Noise, regularizers, and unrealizable scenarios in online learning from restricted training sets

Yuan-Sheng Xiong and David Saad
The Neural Computing Research Group, Aston University, Birmingham B4 7ET, United Kingdom
(Received 18 September 2000; revised manuscript received 5 February 2001; published 27 June 2001)

We study the dynamics of online learning in multilayer neural networks where training examples are sampled with repetition and where the number of examples scales with the number of network weights. The analysis is carried out using the dynamical replica method aimed at obtaining a closed set of coupled equations for a set of macroscopic variables from which both training and generalization errors can be calculated. We focus on scenarios whereby training examples are corrupted by additive Gaussian output noise and regularizers are introduced to improve the network performance. The dependence of the dynamics on the noise level, with and without regularizers, is examined, as well as that of the asymptotic values obtained for both training and generalization errors. We also demonstrate the ability of the method to approximate the learning dynamics in structurally unrealizable scenarios. The theoretical results show good agreement with those obtained from computer simulations.

DOI: 10.1103/PhysRevE.64.011919 PACS number(s): 87.10.+e, 02.50.−r, 05.90.+m

I. INTRODUCTION

Artificial neural networks provide an important tool for tackling nonlinear problems complementary to existing statistical methods (for review see [1,2]). The optimal selection of the network parameters on the basis of examples is termed learning and may be carried out in a variety of methods and techniques. The efficiency and success of the training process are in the heart of the method itself and play a significant part in determining the usefulness of artificial neural networks as a whole.

Significant effort has been invested over the years in optimizing the training methods as well as the choice of training parameters and regularization methods. These have been successfully used in practice, although most of the training methods used as well as the setting of the training coefficients are based on heuristic observations.

One of the most powerful and commonly used approaches to training large layered networks is that of online learning of continuous functions via gradient descent. Online learning refers to the iterative modification of the network parameters according to a predetermined training rule following successive presentations of single training examples, each representing a specific input vector and the corresponding output. This approach has been widely and successfully used for training large networks [3] and is arguably the most efficient technique for these tasks.

Significant progress has been made in analyzing the dynamics of supervised online learning in multilayer networks via methods of statistical physics (reviews can be found in [4,5]). Most of the analyses (e.g., [6–8]) concentrate on the case of infinite training sets, where training examples are sampled without repetition and in which there is no correlation between the network parameters and the examples presented at each training step. They successfully explain the various training phases and the emergence of generalization abilities but lack a vital aspect of the learning process, which may seem insignificant at first sight, assuming that the training set is large. However, the emerging correlations between successive training steps give rise to some of the most harmful effects in neural network training, such as overfitting, to which the above theory is oblivious.

A more realistic scenario is that where the number of training examples scales with the number of free parameters and the examples are sampled with repetition. This gives rise to correlations between the network parameters and the training examples, which clearly affect the learning process. One of the most significant aspects of having a fixed example set is the distinction between the two key performance measures: the *training error* measuring network performance with respect to the restricted training set, and the *test (generalization) error* calculated for all possible inputs sampled from the true distribution. The former may be monitored in practical training scenarios, while the latter (the minimization of which is the true aim of the learning process) can only be assessed up to some confidence level.

The analyses of learning from fixed example sets introduced so far [9–13] have mostly considered single layer systems, focusing on specific (usually simple) learning rules. In addition, most of these studies have been restricted to batch learning, where the network parameters are modified only after the complete example set has been presented.

The current paper builds upon a new approach we recently presented for the case of single layer networks [14], based on the dynamical replica method, which enables one to analyze a broad range of training rules and network configurations that can treat both online and batch learning scenarios. Preliminary analysis of noiseless, realizable, and unrealizable learning scenarios in multilayer networks were briefly described in [15]. Here, we extend the analysis to the case where training examples are corrupted by additive Gaussian output noise and examine the effect of regularization on the training dynamics. We also study the dependence of the asymptotic training and generalization errors on the size of the example set provided, with and without regularization. For brevity we will restrict the analysis to the case of online learning and not consider here the case of batch learning at all.

The paper is organized as follows. Section II provides the general framework and the theoretical basis for the analysis.

In Sec. III we present results obtained for the noiseless realizable case, followed by results obtained for an unrealizable training scenario where the model network is incapable of realizing the underlying rule due to structural limitations in Sec. IV. Section V looks at cases where training examples are corrupted by output Gaussian noise, while Sec. VI examines the impact of regularization on the network performance. We summarize our results and discuss the advantages and drawbacks of the current analysis in Sec. VII.

II. THE FRAMEWORK

We concentrate on information processing tasks in the form of maps from an N-dimensional input space $\boldsymbol{\xi}\in\mathbb{R}^N$ onto a scalar $\zeta\in\mathbb{R}$, realized through a parametrized function $\sigma(\boldsymbol{J},\boldsymbol{\xi})=\sum_{i=1}^{K}g(\boldsymbol{J}_i\cdot\boldsymbol{\xi})$. This function can be viewed as a two layer neural network, where g is the activation function of the hidden units taken here to be the error function $g(x)\equiv\mathrm{erf}(x/\sqrt{2})$; $\mathbf{J}\equiv\{\boldsymbol{J}_i\}_{1\leqslant i\leqslant K}$ is the set of input-to-hidden adaptive weights for the K hidden nodes, and the hidden-to-output weights are set to 1. The activation of hidden node i under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is denoted $x_i^\mu=\boldsymbol{J}_i\cdot\boldsymbol{\xi}^\mu$. This general configuration, usually referred to as the ''soft committee machine'' [7,8], encompasses most of the properties of general multilayer networks. Training examples are drawn from a finite set $\tilde{D}$ and are of the form $(\boldsymbol{\xi}^\mu,\zeta^\mu)$ where $\mu=1,2,\ldots,p$ and $p=\alpha N$. The components of the independently drawn input vectors $\boldsymbol{\xi}^\mu$ are uncorrelated random variables with zero mean and unit variance. The scenarios examined so far [15] focused on realizable and structurally unrealizable cases, where the corresponding output ζ^μ for the various examples is given by a deterministic teacher of an architecture similar to the student, except for a possible difference in the number M of hidden units: $\zeta^\mu=\sum_{n=1}^{M}g(\boldsymbol{B}_n\cdot\boldsymbol{\xi}^\mu)$, where $\boldsymbol{B}\equiv\{\boldsymbol{B}_n\}_{1\leqslant n\leqslant M}$ is the set of input-to-hidden adaptive weights for teacher-hidden nodes. In this paper we will also consider the case of noisy examples, where the teacher output is corrupted by additive Gaussian output noise, denoted as ρ^μ, the components of which are independently drawn uncorrelated random variables of zero mean and variance σ^2, corrupting the different examples. In this more general case the corresponding teacher output is of the form $\zeta^\mu=\sum_{n=1}^{M}g(\boldsymbol{B}_n\cdot\boldsymbol{\xi}^\mu)+\rho^\mu$. The activation of hidden node n under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is denoted $y_n^\mu=\boldsymbol{B}_n\cdot\boldsymbol{\xi}^\mu$. We will use indices $i,j,k,\ldots$ to refer to units in the student network and $n,m,\ldots$ for units in the teacher network. The contribution to the local field due to the noise variable will be denoted as z. Sums over the various indices will be considered from 1 to K or to M, respectively. The general framework [14,15] allows for the analysis of any training rule $\boldsymbol{\mathcal{G}}$ of the form

$$\boldsymbol{J}_j^{l+1}=\boldsymbol{J}_j^l+\frac{\eta}{N}\boldsymbol{\xi}(l)\mathcal{G}_j[\zeta^l,\sigma^l]-\frac{\gamma}{N}\boldsymbol{J}_j^l \tag{1}$$

where l represents the current time step in which a single example is randomly drawn from $\tilde{D}$ and invokes the parameter update. The last term on the right corresponds to a simple quadratic regularization term parametrized by γ, commonly used in regression tasks where examples are corrupted by noise, the usefulness of which will be examined in the current study. Here we concentrate on the most common online learning scenario for regression tasks, where the function $\boldsymbol{\mathcal{G}}$ together with the last term in Eq. (1) is the gradient with respect to the parameters $\boldsymbol{J}$ of the quadratic error measure (per example)

$$\begin{aligned}E(\boldsymbol{J},\boldsymbol{\xi})&=\frac{1}{2}[\sigma(\boldsymbol{J},\boldsymbol{\xi})-\zeta]^2+\frac{1}{2}\frac{\gamma}{\eta}\sum_{i=1}^{K}\boldsymbol{J}_i\cdot\boldsymbol{J}_i\\&=\frac{1}{2}\left[\sum_{i=1}^{K}g(x_i)-\sum_{n=1}^{M}g(y_n)-z\right]^2\\&\quad+\frac{1}{2}\frac{\gamma}{\eta}\sum_{i=1}^{K}\boldsymbol{J}_i\cdot\boldsymbol{J}_i,\end{aligned} \tag{2}$$

and $\boldsymbol{\mathcal{G}}$ is of the explicit form

$$\begin{aligned}\mathcal{G}_i(x_{j=1\ldots K},y_{n=1\ldots M},z)&=\sqrt{\frac{2}{\pi}}e^{-(1/2)x_i^2}\left[\sum_{j=1}^{K}g(x_j)\right.\\&\quad\left.-\sum_{n=1}^{M}g(y_n)-z\right].\end{aligned} \tag{3}$$

In the case of an infinite training set there is no correlation between the current example and those presented previously. As a consequence of that, no correlation between the student vectors and the examples is building up, and the joint probability distribution for the student and teacher node activations $\boldsymbol{x}$ and $\boldsymbol{y}$ (and the noise z) takes a multivariate Gaussian form. This is no longer the case here, when such correlations do exist and the joint probability distribution takes a more general form, which depends on the training patterns and changes dynamically throughout the learning process. In the case of corrupted training examples one should also consider the emerging correlations between the student vectors and the noise corrupting the examples. Due to the pivotal role played by this joint probability distribution it seems natural to define it as one of the macroscopic variables [14],

$$\begin{aligned}P(\mathbf{x},\mathbf{y},z,\boldsymbol{J})&=\frac{1}{p}\sum_{\mu}\prod_{i=1}^{K}\delta(x_i-\boldsymbol{J}_i\cdot\boldsymbol{\xi}^\mu)\prod_{n=1}^{M}\delta(y_n-\boldsymbol{B}_n\cdot\boldsymbol{\xi}^\mu)\\&\quad\times\delta(z-\rho^\mu),\end{aligned} \tag{4}$$

together with the overlaps $R_{in}(\boldsymbol{J})=\boldsymbol{J}_i\cdot\boldsymbol{B}_n$ (between student and teacher weight vectors) and $Q_{ik}(\boldsymbol{J})=\boldsymbol{J}_i\cdot\boldsymbol{J}_k$ (between student weight vectors). An additional macroscopic variable that is worthwhile mentioning, although it is invariant with respect to the learning dynamics, is $T_{nm}=\boldsymbol{B}_n\cdot\boldsymbol{B}_m$, representing the overlap between the various teacher weight vectors. Notice that most of the variables used are not observables but are based on the teacher-student model used. To simplify the calculation we will only examine here the case of orthogonal teacher vectors of unit length $T_{nm}=\delta_{mn}$; extending the results to the general teacher case is straightforward. For

convenience we will also introduce the vector $\mathbf{r}=(\mathbf{x},\mathbf{y},z)$ of dimensionality $K+M+1$, representing student and teacher local fields and the noise contribution.

The main motivation in choosing these macroscopic variables is that in the thermodynamic limit, $N\to\infty$, they are sufficient for calculating the two main performance measures: the generalization error, which corresponds to averaging $\tilde{E}(\boldsymbol{J},\boldsymbol{\xi})=(1/2)[\sigma(\boldsymbol{J},\boldsymbol{\xi})-\zeta]^2$ over the Gaussian input distribution [8]

$$E_g=\frac{1}{\pi}\left[\sum_{i,k}\sin^{-1}\frac{Q_{ik}}{\sqrt{1+Q_{ii}}\sqrt{1+Q_{kk}}}+\sum_{n,m}\sin^{-1}\frac{T_{nm}}{\sqrt{1+T_{nn}}\sqrt{1+T_{mm}}}-2\sum_{i,n}\sin^{-1}\frac{R_{in}}{\sqrt{1+Q_{ii}}\sqrt{1+T_{nn}}}\right]+\frac{1}{2}\sigma^2 \tag{5}$$

and the training error

$$E_t=\left\langle\frac{1}{2}\left[\sum_{i=1}^{K}g(x_i)-\sum_{n=1}^{M}g(y_n)-z\right]^2\right\rangle, \tag{6}$$

using the abbreviation $\langle f(\mathbf{r})\rangle=\int d\mathbf{r}\ P(\mathbf{r})f(\mathbf{r})$. The regularization term has been omitted in both measures as its contribution is limited to the learning dynamics and does not play any role in *measuring* the success of the training process.

To solve the dynamics, one straightforwardly derives a set of coupled differential equations [14,15] describing the evolution of the macroscopic variables in the limit $N\to\infty$,

$$\frac{d}{dt}Q=\eta(V+V^T)+\eta^2Z-2\gamma Q,$$

$$\frac{d}{dt}R=\eta W-\gamma R \tag{7}$$

and

$$\frac{\partial}{\partial t}P(\mathbf{r})=\frac{1}{\alpha}\int d\mathbf{x}'P(\mathbf{x}',\mathbf{y},z)\left\{\prod_i\delta[x_i-x_i'-\eta\mathcal{G}_i(\mathbf{x}',\mathbf{y},z)]-\prod_i\delta(x_i-x_i')\right\}-\sum_i\frac{\partial}{\partial x_i}\left[\eta\int d\mathbf{r}'\mathcal{G}_i(\mathbf{r}')\mathcal{A}(\mathbf{r};\mathbf{r}')-\gamma x_iP(\mathbf{r})\right]+\frac{\eta^2}{2}\sum_{i,k}Z_{ik}\frac{\partial^2P(\mathbf{r})}{\partial x_i\partial x_k}, \tag{8}$$

using a matrix representation for Q and R and defining the matrices

$$V=\langle\boldsymbol{\mathcal{G}}\mathbf{x}^T\rangle,\quad W=\langle\boldsymbol{\mathcal{G}}\mathbf{y}^T\rangle,\quad\text{and}\quad Z=\langle\boldsymbol{\mathcal{G}}\boldsymbol{\mathcal{G}}^T\rangle. \tag{9}$$

This set of equations cannot be closed in general; the difficulties originate in the Green's function

$$\mathcal{A}(\mathbf{r};\mathbf{r}')=\left\langle\left[\int d\boldsymbol{J}\ p_t(\boldsymbol{J}|QRP)\right]^{-1}\int d\boldsymbol{J}\ p_t(\boldsymbol{J}|QRP)\times\delta(\mathbf{x}-\boldsymbol{J}\cdot\boldsymbol{\xi})\,\delta(\mathbf{y}-\boldsymbol{B}\cdot\boldsymbol{\xi})\,\delta(z-\rho)(1-\delta_{\boldsymbol{\xi}\boldsymbol{\xi}'})(\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\times\delta(\mathbf{x}'-\boldsymbol{J}\cdot\boldsymbol{\xi}')\,\delta(\mathbf{y}'-\boldsymbol{B}\cdot\boldsymbol{\xi}')\,\delta(z'-\rho')\right\rangle_{\Xi} \tag{10}$$

where $p_t(\boldsymbol{J}|QRP)$ is the weight probability density conditioned on the values of the macroscopic observables $\{Q,R,P\}$ at time t (the microscopic measure in macroscopic subshells of the ensemble), and $\langle\cdot\rangle_\Xi$ represents averaging over all realizations of the training set. The Kronecker delta comes to filter out the case in which both vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ are identical ($\delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}=1$). We follow the derivation of [14] and employ the dynamical replica theory [16] to close Eqs. (7) and (8) by making two key assumptions.

