M = K = 3$. The teacher is characterized by $T_{nm} = \delta_{nm}$. Initial conditions are $R_{in} = U[0, 10^{-12}]$ and $Q_{ik} = U[0, 0.5]\ \delta_{ik}$.

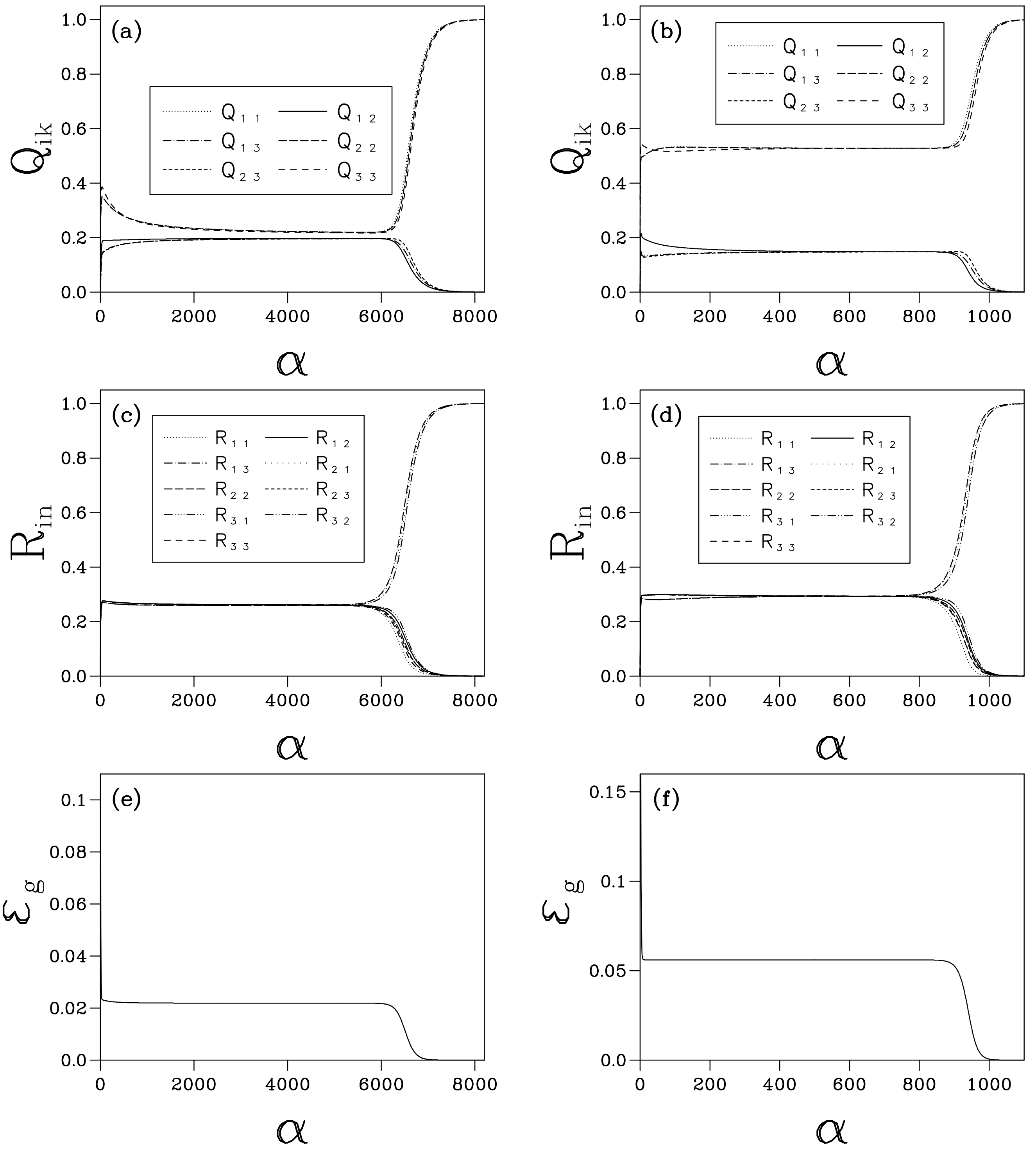

FIG. 9. Evolution of the order parameters and generalization error for small and intermediate learning rates. Results for $M = K = 3$ are shown for $\eta = 0.1$ and $\eta = 0.9$ as follows: student-student overlaps Q_{ik} for (a) $\eta = 0.1$ and (b) $\eta = 0.9$, student-teacher overlaps R_{in} for (c) $\eta = 0.1$ and (d) $\eta = 0.9$, and generalization error for (e) $\eta = 0.1$ and (f) $\eta = 0.9$. The teacher is characterized by $T_{nm} = \delta_{nm}$. Initial conditions are $R_{in} = U[0, 10^{-12}]$ and $Q_{ik} = U[0, 0.5]\ \delta_{ik}$.

