On-line learning of unrealizable tasks

Silvia Scarpetta[1,2] and David Saad[1]
[1]*Department of Computer Science and Applied Mathematics, Aston University, Birmingham B4 7ET, United Kingdom*
[2]*Dipartimento di Scienze Fisiche, Universita' di Salerno, I-84081 Baronissi (SA), Italy and INFM Sezione di Salerno, Salerno, Italy*
(Received 8 June 1999)

The dynamics of on-line learning is investigated for structurally unrealizable tasks in the context of two-layer neural networks with an arbitrary number of hidden neurons. Within a statistical mechanics framework, a closed set of differential equations describing the learning dynamics can be derived, for the general case of unrealizable isotropic tasks. In the asymptotic regime one can solve the dynamics analytically in the limit of a large number of hidden neurons, providing an analytical expression for the residual generalization error, the optimal and critical asymptotic training parameters, and the corresponding prefactor of the generalization error decay. [S1063-651X(99)03211-0]

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

I. INTRODUCTION

Learning in layered neural networks refers to the modification of internal network parameters **J**, so as to bring the map implemented by the network $f_{\mathbf{J}}$ as close as possible to a desired map $f_{\mathbf{B}}$. The resulting performance is monitored through the *generalization error*, a measure of the dissimilarity between $f_{\mathbf{J}}$ and $f_{\mathbf{B}}$. Two-layer feed-forward networks are widely used in classification and regression applications, mainly due to their ability to implement any input-output mapping, in any desired accuracy, provided that the hidden layer has a sufficient number of neurons [1]. The scenario in which the network does *not* have a sufficient number of neurons to implement a certain input-output mapping is termed *structurally unrealizable*; in any other case the task is *realizable*.

Structural unrealizability has been examined, via statistical physics techniques examining the equilibrium distribution of models, mainly for the case of the perceptron [2,3], due to the technical difficulties of examining multilayer networks. In this paper we focus on the analysis of structurally unrealizable tasks in multilayer networks in the *on-line* learning scenario. On-line learning is a popular method for training multilayer feed-forward neural networks, where network parameters are updated according to only the latest in a sequence of training examples. On-line methods can be beneficial in terms of both storage and computational time, and also allow for temporal changes in the task being learned. An overview of on-line learning methods in neural networks can be found in [4]. We analyze unrealizability in *soft committee machine* (SCM) networks [5], in which the hidden units are connected to the output unit with positive couplings of fixed strength, and only the input-to-hidden couplings are adaptive. The learning problem can be formulated in a general student-teacher framework, in which a *student* SCM network with K hidden neurons is trained on examples generated by a *teacher* network of similar configuration, but with M hidden neurons. In unrealizable scenarios, the complexity of the task M is greater then the complexity of the student network K $<M$, and $L=M-K$ measures the degree of structural unrealizability.

We employ a statistical mechanics framework developed in [6] which allows us to describe analytically the learning dynamics, by means of a closed set of differential equations for the order parameters, with the number of examples playing the role of time. The effects of unrealizability on the evolution of the order parameters and the generalization error are studied numerically in all phases of learning process. We focus on the asymptotic phase, which is particularly interesting since here, contrary to realizable scenarios, no prior knowledge of the asymptotic solutions exists. Asymptotically, the system converges towards a stable fixed point which corresponds to a nonzero residual generalization error, whose value increases with the learning rate, and is nonzero even for an asymptotically vanishing learning rate. Although asymptotic solutions cannot be obtained analytically in general, one can obtain analytical solutions in the limit of large student network size K. The dependence of the generalization error decay on the network architecture and parameter choice is then derived, providing the optimal and critical asymptotic learning rate value as a function of the unrealizability measure L, in both standard and normalized SCM architectures defined below.

II. THE FRAMEWORK AND THE DYNAMICAL EQUATIONS

Consider a mapping from an input space $\boldsymbol{\xi}\in\mathbb{R}^N$ onto a scalar $\phi_{\mathbf{J}}(\boldsymbol{\xi})=\gamma\Sigma_{i=1}^{K}g(\mathbf{J}_i^{\mathrm{T}}\boldsymbol{\xi})$, which defines a SCM (termed the ''student'' network), where $\mathbf{J}\equiv\{\mathbf{J}_i\}_{1\leqslant i\leqslant K}$ is the set of input-to-hidden adaptable weights and the hidden-to-output weights are of fixed strength γ. We choose $g(x)$ $\equiv\mathrm{erf}(x/\sqrt{2})$ to be the sigmoidal activation function of the hidden units. The activation of the student hidden unit i under presentation of the input pattern $\boldsymbol{\xi}^{\mu}$ is denoted x_i^{μ} $=\mathbf{J}_i^{\mathrm{T}}\boldsymbol{\xi}^{\mu}$.

Let $(\boldsymbol{\xi}^{\mu},\zeta^{\mu})$ be the μth input-target pair in a sequence of training examples. Components of the input vectors $\boldsymbol{\xi}^{\mu}$ are drawn independently, at each iteration, from a zero mean Gaussian distribution with unitary variance. The corresponding target ζ^{μ} is given by a *teacher network* with the same architecture of the student except for a possible difference in

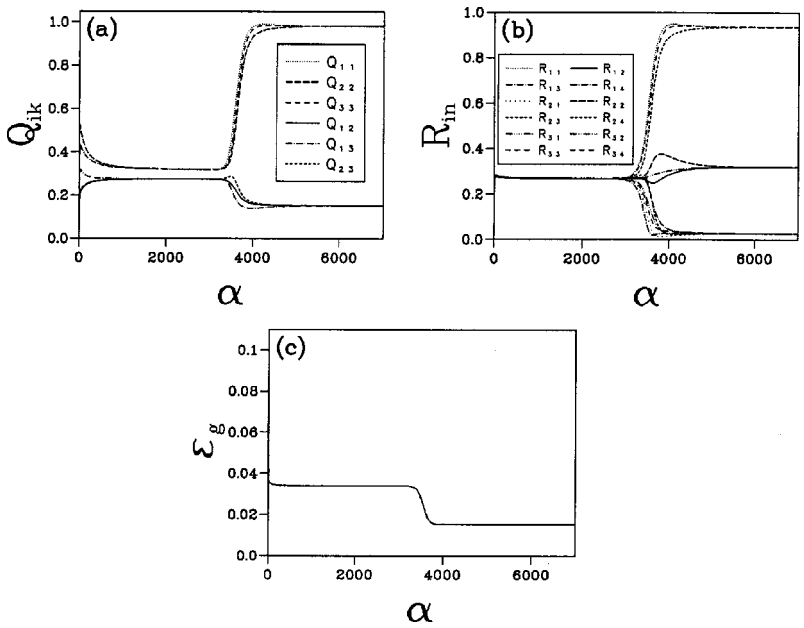

FIG. 1. Evolution of the order parameters and generalization error for the case $M=4$, $K=3$ is shown here for (a) the student-student overlap Q_{ik}, (b) the student-teacher overlap R_{in}, and (c) the generalization error. Initial conditions are $Q=0.5$, $R=U[0,10^{-12}]$.

the number M of hidden units, and is defined by the weight vectors $\mathbf{B}\equiv\{\mathbf{B}_n\}_{1\leqslant n\leqslant M}$. The target mapping is therefore $\zeta(\boldsymbol{\xi}^\mu)=\gamma\Sigma_{n=1}^{M}g(y_n^\mu)$, where $y_n^\mu=\mathbf{B}_n^{\mathrm{T}}\boldsymbol{\xi}^\mu$ is the activation of the teacher hidden unit n. We will use indices i,j,k,l to refer to units in the student network and n,m for units in the teacher network.