(i) For $N\to\infty$ the macroscopic observables obey *closed* dynamic equations; we may thus assume equipartitioning of probability (or maximum entropy) in the macroscopic subshells,

$$p_t(\boldsymbol{J}|QRP)\sim\prod_{i,k}\delta[Q_{ik}-Q_{ik}(\boldsymbol{J})]\prod_{i,n}\delta[R_{in}-R_{in}(\boldsymbol{J})]\times\prod_{\mathbf{r}}\delta[P(\mathbf{r})-P(\mathbf{r}|\boldsymbol{J})]. \tag{11}$$

(ii) The macroscopic equations are self-averaging with respect to the specific realization of $\tilde{D}$; this allows for the averaging of the macroscopic variables over all training sets.

Both assumptions can be regarded as good approximations in general and will be validated against simulation results. They may become exact in some cases (e.g., Hebbian learning); we believe the second assumption to be exact in general. Following the calculation of [14] and employing the replica identity

$$\left\langle\frac{\int d\boldsymbol{J}W[\boldsymbol{J},v]G[\boldsymbol{J},v]}{\int d\boldsymbol{J}W[\boldsymbol{J},v]}\right\rangle_v=\lim_{n\to0}\int d\boldsymbol{J}^1\cdots d\boldsymbol{J}^n\left\langle G[\boldsymbol{J}^1,v]\prod_{\alpha=1}^{n}W[\boldsymbol{J}^\alpha,v]\right\rangle_v, \tag{12}$$

one obtains, under the further assumption of replica symmetry (for details see Appendix A and [14]), a closed form for Eq. (8).

$$\frac{\partial}{\partial t}P(\mathbf{r})=\frac{1}{\alpha}\int d\mathbf{x}'\ \ P(\mathbf{x}',\mathbf{y},z)\left\{\prod_i \delta[x_i-x_i'-\eta\mathcal{G}_i(\mathbf{x}',\mathbf{y},z)]-\prod_i \delta(x_i-x_i')\right\}-\sum_i \frac{\partial}{\partial x_i}[\{\eta[W\mathbf{y}+U(\mathbf{x}-R\mathbf{y})+X(Q-RR^T)\Phi(\mathbf{r})]_i-\gamma x_i\}P(\mathbf{r})]+\frac{\eta^2}{2}\sum_{ik} Z_{ik}\frac{\partial^2 P(\mathbf{r})}{\partial x_i\partial x_k}, \tag{13}$$

where we have introduced the matrices $B=(Q-q)^{-1}L$, $X=(V-WR^T)(Q-RR^T)^{-1}-U$, $LL^T=q-RR^T$, and $U=\langle\mathcal{G}\Phi^T\rangle$, and where

$$\Phi_i(\mathbf{r})=\frac{1}{P[\mathbf{x}|\mathbf{y},z]}\int D\mathbf{v}\ \langle[(Q-q)^{-1}(\mathbf{x}-\mathbf{x}')]_i\rangle_*\langle\delta(\mathbf{x}-\mathbf{x}')\rangle_* \tag{14}$$

using the notation $D\mathbf{v}\equiv\Pi_{i=1}^K 1/\sqrt{2\pi}e^{-(1/2)v_i^2}dv_i$ (used throughout the paper) and

$$\langle f(\mathbf{x},\mathbf{x}')\rangle_*=\frac{\int d\mathbf{x}'\ M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}f(\mathbf{x},\mathbf{x}')}{\int d\mathbf{x}'\ M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}. \tag{15}$$

The $K\times K$ matrix q and the function $M(\mathbf{x}',\mathbf{y},z)$ are derived from the replica symmetric calculation; the former is related to the cross-replica overlap matrix Q while the latter is an effective measure derived from the conjugate variable to the conditional probability $P(\mathbf{r})$. This closed set of equations can be solved iteratively by calculating q and $M(\mathbf{x}',\mathbf{y},z)$ at each step by solving a set of saddle-point equations (for details see Appendix A and [14]).

However, obtaining such a solution is extremely expensive computationally since a large set of nonlinear saddle-point equations should be solved at each time step to obtain a solution to Eqs. (7) and (13). The computation that was just possible in the case of single layer networks would come at a huge computational cost in the case of multilayer networks. We therefore resort to the large α approximation that was shown to provide a highly accurate approximated solution in the single layer case even for low α values (as low as $\alpha=0.5$), and enables one to obtain a simple form for Eq. (13) without solving a set of saddle-point equations at each time step,

$$\frac{\partial}{\partial t}P(\mathbf{r})=\frac{1}{\alpha}\int d\mathbf{x}'\ P(\mathbf{x}',\mathbf{y},z)\left\{\prod_i \delta[x_i-x_i'-\eta\mathcal{G}_i(\mathbf{x}',\mathbf{y},z)]-\prod_i \delta(x_i-x_i')\right\}-\sum_i \frac{\partial}{\partial x_i}[\{\eta\Gamma_i(\mathbf{r})-\gamma x_i\}P(\mathbf{r})]+\frac{\eta^2}{2}\sum_{i,k} Z_{ik}\frac{\partial^2 P(\mathbf{r})}{\partial x_i\partial x_k}, \tag{16}$$

where

$$\Gamma_i(\mathbf{r})=\left[\begin{pmatrix}V\\W\end{pmatrix}^T\begin{pmatrix}Q & R\\R^T & T\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{x}\\\mathbf{y}\end{pmatrix}-\left\{\langle\mathcal{G}\overline{\mathbf{x}}^T\rangle(\mathbf{y},z)-WR^T\times(Q-RR^T)^{-1}\{\overline{\mathbf{x}}(\mathbf{y},z)-R\mathbf{y}\}\right]_i,$$

in which $\overline{\mathbf{x}}(\mathbf{y},z)=\int d\mathbf{x}\,\mathbf{x}P[\mathbf{x}|\mathbf{y},z]$. The large α approximation is particularly suitable to the model examined here since the main features of learning in multilayer networks, such as the breaking of internal symmetries and the asymptotic convergence, can be observed at sensible time scales only for relatively high α values.

To solve the dynamical equations (7) and (16) numerically one should represent the continuous probability distribution using a discrete model. Representing the probability distribution by discrete bins, the method used in the single layer case, can be employed, in principle, here also to provide accurate approximated solutions. However, obtaining solutions in the case of multilayer neural networks comes at a high computational cost, especially as the network size increases; here one should monitor numerically the evolution of a general multivariate probability distribution and solve numerically the differential equations (16) and (7). Using the methods used in the single layer case would require monitoring tens of thousands of variables already in the case of $K=M=2$. We therefore look for a parametric approximated representation of the probability distribution and have considered two different possibilities: a mixture of multivariate Gaussian distributions (described briefly in Appendix B) and the local Gaussian approximation (derived in Appendix C), where the conditional probability $P[\mathbf{x}|\mathbf{y},z]$ is replaced by a Gaussian one with $\mathbf{y}$ and z-dependent mean $\overline{\mathbf{x}}(\mathbf{y},z)$ and covariance matrix $\{\Sigma_{ij}(\mathbf{y},z)\}$. The first representation can, in principle, model any given probability distribution to the desired accuracy, given a sufficient number of Gaussian bases, and provides simple expressions for Eqs. (7) as most of the integrals can be carried out analytically; however, the solution of Eq. (16) requires a continuous update of the various parameters in the representation used, which can be done, in principle but may be computationally difficult due to the variability in sensitivity of the various parameters. The second representation is more limited and assumes a Gaussian distribution with respect to $\mathbf{x}$ for each given $(\mathbf{y},z)$ vector; however, it can be solved analytically and is therefore easier to handle as long as the approximation used is satisfactory. Here we present solutions based on the second representation

$$P[\mathbf{x}|\mathbf{y},z]=\frac{1}{\sqrt{(2\pi)^K|\Sigma(\mathbf{y},z)|}}\times\exp\left\{-\frac{1}{2}[\mathbf{x}-\overline{\mathbf{x}}(\mathbf{y},z)]^T\Sigma^{-1}(\mathbf{y},z)[\mathbf{x}-\overline{\mathbf{x}}(\mathbf{y},z)]\right\}. \tag{17}$$

Using the representation (17) in Eq. (16) results (after some tedious algebra) in the following dynamical equations for $\overline{\mathbf{x}}(\mathbf{y},z)$ and for $\Sigma_{ij}(\mathbf{y},z)$:

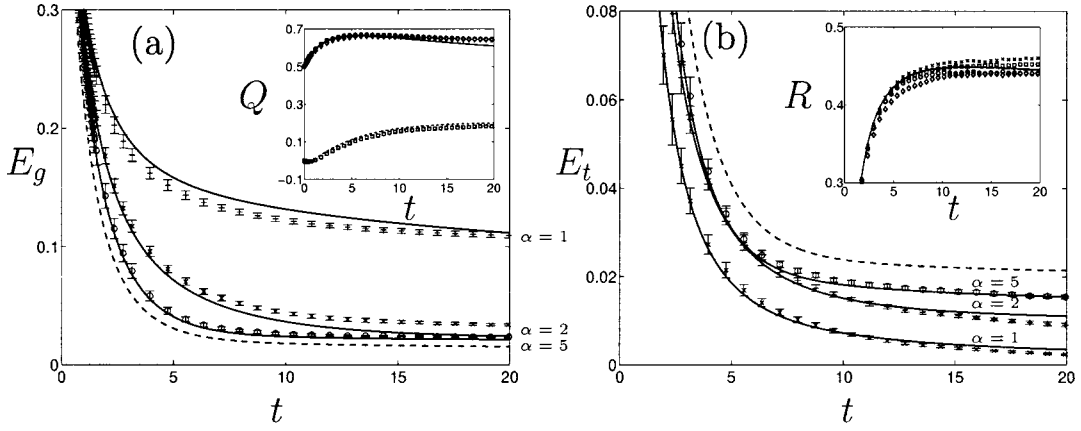

FIG. 1. The evolution of the generalization (a) and training errors (b) as a function of time for $\alpha=1,2,5$. Solid lines represent analytical results while simulation experiments are presented by symbols; both were initialized in a similar manner. Simulation results were averaged over 20 trials; both mean values and error bars are presented. Theoretical results for the training and generalization errors in the case of $\alpha=5$ are presented in (a) and (b), respectively, for comparison (dashed line). The insets in (a) and (b) show the evolution of the various overlaps (Q and R, respectively, different symbols represent the various overlaps) in the case of $\alpha=5$, comparing theoretical results and simulations (mean values). The upper Q lines and symbols correspond to the diagonal values, while the lower lines correspond to the off-diagonal overlaps.

$$\frac{d}{dt}\overline{x_i}(\mathbf{y},z)=\frac{\eta}{\alpha}\overline{\mathcal{G}_i}(\mathbf{y},z)+\eta[W\mathbf{y}+Y(\overline{\mathbf{x}}\{\mathbf{y},z)-R\mathbf{y}\}]_i \quad (18)$$

$$\begin{aligned}\frac{d}{dt}\Sigma_{ik}(\mathbf{y},z)=\frac{1}{\alpha}[\,&\eta\{\overline{V}_{ik}(\mathbf{y},z)+\overline{V}_{ki}(\mathbf{y},z)-\overline{\mathcal{G}_i}(\mathbf{y},z)\overline{x_k}(\mathbf{y},z)\\&-\overline{\mathcal{G}_k}(\mathbf{y},z)\overline{x_i}(\mathbf{y},z)\}+\eta^2\overline{Z}_{ik}(\mathbf{y},z)]\\&+\eta[\{S\Sigma(\mathbf{y},z)\}_{ik}+\{S\Sigma(\mathbf{y},z)\}_{ki}]+\eta^2 Z_{ik},\end{aligned}$$

with the matrices $S=(V-WR^T)(Q-RR^T)^{-1}$ and $Y=(V-\langle\overline{\mathcal{G}\mathbf{x}^T}\rangle)(Q-RR^T)^{-1}$, and with $\overline{\mathcal{G}_i}(\mathbf{y},z)=\int d\mathbf{x}\,\mathcal{G}_i(\mathbf{r})P[\mathbf{x}|\mathbf{y},z]$, $\overline{V}_{ik}(\mathbf{y},z)=\int d\mathbf{x}\,\mathcal{G}_i(\mathbf{r})x_k P[\mathbf{x}|\mathbf{y},z]$, and $\overline{Z}_{ik}(\mathbf{y},z)=\int d\mathbf{x}\,\mathcal{G}_i(\mathbf{r})\mathcal{G}_k(\mathbf{r})P[\mathbf{x}|\mathbf{y},z]$.

Equations (18) and (7) are solved numerically from appropriate initial conditions, providing the theoretical prediction for the evolution of the macroscopic variables, and both generalization [Eq. (5)] and training errors. The latter takes the expression

$$\begin{aligned}E_t=&\frac{1}{2}\int d\mathbf{y}\,dz\,P(\mathbf{y},z)\int d\mathbf{x}\,P[\mathbf{x}|\mathbf{y},z]\\&\times\left[\sum_n g(y_n)+z-\sum_i g(x_i)\right]^2\\=&\frac{1}{2}\int d\mathbf{y}\,dz\,P(\mathbf{y},z)\left[\sum_{ln} g(y_l)g(y_n)-2\sum_{in} g(\theta_i)g(y_n)\right.\\&\left.+\sum_{ij} J_2(i,j)\right]\end{aligned} \quad (19)$$

with $\theta_i=\overline{x_i}/\sqrt{1+\Sigma_{ii}}$ and

$$J_2(i,j)=\int dx\,g(\sqrt{\Sigma_{ii}}x+\overline{x_i})g\left(\frac{\Sigma_{ij}x+\sqrt{\Sigma_{ii}}\;\overline{x_j}}{\sqrt{\Sigma_{ii}(1+\Sigma_{jj})-\Sigma_{ij}^2}}\right). \quad (20)$$

III. THE NOISELESS REALIZABLE CASE

Equations (18) and (7) form the basis to our numerical solutions in the various learning scenarios. Firstly, we validate the analysis in the noiseless realizable scenario by comparing the results to those obtained from numerical simulations. In this section we do not consider the case of noise (i.e, $\sigma=0$) or regularization (i.e., $\gamma=0$).

