On-line learning with adaptive back-propagation in two-layer networks

Ansgar H. L. West[1,2] and David Saad[1]
[1]*Neural Computing Research Group, University of Aston, Birmingham B4 7ET, United Kingdom*
[2]*Department of Physics, University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom*
(Received 28 April 1997)

An adaptive back-propagation algorithm parametrized by an inverse temperature β is studied and compared with gradient descent (standard back-propagation) for on-line learning in two-layer neural networks with an arbitrary number of hidden units. Within a statistical mechanics framework, we analyze these learning algorithms in both the symmetric and the convergence phase for finite learning rates in the case of uncorrelated teachers of similar but arbitrary length T. These analyses show that adaptive back-propagation results generally in faster training by breaking the symmetry between hidden units more efficiently and by providing faster convergence to optimal generalization than gradient descent. [S1063-651X(97)08109-9]

PACS number(s): 87.10.+e, 05.20.−y, 02.50.−r, 02.30.Hq

I. INTRODUCTION

Multilayer feedforward perceptrons are widely used in classification and regression applications mainly due to their ability to learn a wide range of maps [1] from examples. When learning a map f_0 from N-dimensional inputs $\boldsymbol{\xi}$ to scalars ζ the parameters $\{\boldsymbol{W}\}$ of the *student* network are modified according to some training algorithm so that the map defined by these parameters $f_{\boldsymbol{W}}$ approximates the *teacher* f_0 as close as possible. The resulting performance can be measured by the *generalization error* ϵ_g, the average of an error measure ϵ over input space $\epsilon_g = \langle \epsilon \rangle_{\boldsymbol{\xi}}$. The error measure or loss function is often defined as the squared distance between the output of the network and the desired output, i.e.,

$$\epsilon = \frac{1}{2}[f_{\boldsymbol{W}}(\boldsymbol{\xi}) - f_0(\boldsymbol{\xi})]^2. \tag{1}$$

One usually distinguishes between two learning paradigms: *batch learning*, where training algorithms are generally based on minimizing the error on the whole set of given examples, and *on-line learning*, where single examples are presented serially and the training algorithm adjusts the parameters after the presentation of each example. The efficiency of these training algorithms is measured by their speed of convergence to an ''acceptable'' generalization error (in terms of training time or the number of example presentations).

This research has been primarily motivated by recent work [2] investigating an on-line learning scenario of a general two-layer student network trained by gradient descent (which is usually referred to in the neural network literature as *back-propagation*) on a task defined by a teacher network of similar architecture. It has been found that in the early stages of training the student is drawn into a suboptimal symmetric phase, characterized by undifferentiated imitation, by student vectors, of parameter vectors related to the various teacher hidden nodes. Although student node symmetry is eventually broken and student performance converges to the minimal achievable generalization error, a significant part of the training time may be spent with the system trapped in the symmetric subspace. Speeding up the escape from the symmetric phase is likely to improve the training efficiency significantly; in this paper we suggest a simple modification of the basic back-propagation and analyze the resulting expected improvement in training efficiency.

The need for improved neural network training methods is clear as training efficiency is in the heart of the method itself and plays a significant role in determining the usefulness of the method as a whole; new tools may enable us to obtain better performance in shorter training times as well as to expand the envelope of feasible tasks. For batch training there is a variety of efficient training methods available, such as second-order methods (e.g., Newton-Raphson or conjugate gradient). However, as these methods are based on the entire training set they are not applicable to on-line learning. Several different methods have been employed for improving on-line training in both discrete and smooth networks, most of which are based on heuristics or on analysis in the asymptotic regime.

Among the most common modifications to the conventional back-propagation algorithm, for smooth systems, is training with momentum. An analysis using stochastic approximation theory [3] shows that for learning large example sets it merely rescales the learning rate in the convergence phase. Similar trivial effects are also mirrored in the statistical mechanics framework [4], unless different scaling is used for the learning rate term. Its usefulness is so far inconclusive. Other methods aimed at incorporating information about the curvature of the error surface into the learning rule have been proposed recently [3,5]. These rules are expected to be efficient asymptotically, although their effect on earlier stages of the learning process and especially on the length of the symmetric phase is not yet clear.

Several efficient methods have been suggested for on-line learning in discrete networks. Some of the methods are based on a greedy maximization of the local difference in generalization error [6], while others are based on structured learning rules [7,8]. It is, however, unclear whether these methods can be extended to accommodate smooth multilayer networks such as the soft-committee machine [9,2] and whether these extensions would be useful in devising an efficient method

for escaping the symmetric phase, especially since applying local optimization in this phase is likely to fail (as demonstrated in [10]).

A method for breaking the symmetry of the student network in smooth machines by enforcing a weight-ordering penalty term on the space of hidden units has been suggested in [11], showing a considerable improvement in training time for a very simple network architecture. A more detailed numerical investigation, however, shows that this method fails completely in the case of isotropic teacher networks, with uncorrelated teacher weight vectors of similar length, where the student remains indefinitely trapped in a suboptimal symmetric phase [12]. In the case of a soft-committee machine where biases are applied to the hidden layer nodes, as is the case in realistic networks, there is further evidence that the strongest symmetry-breaking effect is provided by the network biases [13], possibly leading to a stagnating competition in breaking the symmetry between biases and the weight-ordering penalty term.

The aim of this paper is twofold. It gives some insight into the reasons for the short-comings of back-propagation and it furthermore investigates possible improvements by introducing an adaptive back-propagation algorithm [14]. This algorithm features, besides the learning rate η, a second adaptable parameter, the inverse temperature β, which improves the ability of the student to distinguish between hidden nodes of the teacher for $\beta>1$. We compare its efficiency with that of gradient descent in training two-layer networks following the framework of [2] and present numerical studies and rigorous analyzes of both the breaking of the symmetric phase and the asymptotic convergence. We note that although these analyzes provide us with optimal values of the user adjustable parameters η and β for different stages of the training process in a range of learning scenarios, it remains an open question how these parameters can be optimized adaptively on-line without *a priori* knowledge of the training task [15]. Within this limitation, we find that the optimized adaptive back-propagation can significantly reduce training time in both regimes by efficiently breaking the symmetry between hidden units and by providing faster exponential convergence asymptotically.

II. DERIVATION OF THE DYNAMICAL EQUATIONS

The student network we consider is a normalized soft-committee machine, consisting of K hidden units, which are connected to N-dimensional inputs $\boldsymbol{\xi}$ by their weight vectors $\boldsymbol{W}=\{\boldsymbol{W}_i\}$ $(i=1,\ldots,K)$. All hidden units are connected to the linear output unit with arbitrary but fixed gain γ by couplings of fixed strength. The activation of any unit is normalized (by the inverse square root of the number of weight connections into the unit) allowing all weights to be of $O(1)$ magnitude, independent of the input dimension or the number of hidden units. Note that this is in contrast to most other on-line learning literature (e.g., [9]); however, as we will see later, this leads to a more intuitive and elegant result for the optimal learning rates. The implemented mapping is therefore

$$f_{\boldsymbol{W}}(\boldsymbol{\xi})=\frac{\gamma}{\sqrt{K}}\sum_{i=1}^{K} g\left(\frac{1}{\sqrt{N}}\boldsymbol{W}_i\cdot\boldsymbol{\xi}\right)=\frac{\gamma}{\sqrt{K}}\sum_{i=1}^{K} g(x_i), \qquad (2)$$

where $x_i=\boldsymbol{W}_i\cdot\boldsymbol{\xi}/\sqrt{N}$ is the student activation and $g()$ is a sigmoidal transfer function. The map f_0 to be learned is defined by a teacher network of the same architecture except for a possible difference in the number of hidden units M and is defined by the weight vectors $\boldsymbol{B}=\{\boldsymbol{B}_n\}$ $(n=1,\ldots,M)$. Training examples are of the form $(\boldsymbol{\xi}^\mu,\zeta^\mu)$, where the components of the input vectors $\boldsymbol{\xi}^\mu$ are drawn independently from a zero mean Gaussian distribution with arbitrary variance σ^2. The targets therefore are

$$\zeta^\mu=\frac{\gamma}{\sqrt{M}}\sum_{n=1}^{M} g\left(\frac{1}{\sqrt{N}}\boldsymbol{B}_n\cdot\boldsymbol{\xi}^\mu\right)=\frac{\gamma}{\sqrt{M}}\sum_{n=1}^{M} g(y_n^\mu), \qquad (3)$$

where $y_n^\mu=\boldsymbol{B}_n\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ is the activation of teacher hidden unit n. Note that we will use indices i,j,k,l to refer to units in the student network and n,m for units in the teacher network.

An on-line training algorithm $\mathcal{A}$ is defined by the update of each weight in response to the presentation of an example $(\boldsymbol{\xi}^\mu,\zeta^\mu)$, which can take the general form

$$\boldsymbol{W}_i^{\mu+1}=\boldsymbol{W}_i^\mu+\mathcal{A}_i(\{\boldsymbol{\gamma}\},\boldsymbol{W}^\mu,\boldsymbol{\xi}^\mu,\zeta^\mu), \qquad (4)$$

where $\{\boldsymbol{\gamma}\}$ defines parameters adjustable by the user. In the case of standard back-propagation, i.e., gradient descent on the error function defined in Eq. (1),

$$\mathcal{A}_i^{\mathrm{GD}}(\eta,\boldsymbol{W}^\mu,\boldsymbol{\xi}^\mu,\zeta^\mu)=\eta\delta_i^\mu\boldsymbol{\xi}^\mu, \qquad (5)$$

with

$$\begin{aligned}\delta_i^\mu&=\delta^\mu g'(x_i^\mu)\\&=[\zeta^\mu-f_{\boldsymbol{W}}(\boldsymbol{\xi}^\mu)]g'(x_i^\mu),\end{aligned} \qquad (6)$$

where the only user adjustable parameter is the learning rate η. One can readily see that each of the three terms in the back-propagation weight update plays a different role. The difference δ^μ between the student output and the target together with the learning rate determines the overall size of the update of all weight parameters by specifying how closely student and teacher are matched. The input vector $\boldsymbol{\xi}^\mu$ discriminates between the weights leading to different inputs. However, only $g'(x_i^\mu)$, i.e., the derivative of the transfer function $g()$, breaks the symmetry between different hidden units. The fact that a prolonged symmetric phase can exist indicates that this term is not significantly different over the hidden units for a typical input in the symmetric phase.

The rationale of the adaptive back-propagation algorithm defined below is therefore to alter the g' term in order to magnify small differences in activation between hidden units. A simple way of enhancing these differences is by altering $g'(x_i)$ to $g'(\beta x_i)$, where β plays the role of an inverse ''temperature.'' Varying β changes the range of hidden unit activations relevant for training, e.g., for $\beta>1$ learning is more confined to small activations, when compared to gradient descent $(\beta=1)$, i.e., the training process is effectively ''frozen'' for larger activations. One could also absorb this modification into gradient descent with a site- and activation-dependent learning rate, making it more obvi-

ous that adaptive back propagation deforms the search space spatially. The adaptive back-propagation learning rule is therefore

$$\mathcal{A}_i^{\mathrm{ABP}}(\eta,\beta,\boldsymbol{W}^\mu,\boldsymbol{\xi}^\mu,\zeta^\mu)=\eta\delta^\mu g'(\beta x_i^\mu)\boldsymbol{\xi}^\mu = \eta\widetilde{\delta_i^\mu}\boldsymbol{\xi}^\mu, \tag{7}$$

with δ^μ as in Eq. (6). To compare the adaptive back-propagation (ABP) algorithm with conventional gradient descent (GD), we follow Ref. [2]. As we are interested in the typical behavior of our training algorithm we average over all possible instances of the examples $\boldsymbol{\xi}$. This average is most conveniently performed implicitly by averaging over the Gaussian distribution of the activations $\boldsymbol{x}=(x_1,\ldots,x_K)$ and $\boldsymbol{y}=(y_1,\ldots,y_M)$. The Gaussian distribution has zero mean as $\langle x_i\rangle_{\boldsymbol{\xi}}=\langle y_n\rangle_{\boldsymbol{\xi}}=0$ and a covariance matrix $\mathcal{C}$ whose components are given by the order parameters describing the overlaps between student and teacher nodes:

$$\langle x_i x_j\rangle_{\boldsymbol{\xi}}=\frac{\sigma^2}{N}\boldsymbol{W}_i\cdot\boldsymbol{W}_j\equiv Q_{ij}, \tag{8a}$$

$$\langle x_i y_n\rangle_{\boldsymbol{\xi}}=\frac{\sigma^2}{N}\boldsymbol{W}_i\cdot\boldsymbol{B}_n\equiv R_{in}, \tag{8b}$$

$$\langle y_n y_m\rangle_{\boldsymbol{\xi}}=\frac{\sigma^2}{N}\boldsymbol{B}_n\cdot\boldsymbol{B}_m\equiv T_{nm}. \tag{8c}$$

The generalization error ϵ_g, measuring the typical performance, can be expressed in these variables only. We can also rewrite the update equations (7) in $\boldsymbol{W}_i$ as equations in these order parameters and the Q_{ij} and R_{in} become the new dynamical variables, which are self-averaging with respect to the randomness in the training data in the thermodynamic limit ($N\to\infty$), whereas the T_{nm} are fixed and given by the task. We note that the variance of the input distribution merely rescales the length of the order parameters and the learning rate by σ^2 and can therefore be set to one without loss of generality.