In standard SCM, the strength of hidden-to-output weights is unitary ($\gamma=1$). The SCM network is referred to as *normalized* if $\gamma=1/$(no. of hidden units); in this case the map implemented by the student and teacher networks is $\phi_{\mathbf{J}}(\boldsymbol{\xi})=(1/K)\Sigma_{i=1}^{K}g(x)$ and $\zeta(\boldsymbol{\xi})=(1/M)\Sigma_{n=1}^{M}g(y_n)$, respectively, so that the output of the teacher and student networks will have the same range $[-1,1]$, even if the number of hidden units is different $K\neq M$ and they implement maps of different complexity.

The case of a perfectly realizable task $K=M$ has been analyzed in [6] (for the standard SCM) and in [7] (for the normalized SCM). We focus here on the unrealizable scenario $M>K$. The error made by a student with weights $\mathbf{J}$ on a given input $\boldsymbol{\xi}$ is provided by the quadratic deviation $\varepsilon(\mathbf{J},\boldsymbol{\xi})=1/2[\zeta^\mu-\phi_{\mathbf{J}}(\boldsymbol{\xi}^\mu)]^2$. The most basic on-line learning rule is to perform gradient descent on this quantity. Then the update of each weight in response to the presentation of the μth example $(\boldsymbol{\xi}^\mu,\zeta^\mu)$ has the form

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

where $\delta_i^\mu\equiv\gamma g'(x_i^\mu)[\zeta^\mu-\phi_{\mathbf{J}}(\boldsymbol{\xi}^\mu)]$ and the learning rate η has been scaled with the input size N. Performance on a typical input defines the generalization error $\varepsilon_g\equiv\langle\varepsilon(\mathbf{J},\boldsymbol{\xi})\rangle_{\{\boldsymbol{\xi}\}}$ through an average over all possible input vectors $\boldsymbol{\xi}$.

We use a statistical mechanics description of the learning process [6] which is exact in the limit of large input dimension N where the dynamics of gradient descent learning in the unrealizable scenario is completely described by a small set of order parameters $\langle x_ix_j\rangle=\mathbf{J}_i^{\mathrm{T}}\mathbf{J}_k\equiv Q_{ik}$, $\langle x_iy_n\rangle=\mathbf{J}_i^{\mathrm{T}}\mathbf{B}_n\equiv R_{in}$, and $\langle y_ny_m\rangle=\mathbf{B}_n^{\mathrm{T}}\mathbf{B}_m\equiv T_{nm}$, measuring overlaps between student and teacher vectors. The order parameters are necessary and sufficient to determine the generalization error $\varepsilon_g=\langle\varepsilon(\mathbf{J},\boldsymbol{\xi})\rangle_{\{\boldsymbol{\xi}\}}$.

If we interpret the normalized number of examples $\alpha=\mu/N$ as a continuous time variable, the update equations (1) gives rise to first-order coupled differential equations of the form

$$\frac{dR_{in}}{d\alpha}=\eta\langle\delta_i^\mu y_n\rangle,$$

$$\frac{dQ_{ik}}{d\alpha}=\eta\langle\delta_i^\mu x_j\delta_j^\mu x_i\rangle+\eta^2\langle\delta_i^\mu\delta_i^\mu\rangle, \tag{2}$$

where the angled brackets denote averages over inputs. Averages in Eq. (2) can be carried out analytically for arbitrary K and $M=K+L$, providing a closed set of equations of motion. Note that δ_i is slightly different for standard or normalized SCM architecture, as well as the corresponding equations of motion.

III. STRUCTURE OF THE SOLUTIONS IN UNREALIZABLE SCENARIOS

In the unrealizable scenario the student does not have enough resources to imitate the teacher units accurately even if an infinite number of examples is provided, so one may expect residual generalization error and a suboptimal mapping of the asymptotic student vectors onto the space spanned by the teacher vectors.

To demonstrate learning in an unrealizable scenario, we show the evolution of the order parameters and the generalization error for a standard SCM with $K=3$ hidden units learning an unrealizable task with $L=1$ $(M=4)$. In the remainder of the paper, we will focus on uncorrelated isotropic teachers of unitary length $T_{nm}=\delta_{nm}$. The dynamical evolution of the overlaps Q_{ik} and R_{in} follows from integrating the equations of motion (2) from initial conditions determined by the (random) initialization of the student weights $\mathbf{J}$; we

initialize Q_{ii} from uniform distributions in the $[0,0.5]$ interval, $Q_{i\neq k}=0$, and R_{in} from $[0,10^{-12}]$.

The time evolution of the various order parameters is shown in Figs. 1(a)–1(c) for $\eta=0.2$. As for realizable scenario [6], the unrealizable dynamics is characterized by two major phases of learning. Initially, the order parameters are trapped in an unstable fixed point characterized by a lack of differentiation between the hidden units of the student where the overlaps of each student unit with all teacher units R_{in} are nearly identical. All the student overlaps $Q_{i\neq k}$ have nearly the same value, which does not differ much from the value of the norms Q_{ii}. Trapping in the symmetric phase for unrealizable scenarios is of the same nature as the one observed and analyzed in the realizable case [6,7]. Eventually, small perturbations introduced by the random initial conditions lead to an escape from this phase and convergence towards the asymptotic (suboptimal) regime [8].

Understanding the evolution of the parameters in the asymptotic phase is particularly important in the study of unrealizable scenarios, where no prior knowledge exists about the asymptotic solutions themselves. The suboptimal mapping that emerges from our numerical solutions suggests that the limited student resources are used mainly to specialize on *certain* teacher vectors, while retaining small correlation with the rest of the teacher vectors. The evolution of the student norms and student-student correlations shown in Fig. 1(a) demonstrates that asymptotically, each one of the student units imitates one of the teacher units ($R_{11}\approx T_{11}$, $R_{24}\approx T_{44}$, and $R_{33}\approx T_{33}$), while ignoring units imitated by other student vectors ($R_{13},R_{14},R_{21},R_{23},R_{31},R_{34}\approx 0$), and retaining some correlation with other teacher units, not imitated by other student units (R_{12},R_{22},R_{32}). The corresponding evolution of the generalization error is shown in Fig. 1(c).

