Exact Solution for On-Line Learning in Multilayer Neural Networks

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.*

Abstract

We present an analytic solution to the problem of on-line gradient-descent learning for two-layer neural networks with an arbitrary number of hidden units in both teacher and student networks.

Phys. Rev. Lett. **74**, 4337-4340 (1995)

Typeset using REVTEX

Layered neural networks are of interest for their ability to implement input-output maps [1]. Classification and regression tasks formulated as a map from an N-dimensional input space $\boldsymbol{\xi}$ onto a scalar ζ are realized through a map $\zeta = f_{\mathbf{J}}(\boldsymbol{\xi})$, which can be modified through changes in the internal parameters $\{\mathbf{J}\}$ specifying the strength of the interneuron couplings [2,3]. Learning refers to the modification of these couplings so as to bring the map $f_{\mathbf{J}}$ implemented by the network as close as possible to a desired map $\tilde{f}$. The degree of success is monitored through the *generalization error*, a measure of the dissimilarity between $f_{\mathbf{J}}$ and $\tilde{f}$.

Learning from examples in layered neural networks is usually formulated as an optimization problem [2,3], based on the minimization of an additive *learning error* defined over a *training set* composed of P independent examples $(\boldsymbol{\xi}^\mu, \zeta^\mu)$, with $\zeta^\mu = \tilde{f}(\boldsymbol{\xi}^\mu)$, $1 \leq \mu \leq P$. Statistical physics tools for investigating the properties of such models, based on the use of the replica method, have been successfully applied to the analysis of single-layer perceptrons [3] and some simplified two-layer structures (e.g., committee machines [4]). 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* [5], 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.

A recently introduced alternative approach investigates *on-line learning* [6]. In this scenario the couplings are adjusted to minimize the error after the presentation of each example. The resulting changes in $\{\mathbf{J}\}$ are described as a dynamical evolution, with the number of examples playing the role of time. The average that accounts for the disorder introduced by the independent random selection of an example at each time step can be performed directly, 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.

While investigating the on-line learning scenario proposed by Biehl and Schwarze [6], we found an unexpected result: the dynamical equations for the order parameters can be obtained *analytically* for a general 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.

Two-layer networks with an arbitrary number of hidden units have been shown to be universal approximators [1] for N-to-one dimensional maps. Our results thus describe the learning of tasks of arbitrary complexity (general M). The complexity of the student network is also arbitrary (general K, independent of M), providing a tool to investigate realizable ($K = M$), over-realizable ($K > M$), and unrealizable ($K < M$) learning scenarios. Such capabilities are to be contrasted with previously available results: the equations provided in [6] can only describe a committee machine with $K = 2$ hidden units learning a linearly separable task ($M = 1$).

In this paper we limit our discussion to the case of the soft-committee machine [6], in which all the hidden units are connected to the output unit with positive couplings of unit strength, and only the input-to-hidden couplings are adaptive. Consider the student network: hidden unit i receives information from input unit r through the weight J_{ir}, and its activation under presentation of an input pattern $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ is $x_i = \mathbf{J}_i \cdot \boldsymbol{\xi}$, with $\mathbf{J}_i = (J_{i1}, \ldots, J_{iN})$ defined as the vector of incoming weights onto the i-th hidden unit. The output of the student network is $\sigma(\mathbf{J}, \boldsymbol{\xi}) = \sum_{i=1}^{K} \, g\,(\mathbf{J}_i \cdot \boldsymbol{\xi})$, where g is the activation function of the hidden units, taken here to be the error function $g(x) \equiv \mathrm{erf}(x/\sqrt{2})$, and $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \leq i \leq K}$ is the set of input-to-hidden adaptive weights.

Training examples are of the form $(\boldsymbol{\xi}^\mu, \zeta^\mu)$. The components of the independently drawn input vectors $\boldsymbol{\xi}^\mu$ are uncorrelated random variables with zero mean and unit variance. The corresponding output ζ^μ is given by a deterministic teacher whose internal structure is that of a network similar to the student except for a possible difference in the number M of hidden units. Hidden unit n in the teacher network receives input information through the weight vector $\mathbf{B}_n = (B_{n1}, \ldots, B_{nN})$, and its activation under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is $y_n^\mu = \mathbf{B}_n \cdot \boldsymbol{\xi}^\mu$. The corresponding output is $\zeta^\mu = \sum_{n=1}^{M} \, g\,(\mathbf{B}_n \cdot \boldsymbol{\xi}^\mu)$. We will use indices $i, j, k, l \ldots$ to refer to units in the student network, and $n, m, \ldots$ for units in the teacher network.