If we interpret the normalized example number $\alpha=\mu/N$ as a continuous time variable, the update equations for the order parameters become first-order coupled differential equations

$$\frac{dR_{in}}{d\alpha}=\eta\langle\widetilde{\delta_i^\mu}y_n\rangle_{\{x,y\}}, \tag{9a}$$

$$\frac{dQ_{ij}}{d\alpha}=\eta\langle\widetilde{\delta_i^\mu}x_j+\widetilde{\delta_i^\mu}x_i\rangle_{\{x,y\}}+\eta^2\langle\widetilde{\delta_i^\mu}\widetilde{\delta_i^\mu}\rangle_{\{x,y\}}. \tag{9b}$$

All the integrals in Eqs. (9) and the generalization error can be calculated explicitly if we choose the error function $g_\nu(x)=\mathrm{erf}(\nu x/\sqrt{2})$ as the sigmoidal activation function with arbitrary gain ν. For the exact form of the dynamical equations and the generalization error, we refer the reader to Appendix A. We only mention in passing that the sigmoidal gain ν merely rescales all order parameters and the learning rate by ν^2, whereas the output gain γ rescales just the learning rate by γ^2. In the following both are therefore set to one without loss of generality.

III. NUMERICAL INTEGRATION OF THE DYNAMICAL EQUATIONS

The differential equations can easily be integrated numerically for any number of K student and M teacher hidden units. For the remainder of the paper, we will, however, focus on the realizable case ($K=M$) and uncorrelated isotropic teachers of arbitrary length $T_{nm}=T\delta_{nm}$.

The dynamical evolution of the overlaps Q_{ij} and R_{in} follows from integrating the equations of motion (9) from initial

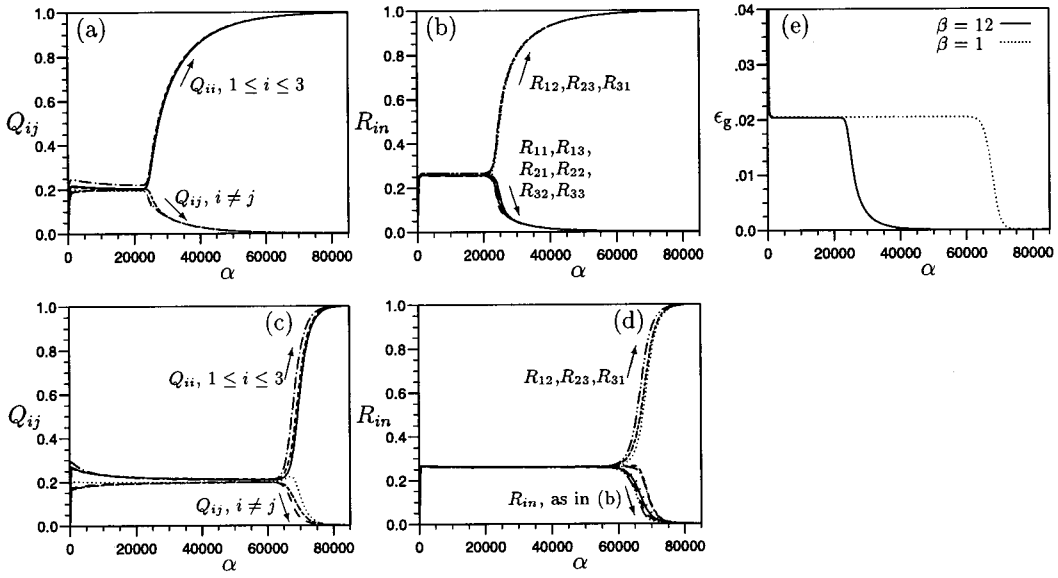

FIG. 1. Dynamical evolution of [(a) and (c)] the student-student overlaps Q_{ij}, [(b) and (d)] the student-teacher overlaps R_{in}, and (e) the generalization error as a function of the normalized example number α for a student with three hidden nodes learning an isotropic three-node teacher ($T_{nm}=\delta_{nm}$). The learning rate $\eta=0.03$ is fixed, but the value of the inverse temperature varies: [(a) and (b)] $\beta=12$ and [(c) and (d)] $\beta=1$ (gradient descent).

conditions determined by the (random) initialization of the student weights $\boldsymbol{W}_i$. For random initialization the resulting norms Q_{ii} of the student vector will be $O(1)$, while the overlaps Q_{ij} between different student vectors, and student-teacher vectors R_{in} will be only $O(1/\sqrt{N})$. A random initialization of the weights and biases can therefore be simulated by initializing the norms Q_{ii}, and the normalized overlaps $\hat{Q}_{ij}=Q_{ij}/\sqrt{Q_{ii}Q_{jj}}$ and $\hat{R}_{in}=R_{in}/\sqrt{Q_{ii}T_{nn}}$ from uniform distributions in the $[0,1]$ and $[-10^{-12},10^{-12}]$ intervals, respectively.

In Fig. 1 we show a typical difference in the evolution of the overlaps and the generalization error for $\beta=12$ and $\beta=1$ (gradient descent) for $K=3$ and $\eta=0.03$. In both cases, the student is drawn quickly into a suboptimal symmetric phase, characterized by a finite generalization error [Fig. 1(e)] and no differentiation between the hidden units of the student. The student norms Q_{ii} and overlaps Q_{ij} are similar [Figs. 1(a) and 1(c)], i.e., the students are highly correlated with each other. The overlaps of each student node with all teacher nodes R_{in} are nearly identical [Figs. 1(b) and 1(d)], i.e., each student unit imitates all teacher units with similar success. The student trained by GD [Figs. 1(c),1(d)] is trapped in this unstable suboptimal solution for most of the training time, whereas ABP [Figs. 1(a) and 1(b)] breaks the symmetry significantly earlier. The convergence phase is characterized by a specialization of each student nodes to a particular teacher node, which corresponds to an evolution of the overlap matrices **Q** and **R** to their optimal value **T**, except for the permutational symmetry due to the arbitrary labeling of the student nodes.

Examining the decay of the generalization error in Fig. 1(e) more closely, one can see that the choice $\beta=12$ is suboptimal in this regime. The student trained with $\beta=1$ converges faster to zero generalization error. In order to optimize both the learning temperature β and the learning rate η simultaneously for both phases of the learning process, the symmetric and the convergence phase, we will examine the equations of motions analytically in the following section.

IV. ANALYSIS OF THE DYNAMICAL EQUATIONS

In the case of a realizable learning scenario ($K=M$) and isotropic teachers ($T_{nm}=T\delta_{nm}$) the order parameter space can be very well characterized by similar diagonal and off-diagonal elements of the overlap matrices **Q** and **R**, justifying the ansatz

$$Q_{ij}=Q\delta_{ij}+C(1-\delta_{ij}), \tag{10a}$$

$$R_{in}=R\delta_{in}+S(1-\delta_{in}) \tag{10b}$$

for the student-student overlaps and (apart from a relabeling of the student nodes) student-teacher overlaps, respectively. As one can see from Fig. 1, this approximation is particularly good in the symmetric phase and during the final convergence to perfect generalization.

The reduction of the number of order parameters from $O(K^2)$ to just four simplifies the differential equations and the generalization error significantly (see Appendix B). This allows us to analyze the learning dynamics exactly as a function of the size of the network K, the length of the teacher hidden units T, and the user adjustable training parameters: the learning rate η and the learning temperature β.

A. Symmetric phase and onset of specialization

Numerical integration of the equations of motion for a range of learning scenarios shows that the length of the symmetric phase depends on the number of hidden units K, the anisotropy in the length of the teacher vectors, the choice of the user adjustable parameters η and β, and the anisotropy of the initial conditions. If we assume that the initial conditions are random and K is fixed, the trapping in the symmetric phase is especially prolonged by isotropic teachers and small learning rates η.

Initially, we will therefore study the dynamics (9) analytically in the symmetric phase for isotropic teachers in the small-η regime, where terms proportional to η^2 can be neglected. Later, the effect of a finite learning rate, i.e., including η^2 terms, will be studied analytically for small η and numerically for arbitrary η.

1. Truncated equations

The truncated equations of motion have only one physical fixed point, given by

$$Q_0^*=C_0^*=\frac{T}{K(1+T)-T}, \tag{11a}$$

$$R_0^*=S_0^*=\sqrt{\frac{Q^*T}{K}}=\frac{T}{\sqrt{K[K(1+T)-T]}}, \tag{11b}$$

which is independent of β and therefore identical to the one obtained in [2] for $T=1$. The fixed point can be understood in geometrical terms: the student weight vectors are confined to the subspace spanned by the teacher weight vectors and their projection onto each teacher weight vector is identical. However, this symmetric solution is an unstable fixed point of the dynamics and the small perturbations introduced by the generically nonsymmetric initial conditions will eventually drive the student towards specialization.

To study the onset of specialization, we expand the truncated differential equations to first order in the deviations $q=Q-Q_0^*$, $c=C-C_0^*$, $r=R-R_0^*$, and $s=S-S_0^*$ from the fixed-point values (11). The linearized equations of motion take the form $d\boldsymbol{v}/d\alpha=\mathbf{M}^T\mathbf{v}$, where $\boldsymbol{v}=(r,s,q,c)$ and **M** is a 4×4 matrix whose elements are the first derivatives of the truncated update equations (B2) at the fixed point with respect to $\boldsymbol{v}$. For the onset of specialization only the modes with positive eigenvalue are relevant, being amplified by the dynamics. For them we can identify the inverse eigenvalue as a typical escape time τ_i from the symmetric phase.

For the truncated equations of motion, we find only one relevant perturbation [see Appendix B 1 a, Eqs. (B6) and (B7)] with an associated eigenvector implying $q=c=0$ and $s=-r/(K-1)$, i.e., a pure rotation of the student weight vectors inside the subspace spanned by the teacher weight vectors towards the teacher unit they will specialize on. This can also be confirmed by a closer look at Fig. 1. The onset of specialization is signaled by the breaking of the symmetry between the student-teacher overlaps, whereas significant

differences from the symmetric fixed-point values of the student norms and overlaps occur later. The escape eigenvalue is

$$\lambda_0(\beta)=\frac{2}{\pi}\frac{\eta\beta T^2}{\sqrt{K(1+T)-T}[K(1+T)+\beta T]^{3/2}}. \quad (12)$$

Maximization of $\lambda_0^{\rm opt}(\beta)$ with respect to β yields

$$\beta^{\rm opt}=2\frac{K(1+T)}{T}, \quad (13)$$

i.e., the optimal β scales with the number of hidden units and also grows proportionally to $1/T$ for small teacher lengths. The optimized escape eigenvalue is

$$\lambda_0^{\rm opt}(\beta^{\rm opt})=\frac{4\sqrt{3}}{9\pi}\frac{\eta T}{\sqrt{K(1+T)}\sqrt{K(1+T)-T}}$$
$$=\lambda_0^{\rm opt}(1)\frac{2\sqrt{3}}{9}\frac{[K(1+T)+T]^{3/2}}{T\sqrt{K(1+T)}}. \quad (14)$$

Trapping in the symmetric phase is therefore for very small learning rates always inversely proportional to the learning rate η. It is interesting to study two limiting cases: $K\to\infty$, i.e., large networks, and $T\to 0$, i.e., small teacher weights or nearly linear functions. In these limits, one finds that the escape eigenvalue is $\lambda\propto 1/K^2$ ($\lambda\propto T^2$) for GD, in contrast to $\lambda\propto 1/K$ ($\lambda\propto T$) for optimized ABP, respectively, i.e., in these limits the time spent in the symmetric phase can be reduced by an order of K or $1/T$.

2. Small-η expansion

Numerical integrations of the differential equations (A4) for larger learning rates indicate a reduced optimal value of β, with the ansatz (10) still valid. It is therefore desirable to analyze the symmetric phase for finite learning rates.

Analytically, we can expand the full set of equations (B2) to first order in $\boldsymbol{v}=(r,s,q,c)$ around the fixed point of zeroth order (11) and find its first-order correction in η by solving the resulting set of linear equations. The new fixed point found is still characterized by $Q^*=C^*$ and $R^*=S^*$ [Eq. (B8)]. This is in contradiction to the numerical results, which predict a fixed point with $Q^*>C^*$ and $R^*=S^*$. This contradiction can be resolved by studying the linear dynamics around the new fixed point. An eigenvalue that was marginal ($\lambda_2=0$) for the truncated equations of motions acquires a positive contribution of $O(\eta^2)$ [Eq. (B9)]. The mode associated with this eigenvalue increases differences between Q and C, leading primarily to a growth of the student weight vectors outside the subspace spanned by the teacher weight vectors (see Appendix B 1 C) and no specialization. As these differences are typically large for random initial conditions (unlike differences in R and S), this mode will drive the student quickly away from the fixed-point characterized by $Q^*=C^*$ to one with $Q^*>C^*$, where the student will be trapped until specialization between R and S will occur eventually. Unfortunately, this fixed point cannot be studied analytically, but can, however, be studied numerically.