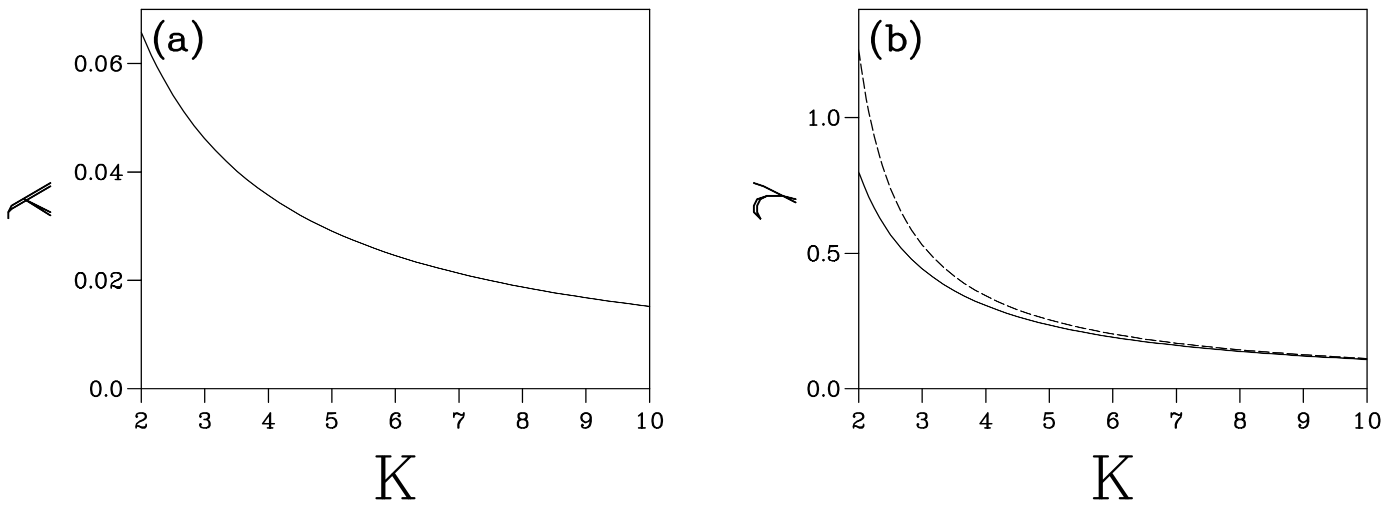

FIG. 10. Dependence of (a) the escape rate $\tilde{\lambda}$, and (b) γ_S (upper curve) and γ_R (lower curve) on the number $K = M$ of hidden units.

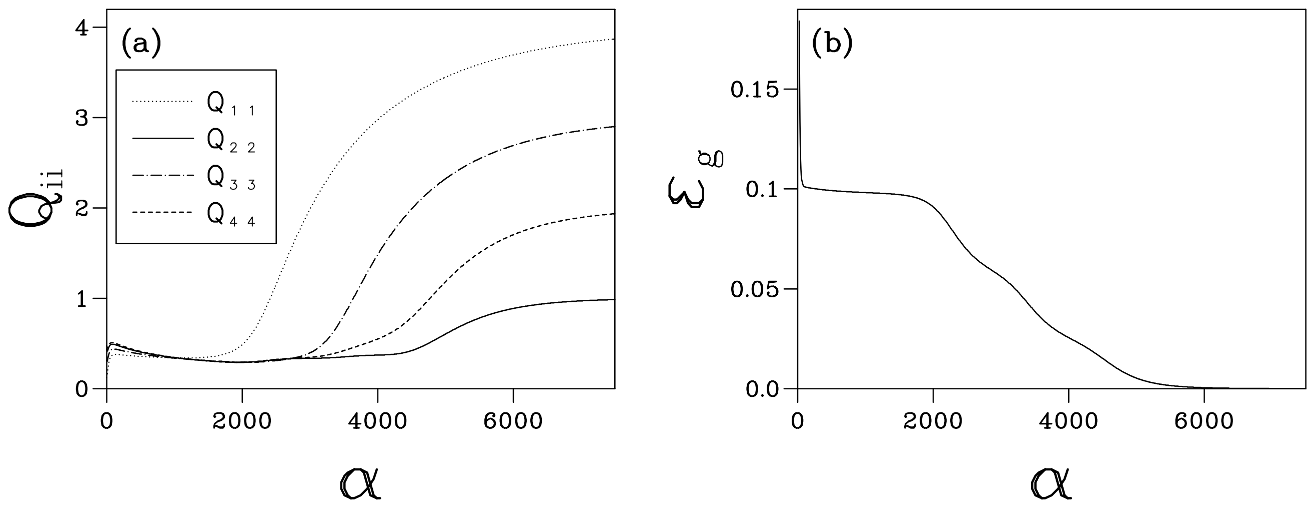

FIG. 11. Dependence of (a) the length of the student vectors and (b) the generalization error on the normalized number of examples α for a graded teacher characterized by $T_{nm} = n\ \delta_{nm}$. Results are shown for $M = K = 4$. Initial conditions are $R_{in} = 0$ and $Q_{ik} = U[0, 0.5]\ \delta_{ik}$.