In structurally unrealizable cases, as for learning with noise [9], suboptimal asymptotic performance will be obtained for any fixed learning rate, suggesting that an annealing schedule should be invoked asymptotically. *Ideally*, one would expect asymptotically the student vectors to be confined to the M-dimensional subspace $\mathcal{S}_B$ spanned by the set of orthogonal unit length teacher vectors, and they can therefore be represented as M ($<N$) dimensional vectors in the *teacher coordinate system*. This is true for vanishing learning rates η. However, 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 $\mathbf{J}_i=\Sigma_{n=1}^{M}R_{in}\mathbf{e}_n+\mathbf{J}_i^{\perp}$, where $\mathbf{J}_i^{\perp}$ indicates the component of $\mathbf{J}_i$ in the orthogonal subspace. The optimal asymptotic solution, with the lowest asymptotic generalization error is characterized by solutions obtained with a vanishing learning rate η and thus a vanishing vector $\mathbf{J}^{\perp}$. In the following section, we present an analysis of the asymptotic solution when the learning rate is annealed.

IV. ASYMPTOTIC REGIME

The number of order parameters in Eq. (2) is $K(K+1)/2+KM$, so that the analysis becomes more and more difficult as K and M grow. However, the symmetric architecture of the teacher network $T_{nm}=\delta_{nm}$ leads to the grouping of the dynamical variables. In the general case of an unrealizable learning scenario and isotropic teachers, the system's dynamics can be described in terms of only five variables, via the ansatz

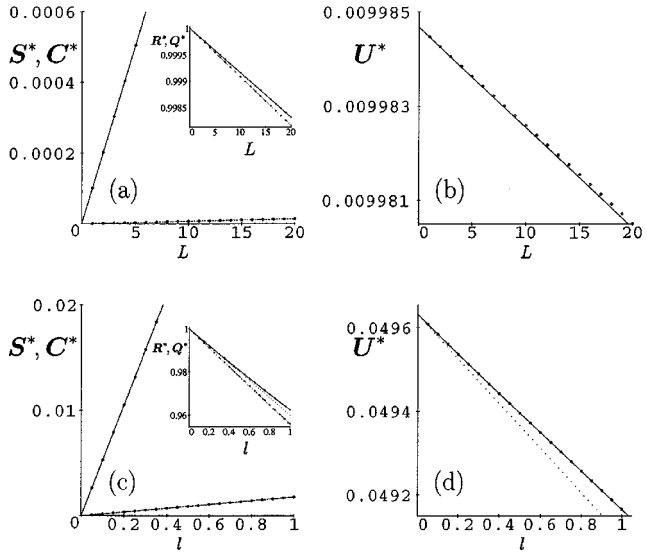

FIG. 2. Theoretical (lines) vs numerical results (circles) for the optimal fixed point Q^*,C^*,R^*,S^*,U^*. (a,b) We plot C^* (upper line), S^* (lower line), in the inset Q^* (upper line) and R^* (lower line), and in (b) U^*, for the case $L\ll K$ as a function of L at $K=100$. (c,d) The overlaps C^* (upper line), S^* (lower line), in the inset Q^* (upper line) and R^* (lower line), and in (d) U^*, for the case $L=lK$ as a function of l at $K=20$; for comparison, dotted lines are analytical results for the $L\ll K$ case at $K=20$.

$$Q_{ik}=Q\delta_{ik}+C(1-\delta_{ik}),$$

$$R_{in}=R\delta_{in}+S(1-\delta_{in})\theta(K-n)+U\theta(n-K), \qquad (3)$$

for the student-student overlaps and (apart from a relabeling of the student hidden units) student-teacher overlaps, respectively, where the step function θ is 0 for negative arguments and 1 otherwise. As one can see from Fig. 1, this approximation (3) is particularly good in the symmetric phase (where also $R\equiv S\equiv U$ holds) and during the final convergence to the asymptotic regime. Asymptotic solutions in the case of an isotropic teacher are characterized by specialized student vectors of similar norms ($Q_{ii}=Q$ for all $1\leqslant i\leqslant K$) and similar correlations among themselves ($Q_{ik}=C$ for all $1\leqslant i,\ k\leqslant K,\ i\neq k$); each one of these vectors specializes on a certain teacher vector ($R_{ii}=R$ for all $1\leqslant i\leqslant K$), while all student vectors have similar correlations with all K teacher vectors imitated by other student vectors ($R_{in}=S$ for all $1\leqslant i,\ n\leqslant K$, and $i\neq n$), as well as with the other $M-K$ teacher vectors on which no student vector specializes ($R_{in}=U$ for all $1\leqslant i\leqslant K$, and $K<n\leqslant M$).

Therefore, the system's dynamics is described asymptotically by only five coupled differential equations derived using the relations (3). In order to find the analytical expression for the optimal fixed point, we solve the truncated equations of motion, neglecting terms of order $O(\eta^2)$ in Eqs. (2). In order to find the asymptotic fixed point of this system of five coupled equations analytically, we exploit the geometrical constraint that hold between the order parameters to simplify the system. Since at the optimal fixed point student vectors are confined to $\mathcal{S}_B$, one may express any vector $\mathbf{J}_i$ as

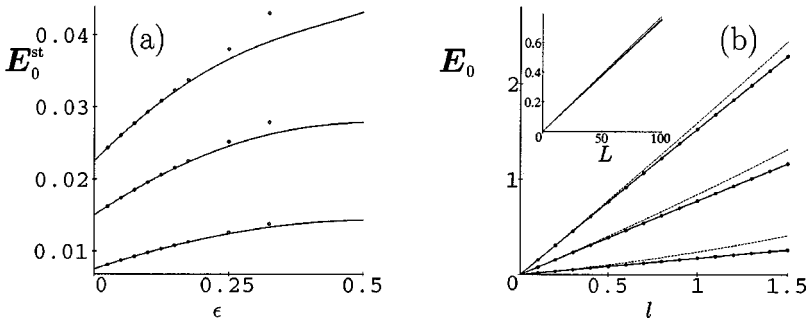

FIG. 3. The residual generalization error in standard SCM—theoretical (lines) vs numerical results (circles). (a) Theoretical value of $E_o(L,K)$ for the case $L=lK$ as a function of ϵ for $L=1,2,3$ from down to up. (b) Theoretical value of $E_o(L,K)$ for the case $L=lK$ (solid line) and the case $L \ll K$ (dashed line) as a function of $l=L/K$. $K=20,100,200$ from down to up. Inset: Theoretical value of $E_o(L,K)$ for the case $L=lK$ are plotted as a function of L for $K=100,500,1000,10\,000$. All the curves, except the one with $K=100$ (dashed line), collapse onto the same straight line.

$$\mathbf{J}_i = S\mathbf{e}_1 + \cdots S\mathbf{e}_{i-1} + R\mathbf{e}_i + S\mathbf{e}_{i+1} + \cdots + S\mathbf{e}_K + U\mathbf{e}_{K+1} + \cdots U\mathbf{e}_M ,$$

where $\mathbf{e}_n$, $n=1,\ldots,M$, are the orthogonal set of teacher vectors. Using this expression for the student vector, one can easily derive a constraint between the order parameters R, S, U, Q, and C:

$$\begin{aligned} Q &= R^2 + (K-1)S^2 + (M-K)U^2, \\ C &= 2RS + (K-2)S^2 + (M-K)U^2. \end{aligned} \tag{4}$$