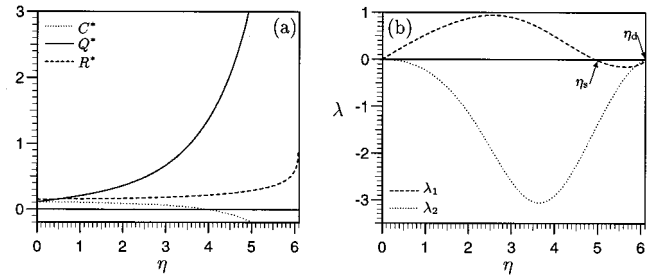

FIG. 2. (a) The symmetric fixed-point values R^*, Q^*, and C^* of the order parameters are shown as a function of the learning rate η at $K=5$ and $T=1$ for $\beta=1$. The values of the order parameters diverge for $\eta\to\eta_d$ (see the text). (b) For the same parameters, the relevant eigenvalues λ_1,λ_2 (see the text) of the linearized dynamics around the (learning-rate-dependent) symmetric fixed point explain the divergent behavior as $\lambda_2(\eta_d)\to 0$. The maximum in λ_1, the eigenvalue that drives the specialization process, defines the optimal learning rate.

3. Numerical finite-η analysis

In Fig. 2(a) we show the order parameter values at the fixed point, which are characterized by $Q^*>C^*$ and $R^*=S^*$ for finite-η values. Whereas R^* is nearly constant over a wide range of learning rates, the value of Q^* increases and C^* decreases rapidly. In fact, as η approaches a certain value, termed here η_d, the values of the order parameters diverge.

This behavior can be understood by linearizing the dynamics around the fixed point and analyzing its eigenvalues [see Fig. 2(b)]. We find two eigenvalues that are always negative and of large magnitude and are therefore irrelevant to the long-term behavior of the dynamics. For the other two eigenvalues one finds that $\lambda_1>0$ and $\lambda_2<0$ for small to intermediate learning rates. The eigenvector associated with λ_1 is in fact identical to the one found for fixed points with $Q^*=C^*$ and corresponds to a pure rotation and instability in R-S space. The eigenvector of λ_2 is also very similar to the eigenvector of the eigenvalue that caused the instability of the $Q^*=C^*$ fixed point in the Q-C space. For increasing learning rate, we first find a global maximum for λ_1 at the optimal learning rate $\eta^{\rm opt}(\beta)$. For even larger learning rates, we find different generic behaviors, depending on the values of the parameters K, T, and β. In general, there are two candidates for a maximal learning rate $\eta_{\rm max}$ identifiable in Fig. 2(b). The first, η_d, corresponds to λ_2 becoming positive, causing an instability in Q-C space and diverging values of the order parameters. The other candidate is given by the learning rate η_s, where λ_1 turns negative and the fixed point becomes attractive. One can identify two phases $\eta_s<\eta_d$ and $\eta_d>\eta_s$ (for which η_s does not actually exist since the fixed point vanishes above η_d). However, in the following we will not distinguish between these two phases, but simply define $\eta_{\rm max}=\min(\eta_d,\eta_s)$.

In order to estimate the potential gain by using ABP in the finite learning rate case, we optimize the dynamics with respect to the learning rate η under the constraint $\beta=1$ (GD) and contrast it with results obtained by optimizing with respect to both the learning rate η and the inverse temperature β (ABP) for a range of K and T values. In Fig. 3 the optimal value of β is shown as a function of both K and T. Figure

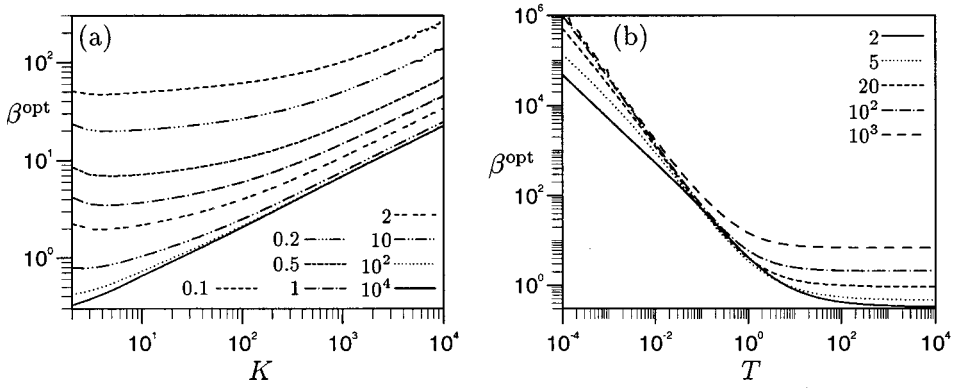

FIG. 3. (a) The optimal inverse temperature $\beta^{\rm opt}$ is shown for various T values (see the legend) as a function of K. For sufficiently large values of TK, $\beta^{\rm opt}$ grows with $\sqrt{K}$. (b) Here $\beta^{\rm opt}$ is shown as a function of T for various K values (see the legend). For small T we find a power-law increase of β with $1/T$ with an exponent that approaches 1 for TK small enough.

3(a) shows that $\beta^{\rm opt}$ increases for growing network size K, as is expected from the small learning rate analysis. However, the size of $\beta^{\rm opt}$ grows significantly slower and becomes dependent on the value of the product TK. For $TK \gg 1$ and $K \to \infty$ one finds $\beta^{\rm opt} \propto \sqrt{K}$, which has to be contrasted with the previously predicted $\beta^{\rm opt} \propto K$ [see Eq. (13)], due to the influence of finite learning rates.

Similarly, as shown in Fig. 3(b), $\beta^{\rm opt}$ grows for decreasing teacher lengths T but remains constant for large T as predicted previously. We find power laws for $T \to 0$, with exponents dependent on the value of TK. For $TK \ll 1$, however, the exponent approaches -1, which is identical to the theoretical prediction in the small-η regime.

Having identified the two interesting regimes where the optimal inverse temperature deviates significantly from its GD value, small teacher weight vectors $T \to 0$ and large networks $K \to \infty$, we investigate the differences in optimal dynamics for GD and ABP further. In Fig. 4 we show the behavior of both the optimal learning rate $\eta^{\rm opt}$ [Figs. 4(a)–4(c)] and the resulting optimal escape eigenvalue $\lambda^{\rm opt}$ [Figs. 4(d)–4(f)] for GD in comparison to ABP for various K-T scenarios.

The optimal learning rate $\eta^{\rm opt}(T)$ of GD, depicted in Fig. 4(a), exhibits a strongly K-dependent limit for large T and a universal limit for small T. In general, $\eta^{\rm opt}(T)$ decreases for increasing T and shows its most volatile behavior in the re-

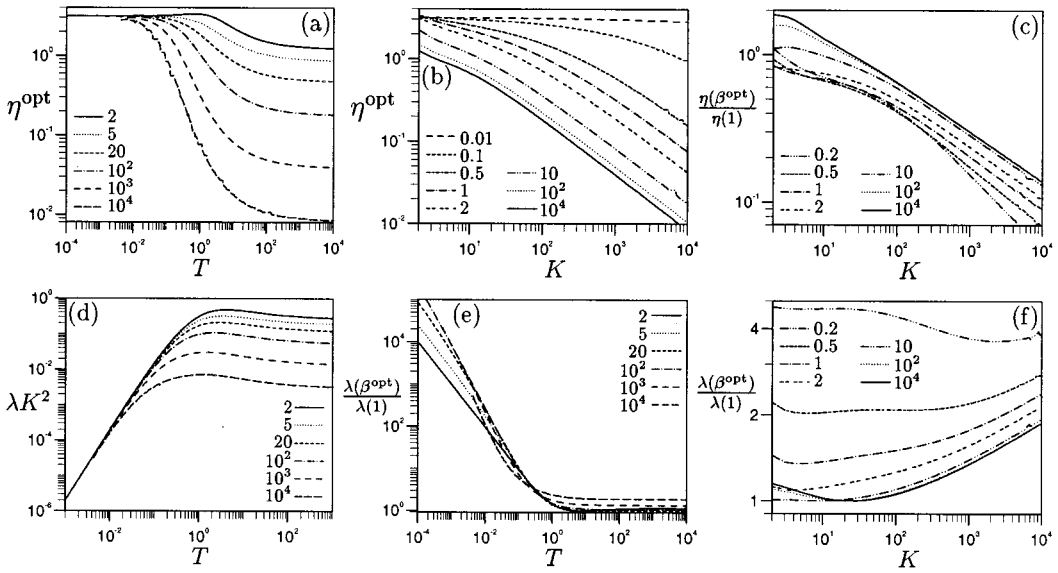

FIG. 4. (a) The optimal learning rate $\eta^{\rm opt}$ for GD descent as a function of T for various K values shows the most volatile behavior for $0.1 \leqslant T \leqslant 10$. (b) $\eta^{\rm opt}(K)$ for several T values shows a power-law decay with exponent $-2/3$ in the large-K limit for $TK \gg 1$. (c) The quotient of the optimal learning rates of ABP and GD as a function of K for various T values shows that $\eta^{\rm opt}(\beta^{\rm opt})$ decays even faster with exponent -1 for large K. (d) The optimal escape eigenvalue for GD multiplied by K^2 as a function of T collapses on a universal (K- independent) curve for small T and decays rapidly with exponent 2. For large T the escape eigenvalue becomes independent of T, but acquire a further K dependence ($\lambda K^2 \propto K^{-2/3}$). (e) The possible gain by using ABP is shown by plotting the quotient of the optimal escape eigenvalue for the two training algorithms. The advantage of ABP is most impressive for small T, where one can gain at least a factor $1/T$ in comparison to GD, depending on the K value (see the legend). (f) The same quotient as a function of K for several T values also shows a power-law gain by using ABP but with a small exponent of 1/6.

gion $0.1 \leqslant T \leqslant 10$ and for large K. These teacher values are the most reasonable for real learning problems, i.e., in practice it will be generally difficult to choose a good learning rate especially for large networks. This picture can be confirmed by examining the influence of K on η^{opt} for GD as shown in Fig. 4(b). For very small T, the learning rate exhibits hardly any dependence on K, whereas for TK large enough, one finds that $\eta^{\mathrm{opt}} \propto K^{-2/3}$.

The behavior of the optimal learning rate for optimized ABP is quite similar to GD. The main difference from GD can be seen in Fig. 4(c), which shows that $\eta^{\mathrm{opt}}(\beta^{\mathrm{opt}})$ decays faster for ABP, with $\eta^{\mathrm{opt}}(\beta^{\mathrm{opt}}) \propto K^{-1}$ for large TK. One also finds that the optimal learning rate saturates for large- and small-T values to K-dependent constants. For large T this may be explained by the fact that the error is limited by the saturation of all units.

The optimized escape eigenvalue, which largely determines the training time spent in the symmetric phase, is shown for GD in Fig. 4(d), where we have multiplied λ^{opt} by K^2 for convenience. For small T, one finds that $\lambda^{\mathrm{opt}}(T)$ collapses on universal curve for all K and we find the same power-law behavior as predicted in the small-η analysis ($\lambda^{\mathrm{opt}} \propto T^2/K^2$) [see Eq. (12)]. For large T, one also finds that λ^{opt} becomes increasingly weakly dependent on T as expected. However, it also shows a further K dependence due to the decay of the optimal learning rate and one finds $\lambda^{\mathrm{opt}} \propto \eta^{\mathrm{opt}}/K^2$.

To highlight the possible gains of using ABP, $\lambda^{\mathrm{opt}}(\beta^{\mathrm{opt}})/\lambda^{\mathrm{opt}}(1)$ is plotted as a function of T and K in Figs. 4(e) and 4(f). In Fig. 4(e) one finds for small T a gain [16] of $1/T$ for $TK \ll 1$, which was predicted from the small-η analysis [see Eq. (14)]. For large K [see Fig. 4(f)] we also find a power-law gain in K for the optimized dynamics, but only for $TK \gg 1$ and with an exponent that is only 1/6, much smaller than the value of 1 predicted previously in Eq. (14). This can be attributed to the slower than predicted increase in β^{opt} and to the smaller optimal learning rate for ABP in this regime.

Of arguably further importance for training is the sensitivity of the choice of the learning rate, especially in the sense of how well the maximal learning rate is separated from its optimal value. Therefore, the normalized difference between the maximal and optimal learning rate $\Delta\eta^{\mathrm{opt}}_{\mathrm{max}} = (\eta_{\mathrm{max}} - \eta^{\mathrm{opt}})/\eta^{\mathrm{opt}}$ is compared for ABP and GD as a function of T for two K values in Fig. 5. Whereas the optimal and maximal learning rates are well separated for all T (and K) for optimized ABP, this is not the case for small T for GD, where one finds a power-law decay of $\Delta\eta^{\mathrm{opt}}_{\mathrm{max}}$ with an exponent that approaches 2/3 for $TK \ll 1$ from above, making an optimal selection of the learning rate increasingly more difficult.