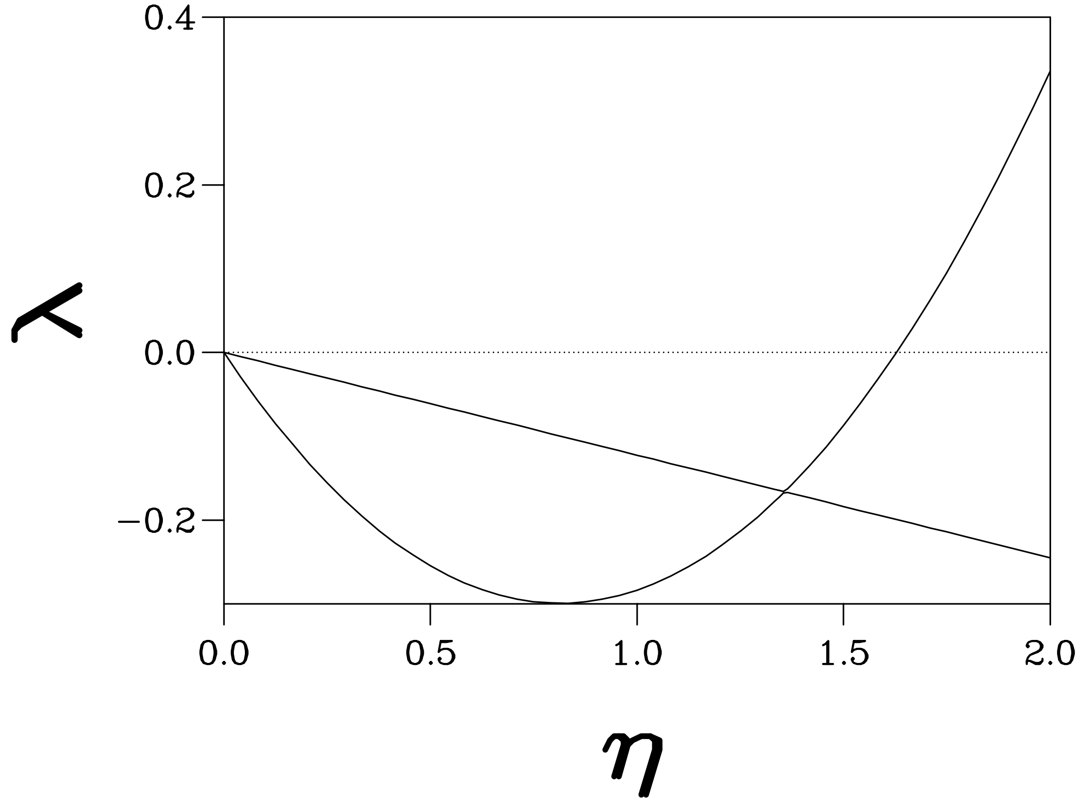

FIG. 12. Dependence of the decay eigenvalues λ_1 (curved line) and λ_2 (straight line) on the learning rate η for $M = K = 3$.

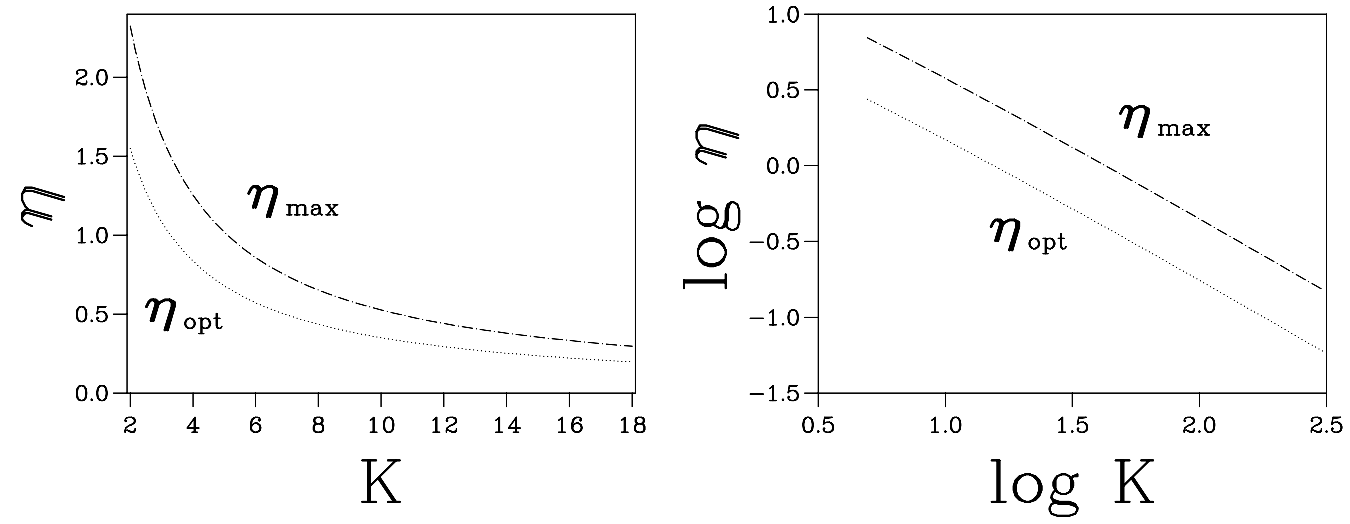

FIG. 13. On the left, dependence of the maximal and optimal learning rates on the number $K = M$ of hidden units. On the right, the same data is shown on logarithmic scales.