Unfortunately, the solutions of the truncated equations of motion, even when using the geometric constraint, still cannot be obtained analytically. However, we can obtain the optimal fixed point in the limit of a large network, when the number of student hidden neurons $K \gg 1$ is large (but still $N \gg K$). We expand both the constraint (4) and the truncated equations of motion in the small parameter $\epsilon \equiv 1/K$. In this scenario we can distinguish two cases: $L \equiv M-K \ll K$ (termed *small unrealizability*) when the excess of teacher hidden neurons L is small compared to the large number of student hidden neurons K [so that L is of $O(\epsilon^0)$], and $L \simeq K$ (termed *strong unrealizability*) when the teacher excess of resources L is of the same order of magnitude of the student resources K, so that $L=lK=l\epsilon^{-1}$, with a finite factor of proportionality l of $O(1)$. In both cases we find the fixed point Q^*, C^*, R^*, S^*, U^* up to $O(\epsilon^3)$. In the following, we discuss the standard SCM architecture. Analytical expressions for the approximated optimal fixed point in the small and strong unrealizability cases are given in Appendix A. The dependence of the order parameters at the fixed point from the unrealizability degree L is shown in Fig. 2. Exact numerical results are included in the figures in order to validate our theoretical predictions. For $L=0$, the realizable case fixed point $Q^*=R^*=1, C^*=S^*=0$ is recovered (U is meaningless for realizable scenarios). The corresponding residual generalization error is

$$E_0^{sm} = \frac{1}{6}\frac{L(-3+\pi)}{\pi} - \frac{3}{2}\frac{L(2474\sqrt{3}-4291)\epsilon}{(-9+8\sqrt{3})^3\pi} + \frac{3}{2}\frac{L(-859\,925+496\,432\sqrt{3}+18\,324L\sqrt{3}-31\,659L)\epsilon^2}{(-9+8\sqrt{3})^4\pi}$$

in the small L case, and

$$E_0^{st} = -\frac{1}{2}\frac{l(273-144\sqrt{3}-91\pi+48\sqrt{3}\pi)}{\pi(-9+8\sqrt{3})^2\epsilon} + \frac{1}{2}\frac{l(-561+326\sqrt{3})}{\pi(-9+8\sqrt{3})^2} - \frac{1}{48}\frac{l(864l-273l^2+29\,658+144\sqrt{3}l^2-472l\sqrt{3}-17\,088\sqrt{3})\epsilon}{\pi(-9+8\sqrt{3})^2}$$

in the strong unrealizability case (with the $L=lK$ scaling assumption). To examine the accuracy of our approximation, theoretical results are compared with values obtained numerically. Dependence of E_0^{st} on K when L is fixed is shown in Fig. 3(a). Both the theoretical predictions of the residual error, E_0^{sm} and E_0^{st}, are shown in Fig. 3(b) as a function of the relative number of teacher units in excess $l=L/K$. We see that the solution obtained for $L \ll K$ (dashed line) becomes more and more inaccurate as l increases, as one expects, while the scaling assumption $L=lK$ gives accurate results also for a very small value of L, where it coincides with the $L \ll K$ solution. It is interesting to note that for large K, the residual error is proportional to L only, giving a direct indication for the number of additional hidden units required to

make the problem realizable. Indeed, all the lines for the residual generalization error corresponding to $K=100,500,1000,10\,000$ collapse onto one straight line if plotted as a function of L, as shown in inset of Fig. 3(b).

In order to describe the approach of the system to the optimal fixed point, we take into consideration terms of order $O(\eta^2)$ in the dynamical equations (2). In this paper we will concentrate on the annealed learning rate $\eta=\eta_0/\alpha$, since this is the optimal annealing schedule, as in the realizable $(K=M)$ noisy case [9]. To solve the asymptotics of the system, we expand the full equations of motion to first order around our estimation of the optimal fixed point Q^*,C^*,R^*,S^*,U^*. We find five linear coupled differential equation for the five order parameters represented by the vector $\mathbf{u}$,

$$\frac{d}{d\alpha}\mathbf{u}=\eta_\alpha M\mathbf{u}+\eta_\alpha^2\mathbf{b}, \tag{5}$$

where

$$\begin{aligned}\mathbf{u}&=(Q-Q^*,C-C^*,R-R^*,S-S^*,U-U^*)^T\\ &\equiv(q,c,r,s,u)^T,\end{aligned} \tag{6}$$

$\eta_\alpha=\eta_0/\alpha$, and both the zero-order term $\mathbf{b}$ and the Jacobian matrix $\mathcal{M}$ are functions of the student network size K and of the degree of unrealizability L. The asymptotic equations of motion (5) are derived by dropping terms of order $O(\eta_\alpha||\mathbf{u}||^2)$ and higher, and terms of order $O(\eta_\alpha^2\mathbf{u})$. The latter are linear in the order parameters $\mathbf{u}$, but are negligible in comparison to the $\eta_\alpha\mathbf{u}$ and $\eta_\alpha^2\mathbf{b}$ terms in Eq. (5) as $\alpha\to\infty$.

Since as our estimation of the optimal fixed point we use an expansion around $\epsilon=0$ truncated at the third order, then also the vector $\mathbf{b}$ and the Jacobian matrix $\mathcal{M}$ of the first derivatives computed at the fixed point are in the form of truncated series in ϵ.

Equations (5) can be exactly solved if one computes analytically the eigenvalues and eigenvectors of the matrix $\mathcal{M}$. Finding analytically exact eigenvalues and eigenvectors of $\mathcal{M}$ is hampered by technical difficulties. We therefore keep the first two orders in the expansion

$$\mathcal{M}=\mathcal{M}_o+\epsilon\mathcal{M}_1+\epsilon^2\mathcal{M}_2+\cdots \tag{7}$$

and use the theory of perturbation for nonsymmetric matrices (e.g., as in [10,11]) in order to compute the eigenvalues and eigenvectors. We stop at the first-order correction in ϵ, $\lambda_i=\lambda_i^0+\epsilon\lambda_i^1$, where the eigenvalue degeneracy which exists in the leading-order terms is removed, to find five different negative eigenvalues:

$$\begin{aligned}\lambda_1&=-\frac{1}{36}\frac{(-9+8\sqrt{3})}{\pi},\quad \lambda_2=-\frac{2}{3}\frac{(-3+2\sqrt{3})}{\pi},\\ \lambda_3&=-\frac{1}{\pi\epsilon}-\frac{1}{3}\frac{-3+2\sqrt{3}}{\pi},\\ \lambda_4&=-\frac{1}{\pi\epsilon}-\frac{1}{12}\frac{-21+8\sqrt{3}}{\pi},\\ \lambda_5&=-\frac{2}{\pi\epsilon}-\frac{2}{3}\frac{-3+2\sqrt{3}}{\pi}\end{aligned} \tag{8}$$