For brevity we will restrict our experiments in this section to the case of $K=M=2$ and orthogonal unit teacher vectors $T_{mn}=\delta_{mn}$ (the Kronecker tensor). To facilitate the comparison between the analytical solutions and the simulation results we introduce fixed initial conditions, breaking the inherent symmetries in the system macroscopically. This is essential for investigating the learning dynamics beyond the symmetric phase as it may take a prohibitively long time to escape the symmetric plateau otherwise, as in the case of infinite training sets [17]. We use the following initial conditions for both theory and simulations: $Q_{11}^0=Q_{22}^0=0.5$, $Q_{12}^0=Q_{21}^0=0$, $R_{11}^0=0.001$, $R_{22}^0=R_{12}^0=R_{21}^0=0$. The initial joint probability $P(\mathbf{r})$ is assumed Gaussian, with the corresponding parameters. The initial conditions for Eq. (18) are $\Sigma(\mathbf{y},z)|_{t=0}=Q^0-R^0(R^0)^T$ and $\overline{\mathbf{x}}(\mathbf{y},z)|_{t=0}=R^0\mathbf{y}$; the learning rate used is $\eta=0.5$. We first investigate the accuracy of our approximation in the case of low α values, where the accuracy of the approximation is expected to be the worst due to the (large α) approximation used. However, in these cases we cannot observe the breaking of the symmetric phase for computationally feasible system sizes. We will therefore concentrate on the prediction accuracy within the symmetric

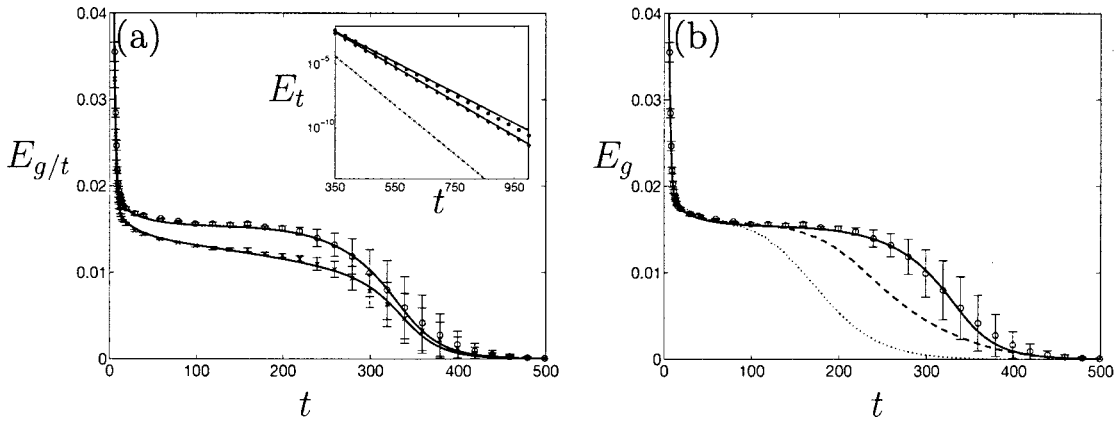

FIG. 2. The evolution of the training and generalization errors in comparison to those obtained from simulations for the case of $K=M=2$, $\alpha=20$. (a) The theoretical values for the training (lower) and generalization (higher) errors are represented by the solid lines; the training error simulation results for system size of $N=5000$ are represented by symbols (mean values and error bars for 10 trials). The inset shows the semilog plot of E_g (solid and circles) and E_t (dashed and crosses) for $t=350,\ldots,500$; theoretical results for the decay of $E_g(\alpha=\infty)$ are also shown for comparison (dashed dotted line). The regression values obtained for the various curves are $E_g(\alpha=20)=60.88\exp[-2.759(1)\times10^{-2}t]$ (theory), $E_g(\alpha=20)=151.34\exp[-2.9(1)\times10^{-2}t]$ (simulations), $E_t(\alpha=20)=181.08\exp[-3.116(1)\times10^{-2}t]$ (theory), $E_t(\alpha=20)=97.65\exp[-3.1(1)\times10^{-2}t]$ (simulations), and $E_g(\alpha=\infty)=224.51\exp[-4.4144(1)\times10^{-2}t]$. Digits in parenthesis indicate the regression error in the last digit; regression has been carried out on the mean values. (b) Finite size effects by plotting simulation results for the generalization error for systems of size $N=1000$ (dashed) and $N=500$ (dotted) lines.

phase, where all vectors of the student system emulate the various vectors in the teacher system with equal success. Figure 1 shows the numerical solutions of the analytical equations in comparison to simulation results obtained for various α values ($\alpha=1,2,5$). The theoretical values are represented by solid lines and the simulation results by symbols. Simulation results were obtained for a similar system of size $N=500$, initialized at random, restricting the overlap values to the ones used for the analytical solutions. Simulation results were averaged over 20 trials and the figure shows both mean values and error bars for all cases ($\alpha=1,2,5$). Figure 1(a) shows the generalization errors as a functions of time, with the training error for the case of $\alpha=5$ added for comparison (dashed line); in all of our experiments, each unit of time corresponds to the presentation of αN examples selected at random. Figure 1(b) focuses on the evolution of the training errors, where the generalization error ($\alpha=5$) is added for comparison. The insets show the evolution of the various overlaps for the case of $\alpha=5$ in comparison to the results obtained from simulations [Q values in Fig. 1(a) and R values in Fig. 1(b)]. We see that the results obtained are in good agreement with the simulations even at these low α values. It is only fair to mention that the discrepancy between the theoretical results and simulations will increase at later times due to the accumulating errors.

However, the main interest of the neural networks community, in the case of multilayer networks, is in the symmetry breaking process whereby specific vectors of the student system specialize, each learning to imitate a specific teacher vector. In addition, one would also like to gain insight into the convergence phase and its dependence on the value of α. In Fig. 2(a) we show the evolution of both the generalization and training errors for the case of $\alpha=20$, which is sufficiently high for observing the symmetry breaking phenomena; the initial conditions and learning rate used are similar to those of Fig. 1. The theoretical values for the training (lower) and generalization (higher) errors are represented by the solid lines; the simulation results for system size of $N=5000$ are represented by symbols (mean values and error bars) and were averaged over 10 trials. In Fig. 2(b) we examine the finite size effects, comparing the theoretical results obtained for the generalization error to the simulation results for $N=500$, 1000, and 5000. Simulation results for lower N values are represented by dashed ($N=1000$) and dotted ($N=500$) lines and were averaged over 30 trials. For brevity, only mean results are presented for smaller N values; errorbars are generally similar to those of $N=5000$.

To examine the decay rate of the training and generalization errors in the asymptotic regime we plotted in the inset of Fig. 1(a) the decay of both errors on a logarithmic scale with respect to the number of training iterations for $t=350,\ldots,1000$; theoretical results for the decay of $E_g(\alpha=\infty)$ are also shown for comparison (dashed dotted line). All three graphs decay exponentially to their asymptotic values although the prefactors and the decay rates seem to differ and probably depend on α. The decay rate for the finite α case is clearly slower than that of the $\alpha\to\infty$ case as expected.

IV. STRUCTURAL UNREALIZABILITY

While interesting academically, realizable training scenarios are very rare in practical online learning applications. We therefore turn to the arguably more interesting case of structural unrealizability, where the number of student vectors is smaller than that of the teacher vectors. It would be particularly important to examine this case due to the approximations taken along the way; we should verify the validity of the theoretical results in this case, which may result in probability distributions quite different from those obtained in the realizable scenario. Also in this section we do

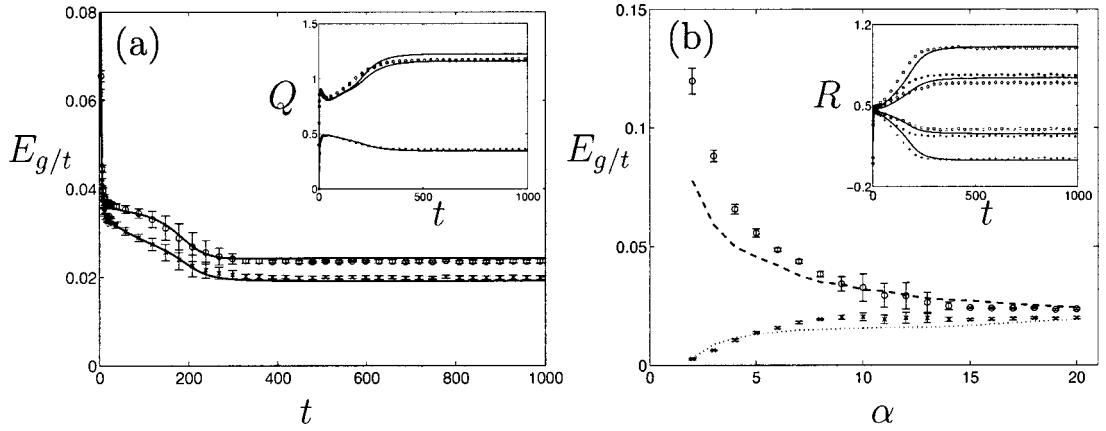

FIG. 3. An unrealizable scenario; a system comprising two student vectors $K=2$ is trained on examples provided by a system comprising three orthonormal teacher vectors $M=3$. The initial conditions used are $R_{11}^0=0.05$, $Q_{11}^0=0.4$, $Q_{22}^0=0.6$, with all other overlaps set to zero, the learning rate is $\eta=1$ and the system size used for simulations is $N=1000$. Simulation results were averaged over 10 trials, presenting both mean values and error bars. (a) The dependence of generalization and training errors on time with $\alpha=20$; the inset shows the corresponding Q values. Lines represent theoretical values and symbols represent simulation results, upper lines correspond to diagonal Q values and the lower lines to off-diagonal values. The inset of (b) shows the corresponding R values, the upper curves represent student vectors that emulate specific teacher vectors while the lower curves represent cross overlaps between student vectors and teacher vectors emulated by other student vectors; the middle curves represent overlaps between student vectors and the teacher vector, which is not emulated by any of the student vectors in particular. (b) The asymptotic ($t=1000$) values of the generalization (dashed line and circles) and training errors (dotted lines and crosses) for different α values, comparing theoretical (lines) and simulation (symbols) results.

not consider the case of noise (i.e, $\sigma=0$) or regularization ($\gamma=0$).

We demonstrate the efficacy of our approach in the case of a two node system ($K=2$) trained on examples provided by a three node teacher system ($M=3$), all orthogonal and of unit length. The equations used are similar to those of the realizable case, Eqs. (18) and (7), but with a modified $M=3$ value. The initial conditions used are $R_{11}^0=0.05$, $Q_{11}^0=0.4$, $Q_{22}^0=0.6$, with all other overlaps set to zero; the learning rate is $\eta=1$, the number of examples is αN, where $\alpha=20$, and the system size used in simulations is $N=1000$. The results presented in Fig. 3(a) show a good agreement between theory and simulations and a qualitatively similar result to the infinite training set case. The insets in Figs. 3(a) and 3(b) show the corresponding Q and R values.

Figure 3(b) describes the asymptotic values of generalization and training errors for different α values, monitored at $t=1000$, once the systems had stabilized (notice that the equilibration of the system at $t=1000$ is not guaranteed due to the spin-glass dynamics). The learning rate used is $\eta=1$. It is easy to see that the agreement between theory and simulations is generally good but deteriorates as α decreases. It is difficult to find the exact manner in which both generalization and training errors decay to their asymptotic values [i.e., $E_g(\alpha=\infty)=E_t(\alpha=\infty)$] as a function of α due to its sensitivity to the inherent numerical errors.

V. ADDITIVE OUTPUT NOISE

Finite α training scenarios are of particular interest in cases where the training data is corrupted by some type of noise, being the most common case in practical training scenarios. This is a particularly important aspect of the current study as it enables one to assess existing methods for alleviating the effect of noise on the model's generalization performance. Similar scenarios have already been examined in the single layer case [18] and discrete learning rules; we will focus here on the multilayer case representing a continuous mapping trained by gradient descent.

The equations used are similar to those of the realizable case, Eqs. (18) and (7), except for the reactivation of the noise term. No regularization is used in the current section setting γ to zero.

In Fig. 4 we demonstrate the effect of additive output noise. We see that the effect is mainly in the length of the symmetric phase and in the convergence to a suboptimal asymptotic solution (a constant learning rate of $\eta=1$ is used). We examine the case of $K=M=2$, using initial conditions of the form: $Q_{ii}^0=0.5$; $Q_{\forall i\neq j}^0$ and R_{in}^0 are set to values samples uniformly $U[0,1/\sqrt{N}]$ according to the system size N used in simulations. The number of examples used is αN with $\alpha=20$ and the noise level (standard deviation of the Gaussian distribution) is $\sigma=0.2$. The system size used in simulations is $N=1000$. Figure 4(a) shows the evolution of the generalization (higher) and training errors as a function of time, while Fig. 4(b) and the inset show the evolution of the order parameters Q and R respectively. The upper Q and R curves correspond to the diagonal overlaps while the lower curves represent the off-diagonal parameters. We see that the analysis is in general consistent with results obtained from simulations, although inconsistencies occur around the transition point between the symmetric and asymptotic regimes.

Next we examine the efficacy of our approximations as the noise level changes shown in Fig. 5(a). We plotted the evolution of the generalization and training (inset) errors as a function of time, comparing them to simulation results averaged over 10 trials each. Initial condition, learning rate, and

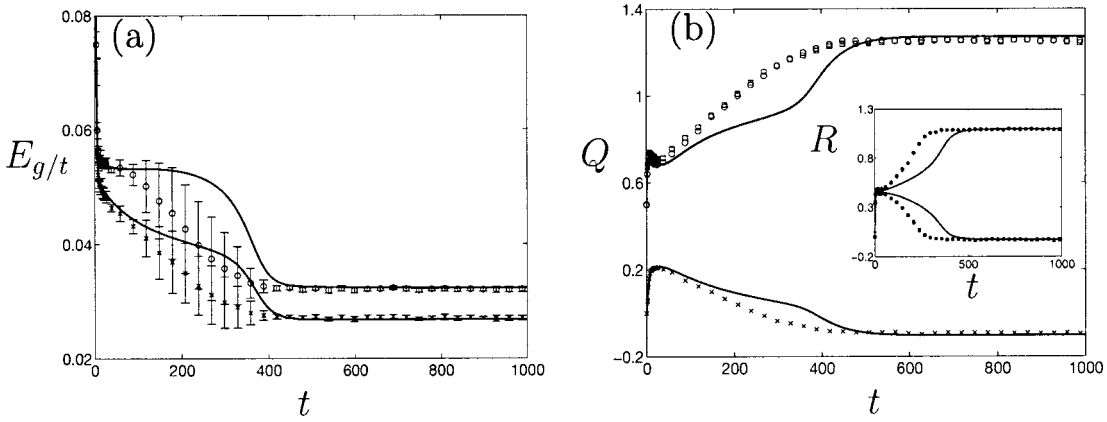

FIG. 4. The effect of additive Gaussian output noise on the evolution of the training and generalization errors and on the macroscopic variables in the case of $K=M=2$. The initial conditions used for the student-vector length are $Q_{ii}^0=0.5$; $Q_{\forall i\neq j}^0$ and R_{in}^0 are set to values sampled uniformly in the range $[0,1/\sqrt{N}]$, corresponding to the system size N used in simulations. The learning rate is $\eta=1$, the examples ratio is $\alpha=20$ and the noise level $\sigma=0.2$. The system size used in simulations is $N=1000$ and the results were averaged over ten trials each.

the ratio of examples α are similar to those of the previous figure. We see that our approximation becomes less accurate as the noise level increases, especially around the breaking of the symmetric phase. This is probably due to the deteriorating accuracy of the local Gaussian approximation as the noise level increases. For low α values, when the inherent system symmetries do not break, our method provides a good approximation to the results obtained in simulations, as shown in Fig. 5(b) for the case of $\alpha=12$. In both cases, the theoretical asymptotic results are in good agreement with the simulations.