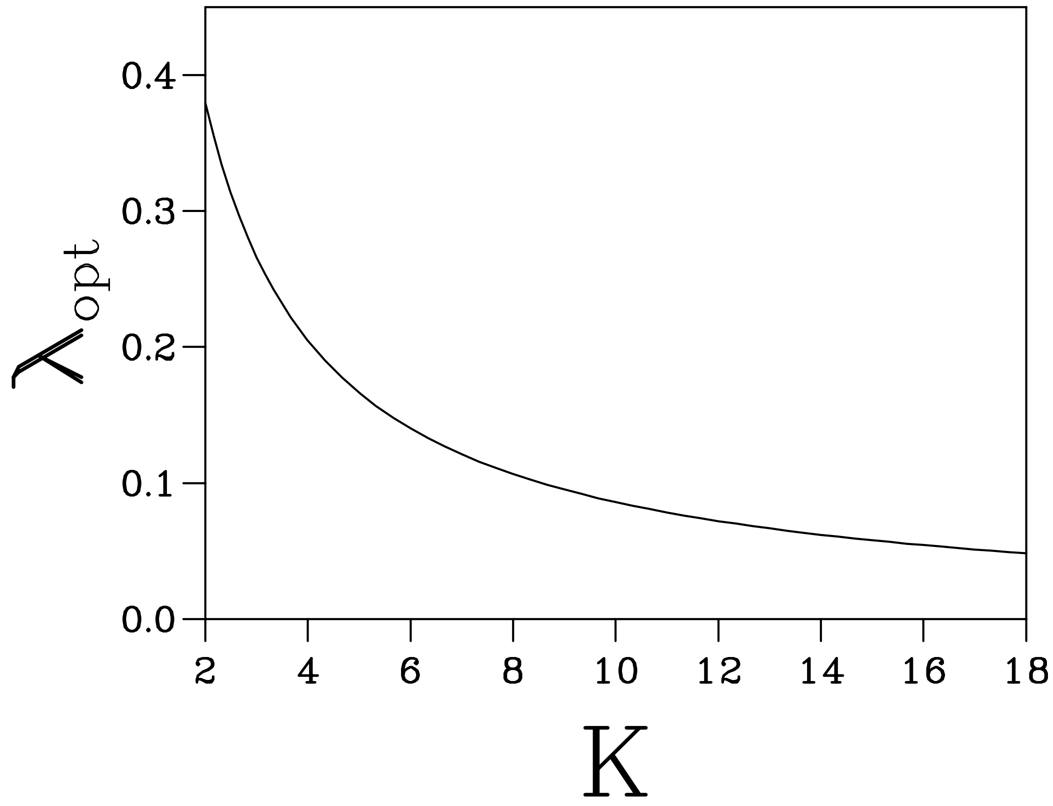

FIG. 14. Dependence of the optimal convergence rate λ_{opt} on the number $K = M$ of hidden units.