for the $L\ll K$ case, and

$$\begin{aligned}\lambda_1&=-\frac{1}{36}\frac{-9+8\sqrt{3}}{\pi},\quad \lambda_2=-\frac{2}{3}\frac{-3+2\sqrt{3}}{\pi},\\ \lambda_3&=-\frac{1}{\pi\epsilon}-\frac{144\sqrt{3}-393}{444\pi}l-\frac{-777+296\sqrt{3}}{444\pi},\\ \lambda_4&=-\frac{1}{\pi\epsilon}-\frac{36\sqrt{3}-15}{111\pi}l-\frac{-111+74\sqrt{3}}{111\pi},\\ \lambda_5&=-\frac{2}{\pi\epsilon}+2\frac{-9+4\sqrt{3}}{\pi(-9+8\sqrt{3})}l+2\frac{-25+14\sqrt{3}}{\pi(-9+8\sqrt{3})}\end{aligned} \tag{9}$$

for the $L=l/\epsilon$ case. Results turn out to be in good agreement, especially for large K, with the exact numerical values of eigenvalues of the Jacobian matrix evaluated around the true optimal fixed point, which can be found numerically. While λ_1 and λ_2 do not depend on ϵ and l, all other eigenvalues do. We find that $\lambda_5<\lambda_4<\lambda_3<\lambda_2<\lambda_1<0$ for all values of $0<l<1/\epsilon$ and $0\leqslant\epsilon\leqslant 0.5$ (i.e., all values of interest $0<L<K^2$ and $K>2$).

If λ_i are the eigenvalues of the matrix $\mathcal{M}$, and D is the matrix of the eigenvectors, such that

$$D^{-1}\mathcal{M}D=\begin{bmatrix}\lambda_1 & 0 & 0 & 0 & 0\\ & \cdots & \cdots & & \\ 0 & 0 & 0 & 0 & \lambda_5\end{bmatrix}, \tag{10}$$

then, following [9], the solution of Eq. (5) is

$$\mathbf{u}(\alpha)=DL(\alpha,\alpha_0)D^{-1}\mathbf{u}(\alpha_0)+D\Theta(\alpha,\alpha_0)D^{-1}\mathbf{b}, \tag{11}$$

where $L(\alpha,\alpha_0)$ and $\Theta(\alpha,\alpha_0)$ are diagonal matrices whose elements take the form

$$L_{ii}(\alpha,\alpha_0)=\left(\frac{\alpha}{\alpha_0}\right)^{\lambda_i\eta_0}$$

and

$$\Theta_{ii}(\alpha,\alpha_0)=\frac{-\eta_0^2}{1+\lambda_i\eta_0}[\alpha^{-1}-\alpha^{\lambda_i\eta_0}\alpha_0^{-1-\lambda_i\eta_0}]. \tag{12}$$

As the first contribution in Eq. (11) depends on the actual initial conditions $\mathbf{u}(\alpha_0)$, and since we are interested mainly in the asymptotic regime, it will be neglected in what follows as it decays more rapidly than the second contribution.

We expand the explicit expression of the generalization error, given in Eq. (B1), around the optimal fixed point to the second order in $\mathbf{u}$, to obtain

$$\epsilon_g^{\mathrm{asy}}=E_o+E_1^T\mathbf{u}+\mathbf{u}^TE_2\mathbf{u}.$$

Elements of both the vector E_1 and the matrix E_2 are truncated series in the small parameter ϵ, since the optimal fixed point is known analytically up to $O(\epsilon^3)$.

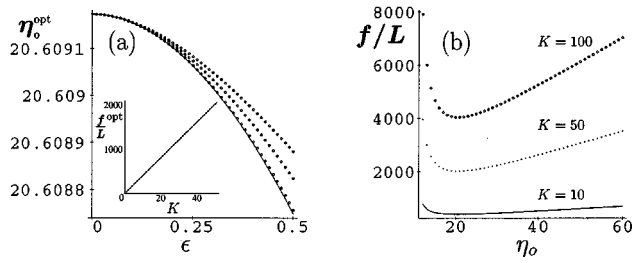

FIG. 4. (a) The optimal learning rate $\eta_0^{\rm opt}(l,K)$ for $l=0.05,0.5,1$ (circled lines, from down to up) as a function of ϵ. Solid line is $\eta_o^{\rm opt}(\epsilon)$ for the case $L \ll K$. In the inset, the optimal prefactor of error decay scaled with L, $f(L,K,\eta_o^{\rm opt})/L$, is shown as a function of K. For all values of $L=1,100,2000$ the plots collapse onto the same curve, corresponding to $f/L=40.3$ K. (b) The scaled prefactor $f(L,K,\eta_0)/L$ as a function of η_0 for student network size $K=10$ (solid line), $K=50$ (dots), and $K=100$ (circles). All values of L give the same result ($L=1,10,100$).

Using the eigenvalues of Eq. (8) or Eq. (9) and the solution (11), the generalization error can then be rewritten as a combination of the modes Θ_{ii}, whose coefficients are functions of ϵ and L.

We find that only two modes, Θ_{22} and Θ_{55}, associated with eigenvalues λ_2 and λ_5, survive in the linear term of the generalization error when we truncate the expansion of $E_{\rm lin}$ to the second leading order in ϵ. We verified numerically that the modes Θ_{11}, Θ_{33}, and Θ_{44} are orthogonal to the first-order term in the generalization error, and therefore do not contribute to its decay *at all orders in* ϵ, but contribute only to the decay of the second-order term with the corresponding eigenvalues $2\lambda_1$, $2\lambda_3$, and $2\lambda_4$.

Therefore, the critical learning rate η_c, above which the generalization $\varepsilon_g^{\rm asy}$ decays as $1/\alpha$, is

$$\eta_c = \max\left\{ -\frac{1}{2\lambda_1}, -\frac{1}{\lambda_2}, -\frac{1}{2\lambda_3}, -\frac{1}{2\lambda_4}, -\frac{1}{\lambda_5} \right\}$$
$$= -\frac{1}{2\lambda_1} = \frac{18\pi}{-9+8\sqrt{3}}$$

in both the $L \ll K$ [Eq. (8)] and $L=lK$ [Eq. (9)] cases.

For $\eta_0 > \eta_c$, the generalization error decays like $1/\alpha$ to the residual error E_o; neglecting second-order terms, since they decay as $1/\alpha^2$, one finds an asymptotic error decay of the form

$$\varepsilon_g^{\rm asy} = E_0 + \eta_0^2 \left(\frac{c_1(L,K)}{(-\lambda_5\eta_0 - 1)} + \frac{c_2(L,K)}{(-\lambda_2\eta_0 - 1)} \right) \alpha^{-1}$$
$$= E_0 + f(L,K,\eta_0)\frac{1}{\alpha}, \tag{13}$$

where c_1 and c_2 for both cases $L \ll K$ and $L=lK$ are given in Appendix B.

For optimal decay of the asymptotic error, one has also to minimize the prefactor $f(L,K,\eta_0)$ in Eq. (13). In the case of $L \ll K$, the optimal value of η_o is independent of L, while in the case of $L=lK$ it shows a rather weak dependence on l. The values of $\eta_0{}^{\rm opt}(L,K)$ for $l=0.05,0.5,1$ as a function of ϵ are shown in Fig. 4(a), where $\eta_o^{\rm opt}(K)$ for the case $L \ll K$ is also included. For large K, the optimal prefactor $\eta_0^{\rm opt}$, for both the small and strong unrealizability case, tends to the same value ($\eta_0^{\rm opt} \sim 20.609$).