In principle, one could obtain from the analytical solutions an estimate to the improvement in performance that can be obtained from employing the early stopping technique as well as an estimate for the optimal point in which early stopping should be applied. However, the disagreement between the results obtained analytically and the simulations is mainly around the point in which the internal symmetries break (and mainly at high noise levels), making such an estimate inaccurate. We assume that employing a refined representation of the conditional probability distribution would enable one to make accurate estimations of this type.

In Fig. 6(a) we examine the dependence of the asymptotic values (measured at $t=1000$, once the system has stabilized) of both generalization and training errors on the value of α, having a fixed noise level $\sigma=0.3$ (in the inset $\sigma=0.1$). We see that our approximation provides a good description for large α values, becoming less accurate for low values as one might expect. In addition, we see that as expected, the gap between training and generalization errors for a given α increases with the noise level. The dependence of generalization error on α for different noise levels $\sigma=0.1$ (lower curve) and 0.3 (higher curve) is shown in Fig. 6(b). As expected, the difference between the asymptotic values decreases as α grows.

One should notice that the asymptotic training and generalization errors do not converge (as α increases) to the optimal value of $\sigma^2/2$; this is due to the fixed learning rate used rather than the decaying rate required for optimal asymptotic results.

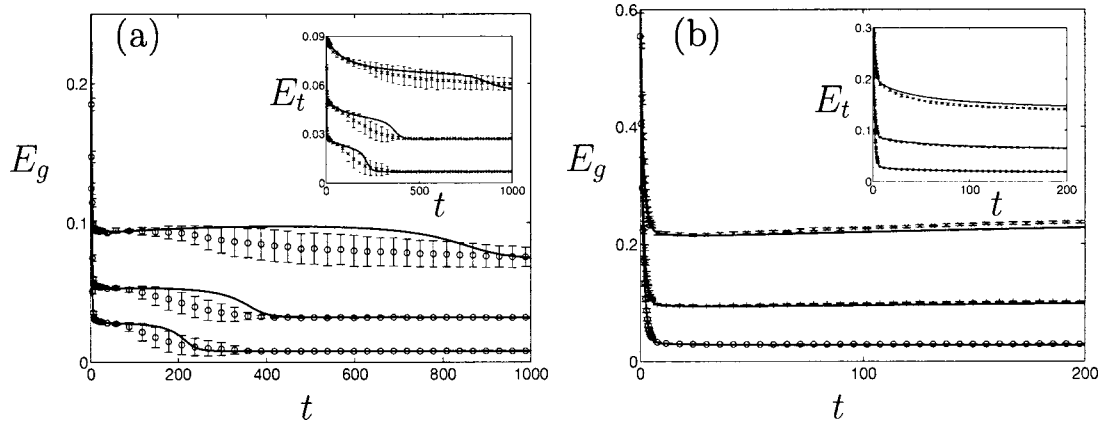

FIG. 5. Additive Gaussian output noise in the case of $K=M=2$; the learning rate used and the initial conditions are as in Fig. 4. The system used for simulations is of size $N=1000$ and results were averaged over ten trials for each point. (a) The dependence of generalization and training (inset) errors on time for different noise levels $\sigma=0.1,0.2,0.3$ (from the bottom up) in the case of $\alpha=20$. (b) The same for the case of $\alpha=12$ and $\sigma=0.1,0.3,0.5$.

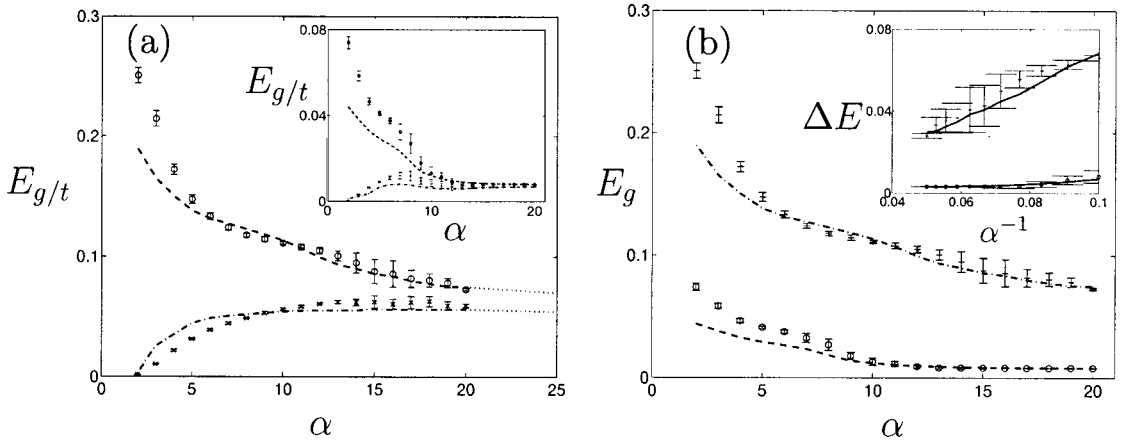

FIG. 6. The asymptotic values of generalization and training errors (measured at $t=1000$) for different α values with a fixed additive Gaussian output noise level; the case considered, the learning rate used, and the initial conditions are as in Fig. 4. The system used for simulations is of size $N=1000$ and results were averaged over ten trials for each point. (a) Generalization (higher curve) and training (lower curve) errors for $\sigma=0.3$, where the dotted line represents the asymptotic value of both training and generalization errors as α becomes infinite and to which both errors converge. The inset shows for comparison the corresponding generalization (higher curve) and training (lower curve) errors for $\sigma=0.1$. (b) The dependence of generalization error on α for different noise levels, $\sigma=0.1$ (lower curve) and 0.3 (higher curve). The inset shows the corresponding dependence of $\Delta E_g=E_g(\alpha)-E_g(\infty)$ on α^{-1} for α values high enough for the system to escape the symmetric phase; the noise levels used are $\sigma=0.1$ (lower curve) and 0.3 (higher curve).

To examine the decay of the generalization error to its asymptotic value we plotted in the inset of Fig. 6(b) the dependence of $\Delta E_g=E_g(\alpha)-E_g(\infty)$ on α^{-1} for α values high enough for the system to escape the symmetric phase. The decay seems to be proportional to α^{-1} [e.g., the power values obtained from regression in the case of $\sigma=0.1$ are 1.0(1) and 0.9(3) from the theoretical results and simulations, respectively] and depends linearly on σ^2; dividing the residual error for the noise levels presented in the figure $\sigma=0.3$ (higher curve) and $\sigma=0.1$ (lower curve), gives approximately a constant value of 9.

To examine the dependence of both training and generalization errors on the noise level σ, we plotted in Fig. 7 the asymptotic values of generalization and training errors (measured once the system has stabilized) for different additive Gaussian output noise levels with fixed $\alpha=20$. Using conventional regression methods we find the following dependence of E_g and E_t on the noise level σ: $E_g \simeq 1.06\sigma^{2.14(1)}$ (theory) and $E_g \simeq 0.94\sigma^{2.082(8)}$ (simulations), and $E_t \simeq 0.63\sigma^{1.957(5)}$ (theory) and $E_t \simeq 0.65\,\sigma^{1.968(3)}$ (simulations). This is in agreement with our assumption of a quadratic σ dependence.

VI. REGULARIZATION

One of the main problems facing practitioners in the field of neural networks is the improvement of generalization ability in trained networks, especially when noisy training data are provided. This is typically done by imposing constraints on the space of solutions (for a general introduction to the problem and the methods used see [2]), reflecting our prior belief in the type of solution we are looking for. One of the most common mechanisms for adding such constraints is the introduction of a quadratic regularization term, as in the last term on the right of Eq. (2), which leads to a modification of the dynamical training equation (1).

Most of the analyses linking the regularization to the noise level corrupting the data are based on single layer systems or on linearizing the system in the asymptotic regime. Ideally, we would have liked to exploit the current analysis to obtain an analytical expression for the optimal regularization term to be used for data corrupted by additive Gaussian noise of a certain variance. However, the current framework, based on Eqs. (18) and (7), is solved numerically making it difficult to provide the desired link analytically. We therefore demonstrate the effect of regularization through numerical

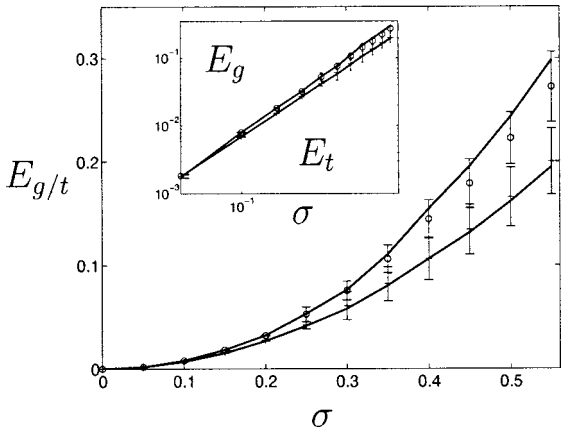

FIG. 7. The asymptotic values of generalization and training errors (measured at $t=1000$) for different additive Gaussian output noise levels σ with a fixed $\alpha=20$; the case considered, the learning rate used, and the initial conditions are as in Fig. 4. The system used for simulations is of size $N=1000$ and results were averaged over ten trials for each point. Using simple regression techniques we find that the asymptotic values of both E_g and E_t depend approximately on σ^2 (for both theory and simulations). The inset shows the log-log plot of the asymptotic values of E_g and E_t versus α.

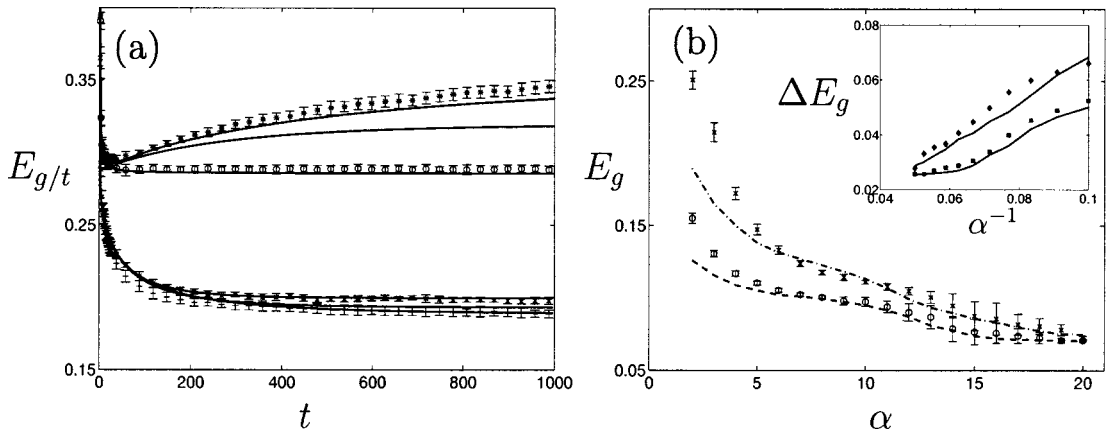

FIG. 8. Training with regularizers. The case considered, the learning rate, the system size used for simulation, and the initial conditions are as in Fig. 4. (a) The dependence of generalization and training errors versus time for different regularizer (γ) values, where generalization errors (the upper three) are for $\gamma=0.01$, $\gamma=0.001$, $\gamma=0.0$ from the bottom to the top and training errors (the lower three) are from the top to the bottom; symbols show the simulation results for $\gamma=0.01$ and $\gamma=0.0$ (simulations for the case of $\gamma=0.001$ have been omitted for brevity). The noise level used is $\sigma=0.6$ and $\alpha=12$. (b) The asymptotic values of the generalization error (measured at $t=1000$) for different α values and fixed noise level $\sigma=0.3$. The upper curve represents the case of no regularization while the lower curve is for $\gamma=0.005$. The inset shows the corresponding dependence of $\Delta E=E_g(\alpha)-E_g(\infty)$ on α^{-1}, where the simulation results are shown by symbols with no error bars for brevity.

solutions obtained in specific cases.

Firstly, to examine the effect of regularization on both the training and generalization errors in the symmetric plateau, we present the training scenario where $K=M=2$, $\alpha=12$ and where training examples are corrupted by additive Gaussian output noise of standard deviation $\sigma=0.6$. Simulations were carried out using a system of size $N=1000$, and simulation results were averaged over 10 trials. Figure 8(a) shows the evolution of the generalization and training errors for different γ values, where generalization errors are for $\gamma=0.01$, $\gamma=0.001$, and $\gamma=0.0$ from the bottom up, while training errors from the top down. Lines represent the theoretical results while symbols represent simulation results. It is clear that while regularization has little effect on the training error in that phase it clearly reduces the generalization error. It should be noted that, although the main significance of regularization is in the asymptotic regime, its effect on the symmetric phase is also important as many practical training sessions are effectively terminated at some suboptimal symmetric plateau.

To examine the effect of regularization asymptotically we plotted in Fig. 8(b) the dependence of the asymptotic generalization error on α, measured at $t=1000$ for fixed $\sigma=0.3$ and a regularization value of $\gamma=0.005$ (lower curve); the upper curve represents values obtained with no regularization.

One should note that in the case of infinite training sets it has been shown that there is no advantage in using a quadratic regularization term with a constant prefactor in the asymptotic regime [19], and in fact, introducing such a term always results in a higher asymptotic (in training steps t) generalization error. Therefore, there must be a value of α, for a given noise level and regularization prefactor, above which the introduction of a quadratic regularization term is detrimental to the asymptotic performance. This critical value of α can be determined, in principle, for a specific scenario using our analysis; however, in practice the numerical inaccuracies reduce the reliability of such a prediction.