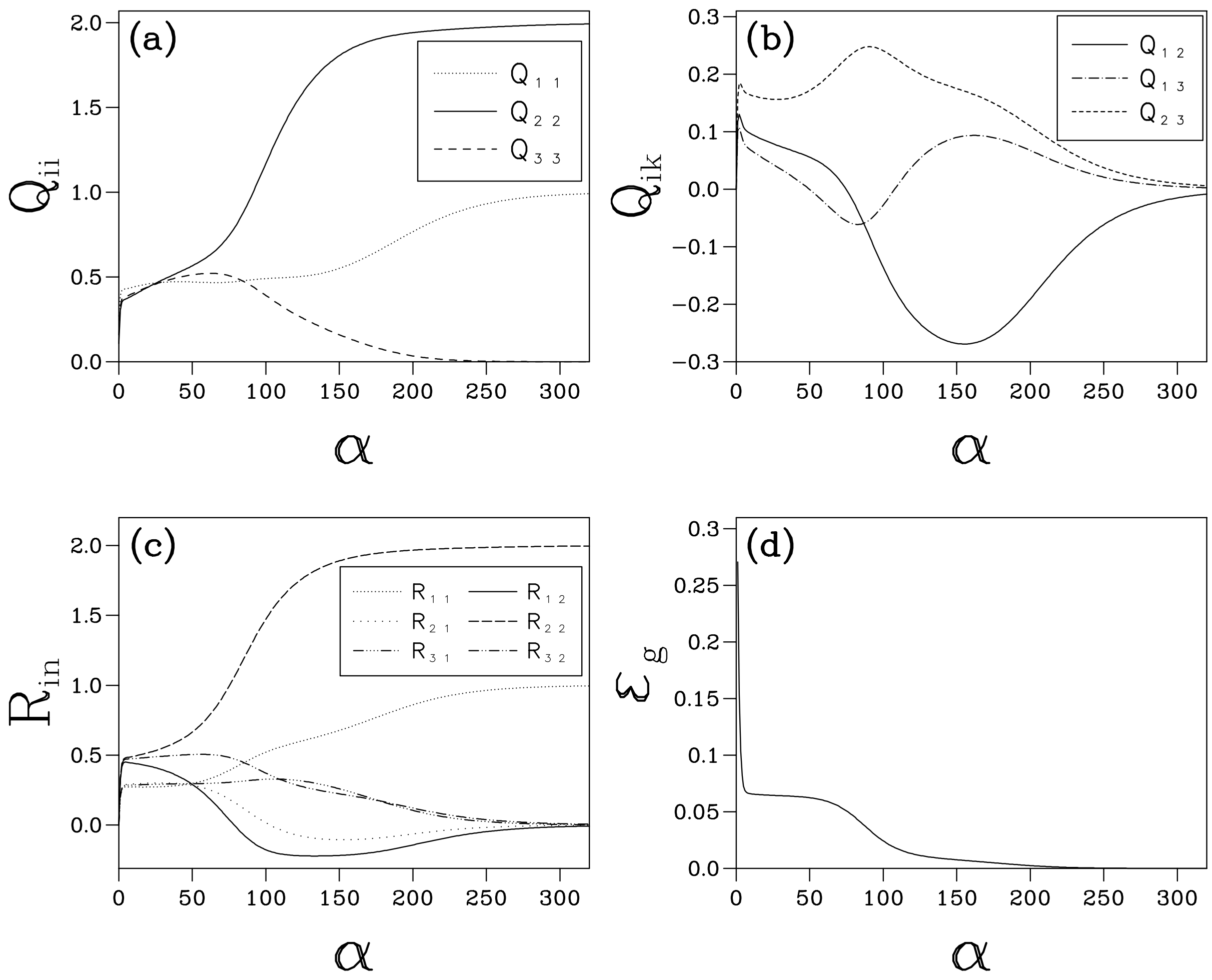

FIG. 15. Dependence of the overlaps and the generalization error on the normalized number of examples α, for a three-node student learning a two-node teacher: (a) the lengths of student vectors, (b) the correlation between student vectors, (c) the overlap between various student and teacher vectors, and (d) the generalization error. The teacher is characterized by $T_{nm} = n\ \delta_{nm}$. Initial conditions are $R_{in} = 0$ and $Q_{ik} = U[0, 0.5]\ \delta_{ik}$.