The sensitivity of the generalization error decay factor $f(L,K,\eta_0)$ to the choice of η_0 is shown in the inset of Fig. 4(b), where $f(L,K,\eta_0)/L$ is plotted as a function of η_0 for $K=10$, 50, 100, and $L=1,100$. Curves for different values of L collapse onto the same line, showing that $f(L,K,\eta_0)/L$ is a function of K and η_0 only. The optimal prefactor $f(L,K,\eta_0^{\rm opt})$ is shown as a function of K in the inset of Fig. 4(a); it seems that $f(L,K,\eta_0^{\rm opt})$ can be well approximated as proportional to the product LK.

V. NORMALIZED SCM ARCHITECTURE

In the standard SCM architecture, the output of the student and teacher network range, respectively, in $[-K,K]$ and $[-M,M]$. Therefore, not only is the complexity of the student and teacher mapping different, but also the range of values that the outputs can assume. We examine in this section unrealizable scenarios for normalized SCM architecture, in which hidden-to-output weights are normalized, so that output values for networks of different sizes always range over the same interval $[-1,1]$.

We look for the optimal asymptotic solution, following the procedure that we have described in the preceding section. Using relations (3), we expand both the equation of motion (2) truncated at order $O(\eta)$ and the constraints (4) in the small parameter ϵ. We find the fixed point solution iteratively for the case $L \ll K$, but unfortunately a solution cannot be found analytically in the $L=lK$ case. Therefore, in the rest of the paper we will focus on the small unrealizability case ($L \ll K$). The optimal fixed point solutions up to order $O(\epsilon^3)$ are given in Appendix A. The dependence of the op-

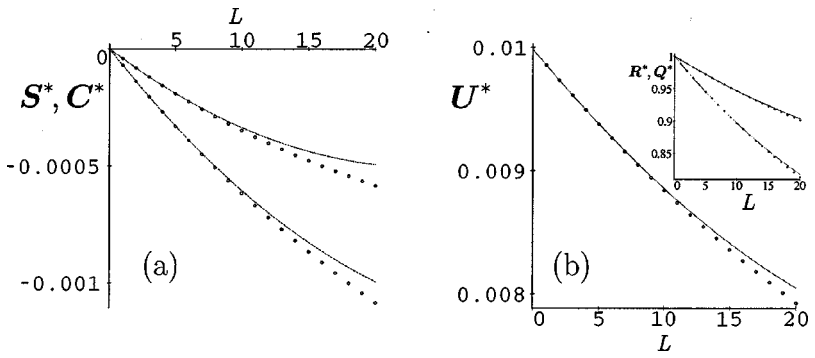

FIG. 5. Theoretical (lines) vs numerical results (circles) for the optimal fixed point Q^*, C^*, R^*, S^*, U^* in normalized SCM architecture, as a function of L at $K=100$. (a) The overlaps C^* (upper line) and S^* (lower line), (b) the overlap U^*, and in the inset Q^* (upper line) and R^* (lower line). This is to be compared with Figs. 2(a) and 2(b) for un-normalized SCM architecture.

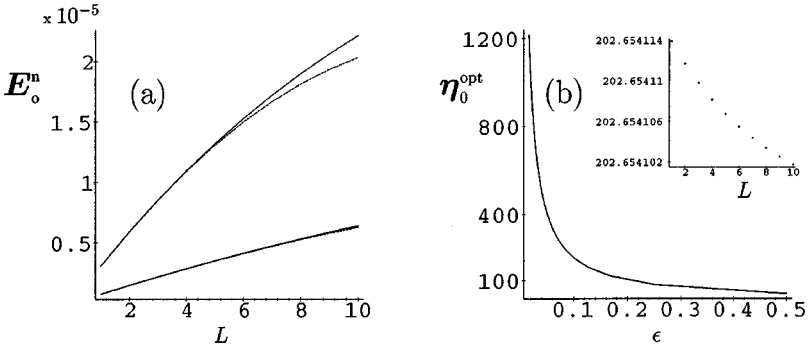

FIG. 6. (a) Residual error $E_0^n(L,K)$ in normalized SCM architecture: theoretical (dashed line) vs numerical results (solid lines) for $K=50,100$ (from up to down). (b) The optimal learning rate $\eta_o^{\mathrm{opt}}(L,K)$ in normalized SCM architectures for $L=1,20,200,1000$ as a function of ϵ. All lines collapse onto the same curve $\eta_0^{\mathrm{opt}}=20.27$ K. In the inset $\eta_o^{\mathrm{opt}}(L,K)$ is shown for $K=10$ as a function of L.

timal fixed point Q^*,C^*,R^*,S^*,U^* on L is shown in Fig. 5, validated by comparison with numerical solutions. Contrary to the un-normalized architecture, here the fixed point produces negative values for the order parameters C and S. Moreover, Q and R decrease with L much faster than in un-normalized architecture. This configuration corresponds to a residual generalization error:

$$E_0^n=\frac{1}{6}\frac{(\pi-3)L\epsilon^2}{\pi}+\left(\frac{1}{2}\frac{(-420\sqrt{3}+750+48\pi\sqrt{3}-91\pi)L^2}{\pi(-9+8\sqrt{3})^2}+\frac{1}{2}\frac{(326\sqrt{3}-561)L}{\pi(-9+8\sqrt{3})^2}\right)\epsilon^3, \tag{14}$$

which, apart from the $1/K^2$ normalization factor, is lower than the one obtained in un-normalized SCM. Numerical values of the residual error are compared with the theoretical results (14) in Fig. 6(a). As we expect, the agreement is good when L is much lower than K, and improves for large K.

In the annealed learning rate $\eta=\eta_0/\alpha$ schedule, the dynamics of the system in the vicinity of the optimal fixed point is described by the linearized equations of motion (5), whose solution is given by Eq. (11). The leading order in the Jacobian matrix, this time, is $O(\epsilon^0)$, in contrast with the non-normalized SCM case, where it was of $O(\epsilon^{-1})$. Keeping only the first two orders in the expansion of $\mathcal{M}$ and using again the perturbation theory for nonsymmetric matrices, one obtains the following approximations for the five eigenvalues:

$$\lambda_1=-\frac{1}{36}\frac{\epsilon(-9+8\sqrt{3})}{\pi},\quad \lambda_2=-\frac{2}{3}\frac{\epsilon(-3+2\sqrt{3})}{\pi},$$

$$\lambda_3=-\frac{1}{\pi}+\epsilon(0.188\,98-0.181\,92L),$$

$$\lambda_4=-\frac{1}{\pi}+\epsilon(-0.049\,12-0.182\,05L),$$

$$\lambda_5=-\frac{2}{\pi}+\epsilon(-0.098\,42-0.364\,70L),$$

where analytical results have been replaced by the numerical equivalent for brevity, and λ_1 and λ_2 are exactly ϵ times the corresponding eigenvalues in the standard SCM [Eq. (8)]. It is again the case that $\lambda_5<\lambda_4<\lambda_3<\lambda_2<\lambda_1<0$ for the range of values K,L which we are interested in (all $L>0$ and $K>1$).