The inset of Fig. 8(b) shows the dependence of $\Delta E_g=E_g(\alpha)-E_g(\infty)$ on α^{-1}, for sufficiently large α such that the system escapes the symmetric plateaus. The theoretical results are in agreement with the simulations, indicating (approximately) a $1/\alpha$ decay in the generalization error to the asymptotic values (the regression power figures obtained numerically from both theory and simulations are generally around the decay power of 1, but have significant error bars).

VII. DISCUSSION

We presented a theoretical framework for the analysis of online learning scenarios in multilayer networks, where the training examples are sampled with repetition from a fixed example set. The framework, being based on rather solid theoretical tools, provides a controlled and unbiased description of the learning dynamics. It is then used for studying realizable and unrealizable scenarios as well as scenarios whereby the data is corrupted by additive Gaussian output noise and where regularizers are employed for improving the network's generalization performance.

To obtain the set of equations representing the network dynamics we employ the dynamical replica method. This is the only *fundamental* approximation used in this analysis, comprising three assumptions: (a) Equipartitioning of the probability (or maximum entropy) in the macroscopic subshells as $N\to\infty$, (b) The macroscopic equations are self-averaging with respect to the specific realization of the data, (c) The replica symmetry ansatz. These assumptions can be regarded as good approximations in general and may become exact in some cases. On the basis of simulation results we believe the self-averaging assumption to hold in general while the equipartitioning and the replica symmetry assumptions may break down in extreme cases such as very low α

values (lower than 1), high over-realizability etc., when the error surface becomes rugged or suffers from multiple minima.

Employing the dynamical replica theory one obtains Eqs. (7) and (13), which are the main result of the analysis and provide a closed set of equations that could be solved at huge computational cost.

To be able to produce results in many different scenarios and under different training conditions we employed two further approximations. These are considered merely for simplifying the numerics and should not be regarded as essential ingredients of the calculation. They have both been employed because they provide a reliable approximation in the relevant parameter range and would have been abandoned otherwise. The first of the two is the high α approximation. This has been shown to provide an excellent approximation even for very low α values [14] and is therefore expected to be highly accurate in the cases we concentrate on here, as most of the interesting phenomena in training multilayer networks appear only when α is sufficiently high (e.g., the breaking of suboptimal symmetric solutions and the asymptotic convergence). This approximation is likely to break down only for very low α (lower than 1), which is outside of the relevant range of the current study.

The second approximation used is the method employed to model the conditional probability distribution of the teacher and student local fields, $P[\mathbf{x}|\mathbf{y}]$; such a model is essential for obtaining numerical solutions to continuous functions in general and may take various forms (e.g., discrete bins, a mixture of Gaussians, etc.). In the current analysis we employed the local Gaussian representation to facilitate the computation as it has been shown to provide a good approximation already for low α values [14]. Also here, the approximation may break down for low α values, specific training rules, high over-realizability, etc., where the field distribution becomes more complex. Of all the approximations used, this is likely to be the most fragile and it may be therefore desirable, in some cases, to replace it by a more accurate model such as the mixture of Gaussians we proposed in Appendix B.

The results obtained are in good agreement with the simulations and support heuristic methods used by practitioners, such as early stopping and regularization. The framework successfully provides a description of the dynamics of both training and generalization errors (and of the various overlaps), some understanding of the link between the value of α and the breaking of internal symmetries, certain asymptotic scaling laws, etc. Unfortunately, due to the complexity of dynamical equations and the computational difficulties we have experienced in solving them, our ability to provide analytical solutions is limited. These are highly desirable for deriving analytical relations between the training and generalization conditions in noisy scenarios, in both the symmetric phase and asymptotically, and to make a quantitative link between the noise level and the optimal regularization to be used.

Other questions that are of interest are to do with the length of the symmetric phase and its dependence on the ratio α, the learning rate, the architecture chosen, and the initial conditions. In addition, it would be desirable to define optimal training parameters and learning rules in a principled manner, similarly to the studies carried out in the case of infinite training sets [20–24].

It is fair to say that it is difficult to see how these objectives could be achieved in the current framework; further simplifications may be required for successful exploitation of the analysis. Nevertheless, the current paper prepares the basis for future studies in this area.

ACKNOWLEDGMENTS

D.S. and Y.-S.X. acknowledge support by EPSRC Grant No. GR/L52093 and the British Council grant: British-German Academic Research Collaboration Programme Project No. 1037. We would like to thank Ton Coolen for his contribution to this work as well as for careful reading of the manuscript.

APPENDIX A: REPLICA CALCULATION OF THE GREEN FUNCTION

The main objective of this appendix is to provide a rough derivation of the Green's function $\mathcal{A}[\ \dots\ ]$ using the dynamic replica theory and following [14] and [15], from which we obtain the macroscopic dynamical equations (13) in an explicit form. We first carry out the disorder averages leading to an effective single-spin problem. The integrations are carried out using saddle-point methods for the replicated order parameters at each time step employing the replica symmetry (RS) ansatz.

1. Disorder averaging

Following the dynamic replica theory in [16], we write the Green function as

$$\mathcal{A}(\mathbf{r};\mathbf{r}')=\lim_{n\to 0}\lim_{N\to\infty}\Big\langle\Big\langle\int\prod_{i\alpha}dJ_i^\alpha\,p_t(\boldsymbol{J}^\alpha|QRP)\times\prod_i\delta(x_i-\boldsymbol{J}_i^1\cdot\boldsymbol{\xi})\prod_n\delta(y_n-\boldsymbol{B}_n\cdot\boldsymbol{\xi})\,\delta(z-\rho)\times(\boldsymbol{\xi}\cdot\boldsymbol{\xi}')(1-\delta_{\xi\xi'})\prod_i\delta(x_i'-\boldsymbol{J}_i^1\cdot\boldsymbol{\xi}')\times\prod_n\delta(y_n'-\boldsymbol{B}_n\cdot\boldsymbol{\xi}')\,\delta(z'-\rho')\Big\rangle_{\tilde{D}\tilde{D}'}\Big\rangle_{\Xi},\quad\text{(A1)}$$

noting that the averages over the data sets already include the noise distribution as well, and that $\langle\cdot\rangle_\Xi$ represents averaging over all realizations of the data set. Using the definition of $P(\mathbf{r};\boldsymbol{J})$ and the integral representations for the δ distributions involving $P(\mathbf{r})$, we obtain

$$\mathcal{A}(\mathbf{r};\mathbf{r}')=\lim_{n\to 0}\lim_{N\to\infty}\int\prod_{\alpha,\mathbf{r}''}d\hat{P}^{\alpha}(\mathbf{r}'')\prod_{\alpha i}dJ_i^{\alpha}$$
$$\times\prod_{\alpha ik}\delta(Q_{ik}-\boldsymbol{J}_i^{\alpha}\cdot\boldsymbol{J}_k^{\alpha})\prod_{\alpha in}\delta(R_{in}-\boldsymbol{J}_i^{\alpha}\cdot\boldsymbol{B}_n)$$
$$\times e^{iN\int d\mathbf{r}''\,\hat{P}(\mathbf{r}'')P_t(\mathbf{r}'')}\int\frac{d\hat{\mathbf{r}}\,d\hat{\mathbf{r}}'}{(2\pi)^{2(K+M+1)}}e^{i\hat{\mathbf{r}}\cdot\mathbf{r}}$$
$$\times\left\langle\frac{1}{p^2}\sum_{\mu\neq\nu}(\boldsymbol{\xi}^{\mu}\cdot\boldsymbol{\xi}^{\nu})\right.$$
$$\times\exp\left[-\frac{i}{\alpha}\sum_{\alpha\lambda}\hat{P}(\boldsymbol{J}^{\alpha}\cdot\boldsymbol{\xi}^{\lambda},\boldsymbol{B}\cdot\boldsymbol{\xi}^{\lambda},\rho^{\lambda})\right]$$
$$\times\exp\left[-i\sum_i\hat{x}_i\boldsymbol{J}_i^1\cdot\boldsymbol{\xi}^{\mu}-i\sum_n\hat{y}_n\boldsymbol{B}_n\cdot\boldsymbol{\xi}^{\mu}-i\hat{z}\rho^{\mu}\right.$$
$$\left.\left.-i\sum_i\hat{x}_i'\boldsymbol{J}_i^1\cdot\boldsymbol{\xi}^{\nu}-i\sum_n\hat{y}_n'\boldsymbol{B}_n\cdot\boldsymbol{\xi}^{\nu}-i\hat{z}'\rho^{\nu}\right]\right\rangle_{\Xi} \tag{A2}$$

with the conjugate function $\hat{P}(\mathbf{r})$.

We first define some relevant functions to facilitate the calculation

$$\mathcal{D}(\hat{\mathbf{r}};\boldsymbol{\xi},\rho)=\exp\left[-\frac{i}{\alpha}\sum_{\alpha}\hat{P}(\boldsymbol{J}^{\alpha}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi},\rho)-i\sum_i\hat{x}_i\boldsymbol{J}_i^1\cdot\boldsymbol{\xi}\right.$$
$$\left.-i\sum_n\hat{y}_n\boldsymbol{B}_n\cdot\boldsymbol{\xi}-i\hat{z}\rho\right],$$
$$\mathcal{D}(\hat{\mathbf{r}})=\langle\mathcal{D}(\hat{\mathbf{r}};\boldsymbol{\xi},\rho)\rangle_{\tilde{D}},$$
$$\mathcal{E}_j(\hat{\mathbf{r}})=\langle\xi_j\mathcal{D}(\hat{\mathbf{r}};\boldsymbol{\xi},\rho)\rangle_{\tilde{D}}=\left\langle\frac{\partial\mathcal{D}(\hat{\mathbf{r}};\boldsymbol{\xi},\rho)}{\partial\xi_j}\right\rangle_{\tilde{D}}. \tag{A3}$$

By using the permutation invariance of the integrations and summations with the pattern labels, we evaluate the training-set average of the expression for $\mathcal{A}[\ \dots\ ]$ in Eq. (A2) in the thermodynamic limit

$$\left\langle\frac{1}{p^2}\sum_{\mu\neq\nu}(\boldsymbol{\xi}^{\mu}\cdot\boldsymbol{\xi}^{\mu})e^{\cdots}\right\rangle_{\Xi}$$
$$=\left\langle\frac{p-1}{p}(\boldsymbol{\xi}^1\cdot\boldsymbol{\xi}^2)e^{\cdots}\right\rangle_{\Xi}$$
$$=\frac{p-1}{p}\sum_j^N\left\langle\exp\left[-\frac{i}{\alpha}\sum_{\alpha}\hat{P}(\boldsymbol{J}^{\alpha}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi},\rho)\right]\right\rangle_{\tilde{D}}^{p-2}$$
$$\times\left\langle\exp\left[-\frac{i}{\alpha}\sum_{\alpha}\hat{P}(\boldsymbol{J}^{\alpha}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi},\rho)-i\sum_i\hat{x}_i\boldsymbol{J}_i^1\cdot\boldsymbol{\xi}\right.\right.$$
$$\left.\left.-i\sum_n\hat{y}_n\boldsymbol{B}_n\cdot\boldsymbol{\xi}-i\hat{z}\rho\right]\right\rangle_{\tilde{D}}$$
$$\times\left\langle\exp\left[-\frac{i}{\alpha}\sum_{\alpha}\hat{P}(\boldsymbol{J}^{\alpha}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi},\rho)-i\sum_i\hat{x}_i'\boldsymbol{J}_i^1\cdot\boldsymbol{\xi}\right.\right.$$
$$\left.\left.-i\sum_n\hat{y}_n'\boldsymbol{B}_n\cdot\boldsymbol{\xi}-i\hat{z}'\rho\right]\right\rangle_{\tilde{D}}$$
$$=\exp\{p\ln[\mathcal{D}(0,0)]\}\ \frac{\mathcal{L}(\hat{\mathbf{r}};\hat{\mathbf{r}}')}{\mathcal{D}^2(\mathbf{0})} \tag{A4}$$

with $\mathcal{L}(\hat{\mathbf{r}};\hat{\mathbf{r}}')=\Sigma_j^N\mathcal{E}_j(\hat{\mathbf{r}})\mathcal{E}_j(\hat{\mathbf{r}}')$. We can then write the Green function in an integral form dominated by saddle points,

$$\mathcal{A}(\mathbf{r};\mathbf{r}')=\int\frac{d\hat{\mathbf{r}}d\hat{\mathbf{r}}'}{(2\pi)^{2(K+M+1)}}\exp[i(\hat{\mathbf{r}}\cdot\mathbf{r}+\hat{\mathbf{r}}'\cdot\mathbf{r}')]$$
$$\times\lim_{n\to 0}\lim_{N\to\infty}\int d\mathbf{q}\,d\mathbf{Q}\,d\hat{\mathbf{q}}\,d\hat{\mathbf{Q}}\,d\hat{\mathbf{R}}\int\prod_{\alpha,\mathbf{r}''}d\hat{P}^{\alpha}(\mathbf{r}'')$$
$$\times\exp(N\Psi[\mathbf{q},\mathbf{Q},\hat{\mathbf{q}},\hat{\mathbf{Q}},\hat{\mathbf{R}},\{\hat{P}\}])\frac{\mathcal{L}(\hat{\mathbf{r}};\hat{\mathbf{r}}')}{\mathcal{D}^2(\mathbf{0})} \tag{A5}$$

with

$$\Psi[\ \dots\ ]=\frac{1}{2}\left[\sum_{\alpha}\mathrm{Tr}(\hat{Q}^{\alpha}Q^{\alpha})-2i\sum_{\alpha}(\mathrm{Tr}\,\hat{R}^{\alpha}R^{\alpha})\right.$$
$$\left.+\sum_{\alpha\beta}\mathrm{Tr}(\hat{q}^{\alpha\beta}q^{\alpha\beta})\right]+i\sum_{\alpha}\int d\mathbf{r}\,\hat{P}^{\alpha}(\mathbf{r})P(\mathbf{r})$$
$$+\alpha\ln\mathcal{D}(\mathbf{0})+\lim_{N\to\infty}\frac{1}{N}\ln\int\prod_{\alpha i}dJ_i^{\alpha}$$
$$\times\exp\left(-\frac{1}{2}\left[\sum_{\alpha ik}\hat{Q}_{ik}^{\alpha}\boldsymbol{J}_i^{\alpha}\cdot\boldsymbol{J}_k^{\alpha}-2i\sum_{\alpha in}\hat{R}_{in}^{\alpha}\boldsymbol{J}_i^{\alpha}\cdot\boldsymbol{B}_n\right.\right.$$
$$\left.\left.+\sum_{\alpha\beta ik}\hat{q}_{ik}^{\alpha\beta}\boldsymbol{J}_i^{\alpha}\cdot\boldsymbol{J}_k^{\beta}\right]\right). \tag{A6}$$