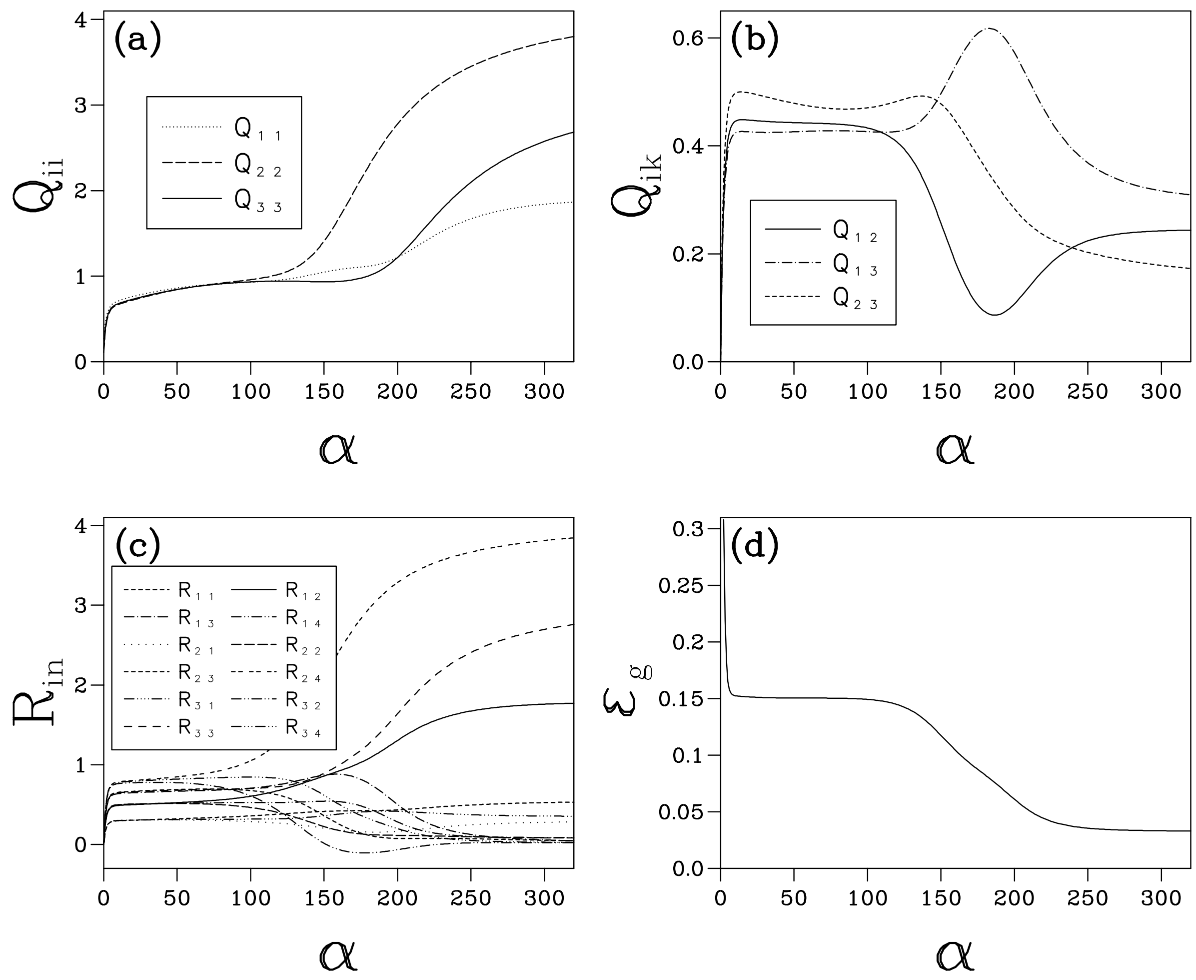

FIG. 16. Dependence of the overlaps and the generalization error on the normalized number of examples α, for a three-node student learning a four-node teacher: (a) the lengths of student vectors, (b) the correlation between student vectors, (c) the overlap between various student and teacher vectors, and (d) the generalization error. The teacher is characterized by $T_{nm} = n\ \delta_{nm}$. Initial conditions are $R_{in} = 0$ and $Q_{ik} = U[0, 0.5]\ \delta_{ik}$.

[1] J. S. Denker, D. B. Schwartz, B. Wittner, S. A. Solla, R. E. Howard, L. D. Jackel, and J. J. Hopfield, *Complex Systems* **1**, 877 (1987).
[2] G. Cybenko, *Math. of Control Signals and Syst.* **2**, 303 (1988).

[3] K. Hornik, M. Stinchcombe and H. White, *Neural Networks* **2**, 359 (1989).
[4] T. L. H. Watkin, A. Rau and M. Biehl, *Rev. Mod. Phys.* **65**, 499 (1993).
[5] H. S. Seung, H. Sompolinsky, and N. Tishby, *Phys. Rev. A* **45**, 6056 (1993).
[6] H. Schwarze, *J. Phys. A* **26**, 5781 (1993).
[7] G. Parisi, *Phys. Rev. Lett.* **43**, 1754 (1979).
[8] T. M. Heskes and B. Kappen, *Phys. Rev. A* **44**, 2718 (1991).
[9] O. Kinouchi and N. Caticha, *J. Phys. A* **25**, 6243 (1992).
[10] M. Biehl and P. Riegler, *Europhys. Lett.* **25**, 525 (1994).
[11] N. Barkai, H. S. Seung, and H. Sompolinsky, NIPS preprint (1994).
[12] Y. Kabashima, *J. Phys. A* **27**, 1917 (1994).
[13] M. Copelli and N. Caticha, *J. Phys. A* **28**, 1615 (1995).
[14] M. Biehl and H. Schwarze, *J. Phys. A* **28**, 643 (1995).
[15] D. Saad and S. A. Solla, *Phys. Rev. Lett.*, in press.
[16] D. Saad and S. A. Solla, in preparation (1995).
[17] P. Sollich, *Phys. Rev. E* **49**, 4637 (1994).

APPENDIX A: MULTIVARIATE GAUSSIAN AVERAGES

Consider an n-dimensional space $\mathbf{x}_n = (x_1, \ldots, x_n)$ and a set of functions $\{f_j\}$, $1 \leq j \leq n$. The goal is to compute an average of the form

$$I_n =< f_1(x_1) \ldots f_{n-1}(x_{n-1}) f_n(x_n) > , \tag{A1}$$

with respect to the multivariate Gaussian distribution

$$\mathcal{P}(\mathbf{x}_n) = \frac{1}{\sqrt{(2\pi)^n |\mathcal{C}_n|}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_n)^T \mathcal{C}_n^{-1} (\mathbf{x}_n) \right\} , \tag{A2}$$

controlled by the covariance matrix $\mathcal{C}_n$ with components $C_{i,j} =< x_i \; x_j >$ for $1 \leq i, j \leq n$. The result of such average will depend on the parameters of the distribution: $I_n = I_n(\mathcal{C}_n)$.