Again, we find that only two modes, Θ_{22} and Θ_{55}, survive in the linear term of the generalization error, while all other modes contribute only to the decay of the second-order term. The critical learning rate is therefore

$$\eta_c^n=\max\left\{-\frac{1}{2\lambda_1},-\frac{1}{\lambda_2},-\frac{1}{2\lambda_3},-\frac{1}{2\lambda_4},-\frac{1}{\lambda_5}\right\}=-\frac{1}{2\lambda_1}=\frac{18\pi}{(-9+8\sqrt{3})\epsilon},$$

exactly K times the critical learning rate for the standard SCM architecture. For optimal decay of the asymptotic error, one has to minimize numerically the prefactor $f(\eta_0,L,K)$ in Eq. (13). The value of $\eta_0^{\mathrm{opt}}(L,K)$, shown in Fig. 6(b), turns out to be almost proportional to K only, with a very weak dependence on L [inset of Fig. 6(b)]. It is to be compared with the corresponding solid line in Fig. 4(a) for non-normalized networks and $L\ll K$.

The optimal error decay prefactor $f(\eta_0^{\mathrm{opt}},L,K)$ is shown in Fig. 7(a). It turns out to be well fitted by $f(\eta_0^{\mathrm{opt}},L,K)=5.83L/K$, i.e., about $7K^2$ times smaller than the optimal prefactor in the un-normalized architecture. The sensitivity of the generalization error decay factor $f(L,K,\eta_0)$ to the choice of η_0 is shown in Fig. 7(b).

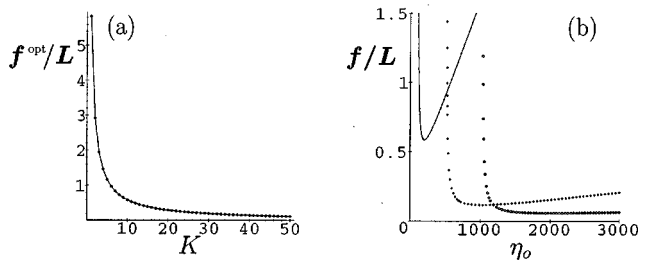

FIG. 7. (a) The optimal prefactor of the asymptotic error decay in normalized SCM scaled with L, $f(L,K,\eta_o^{\mathrm{opt}})/L$, is shown as a function of K. For all values of $L=1,10,20$ the plots collapse on the same curve $f/L=5.83/K$. (b) The prefactor $f(L,K,\eta_0)/L$ as a function of η_0 for student network size $K=10$ (solid line), $K=50$ (dots), and $K=100$ (circles). All values of L give the same result ($L=1,10,100$).

VI. SUMMARY AND DISCUSSION

Solving the dynamical equations numerically in unrealizable scenarios, where the student network does not have enough resources to imitate the teacher mapping, shows that the residual generalization error increases with the learning rate and is therefore minimal when the learning rate is annealed toward zero. The optimal fixed point of the dynamics is found analytically for large network size K. It shows a different behavior in the standard and normalized SCM architectures: In the normalized architecture, the overlap R^* between each student vector and the teacher vector it imitates decreases with L much faster than in the corresponding unnormalized architecture; in addition, contrary to the unnormalized case, each student vector is anticorrelated with all the other student vectors ($C^*<0$) and with the set of teacher vectors on which the other student vectors specialize ($S^*<0$). This configuration also turns out to give a much lower generalization error than that of the un-normalized architecture. In the un-normalized architecture, each student vector also keeps a positive correlation with the set of teacher vectors on which the other student vectors specialize to make up for the disparity in the output ranges. However, the student network is unable to make up completely for the output range differences.

Solving the asymptotic equations analytically for large system size K, one can analyze the approach of the system to the optimal fixed point. It turns out that the generalization error decays to the asymptotic residual error like $1/\alpha$ if the learning rate is annealed as η_0/α and $\eta_0>\eta_0^{\mathrm{crit}}$. We found that the critical learning rate η_o^{crit} is independent of L in both the standard and normalized SCM. The optimal decay of the generalization error is achieved at an optimal learning rate value η_0^{opt}, which shows only a weak dependence on L and K in standard SCM, and is proportional to K in the normalized SCM architecture. The optimal prefactor of the asymptotic error decay turns out to be proportional to the product LK in standard SCM, and is significantly smaller in normalized SCM, where it is proportional to the ratio L/K.

It would be interesting to extend the analysis of unrealizability to general two-layer neural networks in which the hidden-to-output parameters γ are adaptive, and not of fixed strength, as has been considered here.

ACKNOWLEDGMENTS

This work was supported by the EPSRC grant GR/L19232. We would like to thank Sara Solla for her contribution at early stages of this work.

APPENDIX A: THE FIXED POINT

The optimal fixed point is derived for large K. The following approximation is exact up to order $O(\epsilon^3)$. For the standard SCM architecture, in the small unrealizability case, the approximated optimal fixed point is of the form (some analytical results have been replaced by the numerical equivalent for brevity)

$$R^*=1-\frac{9}{2}\frac{L\epsilon^2}{-9+8\sqrt{3}}+\left(-\frac{162}{1369}L\sqrt{3}+\frac{1233}{1369}L\right)\epsilon^3,$$

$$S^*=2\frac{L\sqrt{3}\epsilon^3}{-9+8\sqrt{3}},$$

$$U^*=\epsilon+\left(1-\frac{2}{3}\sqrt{3}\right)\epsilon^2+\frac{1}{24}\frac{(48L\sqrt{3}-108L+760\sqrt{3}-1299)\epsilon^3}{-9+8\sqrt{3}},$$

$$Q^*=1+\left(-9\frac{L}{-9+8\sqrt{3}}+L\right)\epsilon^2+\left(-\frac{6448}{4107}L\sqrt{3}+\frac{5204}{1369}L\right)\epsilon^3,$$

$$C^*=L\epsilon^2+2\frac{L(16\sqrt{3}-25)\epsilon^3}{-9+8\sqrt{3}},$$

while in the strong unrealizability scenario,

$$R^*=1-\frac{9}{2}\frac{l\epsilon}{-9+8\sqrt{3}}+(0.2444l^2+0.6958l)\epsilon^2+(0.1672l^3+0.5275l^2-0.197l)\epsilon^3,$$

$$S^*=2\frac{\sqrt{3}l\epsilon^2}{-9+8\sqrt{3}}+(0.1778l-0.1619l^2)\epsilon^3,$$

$$U^*=\epsilon+\frac{1}{2}\frac{(-9l+4l\sqrt{3}-50+28\sqrt{3})\epsilon^2}{-9+8\sqrt{3}}+(0.148+0.467l+0.0824l^2)\epsilon^3,$$

$$Q^*=1+\left(-9\frac{l}{-9+8\sqrt{3}}+l\right)\epsilon-0.010\,59(-86.9l^2-103.0l)\epsilon^2+(0.092\,00l^3+1.275l^2-0.07l)\epsilon^3,$$

$$C^*=\epsilon l+\frac{l(-9l+4l\sqrt{3}-50+32\sqrt{3})\epsilon^2}{-9+8\sqrt{3}}+(-0.136l^2+0.2104l^3+0.678l)\epsilon^3.$$