Finally, we would like to compare the symmetric fixed point for the optimized dynamics for finite learning rate with the theoretical values (11) for the truncated equations. Instead of illustrating the behavior graphically, we have summarized the results in Table I. We have found it most illuminating to compare the normalized difference $\hat{P}^* = (P^* - P_0^*)/P_0^*$ for all relevant order parameters (note that the identity $R^* = S^*$ is preserved for finite η) in the various limits. In general, one finds for both algorithms that $Q^* > Q_0^*$ and $R^* > R_0^*$. For C^*, however, one finds a T-dependent behavior with $C^* < C_0^*$ for $T < T_s^{\mathrm{crit}}(K)$ and $C^* > C_0^*$ for $T > T_s^{\mathrm{crit}}(K)$, where $T_s^{\mathrm{crit}} \propto K^{1/3}$ for GD and $T_s^{\mathrm{crit}} \propto K^{1/2}$ for ABP. We furthermore find that the optimal symmetric fixed point for ABP is always significantly closer to the zero learning rate fixed point than for GD.

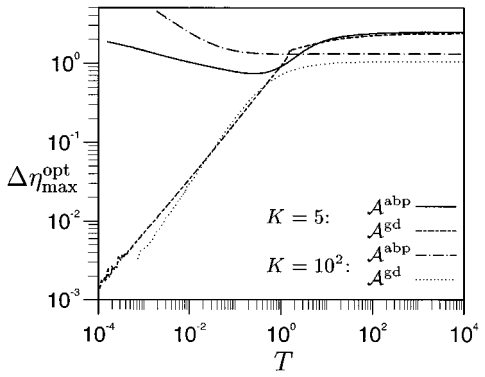

FIG. 5. The normalized difference between the maximal and optimal learning rate $\Delta\eta^{\mathrm{opt}}_{\mathrm{max}} = (\eta_{\mathrm{max}} - \eta^{\mathrm{opt}})/\eta^{\mathrm{opt}}$ is shown for both adaptive back-propagation $\mathcal{A}^{\mathrm{ABP}}$ and gradient descent $\mathcal{A}^{\mathrm{GD}}$ for $K = 5,100$ as a function of T.

Before we turn our attention to the optimization of the dynamics in the convergence phase, we would like to summarize the results obtained so far and put them in the context of previous work. Unlike the small learning rate regime, which has been studied previously for both GD [2] and ABP [14], we find that the amount of training time spent in the symmetric phase actually scales worse than K^2 for the optimal choice of learning parameters (see Table II for an overview of the numerical values of the power laws). This seems to be mainly due to the need for reducing the learning rate η with increasing K. This reduction is arguably caused by the high correlations between student nodes inside and the (mainly uncorrelated) increase of the student lengths Q^* outside the space spanned by the teacher vectors, leading to a discrepency between student and teacher output that increases significantly faster than K for large enough T. For $K \to \infty$ $(TK \gg 0)$, one also finds that the gain, by using the

TABLE I. Symmetric fixed points of the optimized dynamics for both the gradient descent $\mathcal{A}^{\mathrm{GD}}$ and adaptive back-propagation $\mathcal{A}^{\mathrm{ABP}}$ are compared in the limits $T \to 0$ and $K \to \infty$ to the theoretical values for $\eta = 0$ by calculating their normalized difference $\hat{P}^* = (P^* - P_0^*)/P_0^*$. These differences exhibit either power-law behavior, with algorithm-dependent exponents, or saturate at constant limits, whose absolute value may be parameter dependent and are referred to by $c()$. In the limit $T \to \infty$ all parameters exhibit finite limits and are therefore omitted. $T_s^{\mathrm{crit}}(K)$ is defined by $C^* = C_0^*$.

	$T \to 0$ $(TK \ll 1)$		$K \to \infty$ $(TK \gg 1)$	
	$\mathcal{A}^{\mathrm{GD}}$	$\mathcal{A}^{\mathrm{ABP}}$	$\mathcal{A}^{\mathrm{GD}}$	$\mathcal{A}^{\mathrm{ABP}}$
$\hat{Q}^*$	$c(K)$	$T^{0.33\pm3}$	$K^{0.64\pm2}$	$K^{0.48\pm2}$
$\hat{C}^*$	$-c(K)$	$-T^{0.33\pm3}$	$K^{-0.33\pm2}$	$K^{-0.50\pm1}$
$\hat{R}^*$	$T^{1.00\pm1}$	$T^{1.33\pm1}$	$K^{-0.35\pm2}$	$K^{-0.50\pm1}$
T_s^{crit}			$K^{0.31\pm2}$	$K^{0.50\pm1}$

TABLE II. For $T\to 0$ and $K\to\infty$ the optimized dynamics in the symmetric phase show power-law behavior for both the gradient descent $\mathcal{A}^{\mathrm{GD}}$ and adaptive back-propagation $\mathcal{A}^{\mathrm{ABP}}$. The table shows the optimal learning parameters η^{opt} and β, the optimal escape eigenvalue λ^{opt}, and the normalized difference between maximal and optimal learning rate $\Delta\eta^{\mathrm{opt}}_{\mathrm{max}}=(\eta_{\mathrm{max}}-\eta^{\mathrm{opt}})/\eta^{\mathrm{opt}}$. The errors in the exponent are given for the last significant digit only and $c()$ refers to constant limits, whose value is dependent on a parameter.

	$T\to 0$ $(TK\ll 1)$		$K\to\infty$ $(TK\gg 1)$	
	$\mathcal{A}^{\mathrm{GD}}$	$\mathcal{A}^{\mathrm{ABP}}$	$\mathcal{A}^{\mathrm{GD}}$	$\mathcal{A}^{\mathrm{ABP}}$
β^{opt}	1	$T^{-1.00\pm 1}$	1	$K^{0.50\pm 2}$
η^{opt}	π	$c(K)$	$K^{-0.67\pm 3}$	$K^{-1.00\pm 1}$
$\Delta\eta^{\mathrm{opt}}_{\mathrm{max}}$	$T^{0.68\pm 3}$	$c(K)$	$c(T)$	$c(T)$
λ^{opt}	$T^{2.00\pm 1}K^{-2}$	$T^{1.00\pm 1}K^{-2}$	$K^{-2.66\pm 4}$	$K^{-2.50\pm 1}$

optimal ABP choice of $\beta^{\mathrm{opt}}\propto\sqrt{K}$, is only a factor $K^{1/6}$ and not K as predicted previously.

We have furthermore relaxed the constraint $T=1$ used in these works and have found that the optimal learning parameter values change significantly in the most relevant region of teacher lengths, which makes it difficult in practice to choose optimal learning parameters without prior knowledge or estimation of the teacher lengths. For small T, which corresponds to nearly linear (but bounded) rules, one finds that the specialization process is furthermore slowed down by a factor of $1/T^2$ for GD learning. This is arguably due to the fact that the symmetric fixed point is already a very good approximation to the true function and information about the nonlinearities is scarce. In this regime the optimal choice of $\beta^{\mathrm{opt}}\propto 1/T$ helps the student significantly in breaking the symmetry by reducing the region of hidden unit activation relevant for training and favoring rotational over longitudinal changes. The gain achievable in this regime is of order $1/T$.

B. Convergence to optimal generalization

In order to predict the optimal learning rate η^{opt} and inverse temperature β^{opt} for the convergence phase, we linearize the reduced set of equations of motion (B2) in $\{R,Q,C,S\}$ around the zero generalization error fixed point $R^*=Q^*=T$ and $S^*=C^*=0$ (see Appendix).

The matrix $\mathbf{M}$ of the resulting system of four coupled linear differential equations in $r=T-R$, $q=T-Q$, $s=S$, and $c=C$ has two pairs of eigenvalues ($\lambda_{1,2}$ and $\lambda_{3,4}$) that are solutions of quadratic equations (B13). The dependence of these eigenvalues on the learning rate is illustrated in Fig. 6(a) for $K=3$ and $T=1$. The eigenvalues $\lambda_{3,4}$ are linear in η, whereas $\lambda_{1,2}$ have higher orders in η. One further can distinguish between two slow modes associated with eigenvalues λ_1 and λ_3 and two fast modes associated with eigenvalues λ_2 and λ_4, which are negative for all learning rates and whose magnitude is significantly larger in the region of interesting η. The fast modes decay quickly and their influence on the long-time dynamics is negligible. The dependence of the two relevant eigenvalues λ_1 and λ_3 on η and β is more closely illustrated in Fig. 6(b) in the same learning scenario and for two β values. As mentioned, the eigenvalue λ_3 is negative and linear in η, whereas the eigenvalue λ_1 is a nonlinear function of η and negative for small η. For large η, λ_1 becomes positive and training does not converge to the optimal solution defining the maximum learning rate η_{max} as $\lambda_1(\eta_{\mathrm{max}})=0$. For all $\eta<\eta_{\mathrm{max}}$ the generalization error decays exponentionally to $\epsilon_g^*=0$.

In order to identify the optimal convergence eigenvalue λ^{opt}, which is the eigenvalue associated with the slowest decay mode, we expand the generalization error to second order in r, q, s, and c [Eq. (B10)]. We find that the eigenvector (B14) associated with the linear eigenvalue λ_3 is orthogonal to the first-order terms in the generalization error and therefore cannot contribute to their decay, but controls only the decay of second-order terms with $2\lambda_3$. The learning rate η^{opt} that provides the fastest asymptotic decay rate λ^{opt} of the generalization error is therefore given by the condition

$$\lambda^{\mathrm{opt}}=\left|\min_{\eta}[\max(\lambda_1,2\lambda_3)]\right|. \tag{15}$$

This means either $\lambda_1(\eta_r^{\mathrm{opt}})=2\lambda_3(\eta_r^{\mathrm{opt}})$ or $\min_\eta(\lambda_1)$ if $\lambda_1(\eta_m^{\mathrm{opt}})>2\lambda_3(\eta_m^{\mathrm{opt}})$, where η_m^{opt} is the learning rate at the minimum of λ_1. Examples for both two cases can be seen in Fig. 6(b).

For given K, one finds that for GD ($\beta=1$) the optimal learning rate is at the minimum of λ_1 for $T<T_c^{\mathrm{crit}}(K)$ and by $\lambda_1=2\lambda_3$ otherwise, where $T_c^{\mathrm{crit}}(K)$ is a function weakly de-

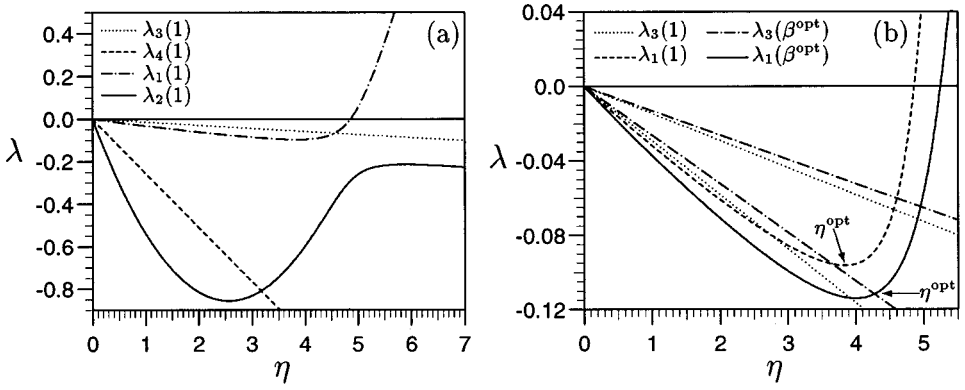

FIG. 6. (a) The four eigenvalues λ_i for gradient descent ($\beta=1$) as a function of the learning rate η at $K=3$ and $T=1$. (b) The two relevant eigenvalues (see the text) λ_1 and λ_3 in the same scenario values of β: $\beta=1$ and $\beta=\beta^{\mathrm{opt}}=1.8314$. For comparison we plot $2\lambda_3$ and find that the optimal learning rate η^{opt} is given by the condition $\lambda_1=2\lambda_3$ for β^{opt}, but by the minimum of λ_1 for $\beta=1$.

TABLE III. For $T\to 0$ and $T\to\infty$ the optimized dynamics in the convergence phase show power-law behavior for both gradient descent $\mathcal{A}^{\mathrm{GD}}$ and adaptive back-propagation $\mathcal{A}^{\mathrm{ABP}}$. The table shows the optimal learning parameters η^{opt} and β, the optimal convergence eigenvalue λ^{opt}, and the normalized difference between maximal and optimal learning rate $\Delta\eta^{\mathrm{opt}}_{\mathrm{max}}=(\eta_{\mathrm{max}}-\eta^{\mathrm{opt}})/\eta^{\mathrm{opt}}$. $c()$ refers to constant limits, whose value is dependent on a parameter.

	$T\to 0$		$T\to\infty$ (K finite)		$T\to\infty$ [$TK^{-1}=O(1)$]	
	$\mathcal{A}^{\mathrm{GD}}$	$\mathcal{A}^{\mathrm{ABP}}$	$\mathcal{A}^{\mathrm{GD}}$	$\mathcal{A}^{\mathrm{ABP}}$	$\mathcal{A}^{\mathrm{GD}}$	$\mathcal{A}^{\mathrm{ABP}}$
β^{opt}	1	T^{-1}	1	$T^{-1/3}$	1	$\frac{1}{3}$
η^{opt}	π	$c(K)$	K^1	K^1	$T^{1/2}$	$T^{1/2}$
$\Delta\eta^{\mathrm{opt}}_{\mathrm{max}}$	T^1	$T^{1/2}$	T^{-1}	T^{-1}	T^{-1}	T^{-1}
λ^{opt}	T^2K^{-1}	T^1K^{-1}	$T^{-3/2}$	$T^{-3/2}$	$T^{-1}K^{-1}$	$T^{-1}K^{-1}$

pendent on K and $T_c^{\mathrm{crit}}(\infty)=1.2780$ [see also Fig. 8(c)]. For optimized ABP, where the decay rate $\lambda^{\mathrm{opt}}(\beta)$ has been maximized with respect to β, the optimal learning rate is given by the root of $\lambda_1-2\lambda_3$ for all values of T.