Similarly, the joint probability distribution can be obtained,

$$P(\mathbf{r})=\int\frac{d\hat{\mathbf{r}}}{(2\pi)^{K+M+1}}e^{i\hat{\mathbf{r}}\cdot\mathbf{r}}$$
$$\times\lim_{n\to 0}\lim_{N\to\infty}\int d\mathbf{q}\,d\mathbf{Q}\,d\hat{\mathbf{q}}\,d\hat{\mathbf{Q}}\,d\hat{\mathbf{R}}\int\prod_{\alpha,\mathbf{r}''}d\hat{P}^{\alpha}(\mathbf{r}'')$$
$$\times\exp(N\Psi[\mathbf{q},\mathbf{Q},\hat{\mathbf{q}},\hat{\mathbf{Q}},\hat{\mathbf{R}},\{\hat{P}\}])\frac{\mathcal{D}(\hat{\mathbf{r}})}{\mathcal{D}(\mathbf{0})}. \tag{A7}$$

Using the normalized expression for $P(\mathbf{r})$ we see that no overall prefactors in the expression of $\mathcal{A}[\mathbf{r};\mathbf{r}']$ or $P(\mathbf{r})$ are to be taken into account. Then we have

$$\mathcal{A}(\mathbf{r};\mathbf{r}')=\int\frac{d\hat{\mathbf{r}}\,d\hat{\mathbf{r}}'}{(2\pi)^{2(K+M+1)}}\exp[i(\hat{\mathbf{r}}\cdot\mathbf{r}+\hat{\mathbf{r}}'\cdot\mathbf{r}')]\frac{\mathcal{L}(\hat{\mathbf{r}};\hat{\mathbf{r}}')}{\mathcal{D}^2(\mathbf{0})}, \tag{A8}$$

with the order parameter values defined at the saddle point, and

$$P(\mathbf{r})=\int \frac{d\hat{\mathbf{r}}}{(2\pi)^{K+M+1}} e^{i\hat{\mathbf{r}}\cdot\mathbf{r}} \frac{\mathcal{D}(\hat{\mathbf{r}})}{\mathcal{D}(\mathbf{0})}. \tag{A9}$$

First, we calculate the explicit expression for $\mathcal{D}(\mathbf{0})$.

$$\begin{aligned}
\mathcal{D}(\mathbf{0})=&\int \prod_{\alpha i} \frac{d\hat{x}_i^\alpha\, dx_i^\alpha}{2\pi} \prod_n \frac{d\hat{y}_n\, dy_n}{2\pi} \frac{d\hat{z}\, dz}{2\pi} \exp\Bigg[i\sum_{\alpha i} \hat{x}_i^\alpha x_i^\alpha \\
&+ i\sum_n \hat{y}_n y_n + i\hat{z}z - \frac{i}{\alpha}\sum_\alpha \hat{P}(\mathbf{x}^\alpha,\mathbf{y},z)\Bigg] \\
&\times \int D\boldsymbol{\xi} \int D(\rho/\sigma) \exp\Bigg[-i\sum_j^N \Bigg(\sum_{\alpha i} \hat{x}_i^\alpha J_{ij}^\alpha \\
&+ \sum_n \hat{y}_n B_{nj}\Bigg)\xi_j - i\hat{z}\rho\Bigg] \\
=&\int \prod_{\alpha i} \frac{d\hat{x}_i^\alpha\, dx_i^\alpha}{2\pi} \prod_n \frac{d\hat{y}_n\, dy_n}{2\pi} D(z/\sigma) \exp\Bigg[i\sum_{\alpha i} \hat{x}_i^\alpha x_i^\alpha \\
&+ i\sum_n \hat{y}_n y_n - \frac{i}{\alpha}\sum_\alpha \hat{P}(\mathbf{x}^\alpha,\mathbf{y},z)\Bigg] \\
&\times \exp\Bigg(-\frac{1}{2}\Bigg[\sum_{\alpha\beta ik} q_{ik}^{\alpha\beta} \hat{x}_i^\alpha \hat{x}_k^\beta \\
&+ 2\sum_{\alpha i n} R_{in} \hat{x}_i^\alpha \hat{y}_n + \sum_n \hat{y}_n^2\Bigg]\Bigg),
\end{aligned} \tag{A10}$$

where Dv is the Gaussian measure as defined before, and where the spin-glass order parameters and the overlaps R_{in}^α between the student and teacher weights are defined as

$$q_{ik}^{\alpha\beta}=\boldsymbol{J}_i^\alpha\cdot\boldsymbol{J}_k^\beta, \qquad R_{in}^\alpha=\boldsymbol{J}_i^\alpha\cdot\boldsymbol{B}_n. \tag{A11}$$

We now employ the RS ansatz: $q_{ik}^{\alpha\beta}=\{Q_{ik}(\alpha=\beta), q_{ik}(\alpha\neq\beta)\}$, $R_{in}^\alpha=R_{in}$, and $\hat{P}^\alpha(\mathbf{r})=i\chi(\mathbf{r})$. Then $\mathcal{D}(\mathbf{0})$ can be further simplified

$$\begin{aligned}
\mathcal{D}(\mathbf{0})=&\int \prod_{\alpha i} \frac{d\hat{x}_i^\alpha\, dx_i^\alpha}{2\pi} \prod_n \frac{d\hat{y}_n\, dy_n}{2\pi} D(z/\sigma) \exp\Bigg[i\sum_{\alpha i} \hat{x}_i^\alpha x_i^\alpha \\
&+ i\sum_n \hat{y}_n y_n + \frac{1}{\alpha}\sum_\alpha \chi(\mathbf{x}^\alpha,\mathbf{y},z)\Bigg] \\
&\times \exp\Bigg\{ -\frac{1}{2}\Bigg[\sum_{\alpha ik} (Q_{ik}-q_{ik}) \hat{x}_i^\alpha \hat{x}_k^\alpha \\
&+ \sum_{ik} q_{ik}\Bigg(\sum_\alpha \hat{x}_i^\alpha\Bigg)\Bigg(\sum_\alpha \hat{x}_k^\alpha\Bigg) + 2\sum_{\alpha i n} R_{in} \hat{x}_i^\alpha \hat{y}_n \\
&+ \sum_n \hat{y}_n^2\Bigg]\Bigg\} \\
=&\int \prod_{\alpha i} \frac{d\hat{x}_i^\alpha\, dx_i^\alpha}{2\pi} \prod_n \frac{dy_n}{\sqrt{2\pi}} D(z/\sigma) \\
&\times \exp\Bigg\{ i\sum_{\alpha i} \hat{x}_i^\alpha [x_i^\alpha - (R\mathbf{y})_i] + \frac{1}{\alpha}\sum_\alpha \chi(\mathbf{x}^\alpha,\mathbf{y},z)\Bigg\} \\
&\times \exp\Bigg\{ -\frac{1}{2}\Bigg[\sum_{\alpha ik} \hat{x}_i^\alpha (Q-q)_{ik} \hat{x}_k^\alpha \\
&+ \sum_{ik} \Bigg(\sum_\alpha \hat{x}_i^\alpha\Bigg)(q-RR^T)_{ik}\Bigg(\sum_\alpha \hat{x}_k^\alpha\Bigg) + \sum_n y_n^2\Bigg]\Bigg\} \\
=&\frac{1}{\sqrt{|q-RR^T|}} \int D\mathbf{y}\, D(z/\sigma) \int \prod_i \frac{dv_i}{\sqrt{2\pi}} \\
&\times \exp\Bigg[-\frac{1}{2}\mathbf{v}^T (q-RR^T)^{-1}\mathbf{v}\Bigg] \int \prod_{\alpha i} \frac{d\hat{x}_i^\alpha\, dx_i^\alpha}{2\pi} \\
&\times \exp\Bigg\{ -\frac{1}{2}\sum_{\alpha ik} \hat{x}_i^\alpha (Q-q)_{ik} \hat{x}_k^\alpha \\
&+ i\sum_{\alpha i} \hat{x}_i^\alpha [x_i^\alpha + u_i - (R\mathbf{y})_i] + \frac{1}{\alpha}\sum_\alpha \chi(\mathbf{x}^\alpha,\mathbf{y},z)\Bigg\} \\
=&\frac{1}{\sqrt{|q-RR^T|}} \int D\mathbf{y} D(z/\sigma) \int \prod_i \frac{dv_i}{\sqrt{2\pi}} \\
&\times \exp\Bigg[-\frac{1}{2}\mathbf{v}^T (q-RR^T)^{-1}\mathbf{v}\Bigg] \\
&\times \Bigg[\frac{1}{\sqrt{|Q-q|}} \int \prod_i \frac{dx_i}{\sqrt{2\pi}} \exp\Bigg\{ \frac{1}{\alpha}\chi(\mathbf{r}) \\
&- \frac{1}{2}(\mathbf{x}-R\mathbf{y}-\mathbf{v})^T (Q-q)^{-1} (\mathbf{x}-R\mathbf{y}-\mathbf{v})\Bigg\}\Bigg]^n \\
=&\int D\mathbf{y}\, D(z/\sigma) \int D\mathbf{v} \Bigg[\int d\mathbf{x}\ \ \Omega(\mathbf{r};\mathbf{v})\Bigg]^n
\end{aligned} \tag{A12}$$

with

$$\begin{aligned}
\Omega(\mathbf{r};\mathbf{v})=&\frac{1}{\sqrt{|Q-q|(2\pi)^K}} \exp\Bigg[\frac{1}{\alpha}\chi(\mathbf{r}) - \frac{1}{2}(\mathbf{x}-R\mathbf{y}-L\mathbf{v})^T \\
&\times (Q-q)^{-1} (\mathbf{x}-R\mathbf{y}-L\mathbf{v})\Bigg],
\end{aligned} \tag{A13}$$

$LL^T=q-RR^T$, and $B=(Q-q)^{-1}L$.

Second, the integration on $\boldsymbol{J}_i^\alpha$ can be carried out and the corresponding expression can be evaluated explicitly using the RS ansatz (in the limit $n\to 0$)

$$\lim_{N\to\infty}\frac{1}{N}\ln\int\prod_{\alpha i}dJ_i^\alpha\exp\Bigg(-\frac{1}{2}\Bigg[\sum_{\alpha ik}\hat{Q}_{ik}^\alpha J_i^\alpha\cdot J_i^\alpha - 2i\sum_{\alpha in}\hat{R}_{in}^\alpha J_i^\alpha\cdot B_n+\sum_{\alpha\beta ik}\hat{q}_{ik}^{\alpha\beta}J_i^\alpha\cdot J_k^\beta\Bigg]\Bigg)$$

$$\sim-\frac{1}{2}\{(n-1)\ln|\hat{Q}-\hat{q}|+\ln|\hat{Q}+(n-1)\hat{q}| + n\,\mathrm{Tr}[\hat{R}^T(\hat{Q}-\hat{q})^{-1}\hat{R}]+O(n^2)\}. \tag{A14}$$

Together with the rest of the terms in $\Psi[\ \dots\ ]$, we have

$$\lim_{n\to0}\frac{\Psi}{n}=\frac{1}{2}\{\mathrm{Tr}(\hat{Q}Q)-2i\,\mathrm{Tr}(\hat{R}R)-\mathrm{Tr}(\hat{q}q)-\ln|\hat{Q}-\hat{q}| - \mathrm{Tr}[(\hat{Q}-\hat{q})^{-1}\hat{q}]-\mathrm{Tr}[\hat{R}^T(\hat{Q}-\hat{q})^{-1}\hat{R}]\} - \int d\mathbf{r}\,\chi(\mathbf{r})P(\mathbf{r}) + \alpha\int D\mathbf{y}\,D(z/\sigma)\int D\mathbf{v}\ln\Big[\int d\mathbf{x}\ \ \Omega(\mathbf{r};\mathbf{v})\Big]. \tag{A15}$$

2. Derivation of the RS saddle-point equations

We then work out the saddle-point equations with respect to $\hat{Q},\hat{R},\hat{q}$

$$\hat{r}=\hat{Q}-\hat{q}=(Q-q)^{-1},\qquad \hat{R}=-i(Q-q)^{-1}R,$$

$$\hat{q}=-(Q-q)^{-1}(q-RR^T)(Q-q)^{-1}, \tag{A16}$$

which allow us to eliminate most variational parameters. Then the Ψ can be simplified as

$$\Psi=\frac{1}{2}\,\mathrm{Tr}[(Q-RR^T)(Q-q)^{-1}]+\frac{1}{2}\ln|Q-q| - \int d\mathbf{r}\,\chi(\mathbf{r})P(\mathbf{r}) + \alpha\int D\mathbf{y}\,D(z/\sigma)\int D\mathbf{v}\ln\Big[\int d\mathbf{x}\ \ \Omega(\mathbf{x},\mathbf{y},z;\mathbf{v})\Big]. \tag{A17}$$

The saddle-point equation for $\chi(\mathbf{r})$ results in

$$P(\mathbf{r})=\frac{e^{-(1/2)\mathbf{y}^2}}{\sqrt{(2\pi)^M}}\frac{e^{-z^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\int D\mathbf{v}\left[\frac{\Omega(\mathbf{r};\mathbf{v})}{\int d\mathbf{x}'\ \ \Omega(\mathbf{x}',\mathbf{y},z;\mathbf{v})}\right] \equiv P(\mathbf{y},z)P[\mathbf{x}|\mathbf{y},z] \tag{A18}$$

where we have defined $P(\mathbf{y},z)$ and conditional probability $P[\mathbf{x}|\mathbf{y},z]$, respectively, as

$$P(\mathbf{y},z)=\frac{e^{-(1/2)\mathbf{y}^2}}{\sqrt{(2\pi)^M}}\frac{e^{-z^2/2\sigma^2}}{\sqrt{2\pi}\sigma},$$

$$P[\mathbf{x}|\mathbf{y},z]=\int D\mathbf{v}\left[\frac{M(\mathbf{r})e^{\mathbf{x}^T B\mathbf{v}}}{\int d\mathbf{x}'\,M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right] \tag{A19}$$

with

$$M(\mathbf{r})=\exp\left[\frac{1}{\alpha}\chi(\mathbf{r})-\frac{1}{2}(\mathbf{x}-R\mathbf{y})^T(Q-q)^{-1}(\mathbf{x}-R\mathbf{y})\right]. \tag{A20}$$