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

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

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

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

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

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

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

A gradient descent rule for the update of the student weights results in $\mathbf{J}_i^{\mu+1} = \mathbf{J}_i^{\mu} + \frac{\eta}{N}\ \delta_i^{\mu}\ \boldsymbol{\xi}^{\mu}$, where the learning rate η has been scaled with the input size N, and

$$\delta_i^{\mu} \equiv g'(x_i^{\mu}) \left[\sum_{n=1}^{M} g(y_n^{\mu}) - \sum_{j=1}^{K} g(x_j^{\mu}) \right] \tag{4}$$

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

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

$$\frac{dR_{in}}{d\alpha} = \eta < \delta_i \; y_n > ,$$

$$\frac{dQ_{ik}}{d\alpha} = \eta < \delta_i \; x_k > + \; \eta < \delta_k \; x_i > + \; \eta^2 < \delta_i \; \delta_k > \quad . \tag{5}$$

The averages in Eq. (5) can be performed analytically for the choice $g(x) = \mathrm{erf}(x/\sqrt{2})$.

We find that after considerable amount of algebra, the equations of motion can be proven to reduce to a surprisingly compact form in terms of only two multivariate Gaussian integrals: $I_3 \equiv < g'(u) \; v \; g(w) >$ and $I_4 \equiv < g'(u) \; g'(v) \; g(w) \; g(z) >$. Arguments assigned to I_3 and I_4 are to be interpreted following our convention to distinguish student from teacher activations. For example, $I_3(i,n,j) \equiv < g'(x_i) \; y_n \; g(x_j) >$, and the average is performed using the three-dimensional covariance matrix C_3 which results from projecting the full covariance matrix $\mathcal{C}$ of Eq. (2) onto the relevant subspace. For $I_3(i,n,j)$ the corresponding matrix is:

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

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

$$I_3 = \frac{2}{\pi} \, \frac{1}{\sqrt{\Lambda_3}} \, \frac{C_{23}(1+C_{11}) - C_{12}C_{13}}{1+C_{11}} \, , \tag{6}$$

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

$$I_4 = \frac{4}{\pi^2} \, \frac{1}{\sqrt{\Lambda_4}} \; \arcsin\left(\frac{\Lambda_0}{\sqrt{\Lambda_1}\sqrt{\Lambda_2}}\right) , \tag{7}$$

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

$$\Lambda_0 = \Lambda_4 C_{34} - C_{23}C_{24}(1+C_{11}) - C_{13}C_{14}(1+C_{22}) + C_{12}C_{13}C_{24} + C_{12}C_{14}C_{23} \ ,$$

$$\Lambda_1 = \Lambda_4(1+C_{33}) - C_{23}^2(1+C_{11}) - C_{13}^2(1+C_{22}) + 2C_{12}C_{13}C_{23} \ ,$$

$$\Lambda_2 = \Lambda_4(1+C_{44}) - C_{24}^2(1+C_{11}) - C_{14}^2(1+C_{22}) + 2C_{12}C_{14}C_{24} \ .$$

The final form of the equations of motion for the overlaps is:

$$\frac{dR_{in}}{d\alpha} = \eta \left\{ \sum_m I_3(i,n,m) - \sum_j I_3(i,n,j) \right\} \ ,$$

$$\frac{dQ_{ik}}{d\alpha} = \eta \left\{ \sum_m I_3(i,k,m) - \sum_j I_3(i,k,j) \right\} + \eta \left\{ \sum_m I_3(k,i,m) - \sum_j I_3(k,i,j) \right\} +$$

$$\eta^2 \left\{ \sum_{n,m} I_4(i,k,n,m) - 2\sum_{j,n} I_4(i,k,j,n) + \sum_{j,l} I_4(i,k,j,l) \right\} . \qquad (8)$$

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

A detailed investigation of the realizable case ($K = M$) includes the dependence of the learning process on the learning rate η, the dynamical trapping on a symmetric subspace which exhibits no differentiation among student hidden units, the emergence of specialization, and the exponential convergence to perfect generalization. We postpone the discussion of these effects to a subsequent report [7], and conclude this presentation with two simple examples which demonstrate the power of the approach developed here when applied to the analysis of over-realizable and unrealizable learning scenarios.

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

The second example reveals the learning strategy in an unrealizable scenario, in which the student does not have enough resources to implement the task and cannot achieve perfect generalization. Numerical results are shown in Fig. 2 for $K = 3$, $M = 4$, and $\eta = 0.6$. The characterization of the teacher and initialization of the order parameters is as before. Examination of the time evolution of the order parameters shows that trapping in the symmetric subspace is followed by a process in which each student unit specializes on one of the three dominant teacher nodes ($i = 1 \to n = 2$, $i = 2 \to n = 4$, $i = 3 \to n = 3$) while ignoring the other two, and all three student nodes retain some overlap with the less dominant teacher node, $n = 1$. This residual overlap generates nonvanishing correlations among the student vectors, as shown in Fig. 2b. Note that the specialization of the student hidden units is ordered according to the relevance of the corresponding teacher nodes, resulting in a cascade of specialization transitions. The evolution of the generalization error is shown in Fig. 2d. .