Both these optimizations are analytically unfeasible for arbitrary K and T. However, for some special cases further analytical progress can be made: $K\to\infty$, $T\to\infty$, and $T\to 0$. These cases are studied in detail in Appendices B 2 a–B 2 c. The resulting power laws will be referred to in the discussion of the appropriate figures and are summarized for all relevant scenarios in Table III.

As in the symmetric phase, one expects the largest gains by using ABP in regions of T-K space, where β^{opt} deviates significantly from 1. In Fig. 7 the optimal value of β is shown as a function of both K and T. Figure 7(a) shows that β^{opt} is only a weak function of K and does not change its order for $K\to\infty$ unlike in the symmetric phase. The only significant K dependence is found for large T and small K.

This should be contrasted to the strong T dependence of β^{opt} depicted in Fig. 7(b), where the theoretical results for $K\to\infty$ are included as well. For small T one finds to leading order $\beta^{\mathrm{opt}}=2/T$, independent of K, whereas a strong dependence of K on β^{opt} is found for large T. For finite K or $T/K\gg 1$, one finds $\beta^{\mathrm{opt}}\propto T^{-1/3}$, whereas $\beta^{\mathrm{opt}}\approx 1/3$ for $T/K\leqslant O(1)$. The qualitative difference of learning for finite and infinite K in the large-T limit will become clear later.

Again, we would like to assess the potential benefits of ABP over GD. Note the discrepency between our results and those previously presented [2] for GD in the convergence phase for the special case $T=1$, where an approximation by reducing the dynamics to the q-r space was employed, producing inaccurate results.

In Fig. 8 we therefore show the behavior of both the optimal learning rate η^{opt} [Figs. 8(a) and 8(b)] and the resulting optimal convergence eigenvalue λ^{opt} [Figs. 8(d) and 8(e)] for GD in comparison to ABP as a function of T for several values of K, including the dominant term for $K\to\infty$. The optimal learning rate $\eta^{\mathrm{opt}}(T)$ of GD depicted in Fig. 8(a) has a universal limit of π for small T identical to the symmetric phase. For large T the limit becomes strongly dependent on K. Again, there exists a qualitative difference between finite K, where one finds analytically $\eta^{\mathrm{opt}}\propto K$ for $T\to\infty$ and infinite K, where $\eta^{\mathrm{opt}}\propto\sqrt{T}$.

The quotient between the optimal learning rates of ABP and GD in Fig. 8(b) shows no significant difference, in stark contrast to results in the symmetric phase. In general, one finds that the learning rate for ABP is larger than for GD when $\beta^{\mathrm{opt}}>1$ and vice versa. For small T the optimal learning rate approaches $\sqrt{3}\pi$ for infinite K [Eq. (B22c)] with minor corrections for finite K [Eq. (B26c)]. For large T, the difference is a factor of $1/\sqrt{2}$ for infinite K, whereas they are identical for finite K.

The kink in the curves around $T\approx 1$ can be explained by the fact that the condition that defines η^{opt} for GD changes at that point (see above). The corresponding critical teacher value $T_c^{\mathrm{crit}}(K)$ is shown in Fig. 8(c).

The optimized convergence eigenvalue, which largely determines the training time spent achieving an acceptable generalization error, is shown for GD in Fig. 8(d), where we have multiplied λ^{opt} by K for convenience. For small T, one finds that λ^{opt} collapses on a universal curve ($\lambda^{\mathrm{opt}}\propto T^2/K$), similar to its symmetric phase behavior. For large T, the behavior for λ^{opt} depends significantly on the order of K as that of the learning rate. Analytically, one finds for K finite

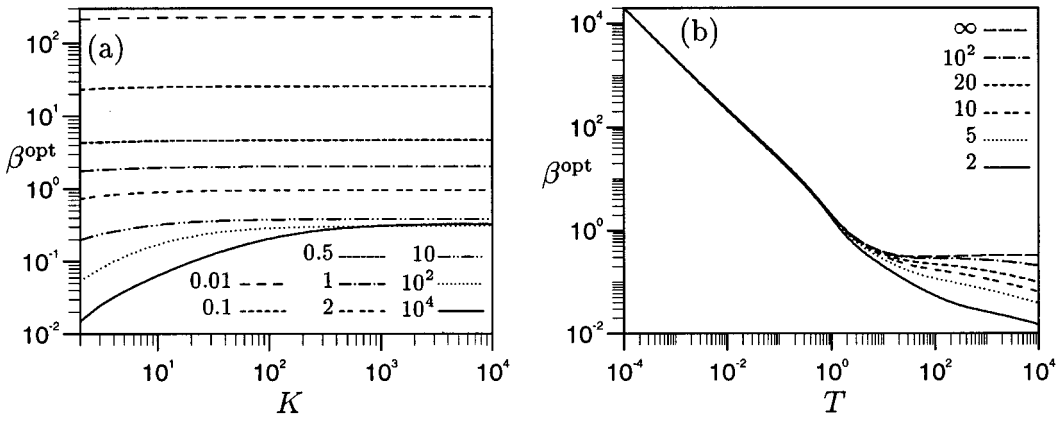

FIG. 7. (a) The optimal inverse temperature β^{opt} is shown for various T values (see the legend) as a function of K. It exhibits only a significant K dependence for large T. (b) β^{opt} is shown as a function of T for various K values (see the legend), including the dominant term for $K\to\infty$. For small T, we find a power-law increase of β with $1/T$ independent of K. For large T, the behavior of β strongly depends on K.

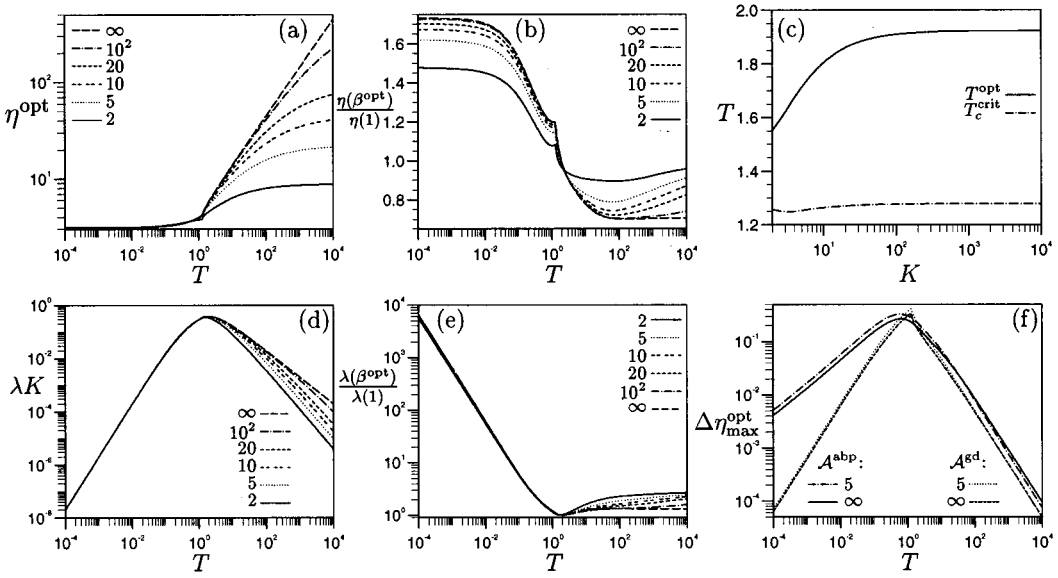

FIG. 8. (a) The optimal learning rate $\eta^{\rm opt}$ for GD as a function of T for various K values shows a significant increase for large T and K. (b) The quotient of the optimal learning rates of ABP and GD as a function of T for various K shows no significant difference in the learning rates of the two algorithms. (c) The teacher length $T_c^{\rm crit}(K)$, where the optimal learning rate changes from the minimum of λ_1 to the root of $\lambda_1 - 2\lambda_3$, and the teacher length $T^{\rm opt}(K)$, where the convergence rate λ takes its global minimum. The latter coincides with $\beta^{\rm opt}=1$ for all K. (d) The optimal convergence rate for GD multiplied by K as a function of T collapses on a universal (K-independent) curve for small T and decays rapidly with exponent 2 as in the symmetric phase. For large T, the convergence rate also decays in T, but with an exponent that seems to be K dependent. (e) The possible gain by using ABP is shown by plotting the quotient of the optimal convergence eigenvalue for the two training algorithms. The advantage of ABP is most impressive for small T, where one can gain a K-independent factor $1/T$ in comparison to gradient descent. For large T the gain is K dependent but constant in T. (f) The normalized difference between the maximal and optimal learning rate $\Delta\eta_{\rm max}^{\rm opt}$ is shown for both adaptive back-propagation $\mathcal{A}^{\rm ABP}$ and gradient descent $\mathcal{A}^{\rm GD}$ for $K=5,\infty$ as a function of T. For both small and large T one finds power-law behavior.

and $TK \gg 1$ that $\lambda^{\rm opt}$ is actually independent of K and decreases proportionally to $T^{3/2}$. For large T and $T/K = O(1)$, on the other hand, the scaling is $\lambda \propto 1/({\rm TK})$.

To highlight the possible gains from using ABP, $\lambda^{\rm opt}(\beta^{\rm opt})/\lambda^{\rm opt}(1)$ is plotted as a function of T in Fig. 8(e). For small T, one finds as in the symmetric phase that ABP gains a factor $1/T$, with only a very weak K dependence due to corrections in the $1/K$ dependence for ABP. For large T, one finds only a constant gain for ABP, which ranges between 1.299 and 2.828 depending on the values of T and K, although $\beta^{\rm opt}$ deviates significantly from 1 for finite K.

A question one could ask is which teacher length $T^{\rm opt}$ maximized $\lambda^{\rm opt}$ for given K. This turns out to be identical for both algorithms [$\beta^{\rm opt}(T^{\rm opt})=1$] and its dependence on K is shown in Fig. 8(c). Although only of academic interest as T is given by the rule to be learned, it nevertheless presents some interesting insights. ABP effectively deforms the search space via the single parameter β to compensate for the anisotropy of the generalization error surface. At $T^{\rm opt}$ no useful deformation can be obtained by using $\beta \neq 1$, leaving room for speculation whether isotropy is recovered. Other methods for deforming the search space based on information geometry have been introduced recently and involve more complicated learning rules, which may not always be tractable [5].

In Fig. 8(f) the normalized separation between the maximal and optimal learning rate shows for both algorithms only a very weak dependence on K in comparison to T. The gap is largest for $T = O(1)$, the region of most likely T values, with a maximal separation around 30% for both algorithms, which is significantly smaller than the separation in the symmetric phase. For both large and small T, we find decays of the normalized gap in T. For large T, the decay is proportional to $1/T$ for both algorithm, with slight differences in the constant prefactor. For small T, however, the behavior is algorithm dependent, with a decay proportional to T for GD proportional to $\sqrt{T}$ for ABP.

As in the symmetric phase, the extension of the analysis to the full R-Q-S-C space and arbitrary T values has revealed several insights. The normalization for the soft-committee machine chosen here leads to the optimal learning rate for both algorithms (and the optimal inverse temperature for ABP) being only weakly dependent on K in most practical learning scenarios, suggesting a similar scaling for applied networks. For large K one finds furthermore that the training time scales with K in almost all cases, in contrast to the symmetric phase, reflecting the fact that the student hidden units have already specialized on a particular teacher hidden unit.

For extreme values of T, one finds further interesting effects. For small T, GD training is slowed down by a further factor of $1/T^2$, which can be reduced to a factor of $1/T$ by the optimal choice of $\beta^{\rm opt} \propto 1/T$, similar to the symmetric phase.

For large T, one has to distinguish between two regimes. For finite K, both the mapping of the network and the error signal become increasingly discrete in this limit, leading to an architecture similar to a committee machine. In this case,

the error signal is of $O(1/K)$ leading to a rescaling of the learning rate with K, in order to keep the weight update constant for all network sizes, making the convergence rate independent of K. The increasingly discrete nature of the error signal, however, seems responsible for the decrease in the convergence rate by $T^{-3/2}$ for both algorithms. The possible gain of ABP stays constant in this limit, in spite of the significant scaling of $\beta^{\mathrm{opt}} \propto T^{-1/3}$.

In the limit where K grows simultaneously with T, one finds a qualitatively different behavior. This can be explained by the smoothness of the network output and the error signal in this case due to the fact that hidden units outputs are discrete but uncorrelated, giving rise to a Gaussian output distribution (central limit theorem).

V. SUMMARY AND DISCUSSION

This research has been initially motivated by the dominance of the suboptimal symmetric phase in on-line learning of two-layer feedforward networks trained by gradient descent [2]. We proposed an adaptive back-propagation training algorithm [Eq. (7)] parametrized by an inverse temperature β. For $\beta=1$ standard back-propagation or GD is recovered, whereas $\beta=0$ corresponds to a generalized Hebb rule.