3. Explicit expression for the Green function

In order to work out the explicit expression for the Green function (A8) we need to calculate the function $\mathcal{L}(\hat{\mathbf{r}};\hat{\mathbf{r}}')$. First we take the $n\to0$ limit of $\mathcal{D}(\hat{\mathbf{r}},\boldsymbol{\xi},\rho)$ [Eq. (A3)] and simplify the result using the saddle-point equation (A18)

$$\mathcal{D}(\hat{\mathbf{r}},\boldsymbol{\xi},\rho)=\lim_{n\to0}\int D\mathbf{y}\,D(z/\sigma)\int D\mathbf{v}\ \left[\int d\mathbf{x}\,\Omega(\mathbf{r};\mathbf{v})e^{-i\hat{\mathbf{r}}\cdot\mathbf{r}}\right]\times\left[\int d\mathbf{x}\,\Omega(\mathbf{r};\mathbf{v})\right]^{n-1}$$

$$=\int D\mathbf{y}\,D(z/\sigma)\int D\mathbf{v}\ \left[\frac{\int d\mathbf{x}\,\Omega(\mathbf{r};\mathbf{v})e^{-i\hat{\mathbf{r}}\cdot\mathbf{r}}}{\int d\mathbf{x}\,\Omega(\mathbf{r};\mathbf{v})}\right]$$

$$=\int d\mathbf{r}\,P(\mathbf{r})e^{-i\hat{\mathbf{r}}\cdot\mathbf{r}}. \tag{A21}$$

Next we evaluate the $\mathcal{E}_j(\hat{\mathbf{r}})$ by working out the partial derivative on ξ_j and separating the summation over replica indices into two groups: $\alpha=1$ and $\alpha>1$,

$$\mathcal{E}_j(\hat{\mathbf{r}})=\left\langle\left[\frac{1}{\alpha}\sum_{\alpha i}\partial_{\alpha i}\chi^\alpha J_{ij}^\alpha+\frac{1}{\alpha}\sum_{\alpha n}\partial_{\alpha n}\chi^\alpha B_{nj}-\sum_i i\hat{x}_i J_{ij}^1 - \sum_n i\hat{y}_n B_{nj}\right]\mathcal{D}(\hat{\mathbf{r}},\boldsymbol{\xi},\rho)\right\rangle_{\tilde{D}}$$

$$=\Bigg[\sum_i\hat{\mathcal{F}}_i(\hat{\mathbf{r}})J_{ij}^1+\sum_n\hat{\mathcal{F}}_n(\hat{\mathbf{r}})B_{nj}+\sum_{i,\alpha>1}\hat{\mathcal{K}}_i(\hat{\mathbf{r}})J_{ij}^\alpha + \sum_{n,\alpha>1}\hat{\mathcal{K}}_n(\hat{\mathbf{r}})B_{nj}\Bigg], \tag{A22}$$

where the RS ansatz is used,

$$\hat{\mathcal{F}}_l^\alpha(\hat{\mathbf{r}})=\delta_{\alpha1}\hat{\mathcal{F}}_l(\hat{\mathbf{r}})+(1-\delta_{\alpha1})\hat{\mathcal{K}}_l(\hat{\mathbf{r}}), \tag{A23}$$

with

$$\hat{\mathcal{F}}_l(\hat{\mathbf{r}}) = \frac{1}{\alpha}\langle [\partial_{1,l}\chi^{(1)}(\mathbf{r})]\mathcal{D}(\hat{\mathbf{r}},\boldsymbol{\xi},\rho)\rangle_{\tilde{D}} - i\hat{x}_l\mathcal{D}(\hat{\mathbf{r}}),$$

$$\hat{\mathcal{K}}_l(\hat{\mathbf{r}}) = \frac{1}{\alpha}\langle [\partial_{2,l}\chi^{(2)}(\mathbf{r})]\mathcal{D}(\hat{\mathbf{r}};\boldsymbol{\xi},\rho)\rangle_{\tilde{D}}, \tag{A24}$$

and the index l runs through all student and teacher indices. We express $\mathcal{L}(\hat{\mathbf{r}};\hat{\mathbf{r}}')$ in terms of Eq. (A22), performing the summation over the replica indices and taking the limit of $n\to 0$. We then obtain

$$\begin{aligned}\mathcal{L}(\hat{\mathbf{r}};\hat{\mathbf{r}}') = &\sum_{ik}\ [\hat{\mathcal{F}}_i\hat{\mathcal{F}}'_k - \hat{\mathcal{K}}_i\hat{\mathcal{K}}'_k](Q_{ik}-q_{ik})\\ &+\sum_{ik}\ (\hat{\mathcal{F}}_i-\hat{\mathcal{K}}_i)(\hat{\mathcal{F}}'_k-\hat{\mathcal{K}}'_k)q_{ik}\\ &+\sum_{in}\ [(\hat{\mathcal{F}}_i-\hat{\mathcal{K}}_i)(\hat{\mathcal{F}}'_n-\hat{\mathcal{K}}'_n)\\ &+(\hat{\mathcal{F}}'_i-\hat{\mathcal{K}}'_i)(\hat{\mathcal{F}}_n-\hat{\mathcal{K}}_n)]R_{in}\\ &+\sum_{n}\ (\hat{\mathcal{F}}_n-\hat{\mathcal{K}}_n)(\hat{\mathcal{F}}'_n-\hat{\mathcal{K}}'_n).\end{aligned} \tag{A25}$$

The Green function becomes

$$\begin{aligned}\mathcal{A}(\mathbf{r};\mathbf{r}') = &\sum_{ik}\ [\mathcal{F}_i\mathcal{F}'_k - \mathcal{K}_i\mathcal{K}'_k](Q_{ik}-q_{ik})\\ &+\sum_{ik}\ (\mathcal{F}_i-\mathcal{K}_i)(\mathcal{F}'_k-\mathcal{K}'_k)q_{ik}\\ &+\sum_{in}\ [(\mathcal{F}_i-\mathcal{K}_i)(\mathcal{F}'_n-\mathcal{K}'_n)\\ &+(\mathcal{F}'_i-\mathcal{K}'_i)(\mathcal{F}_n-\mathcal{K}_n)]R_{in}\\ &+\sum_{n}\ (\mathcal{F}_n-\mathcal{K}_n)(\mathcal{F}'_n-\mathcal{K}'_n),\end{aligned} \tag{A26}$$

using the inverse Fourier transforms of $\hat{\mathcal{F}}_l(\hat{\mathbf{r}})$ and $\hat{\mathcal{K}}_l(\hat{\mathbf{r}})$

$$\mathcal{F}_l(\mathbf{r}) = \int \frac{d\mathbf{r}}{(2\pi)^{K+M+1}}\hat{\mathcal{F}}_l(\hat{\mathbf{r}})e^{i\hat{\mathbf{r}}\cdot\mathbf{r}}, \tag{A27}$$

$$\mathcal{K}_l(\mathbf{r}) = \int \frac{d\mathbf{r}}{(2\pi)^{K+M+1}}\hat{\mathcal{K}}_l(\hat{\mathbf{r}})e^{i\hat{\mathbf{r}}\cdot\mathbf{r}}. \tag{A28}$$

Making use of the saddle-point equation for $\chi(\mathbf{r})$, Eq. (A18), and the expression for $\mathcal{D}(\hat{\mathbf{r}},\boldsymbol{\xi},\rho)$, Eq. (A21), we can work out the explicit expressions of the functions $\mathcal{F}_l(\mathbf{r})$ and $\mathcal{K}_l(\mathbf{r})$,

$$\mathcal{F}_l(\mathbf{r}) = \frac{1}{\alpha}P(\mathbf{r})[\partial_l\chi(\mathbf{r})] - [\partial_l P(\mathbf{r})], \tag{A29}$$

$$\begin{aligned}\mathcal{K}_l(\mathbf{r}) = &\frac{1}{\alpha}P(\mathbf{y},z)\int D\mathbf{v}\ \left[\frac{\Omega(\mathbf{r};\mathbf{v})}{\int d\mathbf{x}'\ \Omega(\mathbf{x}',\mathbf{y},z;\mathbf{v})}\right]\\ &\times\left[\frac{\int d\mathbf{x}'\ \Omega(\mathbf{x}',\mathbf{y},z;\mathbf{v})[\partial_l\chi(\mathbf{r})]}{\int d\mathbf{x}'\ \Omega(\mathbf{x}',\mathbf{y},z;\mathbf{v})}\right].\end{aligned} \tag{A30}$$

Separating the index l to the student (labeled by i) and teacher (labeled by n) indices, we obtain four different functions

$$\begin{aligned}\mathcal{F}_i(\mathbf{r}) = &[(Q-q)^{-1}(\mathbf{x}-R\mathbf{y})]_i P(\mathbf{r}) + [\partial_i \ln M(\mathbf{r})]P(\mathbf{r})\\ &-\partial_i P(\mathbf{r})\\ \mathcal{F}_n(\mathbf{r}) = &-[R^T(Q-q)^{-1}(\mathbf{x}-R\mathbf{y})]_n P(\mathbf{r}) + [\partial_n \ln M(\mathbf{r})]P(\mathbf{r})\\ &-\partial_n P(\mathbf{r})\\ = &-[R^T(Q-q)^{-1}(\mathbf{x}-R\mathbf{y})]_n P(\mathbf{r}) + y_n P(\mathbf{r})\\ &+P(\mathbf{y},z)\int D\mathbf{v}\ \left[\frac{M(\mathbf{r})e^{\mathbf{x}^T B\mathbf{v}}}{\int d\mathbf{x}'\ M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right]\\ &\times\left\{\frac{\int d\mathbf{x}'[\partial_n M(\mathbf{x}',\mathbf{y},z)]e^{\mathbf{x}'^T B\mathbf{v}}}{\int d\mathbf{x}'\ M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right\},\end{aligned} \tag{A31}$$

$$\begin{aligned}\mathcal{K}_i(\mathbf{r}) = &-[(Q-q)^{-1}R\mathbf{y}]_i P(\mathbf{r}) - \partial_i P(\mathbf{r}) + [\partial_i \ln M(\mathbf{r})]P(\mathbf{r})\\ &+P(\mathbf{y},z)\int D\mathbf{v}\ \left[\frac{M(\mathbf{r})e^{\mathbf{x}^T B\mathbf{v}}}{\int d\mathbf{x}'\ M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right]\\ &\times\left\{\frac{\int d\mathbf{x}'[(Q-q)^{-1}\mathbf{x}']_i M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}{\int d\mathbf{x}'\ M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right\},\\ \mathcal{K}_n(\mathbf{r}) = &[R^T(Q-q)^{-1}R\mathbf{y}]_n P(\mathbf{r})\\ &-P(\mathbf{y},z)\int D\mathbf{v}\ \left[\frac{M(\mathbf{r})e^{\mathbf{x}^T B\mathbf{v}}}{\int d\mathbf{x}'\ M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right]\\ &\times\left\{\frac{\int d\mathbf{x}'[R^T(Q-q)^{-1}\mathbf{x}']_n M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}{\int d\mathbf{x}'\ M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right\}\end{aligned}$$

$$+P(\mathbf{y},z)\int D\mathbf{v}\left[\frac{Me^{\mathbf{x}^T B\mathbf{v}}}{\int d\mathbf{x}'\,M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right]$$

$$\times\left\{\frac{\int d\mathbf{x}'[\partial_n M(\mathbf{x}',\mathbf{y},z)]e^{\mathbf{x}'^T B\mathbf{v}}}{\int d\mathbf{x}'\,M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}\right\}. \quad \text{(A32)}$$

Rescaling the above functions by $P(\mathbf{r})$: $\tilde{\mathcal{F}}_l(\mathbf{r})=\mathcal{F}_l(\mathbf{r})/P(\mathbf{r})$ and $\tilde{\mathcal{K}}_l(\mathbf{r})=\mathcal{K}_l(\mathbf{r})/P(\mathbf{r})$, and defining the function

$$\Phi_i(\mathbf{r})=\tilde{\mathcal{F}}_i(\mathbf{r})-\tilde{\mathcal{K}}_i(\mathbf{r})$$
$$=\frac{1}{P[\mathbf{x}|\mathbf{y},z]}\int D\mathbf{v}\langle[(Q-q)^{-1}(\mathbf{x}-\mathbf{x}')]_i\rangle_*$$
$$\times\langle\delta(\mathbf{x}-\mathbf{x}')\rangle_*, \quad \text{(A33)}$$

with the abbreviation

$$\langle f(\mathbf{x},\mathbf{x}')\rangle_*=\frac{\int d\mathbf{x}'\,M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}f(\mathbf{x},\mathbf{x}')}{\int d\mathbf{x}'\,M(\mathbf{x}',\mathbf{y},z)e^{\mathbf{x}'^T B\mathbf{v}}}, \quad \text{(A34)}$$

we obtain the following compact forms for $\tilde{\mathcal{F}}(\mathbf{r})$ and $\tilde{\mathcal{K}}(\mathbf{r})$

$$\tilde{\mathcal{F}}_i(\mathbf{r})=[(Q-q)^{-1}(\mathbf{x}-R\mathbf{y})]_i$$
$$-[(Q-q)^{-1}(q-RR^T)\mathbf{\Phi}(\mathbf{r})]_i,$$

$$\tilde{\mathcal{K}}_i(\mathbf{r})=\tilde{\mathcal{F}}_i(\mathbf{r})-\Phi_i(\mathbf{r}),$$

$$\tilde{\mathcal{F}}_n(\mathbf{r})-\tilde{\mathcal{K}}_n(\mathbf{r})=y_n-[R^T\mathbf{\Phi}(\mathbf{r})]_n. \quad \text{(A35)}$$

Inserting Eqs. (A33) and (A35) into Eq. (A26), we finally obtain the rescaled Green function

$$\tilde{\mathcal{A}}(\mathbf{r};\mathbf{r}')=\frac{\mathcal{A}(\mathbf{r};\mathbf{r}')}{P(\mathbf{r})P(\mathbf{r}')}$$
$$=\mathbf{y}^T\mathbf{y}'+(\mathbf{x}-R\mathbf{y})^T\mathbf{\Phi}(\mathbf{r}')+\mathbf{\Phi}^T(\mathbf{r})(\mathbf{x}'-R\mathbf{y}')$$
$$-\mathbf{\Phi}^T(\mathbf{r})(Q-RR^T)\mathbf{\Phi}(\mathbf{r}') \quad \text{(A36)}$$

with $\mathbf{\Phi}(\mathbf{r})$ given in Eq. (A33). Working out the integration

$$\int d\mathbf{r}'\,\mathcal{A}(\mathbf{r};\mathbf{r}')\mathcal{G}(\mathbf{r}')=P(\mathbf{r})\int d\mathbf{r}'\,P(\mathbf{r}')\mathcal{G}(\mathbf{r}')\tilde{\mathcal{A}}(\mathbf{r};\mathbf{r}')$$
$$=P(\mathbf{r})\mathbf{\Gamma}(\mathbf{r}) \quad \text{(A37)}$$

with

$$\mathbf{\Gamma}(\mathbf{r})=W\mathbf{y}+U(\mathbf{x}-R\mathbf{y})+X(Q-RR^T)\mathbf{\Phi}(\mathbf{r}) \quad \text{(A38)}$$

and

$$X=(V-WR^T)(Q-RR^T)^{-1}-U, \qquad U=\langle\mathcal{G}\mathbf{\Phi}^T\rangle, \quad \text{(A39)}$$

we finally obtain the equation for probability distribution under RS ansatz, which is Eq. (13).