For the same set of n functions $\{f_j\}$, $1 \leq j \leq n$ consider now the average

$$I_{n-1} =< f_1(x_1) \ldots f_{n-1}(x_{n-1}) f_n(x_{n-1}) > , \tag{A3}$$

to be performed in an $(n-1)$-dimensional space $\mathbf{x}_{n-1} = (x_1, \ldots, x_{n-1})$ with respect to the multivariate Gaussian distribution

$$\mathcal{P}(\mathbf{x}_{n-1}) = \frac{1}{\sqrt{(2\pi)^{n-1} |\mathcal{C}_{n-1}|}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_{n-1})^T \mathcal{C}_{n-1}^{-1} (\mathbf{x}_{n-1}) \right\} , \tag{A4}$$

controlled by the covariance matrix $\mathcal{C}_{n-1}$ with components $C_{i,j} =< x_i \; x_j >$ for $1 \leq i, j \leq (n-1)$. The results will depend on the elements of this reduced-dimensionality matrix: $I_{n-1} = I_{n-1}(\mathcal{C}_{n-1})$.

The claim is that the reduced dimensionality integral I_{n-1} requires no independent evaluation. The corresponding result follows from specializing the result for $I_n(\mathcal{C}_n)$ to the singular covariance matrix $\widetilde{\mathcal{C}_n}$ defined as follows:

$$\begin{aligned}
\widetilde{C_{i,j}} &= C_{i,j} \quad \text{for} \quad 1 \leq i, j \leq (n-1), \\
\widetilde{C_{i,n}} &= C_{i,n-1} \quad \text{for} \quad 1 \leq i \leq n, \\
\widetilde{C_{n,i}} &= C_{n-1,i} \quad \text{for} \quad 1 \leq i \leq n.
\end{aligned} \tag{A5}$$

Note that $\widetilde{\mathcal{C}_n}$ follows from $\mathcal{C}_n$ by imposing the coordinate constraint $x_n = x_{n-1}$. The identity

$$I_{n-1}(\mathcal{C}_{n-1}) = I_n(\widetilde{\mathcal{C}_n}) . \tag{A6}$$

makes it unnecessary to compute the reduced-dimensionality integrals.

We now summarize a proof of Eq.(A6), based on the identity

$$\frac{1}{\sqrt{|\mathcal{C}_n|}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_n)^T \mathcal{C}_n^{-1} (\mathbf{x}_n) \right\} = \int \frac{dy_1 \ldots dy_n}{\sqrt{(2\pi)^n}} \exp\left\{ -\frac{1}{2} (\mathbf{y}_n)^T \mathcal{C}_n (\mathbf{y}_n) + i (\mathbf{x}_n)^T (\mathbf{y}_n) \right\} . \tag{A7}$$

Consider the average (A1) with respect to the distribution (A2):

$$I_n(\mathcal{C}_n) = \int \frac{dx_1 \ldots dx_n}{\sqrt{(2\pi)^n |\mathcal{C}_n|}} \exp\left\{ -\frac{1}{2} (\mathbf{x}_n)^T \mathcal{C}_n^{-1} (\mathbf{x}_n) \right\} f_1(x_1) \ldots f_n(x_n) . \tag{A8}$$

The expression (A7) is now substituted into (A8), to obtain:

$$I_n(\mathcal{C}_n) = \int \frac{dy_1 \ldots dy_n}{\sqrt{(2\pi)^n}} \exp\left\{ -\frac{1}{2} (\mathbf{y}_n)^T \mathcal{C}_n (\mathbf{y}_n) \right\} \tilde{f}_1(y_1) \ldots \tilde{f}_n(y_n) , \tag{A9}$$

expressed in terms of the Fourier transforms

$$\tilde{f}_j(y) = \int \frac{dx}{\sqrt{2\pi}} f_j(x) \, e^{iyx} . \tag{A10}$$

Note that (A9) is a particularly suitable form for the evaluation of $I_n(\widetilde{\mathcal{C}_n})$, since the singularity due to $|\widetilde{\mathcal{C}_n}|$ has been eliminated, and the constraints (A5) due to dimensionality reduction are easily imposed on some of the components of $\widetilde{\mathcal{C}_n}$ itself, while they affect in a complicated way all the components of $\widetilde{\mathcal{C}_n}^{-1}$.

To prepare for the replacement of $\mathcal{C}_n$ by $\widetilde{\mathcal{C}}_n$ it is convenient to rewrite Eq. (A9) by decoupling y_n from the other components:

$$I_n(\mathcal{C}_n) = \int \frac{dy_1 \ldots dy_{n-1}}{\sqrt{(2\pi)^n}} \; \exp\left\{-\frac{1}{2}(\mathbf{y}_{n-1})^T \widehat{\mathcal{C}}_{n-1}(\mathbf{y}_{n-1})\right\} \tilde{f}_1(y_1)\ldots\tilde{f}_{n-1}(y_{n-1})$$
$$\times \int \frac{dy_n}{\sqrt{C_{n,n}}} \; \exp\left\{-\frac{1}{2}(y_n)^2\right\} \tilde{f}_n\left(\frac{y_n}{\sqrt{C_{n,n}}} - \frac{1}{C_{n,n}}\sum_{i=1}^{n-1} C_{i,n}\; y_i\right) , \tag{A11}$$

where

$$\widehat{C}_{i,j} = C_{i,j} - \frac{C_{i,n}\; C_{j,n}}{C_{n,n}} \tag{A12}$$

for all $1 \leq i, j \leq (n-1)$.