In the normalized SCM architecture, for $L\ll K$ we find

$$R^*=1-6\frac{L(-3+2\sqrt{3})\epsilon}{-9+8\sqrt{3}}+(0.526L^2-1.336L)\epsilon^2,$$

$$S^*=-2\frac{L\sqrt{3}\epsilon^2}{-9+8\sqrt{3}}+0.000\,013\,16L(23\,930.0+80\,930.0L)\epsilon^3,$$

$$U^*=\epsilon-\frac{(-18L+14L\sqrt{3}-14\sqrt{3}+25)\epsilon^2}{-9+8\sqrt{3}}+(1.591L^2+0.001L+0.149)\epsilon^3,$$

$$Q^*=1-12\frac{L(-3+2\sqrt{3})\epsilon}{-9+8\sqrt{3}}-0.006\,869L(-200.0L+243.0)\epsilon^2+(-0.602L^3-0.533L^2-0.309L)\epsilon^3,$$

$$C^*=\frac{L(4\sqrt{3}-9)\epsilon^2}{-9+8\sqrt{3}}+0.000\,039\,48\,L(8110.0+22\,380.0L)\epsilon^3.$$

APPENDIX B: GENERALIZATION ERROR ASYMPTOTIC DECAY

Explicit expressions obtained for the generalization error $\varepsilon_g \equiv \langle \varepsilon(\mathbf{J},\boldsymbol{\xi})\rangle_{\{\boldsymbol{\xi}\}}$ are

$$\varepsilon_g=\Bigg[K\arcsin\left(\frac{Q}{1+Q}\right)+(K-1)K\arcsin\left(\frac{C}{1+Q}\right)+\frac{1}{6}(L+K)\pi-2(K-1)K\arcsin\left(\frac{S}{\sqrt{2+2Q}}\right)-2K\arcsin\left(\frac{R}{\sqrt{2+2Q}}\right)-2LK\arcsin\left(\frac{U}{\sqrt{2+2Q}}\right)\Bigg/\pi \tag{B1}$$

for the standard SCM architecture and

$$\varepsilon_g=\Bigg(\frac{\arcsin\left(\frac{Q}{1+Q}\right)}{K}+\frac{(K-1)\arcsin\left(\frac{C}{1+Q}\right)}{K}+\frac{1}{6}\frac{\pi}{L+K}-2\frac{(K-1)\arcsin\left(\frac{S}{\sqrt{2+2Q}}\right)}{L+K}-2\frac{\arcsin\left(\frac{R}{\sqrt{2+2Q}}\right)}{L+K}-2\frac{L\arcsin\left(\frac{U}{\sqrt{2+2Q}}\right)}{L+K}\Bigg)\Bigg/\pi \tag{B2}$$

for the normalized SCM network.

When the learning rate is annealed as $\eta=\eta_o/\alpha$ and $\eta_o > \eta_c$, then the generalization error decays proportionally to $1/\alpha$, as in Eq. (13), to the residual error E_0 corresponding to the optimal fixed point.

In standard SCM architecture, in the case $L \ll K$ we find the following form for the factors c_1 and c_2 in Eq. (13) for the asymptotic error decay:

$$c_1=-\frac{1}{6}\frac{L(9744\%1+1218-406\pi+227\sqrt{3}\pi-681\sqrt{3}-5448\sqrt{3}\%1)}{\pi^3(131\sqrt{3}-144)\epsilon}-\frac{1}{6}\frac{L(-432+393\sqrt{3}-131\sqrt{3}\pi+144\pi+3144\sqrt{3}\%1-3456\%1)}{\pi^3(131\sqrt{3}-144)\epsilon^2},$$

$$\%1=\arcsin\left(\frac{1}{6}\sqrt{3}\right),$$

$$c_2=-\frac{1}{6}L\Bigg[1152\arcsin\left(\frac{3}{4}\right)-236\sqrt{3}-96\sqrt{3}\pi+262\pi-6288\arcsin\left(\frac{1}{6}\sqrt{3}\right)+2304\sqrt{3}\arcsin\left(\frac{1}{6}\sqrt{3}\right)-210-1048\arcsin\left(\frac{3}{4}\right)\sqrt{3}\Bigg]\Bigg/[\pi^3(131\sqrt{3}-144)\epsilon],$$

while in the case $L=lK$ it is

$$c_1=-\frac{1}{6}\frac{(-1017\sqrt{3}+19\,296\%1-804\pi-8136\sqrt{3}\%1+2412+339\sqrt{3}\pi)l^2}{\pi^3(5\sqrt{3}+96)(-9+8\sqrt{3})\epsilon^2}+\Bigg(-\frac{1}{6}\frac{-29\,256\sqrt{3}\%1+6570-2190\pi+52\,560\%1-3657\sqrt{3}+1219\sqrt{3}\pi}{\pi^3(5\sqrt{3}+96)(-9+8\sqrt{3})\epsilon^2}-\frac{1}{6}\frac{17\,352\sqrt{3}\%1-723\sqrt{3}\pi-2232+744\pi-17\,856\%1+2169\sqrt{3}}{\pi^3(5\sqrt{3}+96)(-9+8\sqrt{3})\epsilon^3}\Bigg)l,$$

$$\%1 = \arcsin\left(\frac{1}{6}\sqrt{3}\right),$$

$$c_2 = -\frac{1}{6}\left[1446\pi - 34\,704\arcsin\left(\frac{1}{6}\sqrt{3}\right) + 5952\arcsin\left(\frac{3}{4}\right) - 1404\sqrt{3} - 1362 - 5784\arcsin\left(\frac{3}{4}\right)\sqrt{3} - 496\sqrt{3}\pi + 11\,904\sqrt{3}\arcsin\left(\frac{1}{6}\sqrt{3}\right)\right] l \Big/ \left[\pi^3(5\sqrt{3}+96)(-9+8\sqrt{3})\epsilon^2\right].$$

[1] G. Cybenko, Math. Control Signals and Syst. **2**, 303 (1988).

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

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

[4] *On-Line Learning in Neural Networks*, edited by D. Saad (Newton Institute, Cambridge University Press, Cambridge, 1998).

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

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

[7] A. West and D. Saad, Phys. Rev. E **56**, 3426 (1997).

[8] M. Biehl, P. Riegler, and C. Wohel, J. Phys. A **29**, 4767 (1996).

[9] T. K. Leen, B. Schottky, and D. Saad, *Advances in Neural Information Systems*, edited by M. I. Jordan, M. J. Kearns, and S. A. Solla (MIT Press, Cambridge, MA, 1998), Vol. 10, p. 301; T. K. Leen, B. Schottky, and D. Saad, Phys. Rev. E **59**, 985 (1999).

[10] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators* (Springer-Verlag, Berlin, 1982).

[11] C. Cohen-Tannoudji, *Quantum Mechanics* (Wiley, New York, 1977).