ABP is designed to deform search space using the single parameter β. For $\beta>1$, the specialization of the student nodes is improved by enhancing differences in the activation between hidden units. In this region, the achievable learning rate is usually higher than for GD, leading effectively to favoring rotational changes of the weight vector over length changes. For $0<\beta<1$, we find the opposite effect, as the activation region of the student relevant for training is increased and the learning rate decreased, causing an enhancement of length changes. Its performance has been compared to GD for a normalized soft-committee student network with K hidden units learning a rule defined by an isotropic teacher ($T_{nm}=T\delta_{nm}$) of the same architecture. Furthermore, the introduction of a natural normalization of the soft-committee machine leads to more elegant results as it eliminates the unnatural scaling of the learning rate with the input dimension N and, in many cases, with the number of hidden units K, which is a feature of the unnormalized model and suggests a similar approach for real world networks.

For both relevant phases of learning, the symmetric and convergence phase, this work extends previous results [2,14] substantially by addressing the influence of finite learning rates in the symmetric phase and the influence of the teacher length T on the dynamics. The analysis identifies three interesting regimes: large K, small T, and large T.

A. Large K

For large K, the linear analysis of the equations of motion around the symmetric fixed point for small learning rates suggests that the trapping time is inversely proportional to the learning rate and grows $\tau \propto K^2$ for GD [17] and $\tau \propto K$ for optimized ABP with $\beta^{\mathrm{opt}} \propto K$. This suggests that for increasing network size it seems to become harder for a student node to distinguish between the many teacher nodes and to specialize on one of them. This is reflected by the decrease in the squared student length $Q^* \propto 1/K$ at the symmetric fixed point, pushing the student hidden nodes into the linear region of the sigmoidal activation function, where differentiation is more difficult.

This picture is altered significantly when accounting for finite learning rate effects, due to the decrease in the optimal learning rate η^{opt} with K, beyond the rescaling implicit in the network normalization. This rescaling assumes an unnormalized network output of $O(\sqrt{K})$ and a typical squared error of $O(K)$, which is appropriate in the case when the hidden units of both the student and the teacher network are uncorrelated. However, in the symmetric phase this is not the case for the student network leading to errors that grow faster than $O(K)$ and making a decrease in the learning rate necessary. The significant reduction of the learning rate may also be associated with the need to limit the proportion of the student length outside the space spanned by the teacher for large K.

The actual training time spent in the symmetric phase therefore scales $\tau \propto K^{8/3}$ for GD and $\tau \propto K^{5/2}$ for ABP, reducing the benefit of an adjustable temperature to $K^{1/6}$. One also finds that the scaling for the optimal temperature changes to $\beta^{\mathrm{opt}} \propto \sqrt{K}$ in this limit.

For the convergence phase one finds that the training time scales with K in almost all cases, reflecting the fact that the learning rate must (implicitly) be rescaled by $1/K$ as the typical quadratic deviation between teacher and student output increases proportionally to K. The optimal inverse temperature and the optimal gain of using ABP in this regime are dependent on T but remain constant for large K due to the fact that each student hidden unit is already specialized on one teacher unit and the effect of other units in inhibiting further specialization is negligible.

These results mean that most of the training time is spent in the symmetric phase (or search regime) for large networks, at least in learning scenarios with a certain amount of symmetry. This suggests that considerably more effort should be directed towards developing algorithms, which can significantly reduce the training time in this phase, than towards fine tuning of the asymptotic convergence.

B. Small T

In the small-T limit, one finds very similar results for both the symmetric and the convergence phases, e.g., the optimal learning rate is universally π for GD, the optimal inverse temperature has the same scaling behavior ($\beta^{\mathrm{opt}} \propto 1/T$), and the optimal escape and the optimal convergence eigenvalue scale with T^2 for GD and with T for ABP in both learning phases. This results in a gain of order $1/T$, in using ABP, for the whole training process.

The universal slowdown of learning in the small-T limit may be explained by the fact that the learning rule becomes increasingly linear, resulting in a very flat (generalization) error surface between the symmetric and the zero-generalization error fixed point. The major difference is the scaling of the relevant eigenvalue with the number of hidden units K, reflecting the lesser degree of confusion once the hidden unit symmetry is broken.

C. Large T

For large T the picture is not as coherent, which can be explained by the increasingly binary nature of the hidden

unit outputs. In the symmetric phase, the outputs of the student hidden units are highly correlated, whereas the outputs of the teacher hidden units are uncorrelated, leading to large errors between the student and teacher network output that scale with K but saturate for large T, explaining the large changes in the optimal learning parameters for medium T but also their indifference to further increases in T once T is sufficiently large.

In the convergence phase, a significantly different behavior is observed for the two cases of finite K and infinite K, where the network output is discrete and continuous, respectively. For infinite K, the error remains smooth and actually decreases for large T due to the increasingly binary hidden unit output, giving rise to an increase of $\eta^{\mathrm{opt}} \propto T^{1/2}$. For finite K, one typically finds that at most one student hidden unit ''misclassifies'' the output of the corresponding hidden unit of the teacher, causing a discrete error of either 0 or $1/K$ and leading to a rescaling of the learning rate proportional to K.

It would be quite interesting to study this limit more closely due to its similarity to the committee machine. The possibility of tuning the weight function with β between a Hebb-like form for $\beta=0$ and a Gaussian form for finite β may give some idea about successful training algorithms for binary networks.

However, throughout our analyses we have implicitly assumed that the decay or increase in the exponential terms outstrips any algebraic variation in the prefactors and all optimizations were carried out under this assumption. This is reasonable at least for medium values of T, which are most likely to be encountered practically, but probably also for any finite values of T. For infinite T, i.e., networks with discrete hidden units, this ansatz is, however, insufficient as the exponential term vanishes and the dynamics become algebraic in α.

In principle, one could encompass these limiting cases by incorporating second-order terms of the Taylor series around the fixed points and solving the resulting set of nonlinear differential equations by transforming them into matrix Riccati equations. Although this is in principle feasible, it goes beyond the scope of this paper.

D. Conclusions

This paper has shown the learning performance limitations of gradient descent in the on-line learning paradigm. Within the model studied one finds severe drawbacks of GD, especially in the symmetric phase, which dominates the learning process for large networks. The suggested adaptive back-propagation algorithm generally speeds up the training process considerably if its extra parameter, the inverse temperature β, is chosen optimally. It has provided us also with some insight into the shortcomings of GD and has outlined possible further research directions.

The relaxation of the constraint $T=1$ has shown that the optimal learning parameter values change significantly in the region of usually relevant teacher lengths and between the symmetric and the convergence phase, making it difficult to choose good learning parameters, i.e., the learning rate η and the inverse temperature β, in practice without prior knowledge or estimation of the teacher lengths and the progress made in learning. This should encourage more research into reliable on-line estimation of optimal learning parameters. It further suggests that the selection of individual learning parameters for each hidden node of the network could potentially be hugely beneficial [10]. We therefore hope that this work will motivate further research into the efficiency of on-line learning training algorithms and their systematic improvement.

ACKNOWLEDGMENTS

A.H.L.W. would like to acknowledge gratefully financial support by the EPSRC, a research scholarship from the Department of Physics of the University of Edinburgh, and the financial support and hospitality of the Neural Computing Research Group at Aston University, where part of this research was carried out. This research was further supported by EPSRC Grant No. GR/L19232.

APPENDIX A: DYNAMICAL EQUATIONS

The generalization error is calculated by averaging the quadratic loss function (1) explicitly over the activations $\{\boldsymbol{x},\boldsymbol{y}\}$ (and implicitly over all inputs), which are multivariate Gaussian distributed with zero mean and covariance matrix $\mathcal{C}$ given by

$$\mathcal{C}=\begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{R}^T & \mathbf{T} \end{bmatrix}. \tag{A1}$$

In the following all averages are taken with respect to this distribution and making use of the convention that indices i,j,k,l and n,m label student and teacher nodes, respectively.

The generalization error then takes the form

$$\epsilon_g=\frac{\gamma^2}{2K}\left\{\frac{K}{M}\sum_{n,m=1}^{M} J_2(n,m)-2\sqrt{\frac{K}{M}}\sum_{i,n=1}^{K,M} J_2(i,n) + \sum_{i,j=1}^{K} J_2(i,j)\right\}, \tag{A2}$$

with the integral $J_2(1,2)=\langle g(u_1)g(u_2)\rangle$, where u_i represent members of $\{\boldsymbol{x},\boldsymbol{y}\}$ and we denote with I_d, J_d averages over d variables with one and two g terms, respectively. The integral $J_2()$ can be calculated analytically for the generalized error function $g_\nu(u)=\mathrm{erf}(\nu u/\sqrt{2})$ giving

$$J_2(1,2)=\frac{2}{\pi}\arcsin\left(\frac{\nu^2 C_{12}}{\sqrt{1+\nu^2 C_{11}}\sqrt{1+\nu^2 C_{22}}}\right). \tag{A3}$$

The dependence of the integral on the sigmoidal gain ν can be absorbed by redefining

$$\widetilde{C}_{ij}=\nu^2 C_{ij},$$

a rescaling that also holds for the other integrals below. To evaluate an integral explicitly, the full covariance matrix $\mathcal{C}$ is projected into the relevant subspace. For example, the relevant elements for $J_2(i,n)$ are $C_{11}=Q_{ii}$, $C_{12}=R_{in}$, and $C_{22}=T_{nn}$. It is a property of multivariate Gaussian distributions [2] that integrals of reduced dimensionality such as

$J_2(1,1)$ are generated from the general form $J_2(1,2)$ by the appropriate constraints (in this case $C_{11}=C_{12}=C_{22}$).

The differential equations for **R** and **Q** are calculated similarly and take the form

$$\frac{dR_{in}}{d\alpha}=\frac{\eta\gamma^2}{K}\left\{\sqrt{\frac{K}{M}}\sum_{m=1}^{M} I_3(i,n,m)-\sum_{k=1}^{K} I_3(i,n,k)\right\}, \quad \text{(A4a)}$$

$$\begin{aligned}\frac{dQ_{ij}}{d\alpha}=&\frac{\eta\gamma^2}{K}\left\{\sqrt{\frac{K}{M}}\sum_{m=1}^{M} I_3(i,j,m)+I_3(j,i,m)\right.\\&\left.-\sum_{k=1}^{K} I_3(i,j,k)+I_3(j,i,k)\right\}\\&+\left(\frac{\eta\gamma^2}{K}\right)^2\left\{\frac{K}{M}\sum_{n,m=1}^{M} J_4(i,j,n,m)\right.\\&\left.-2\sqrt{\frac{K}{M}}\sum_{k,n=1}^{K,M} J_4(i,j,k,n)+\sum_{k,l=1}^{K} J_4(i,j,k,l)\right\},\end{aligned} \quad \text{(A4b)}$$

with the integrals $I_3(1,2,3)=\langle g'(u_1)u_2 g(u_3)\rangle$ and $J_4(1,2,3,4)=\langle g'(u_1)g'(u_2)g(u_3)g(u_4)\rangle$. Again for the above choice of sigmoidal transfer function, these integrals can be calculated analytically. We find

$$I_3(1,2,3)=\frac{2}{\pi}\frac{\Psi_{12}(\beta)}{\sqrt{\Psi_{13}(1)}}\frac{\Gamma_3}{\psi_1(\beta)}, \quad \text{(A5a)}$$

$$J_4(1,2,3,4)=\left(\frac{2}{\pi}\right)^2\frac{\nu^2}{\sqrt{\Psi_{12}(\beta)}}\arcsin\left(\frac{\widetilde{C}'_{34}}{\sqrt{1+\widetilde{C}'_{33}}\sqrt{1+\widetilde{C}'_{44}}}\right), \quad \text{(A5b)}$$

where we have conveniently defined

$$\psi_i(\beta)=1+\beta\widetilde{C}_{ii}, \quad \psi_{ij}(\beta)=\beta\widetilde{C}_{ij}$$

$$\Psi_{ij}(\cdots)=\psi_i(\beta)\psi_j(\cdots)-\psi_{ij}(\beta)\psi_{ij}(\cdots)$$

$$\Phi_i=\frac{\psi_2(\beta)\widetilde{C}_{1i}-\psi_{12}(\beta)\widetilde{C}_{2i}}{\Psi_{12}(\beta)},$$

$$\Gamma_i=\frac{\psi_1(\beta)\widetilde{C}_{2i}-\psi_{12}(\beta)\widetilde{C}_{1i}}{\Psi_{12}(\beta)},$$

$$\widetilde{C}'_{ij}=\widetilde{C}_{ij}-\beta[\widetilde{C}_{1i}\Phi_j+\widetilde{C}_{2i}\Gamma_j],$$

with $(\cdots)$ representing either β or 1. Again, one infers the elements of the reduced covariance matrix using the unit labeling convention and the appropriate dimensionality reduction.

One can see that the only role of the gain ν is an explicit rescaling of all order parameters by a factor ν^2 and an implicit rescaling of the learning rate η by ν^2 in the differential equations (A4). The learning rate is further rescaled by the linear output gain by γ^2. In combination with the input variance σ^2, the overall rescaling for any order parameter P and the learning rate η becomes

$$\widetilde{P}=\nu^2\sigma^2 P, \quad \widetilde{\eta}=\frac{\nu^2\gamma^2\sigma^2}{K}\eta. \quad \text{(A6)}$$

In the remainder of the paper we will therefore set $\nu=\gamma=\sigma=1$ without loss of generality.