An expression for $I_n(\widetilde{\mathcal{C}_n})$ follows from substituting the matrix components defined in Eq. (A5) onto Eq. (A11):

$$I_n(\widetilde{\mathcal{C}_n}) = \int \frac{dy_1 \ldots dy_{n-2}}{\sqrt{(2\pi)^{n-1}}} \; \exp\left\{-\frac{1}{2}(\mathbf{y}_{n-2})^T \widehat{\mathcal{C}}_{n-2}(\mathbf{y}_{n-2})\right\} \tilde{f}_1(y_1)\ldots\tilde{f}_{n-2}(y_{n-2})$$
$$\times \int \frac{dy_n}{\sqrt{C_{n-1,n-1}}} \; \exp\left\{-\frac{1}{2}(y_n)^2\right\} \tilde{F}\left(\frac{y_n}{\sqrt{C_{n-1,n-1}}} - \frac{1}{C_{n-1,n-1}}\sum_{i=1}^{n-2} C_{i,n-1}\; y_i\right) , \tag{A13}$$

expressed in terms of the convolutionary Fourier transform

$$\tilde{F}(y) = \int \frac{dx}{\sqrt{2\pi}} \; f_{n-1}(x)\; f_n(x)\; e^{iyx} \; . \tag{A14}$$

Note that under the transformation (A5) the components of $\widehat{\mathcal{C}}$ in Eq. (A12) become

$$\widehat{C}_{i,j} = C_{i,j} - \frac{C_{i,n-1}\; C_{j,n-1}}{C_{n-1,n-1}}, \tag{A15}$$

for all $1 \leq i, j \leq (n-2)$.

The computation of $I_{n-1}(\mathcal{C}_{n-1})$ proceeds along similar lines. The $(n-1)$-dimensional version of Eq. (A7) is substituted into Eq. (A4) to obtain

$$I_{n-1}(\mathcal{C}_{n-1}) = \int \frac{dy_1 \ldots dy_{n-1}}{\sqrt{(2\pi)^{n-1}}} \; \exp\left\{-\frac{1}{2}(\mathbf{y}_{n-1})^T \mathcal{C}_{n-1}(\mathbf{y}_{n-1})\right\} \tilde{f}_1(y_1)\ldots\tilde{f}_{n-2}(y_{n-2})\tilde{F}(y_{n-1}) \; , \tag{A16}$$

expressed in terms of the Fourier transforms of Eqs. (A10) and (A14). The component y_{n-1} is now decoupled from the other components to obtain:

$$I_{n-1}(\mathcal{C}_{n-1}) = \int \frac{dy_1 \ldots dy_{n-2}}{\sqrt{(2\pi)^{n-1}}} \; \exp\left\{-\frac{1}{2}(\mathbf{y}_{n-2})^T \widehat{\mathcal{C}}_{n-2}(\mathbf{y}_{n-2})\right\} \tilde{f}_1(y_1)\ldots\tilde{f}_{n-2}(y_{n-2})$$
$$\times \int \frac{dy_{n-1}}{\sqrt{C_{n-1,n-1}}} \; \exp\left\{-\frac{1}{2}(y_{n-1})^2\right\} \tilde{F}\left(\frac{y_{n-1}}{\sqrt{C_{n-1,n-1}}} - \frac{1}{C_{n-1,n-1}}\sum_{i=1}^{n-2} C_{i,n-1}\; y_i\right) , \tag{A17}$$

where the components of $\widehat{\mathcal{C}}_{n-2}$ are given in Eq. (A15).

Comparison of Eq. (A13) to Eq. (A17) establishes the identity (A6). As an example of the application of this identity in the context of our paper, consider the evaluation of $I_3(i,n,j) \equiv < g'(x_i)\; y_n\; g(x_j) >$ discussed in Section II. A two-dimensional integral such as $I_2(i,n) \equiv < g'(x_i)\; y_n\; g(x_i) >$ does not need to be evaluated independently once a general result for I_3 has been obtained, since $I_2(i,n) = I_3(i,n,i)$. An expression for $I_2(i,n)$ is obtained by applying the solution (15)-(16) for I_3 to the singular covariance matrix

$$\widetilde{C_3} = \begin{pmatrix} Q_{ii} & R_{in} & Q_{ii} \\ R_{in} & T_{nn} & R_{in} \\ Q_{ii} & R_{in} & Q_{ii} \end{pmatrix} \; .$$