4. The large α approximation

In the large α limit, the order parameter matrix q takes the value RR^T and the elements of matrix B are very small. We can therefore use the cumulant expansion up to the second order to obtain

$$M(\mathbf{r})=P[\mathbf{x}|\mathbf{y},z]\exp\left\{-\frac{1}{2}[\mathbf{x}-\overline{\mathbf{x}}(\mathbf{y},z)]^T B'[\mathbf{x}-\overline{\mathbf{x}}(\mathbf{y},z)]\right.$$
$$\left.+\frac{1}{2}\overline{[\mathbf{x}^T B'\mathbf{x}-\overline{\mathbf{x}}^T}(\mathbf{y},z)B'\overline{\mathbf{x}}(\mathbf{y},z)]\right\}+\cdots, \quad \text{(A40)}$$

the overline denotes averages with respect to $P[\mathbf{x}|\mathbf{y},z]$ and the matrix B' is of the form $B'=(Q-q)^{-1}(q-RR^T)(Q-q)^{-1}$. Furthermore, we have $(Q-q)\simeq(Q-RR^T)$, the function $\mathbf{\Phi}(\mathbf{r})$ in Eq. (A33) and the matrix U in Eq. (A39) become

$$\mathbf{\Phi}(\mathbf{r})\simeq(Q-RR^T)^{-1}(\mathbf{x}-\overline{\mathbf{x}}),$$

$$U=[V-\langle\mathcal{G}\overline{\mathbf{x}}^T(\mathbf{y},z)\rangle](Q-RR^T)^{-1}. \quad \text{(A41)}$$

Finally, the dynamical equation for the probability distribution in Eq. (13) becomes equivalent to Eq. (16) with the explicit form of $\mathbf{\Gamma}(\mathbf{r})$.

APPENDIX B: THE MIXTURE OF GAUSSIAN REPRESENTATION

A mixture of Gaussians can represent an arbitrary probability distribution given a sufficient number of basis functions. Using a mixture of Gaussian representations for the probability distribution (in the noiseless case)

$$\mathcal{Q}(\mathbf{x},\mathbf{y})=\sum_{\rho=1}^{L}\frac{w_\rho}{\sqrt{(2\pi)^{K+M}|A_\rho|}}$$
$$\times\exp\left[-\frac{1}{2}\begin{pmatrix}\mathbf{x}-\overline{x}_\rho\\ \mathbf{y}\end{pmatrix}^T A_\rho^{-1}\begin{pmatrix}\mathbf{x}-\overline{x}_\rho\\ \mathbf{y}\end{pmatrix}\right] \quad \text{(B1)}$$

and the parameter set $\theta=[w_\rho,\overline{x}_\rho,A_\rho]$, from which the equations for R and Q follow directly,

$$\frac{dR_{in}}{dt}=\eta\sum_\rho w_\rho\left[\sum_m I_3^\rho(i,n,m)-\sum_j I_3^\rho(i,n,j)\right]-\gamma R_{in} \quad \text{(B2)}$$

and

$$\frac{dQ_{ik}}{dt}=\eta\sum_{\rho}\ w_{\rho}\Bigg\{\sum_{m}\ [I_3^{\rho}(i,k,m)+I_3^{\rho}(k,i,m)]-\sum_{j}\ [I_3^{\rho}(i,k,j)+I_3^{\rho}(k,i,j)]\Bigg\}+\eta^2\sum_{\rho}\ w_{\rho}Z_{ik}^{\rho}-2\gamma Q_{ik}, \tag{B3}$$

where

$$Z_{ik}^{\rho}=\sum_{jl}\ I_4(i,k,j,l)-2\sum_{jm}\ J_4(i,k,j,m)+\sum_{mn}\ K_4(i,k,m,n).$$

The integrals I_3, I_4, J_4, and K_4 are defined in Appendix C.

The difficulty is in obtaining a set of equations for the evolution of the parameter set θ. This can be done in principle by minimizing some distance measure between the updated distribution $P(\mathbf{x},\mathbf{y})$ and the approximation $Q(\mathbf{x},\mathbf{y})$. We experienced computational difficulties in carrying it out using a quadratic distance measure mainly due to the different sensitivities of the various parameters. Nevertheless, being capable of representing any probability distribution, we believe that this representation may allow one to obtain more accurate results where the local Gaussian approximation breaks down.

APPENDIX C: LOCAL GAUSSIAN REPRESENTATION FOR THE CASE OF OUTPUT NOISE AND REGULARIZER

For a locally Gaussian approximation, the conditional probability has a form

$$P[\mathbf{x}|\mathbf{y},z]=\frac{1}{\sqrt{(2\pi)^K|\Sigma(\mathbf{y},z)|}}\times\exp\Bigg\{-\frac{1}{2}[\mathbf{x}-\overline{\mathbf{x}}(\mathbf{y},z)]^T\Sigma^{-1}(\mathbf{y},z)[\mathbf{x}-\overline{\mathbf{x}}(\mathbf{y},z)]\Bigg\}. \tag{C1}$$

The main advantages of this approximation are that the integration over the student field $\mathbf{x}$ can be carried out analytically and the partial differential equation for $P(\mathbf{r})$ in Eq. (16) can be simplified to a set of differential equations for the parameters $\Sigma(\mathbf{y},z)$,$\overline{\mathbf{x}}(\mathbf{y},\mathbf{z})$ as described in Eq. (18).

1. The equations for the parameters *Q* and *R*

Under this approximation, the equations for the macroscopic parameters Q and R in Eqs. (7) become

$$\frac{dR}{dt}=\eta\int d\mathbf{y}\,dz\,P(\mathbf{y},z)\overline{W}(\mathbf{y},z)-\gamma R,$$

$$\frac{dQ}{dt}=\eta\int d\mathbf{y}\,dz\,P(\mathbf{y},z)[\overline{V}(\mathbf{y},z)+\overline{V}^T(\mathbf{y},z)]+\eta^2\int d\mathbf{y}\,dz\,P(\mathbf{y},z)\overline{Z}(\mathbf{y},z)-2\gamma Q \tag{C2}$$

with

$$\overline{V}_{ik}(\mathbf{y},z)=\sum_{l}\ I_3(i,k,l)-\sum_{j}\ J_3(i,k,j),$$

$$\overline{W}_{in}(\mathbf{y},z)=\sum_{l}\ K_3(i,n,l)-\sum_{j}\ L_3(i,n,j),$$

$$\overline{Z}_{ik}(\mathbf{y},z)=\sum_{jl}\ I_4(i,k,j,l)-2\sum_{jm}\ J_4(i,k,j,m)+\sum_{mn}\ K_4(i,k,m,n) \tag{C3}$$

where the integrals on the right-hand side depend on $\mathbf{y}$ and z through $\Sigma(\mathbf{y},z)$ and $\overline{\mathbf{x}}(\mathbf{y},z)$.

2. Three-dimensional integrals

The three-dimensional integrals in Eq. (C3) are given by

$$I_3(1,2,3)=\sqrt{\frac{2}{\pi}}\langle e^{-(1/2)x_1^2}x_2g(y_3)\rangle=I_1\Gamma_{12}g(y_3),$$

$$J_3(1,2,3)=\sqrt{\frac{2}{\pi}}\langle e^{-(1/2)x_1^2}x_2g(x_3)\rangle=I_1\Bigg[\Gamma_{12}g(\Theta_{13})+\sqrt{\frac{2}{\pi}}\Delta_c e^{-(1/2)\Theta_{13}^2}\Bigg],$$

$$K_3(1,2,3)=\sqrt{\frac{2}{\pi}}\langle e^{-(1/2)x_1^2}y_2g(y_3)\rangle=I_1y_2g(y_3),$$

$$L_3(1,2,3)=\sqrt{\frac{2}{\pi}}\langle e^{-(1/2)x_1^2}y_2g(x_3)\rangle=I_1y_2g(\Theta_{13}), \tag{C4}$$

with $\langle\ldots\rangle=\int d\mathbf{x}\ \ P[\mathbf{x}|\mathbf{y},z]\ldots$ and

$$I_i=\sqrt{\frac{2}{\pi}}\langle e^{-(1/2)x_i^2}\rangle=\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\phi_i}}e^{-(1/2)(\overline{x}_i^2/\phi_i)}, \tag{C5}$$

$$\Theta_{13}=(\overline{x}_3\phi_1-\overline{x}_1\Sigma_{13})\phi_{13},\qquad \Delta_c=(\phi_1\Sigma_{23}-\Sigma_{13}\Sigma_{12})\phi_{13},$$

$$\phi_i=1+\Sigma_{ii},\qquad \phi_{13}=1/\sqrt{\phi_1(\phi_1\phi_3-\Sigma_{13}^2)},$$

$$\Gamma_{12}=\overline{x}_2-\Sigma_{12}\overline{x}_1/\phi_1.$$

3. Four-dimensional integrals

The four-dimensional integrals in Eq. (C3) are given by

$$I_4(1,2,3,4)=\frac{2}{\pi}\langle e^{-(1/2)x_1^2-(1/2)x_2^2}g(y_3)g(y_4)\rangle=I_2(1,2)g(y_3)g(y_4),$$

$$J_4(1,2,3,4) = \frac{2}{\pi}\langle e^{-(1/2)x_1^2-(1/2)x_2^2} g(x_3)g(y_4)\rangle$$
$$= I_2(1,2)g(\Theta_{123})g(y_4),$$
$$K_4(1,2,3,4) = \frac{2}{\pi}\langle e^{-(1/2)x_1^2-(1/2)x_2^2} g(x_3)g(x_4)\rangle,$$
$$= I_2(1,2)\int Dx\, g(\sqrt{\Delta_{11}}x+\Theta_3) \times g\left(\frac{\Delta_{12}x+\sqrt{\Delta_{11}}\Theta_4}{\sqrt{|\Delta|}}\right), \quad \text{(C6)}$$

where the two-dimensional integral is defined as

$$I_2(1,2) = \left\langle \frac{2}{\pi} e^{-(1/2)x_1^2-(1/2)x_2^2}\right\rangle = \frac{2}{\pi}\frac{1}{\sqrt{|C|}}\exp\left[-\frac{1}{2}\begin{pmatrix}\overline{x_1}\\ \overline{x_2}\end{pmatrix}^T C^{-1}\begin{pmatrix}\overline{x_1}\\ \overline{x_2}\end{pmatrix}\right], \quad \text{(C7)}$$

with the matrix

$$C = \begin{pmatrix}\phi_1 & \Sigma_{12}\\ \Sigma_{12} & \phi_2\end{pmatrix}$$

and the arguments are defined as

$$\begin{pmatrix}\Theta_3\\ \Theta_4\end{pmatrix} = \begin{bmatrix}\overline{x_3}-(\overline{x_1}D_{11}+\overline{x_2}D_{21})\\ \overline{x_4}-(\overline{x_1}D_{12}+\overline{x_2}D_{22})\end{bmatrix},$$

$$\Theta_{123} = \frac{\overline{x_3}-(\overline{x_1}T_1+\overline{x_2}T_2)}{\sqrt{\phi_3-(T_1\Sigma_{13}+T_2\Sigma_{23})}} \quad \text{with} \quad \begin{pmatrix}T_1\\ T_2\end{pmatrix} = C^{-1}\begin{pmatrix}\Sigma_{13}\\ \Sigma_{23}\end{pmatrix},$$

and

$$\Delta = \begin{pmatrix}\Sigma_{33}-E_{11}, & \Sigma_{34}-E_{12}\\ \Sigma_{34}-E_{21}, & \phi_4-E_{22}\end{pmatrix},$$

$$E = \begin{pmatrix}\Sigma_{13}D_{11}+\Sigma_{23}D_{21}, & \Sigma_{13}D_{12}+\Sigma_{23}D_{22}\\ \Sigma_{14}D_{11}+\Sigma_{24}D_{21}, & \Sigma_{14}D_{12}+\Sigma_{24}D_{22}\end{pmatrix},$$

$$D = \frac{1}{|C|}\begin{pmatrix}\phi_2\Sigma_{13}-\Sigma_{12}\Sigma_{23}, & \phi_2\Sigma_{14}-\Sigma_{12}\Sigma_{24}\\ \phi_1\Sigma_{23}-\Sigma_{12}\Sigma_{13}, & \phi_1\Sigma_{24}-\Sigma_{12}\Sigma_{14}\end{pmatrix}. \quad \text{(C8)}$$

[1] J.A. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1990).
[2] C.M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995).
[3] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel, Neural Comput. **1**, 541 (1989).
[4] C.W.H. Mace and A.C.C. Coolen, Stat. Comput. **8**, 55 (1998).
[5] *On-line Learning in Neural Networks*, edited by D. Saad (Cambridge University Press, Cambridge, 1998).
[6] W. Kinzel and P. Rujan, Europhys. Lett. **13**, 473 (1990).
[7] M. Biehl and H. Schwarze, J. Phys. A **28**, 643 (1995).
[8] D. Saad and S.A. Solla Phys. Rev. Lett. **74**, 4337 (1995); Phys. Rev. E **52**, 4225 (1995).
[9] A. Krogh and J.A. Hertz, J. Phys. A **25**, 1135 (1992).
[10] H. Horner, Z. Phys. B: Condens. Matter **86**, 291 (1992); **87**, 371 (1992).
[11] P. Sollich and D. Barber, in *On-line Learning in Neural Net-Works* (Ref. [5]), p. 279.
[12] B. Lopez and M. Opper, Europhys. Lett. **49**, 275 (2000).
[13] S. Lee and K.Y.M. Wong, in *Advances in Neural Information Processing Systems*, edited by S.A. Solla, T.K. Leen, and K. Müller (MIT Press, Cambridge, 2000), Vol. 12, p. 286.
[14] A.C.C. Coolen and D. Saad, in *On-line Learning in Neural Networks* (Ref. [5]), p. 303; Phys. Rev. E **62**, 5444 (2000).
[15] A.C.C. Coolen, D. Saad, and Y. Xiong, Europhys. Lett. **51**, 691 (2000).
[16] A.C.C. Coolen, S.N. Laughton, and D. Sherrington, Phys. Rev. B **53**, 8184 (1996).
[17] M. Biehl, P. Riegler, and C. Wohel, J. Phys. A **29**, 4767 (1996).
[18] C.W.H. Mace and A.C.C. Coolen, in *Advances in Neural Information Processing Systems* (Ref. [13]) , Vol. 12, p. 237.
[19] D. Saad and M. Rattray, Phys. Rev. E **57**, 2170 (1998).
[20] D. Saad and M. Rattray, Phys. Rev. Lett. **79**, 2578 (1997).
[21] M. Rattray and D. Saad, J. Phys. A **30**, L771 (1997).
[22] M. Rattray and D. Saad, Phys. Rev. E **58**, 6379 (1998).
[23] M. Rattray, D. Saad, and S. Amari, Phys. Rev. Lett. **81**, 5461 (1998).
[24] M. Rattray and D. Saad, Phys. Rev. E **59**, 4523 (1999).