APPENDIX B: REDUCED EQUATIONS

Reducing the free parameters for $K=M$ and $T_{nm}=T\delta_{nm}$ with the ansatz (10) to just R, S, Q, and C simplifies the generalization error (A2) to

$$\begin{aligned}\epsilon_g=\frac{1}{\pi}&\left\{\arcsin\left(\frac{T}{1+T}\right)-2\arcsin\left(\frac{R}{\sqrt{1+Q}\sqrt{1+T}}\right)-2(K-1)\arcsin\left(\frac{S}{\sqrt{1+Q}\sqrt{1+T}}\right)\right.\\&\left.+(K-1)\arcsin\left(\frac{C}{1+Q}\right)+\arcsin\left(\frac{Q}{1+Q}\right)\right\}.\end{aligned} \quad \text{(B1)}$$

The differential equations for R, S, Q and C are determined from Eq. (A4) similarly and take the form

$$\frac{dR}{d\alpha}=\frac{2}{\pi}\frac{\eta}{K}\frac{1}{\gamma_1}\left\{\frac{\mathcal{R}_0-\gamma_1}{\sqrt{\mathcal{R}_0}}-\frac{R}{\sqrt{\mathcal{Q}_0}}-(K-1)\left[\frac{\beta RS}{\sqrt{\mathcal{S}_0}}+\frac{S\gamma_1-\beta RC}{\sqrt{\mathcal{C}_0}}\right]\right\}, \quad \text{(B2a)}$$

$$\frac{dS}{d\alpha}=\frac{2}{\pi}\frac{\eta}{K}\frac{1}{\gamma_1}\left\{\frac{\mathcal{S}_0-\gamma_1}{\sqrt{\mathcal{S}_0}}-\frac{R\gamma_1-\beta SC}{\sqrt{\mathcal{C}_0}}-\frac{\beta RS}{\sqrt{\mathcal{R}_0}}-\frac{S}{\sqrt{\mathcal{Q}_0}}-(K-2)\left[\frac{\beta S^2}{\sqrt{\mathcal{S}_0}}+\frac{S\gamma_1}{\sqrt{\mathcal{C}_0}}\right]\right\}, \quad \text{(B2b)}$$

$$\frac{dQ}{d\alpha}=\frac{4}{\pi}\frac{\eta}{K}\frac{1}{\gamma_1}\left\{\frac{R}{\sqrt{\mathcal{R}_0}}-\frac{Q}{\sqrt{\mathcal{Q}_0}}+(K-1)\left[\frac{S}{\sqrt{\mathcal{S}_0}}-\frac{C}{\sqrt{\mathcal{C}_0}}\right]\right\}+\frac{4}{\pi^2}\frac{\eta^2}{K^2}\frac{1}{\gamma_2}\left\{\arcsin\left(\frac{\mathcal{R}_1-\gamma_2}{\mathcal{R}_1}\right)\right.$$

$$-2\arcsin\left(\frac{R}{\sqrt{\mathcal{Q}_1\mathcal{R}_1}}\right)+\arcsin\left(\frac{Q}{\mathcal{Q}_1}\right)+(K-1)\left[2\arcsin\left(\frac{C}{\sqrt{\mathcal{Q}_1\mathcal{C}_1}}\right)-2\arcsin\left(\frac{S}{\sqrt{\mathcal{Q}_1\mathcal{S}_1}}\right)-2\arcsin\left(\frac{S\gamma_2-2\beta RC}{\sqrt{\mathcal{R}_1\mathcal{C}_1}}\right)\right.$$
$$\left.-2\arcsin\left(\frac{R\gamma_2-2\beta SC}{\sqrt{\mathcal{S}_1\mathcal{C}_1}}\right)-2\arcsin\left(\frac{2\beta RS}{\sqrt{\mathcal{R}_1\mathcal{S}_1}}\right)+\arcsin\left(\frac{\mathcal{C}_1-\gamma_2}{\mathcal{C}_1}\right)+\arcsin\left(\frac{\mathcal{S}_1-\gamma_2}{\mathcal{S}_1}\right)\right]$$
$$+(K-1)(K-2)\left[\arcsin\left(\frac{C(\gamma_2-2\beta C)}{\mathcal{C}_1}\right)-2\arcsin\left(\frac{S(\gamma_2-2\beta C)}{\sqrt{\mathcal{S}_1\mathcal{C}_1}}\right)-\arcsin\left(\frac{2\beta S^2}{\mathcal{S}_1}\right)\right], \tag{B2c}$$

$$\frac{dC}{d\alpha}=\frac{4}{\pi}\frac{\eta}{K}\frac{1}{\gamma_1}\left\{\frac{R\gamma_1-\beta SC}{\sqrt{\mathcal{S}_0}}-\frac{Q\gamma_1-\beta C^2}{\sqrt{\mathcal{C}_0}}+\frac{S\gamma_1-\beta RC}{\sqrt{\mathcal{R}_0}}-\frac{C}{\sqrt{\mathcal{Q}_0}}+(K-2)\left[\frac{S\gamma_3}{\sqrt{\mathcal{S}_0}}+\frac{C\gamma_3}{\sqrt{\mathcal{C}_0}}\right]\right\}$$
$$+\frac{4}{\pi^2}\frac{\eta^2}{K^2}\frac{1}{\sqrt{\gamma_3\gamma_4}}\left\{2\arcsin\left(\frac{\mathcal{Q}_2-\gamma_3\gamma_4}{\mathcal{Q}_2}\right)-4\arcsin\left(\frac{R\gamma_1-\beta SC}{\sqrt{\mathcal{Q}_2\mathcal{R}_2}}\right)+2\arcsin\left(\frac{\mathcal{R}_2-\gamma_3\gamma_4}{\mathcal{R}_2}\right)\right.$$
$$+2\arcsin\left(\frac{C}{\mathcal{Q}_2}\right)-4\arcsin\left(\frac{S\gamma_1-\beta RC}{\sqrt{\mathcal{Q}_2\mathcal{R}_2}}\right)+2\arcsin\left(\frac{\beta^2(R^2+S^2)-2\gamma_1\beta RS}{\mathcal{R}_2}\right)$$
$$+(K-2)\left[4\arcsin\left(\frac{C\sqrt{\gamma_3}}{\sqrt{\mathcal{Q}_2\mathcal{C}_2}}\right)-4\arcsin\left(\frac{(S\gamma_1-\beta RC)\sqrt{\gamma_3}}{\sqrt{\mathcal{R}_2\mathcal{C}_2}}\right)-4\arcsin\left(\frac{\beta S(S+R)\sqrt{\gamma_3}}{\sqrt{\mathcal{R}_2\mathcal{S}_2}}\right)\right.$$
$$\left.-4\arcsin\left(\frac{S\sqrt{\gamma_3}}{\sqrt{\mathcal{Q}_2\mathcal{S}_2}}\right)+\arcsin\left(\frac{\mathcal{C}_2-\gamma_4}{\mathcal{C}_2}\right)-2\arcsin\left(\frac{R\gamma_4-2\beta SC}{\sqrt{\mathcal{S}_2\mathcal{C}_2}}\right)+\arcsin\left(\frac{\mathcal{S}_2-\gamma_4}{\mathcal{S}_2}\right)\right]$$
$$\left.+(K-2)(K-3)\left[\arcsin\left(\frac{C\gamma_3}{\mathcal{C}_2}\right)-2\arcsin\left(\frac{S\gamma_3}{\sqrt{\mathcal{S}_2\mathcal{C}_2}}\right)-\arcsin\left(\frac{2\beta S^2}{\mathcal{S}_2}\right)\right]\right\}, \tag{B2d}$$

where we have for convenience defined

$$\gamma_1=1+\beta Q,\quad \gamma_2=1+2\beta Q,\quad \gamma_3=1+\beta(Q-C),$$
$$\gamma_4=1+\beta(Q+C),\quad \mathcal{Q}_0=\gamma_1+Q,\quad \mathcal{Q}_1=\gamma_2+Q,$$
$$\mathcal{Q}_2=\gamma_3\gamma_4+Q\gamma_1-\beta C^2,\quad \mathcal{C}_0=(1+Q)\gamma_1-\beta C^2,$$
$$\mathcal{C}_1=(1+Q)\gamma_2-2\beta C^2,\quad \mathcal{C}_2=(1+Q)\gamma_4-2\beta C^2,$$
$$\mathcal{S}_0=(1+T)\gamma_1-\beta S^2,\quad \mathcal{S}_1=(1+T)\gamma_2-2\beta S^2,$$
$$\mathcal{S}_2=(1+T)\gamma_4-2\beta S^2,\quad \mathcal{R}_0=(1+T)\gamma_1-\beta R^2,$$
$$\mathcal{R}_1=(1+T)\gamma_2-2\beta R^2,$$
$$\mathcal{R}_2=(1+T)\gamma_3\gamma_4-\beta\gamma_1(R^2+S^2)+2\beta^2RSC.$$

1. Symmetric fixed-point dynamics

For a linear theory of the dynamics around their fixed point, we need to expand the differential equations (B2) in a Taylor series to first order

$$\frac{dp_i}{d\alpha}=m_{i0}+\sum_{j=1}^{4}m_{ij}p_j,$$

where $p_i=P_i-P_i^*$ and P_i are generic order parameters. For a fixed point the zeroth-order terms vanish and the eigenvalues and eigenvectors of the Jacobian matrix $\mathbf{M}$ of first derivatives determine the solution of the linearized differential equation.

Under the constraints $Q=C$ and $R=S$, which are characteristic for the symmetric fixed points studied analytically, one finds that the zeroth-order terms and the entries of the Jacobian matrix $\mathbf{M}$ obey the relations (here $P_1=R$, $P_2=S$, $P_3=Q$, and $P_4=C$)

$$m_{10}=m_{20},\quad m_{30}=m_{40},\quad m_{12}=(K-1)m_{21},$$
$$m_{24}=m_{14},\quad m_{22}=m_{11}+(K-2)m_{21},\quad m_{23}=m_{13},$$
$$m_{32}=m_{42}=(K-1)m_{31},\quad m_{32}=(K-1)m_{31},$$
$$m_{44}=m_{33}+m_{34}-m_{43},\quad m_{41}=m_{31}. \tag{B3}$$

We omit the exact form of the remaining free parameters of the matrix as they are extremely tedious but easily derivable from (B2). The eigenvalues of such a Jacobian matrix are given by

$$\lambda_1=m_{11}-m_{21},\quad \lambda_2=m_{33}-m_{43}, \tag{B4}$$
$$\lambda_{3,4}=\frac{1}{2}[A_0+B_0\pm\sqrt{(A_0-B_0)^2+4Km_{31}C_0}],$$

with $A_0=m_{11}+(K-1)m_{21}$, $B_0=m_{33}+m_{34}$, and $C_0=m_{13}+m_{14}$. The corresponding (unnormalized) eigenvectors $\boldsymbol{v}_i$ are given by

$$\begin{aligned}\boldsymbol{v}_1&=((K-1)\quad -1\quad 0\quad 0),\\ \boldsymbol{v}_2&=(1\quad 1\quad v_{23}\quad v_{24}),\\ \boldsymbol{v}_{3,4}&=(v_{(3,4);(1/2)}\quad v_{(3,4);(1/2)}\quad 1\quad 1),\end{aligned}\tag{B5a}$$

with

$$\begin{aligned}v_{23}&=\frac{m_{34}(m_{33}-m_{43}-A_0)+Km_{14}m_{31}}{m_{13}m_{34}-m_{14}m_{43}},\\ v_{24}&=\frac{m_{43}(A_0+m_{43}-m_{33})-Km_{13}m_{31}}{m_{13}m_{34}-m_{14}m_{43}},\end{aligned}\tag{B5b}$$

$$v_{(3,4);(1/2)}=\frac{\lambda_{3,4}-B_0}{Km_{31}},$$

where the first digit indicates the eigenvalue number and the second indicates the component index.

a. Truncated equations

For the truncated differential equations, where η^2 are neglected, the onset of specialization is characterized by the eigenvalues

$$\lambda_1^0=\frac{2}{\pi}\frac{\eta\beta T^2}{\sqrt{K(1+T)-T}[K(1+T)+\beta T]^{3/2}},\tag{B6a}$$

$$\lambda_2^0=0\tag{B6b}$$

$$\lambda_3^0=-\frac{2}{\pi}\eta\left[\frac{K(1+T)-T}{K(1+T)+\beta T}\right]^{3/2},\tag{B6c}$$

$$\lambda_4^0=-\frac{4}{\pi}\eta\sqrt{\frac{K(1+T)-T}{K(1+T)+\beta T}},\tag{B6d}$$

i.e., one finds only one relevant eigenvalue λ_1^0 (and one marginal eigenvalue λ_2^0). If one takes a closer look at the eigenvectors, whose nonconstant terms take the form

$$v_{23}^0=\frac{2K^{3/2}(1+T)}{T\sqrt{K(1+T)-T}},\tag{B7a}$$

$$v_{24}^0=-\frac{2K^{3/2}}{(K-1)T\sqrt{K(1+T)-T}},\tag{B7b}$$

$$v_{3;(1/2)}^0=\frac{2\sqrt{K}}{\sqrt{K(1+T)-T}},\tag{B7c}$$

$$v_{4;(1/2)}^0=-\frac{2K^{3/2}(1+T)}{T(1+2\beta)\sqrt{K(1+T)-T}},\tag{B7d}$$

one can see that the positive eigenvalue λ_1^0 acts solely in the student-teacher overlap space. This eigenvalue is associated with a pure rotation of the weight vectors towards the teacher unit they will specialize on. The marginal eigenvalue λ_2^0 (which will be important in the case where η^2 terms are not neglected) shows an increase in the squared norm Q of the student weight vectors of $O(K)$, but a decrease in their correlations C of $O(1)$, which corresponds primarily to a growth of the student weight vectors outside the subspace spanned by the teacher weight vectors.