These examples illustrate the first step in a systematic application of the equations of motion (8) to the analysis of a given learning scenario: numerical studies of small networks that reveal the essential features are followed by studies of increasingly large networks to capture trends and regularities, which are then quantified through analytic investigation of the relevant fixed points and their stability [7]. We have applied this program to the analysis of learning scenarios of the type described here, including correlations among the teacher hidden nodes, and extended it to include the effect of noise at both input and output level, as well as regularization through weight-decay.

The equations of motion (8) are easily extended to the case where the hidden-to-output teacher weights are not restricted to be positive and of unit strength, and are adaptive in the student network. The requirement of linear output units can be relaxed to allow for continuous activation functions at the output level. We have presented here the only available tool for studying such general aspects of learning and the emergence of generalization ability in multilayer networks.

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.

References

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

[2] T. L. H. Watkin, A. Rau and M. Biehl, *Rev. Mod. Phys.* **65**, 499 (1993).

[3] H. S. Seung, H. Sompolinsky, and N. Tishby, *Phys. Rev. A* **45** 6056 (1993).

[4] H. Schwarze, *J. Phys. A* **26**, 5781 (1993).

[5] G. Parisi, *Phys. Rev. Lett.* **43**, 1754 (1979).

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

[7] D. Saad and S. A. Solla, to appear in *Phys. Rev. E.*

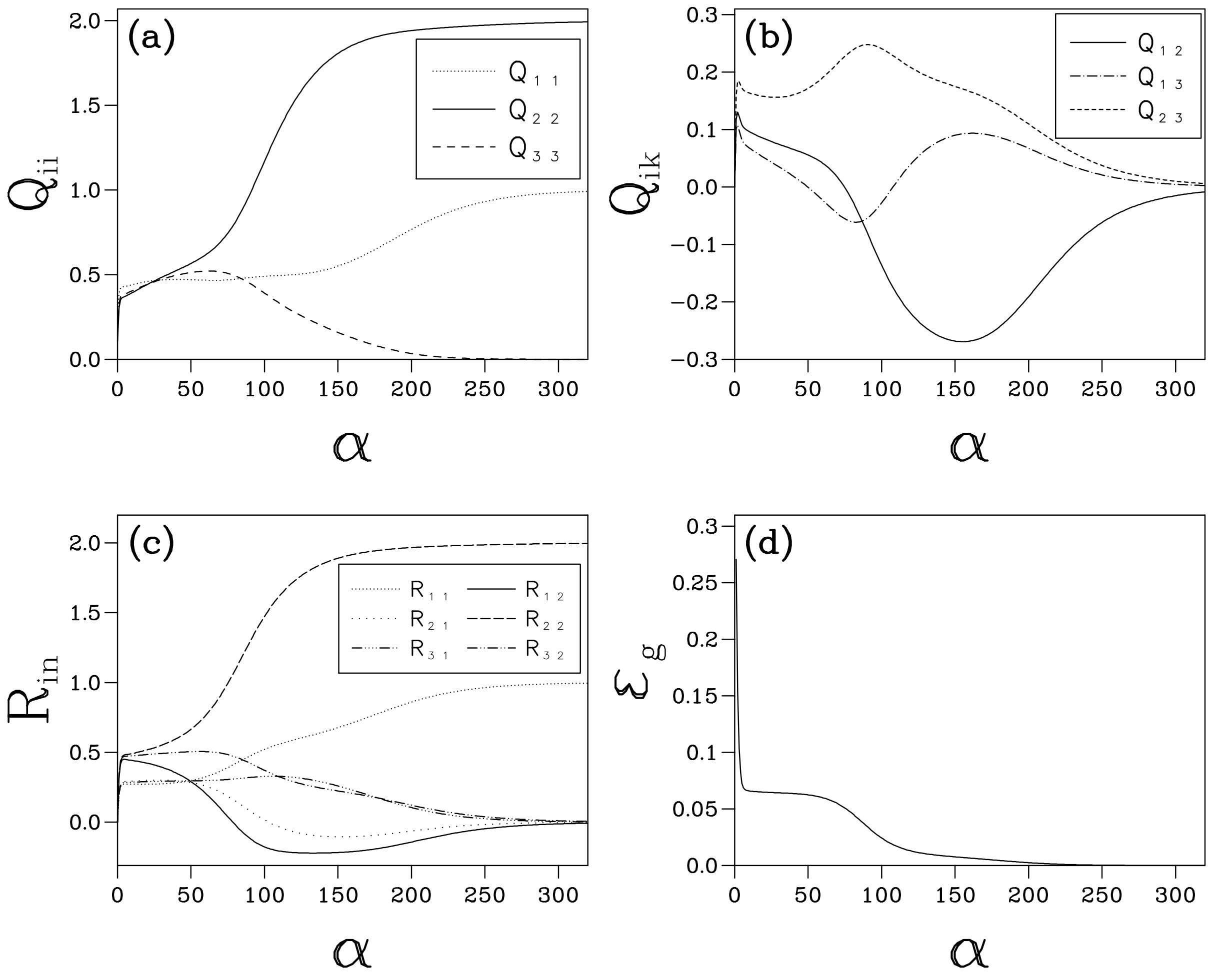

FIG. 1. Dependence of the overlaps and the generalization error on the normalized number of examples α, for a three-node student learning a two-node teacher: (a) the lengths of student vectors, (b) the correlation between student vectors, (c) the overlap between various student and teacher vectors, and (d) the generalization error.

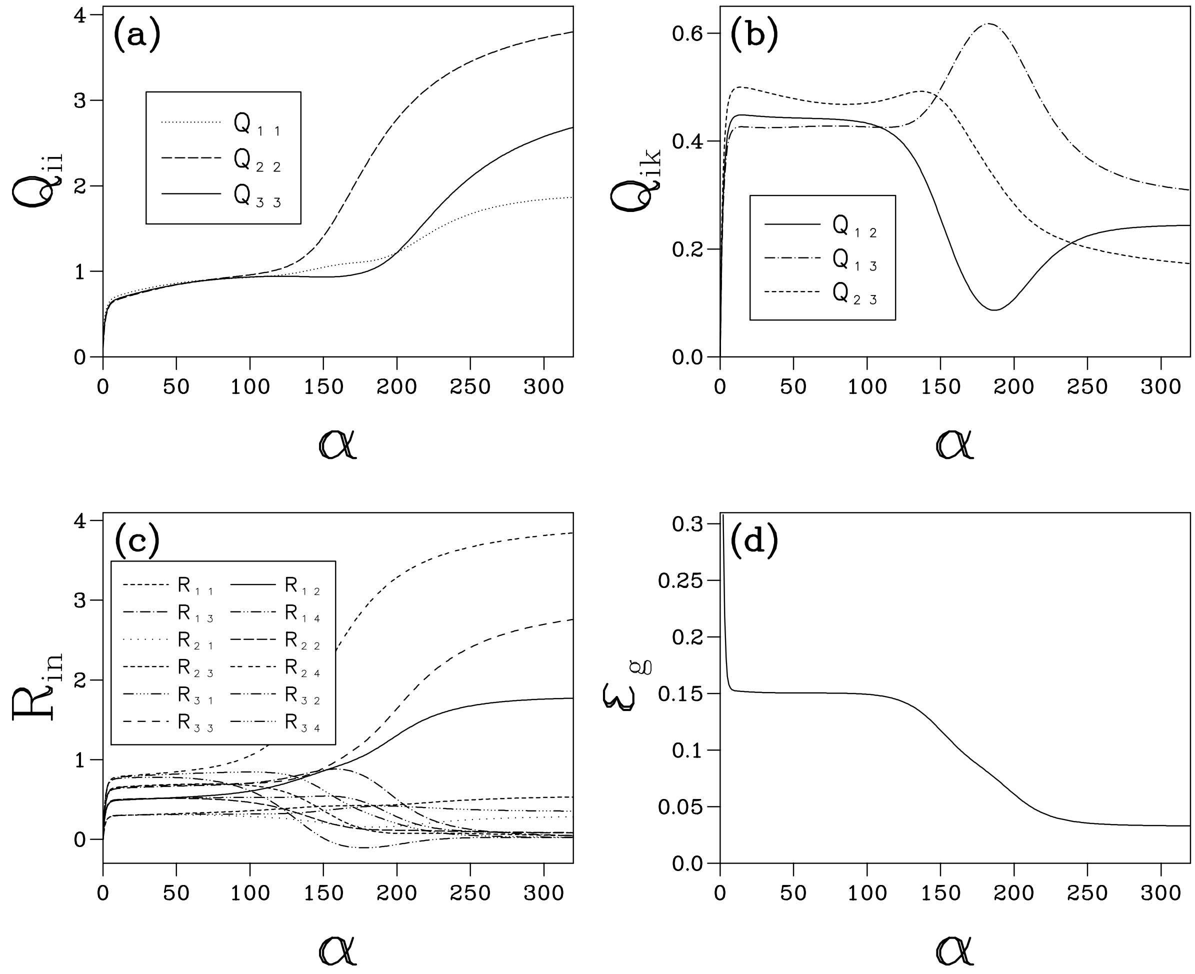

FIG. 2. Dependence of the overlaps and the generalization error on the normalized number of examples α, for a three-node student learning a four-node teacher: (a) the lengths of student vectors, (b) the correlation between student vectors, (c) the overlap between various student and teacher vectors, and (d) the generalization error.