b. Small-η fixed point

To calculate the first-order correction in η to the fixed point of the truncated equations (11), we expand the full differential equations (B2) to first order around Eqs. (11) and find the zeros of the resulting set of linear equations in (r,s,q,c). Examining the relations (B3) more closely, one can see that the solution is characterized by $r=s$ and $q=c$, and we find for the new symmetric fixed point $Q^*=C^*=Q_0^*+Q_1^*$ and $R^*=S^*=R_0^*+R_1^*$, ignoring terms of $O(\eta^2)$,

$$Q_1^*=\frac{1}{\pi}\frac{[K(1+T)+2\beta T]}{[K(1+T)-T]}\mathcal{GF}\frac{\eta}{K},\tag{B8a}$$

$$R_1^*=\frac{1}{2\pi}\frac{T(1+2\beta)}{\sqrt{K(1+T)-T}}\mathcal{GF}\frac{\eta}{K^{3/2}},\tag{B8b}$$

with

$$\mathcal{G}=\frac{\sqrt{K(1+T)+\beta T}}{\sqrt{K(1+T)+(2\beta-1)T}},\tag{B8c}$$

$$\begin{aligned}\mathcal{F}=&\arcsin\left(\frac{T\{K[K(1+T)-T]+(K-1)2\beta T\}}{[K(1+T)-T][K(1+T)+2\beta T]}\right)\\ &-K\arcsin\left(\frac{T}{K(1+T)K+2\beta T}\right)-(K-1)\\ &\times\arcsin\left(\frac{2\beta T^2}{[K(1+T)-T][K(1+T)+2\beta T]}\right).\end{aligned}\tag{B8d}$$

For the expansion to be valid, η has to be chosen to ensure $Q_1^*\ll Q_0^*$ and $R_1^*\ll R_0^*$. For large K, this implies $\eta\leqslant O(K^{-1})$. We further note that the new fixed point is no longer confined to the subspace spanned by the teacher weight vectors as $R^*<\sqrt{Q^*T/K}$. However, the symmetries $Q=C$ and $R=S$ are not broken to first order. This is in contrast to the numerical results from integrating the full dynamics (A4), where we observe that the symmetric phase for finite learning rates is characterized by $Q>C$ (and $R=S$).

c. Small-η dynamics

To study the onset of specialization, we expand the differential equations (B2) around the new fixed point, which is again characterized by $Q=C$ and $R=S$, and the matrix relations (B3) hold. Ignoring terms of $O(\eta^3)$, we find that the eigenvalues (eigenvectors) of the Jacobian have acquired $O(\eta^2)$ [$O(\eta)$] corrections to their values in Eq. (B6) [Eq. (B7)]. In particular

$$\lambda_2^1=\frac{4}{\pi^2}\frac{\sqrt{K(1+T)-T}}{K(1+T)+(2\beta-1)T}\beta\eta^2\left\{\frac{K(1+T)+(3\beta-1)T}{K\sqrt{K(1+T)+(2\beta-1)T}}\mathcal{F}-\frac{2\beta T^2}{\sqrt{K}[K(1+T)+2\beta T]}\left[\frac{\sqrt{K}}{\sqrt{K(1+T)+(2\beta+1)T}}\right.\right.$$
$$\left.\left.-\frac{1}{\sqrt{K^2(1+T)(1+2T)+K(1+2T)T(2\beta-1)-4\beta T^2}}-\frac{(K-1)\sqrt{1+T}}{\sqrt{K^2(1+T)^2+K(1+T)T(2\beta-1)-4\beta T^2}}\right]\right\}, \tag{B9}$$

which is, in general, positive and dominated by the $\mathcal{F}$ term, i.e., the marginal eigenvalue now becomes relevant to the dynamics. As mentioned in Appendix B 1, the associated eigenvector (whose η dependence can be ignored as it constitutes only a minor correction) shows an increase in Q of $O(K)$ and a decrease in C of $O(1)$. As the increases in R and S are equal, this mode does not contribute to the specialization process but corresponds primarily to a growth of the student weight vectors outside the subspace spanned by the teacher weight vectors. Since the initial differences between Q and C are typically large, this eigenvalue will actually dominate the dynamics and quickly drive the student away from this particular fixed point. We therefore conclude that the fixed point associated with $Q=C$ is relevant only for $\eta=0$ and that a fixed point characterized by $Q>C$ leads to the long symmetric phase for $\eta>0$, which is not accessible by first-order correction to the fixed point studied in Appendix B 1 b. An analytic study of that fixed point necessitates an expansion to second order and the subsequent solution of a set of quadratic equations, which we have found to be infeasible.

2. Convergence fixed-point dynamics

As for the symmetric fixed point, we expand the differential equations (B2) to first order around the zero generalization error fixed point $Q^*=R^*=T$ and $C^*=S^*=0$, where we use the ordering $P_1=R$, $P_2=Q$, $P_3=S$, and $P_4=C$ for the convergence phase (again following the convention of earlier work [2]). Similarly, we also expand the generalization error (B1) to second order. Explicitly, one finds for the generalization error

$$\epsilon_g=\frac{1}{\pi}\left\{\frac{2r-q}{\sqrt{1+2T}}-\frac{1}{4}\frac{T(2r-q)^2}{(1+2T)^{3/2}}+\frac{q(r-q)}{(1+2T)^{3/2}}\right.$$
$$\left.-\frac{K-1}{1+T}\left[(2s-c)+\frac{q(s-c)}{1+T}\right]\right\}. \tag{B10}$$

The elements of the Jacobian matrix are given by

$$c_{11}=-\frac{2}{\pi}\frac{\eta}{K}\frac{1+(1+2\beta)T}{[1+(1+\beta)T]^{3/2}}, \tag{B11a}$$

$$c_{12}=\frac{1}{\pi}\frac{\eta}{K}\frac{T(1+2\beta)}{[1+(1+\beta)T]^{3/2}}, \tag{B11b}$$

$$c_{13}=\frac{2}{\pi}\frac{\eta}{K}\frac{(K-1)(1+2\beta T)}{\sqrt{1+T}(1+\beta T)^{3/2}}, \tag{B11c}$$

$$c_{14}=-\frac{2}{\pi}\frac{\eta}{K}\frac{(K-1)\beta T}{\sqrt{1+T}(1+\beta T)^{3/2}}, \tag{B11d}$$

$$c_{21}=\frac{4}{\pi}\frac{\eta}{K}\left\{\frac{1+T}{[1+(1+\beta)T]^{3/2}}-\frac{2}{\pi}\frac{\eta}{K}\left[\frac{1}{\sqrt{1+2(1+\beta)T}}\right.\right.$$
$$\left.\left.+\frac{(K-1)}{\sqrt{(1+2\beta T)(1+2T)}}\right]\right\}, \tag{B11e}$$

$$c_{23}=-\frac{4}{\pi}\frac{\eta}{K}\frac{(K-1)}{\sqrt{1+T}}\left\{\frac{1}{(1+\beta T)^{3/2}}-\frac{2}{\pi}\frac{\eta}{K}\left[\frac{2}{\sqrt{1+(1+2\beta)T}}\right.\right.$$
$$\left.\left.+\frac{(K-2)}{\sqrt{(1+2\beta T)(1+T)}}\right]\right\}, \tag{B11f}$$

$$c_{31}=\frac{2}{\pi}\frac{\eta}{K}\frac{1}{\sqrt{(1+\beta T)(1+T)}}, \tag{B11g}$$

$$c_{32}=-\frac{1}{\pi}\frac{\eta}{K}\frac{T}{\sqrt{1+\beta T}(1+T)^{3/2}}, \tag{B11h}$$

$$c_{33}=-\frac{2}{\pi}\frac{\eta}{K}\left[\frac{1}{\sqrt{1+(1+\beta)T}}\frac{(K-2)}{\sqrt{(1+\beta T)(1+T)}}\right], \tag{B11i}$$

$$c_{34}=0, \tag{B11j}$$

$$c_{41}=-\frac{4}{\pi}\frac{\eta}{K}\frac{1}{\sqrt{1+\beta T}}\left\{\frac{1}{\sqrt{1+T}}-\frac{2}{\pi}\frac{\eta}{K}\left[\frac{2}{\sqrt{1+(2+\beta)T}}\right.\right.$$
$$\left.\left.+\frac{(K-2)}{\sqrt{(1+\beta T)(1+2T)}}\right]\right\}, \tag{B11k}$$

$$c_{43}=\frac{4}{\pi}\frac{\eta}{K}\Bigg\{\frac{1}{\sqrt{1+(1+\beta)T}}+\frac{(K-2)}{\sqrt{(1+\beta T)(1+T)}}$$
$$-\frac{2}{\pi}\frac{\eta}{K}\Bigg[\frac{2}{1+(1+\beta)T}+\frac{(K-2)}{(1+\beta T)(1+T)}$$
$$\times\left(4\frac{\sqrt{(1+\beta T)(1+T)}}{\sqrt{1+(1+\beta)T}}+(K-3)\right)\Bigg]\Bigg\}. \quad \text{(B11l)}$$

The remaining elements can be deduced by the matrix relations [18]

$$c_{11}-\frac{1}{2}c_{21}=c_{22}-2c_{12}, \quad \text{(B12a)}$$

$$c_{33}-\frac{1}{2}c_{43}=c_{44}-2c_{34}, \quad \text{(B12b)}$$

$$c_{13}-\frac{1}{2}c_{23}=c_{24}-2c_{14}, \quad \text{(B12c)}$$

$$c_{31}-\frac{1}{2}c_{41}=c_{42}-2c_{32}. \quad \text{(B12d)}$$

The eigenvalues of such a Jacobian matrix are given by the solutions to two quadratic equations

$$\lambda_{1,2}=\frac{1}{2}[A_1+B_1\pm\sqrt{(A_1-B_1)^2+4C_1D_1}], \quad \text{(B13a)}$$

$$\lambda_{3,4}=\frac{1}{2}[A_2+B_2\pm\sqrt{(A_2-B_2)^2+4C_2D_2}], \quad \text{(B13b)}$$

with

$$A_1=c_{11}-\frac{1}{2}c_{21},\quad B_1=c_{44}-2c_{34},\quad C_1=c_{31}-\frac{1}{2}c_{41},$$
$$D_1=c_{24}-2c_{14},\quad A_2=c_{11}+2c_{12},\quad B_2=c_{44}+\frac{1}{2}c_{43},$$
$$C_2=c_{31}+2c_{32},\quad D_2=c_{24}+\frac{1}{2}c_{23}.$$

The corresponding (unnormalized) eigenvectors $\boldsymbol{v}_i$ are given by

$$\boldsymbol{v}_{1,2}=(v_{(1,2);1}\quad v_{(1,2);2}\quad v_{(1,2);3}\quad v_{(1,2);4}), \quad \text{(B14a)}$$

$$\boldsymbol{v}_{3,4}=(1\quad 2\quad v_{(3,4);(3/4)}\quad 2v_{(3,4);(3/4)}), \quad \text{(B14b)}$$

with (using $c_{34}=0$)

$$v_{(3,4);(3/4)}=\frac{\lambda_{3,4}-A_2}{D_2}, \quad \text{(B14c)}$$

$$v_{(1,2);1}=-\{2D_1[c_{14}C_1+c_{12}(B_2-\lambda_{1,2})+c_{32}D_2]$$
$$+c_{43}c_{14}(A_1-\lambda_{1,2})\}, \quad \text{(B14d)}$$

$$v_{(1,2);2}=c_{21}D_1(\lambda_{1,2}-c_{44})+c_{43}c_{24}(\lambda_{1,2}-c_{11})$$
$$+D_1(c_{31}c_{23}+c_{41}c_{24})+c_{43}c_{21}c_{14}, \quad \text{(B14e)}$$

$$v_{(1,2);3}=2c_{31}c_{14}(A_2-\lambda_{1,2})+2c_{32}c_{24}(c_{11}-\lambda_{1,2})$$
$$-c_{14}c_{21}C_2-2c_{24}c_{12}c_{31}, \quad \text{(B14f)}$$

$$v_{(1,2);4}=\frac{1}{C_1}(\lambda_{1,2}-A_1)\{2(c_{21}c_{32}-c_{12}c_{41})(\lambda_{1,2}-c_{44})$$
$$+C_1[c_{21}(\lambda_{1,2}-c_{44})+c_{43}(\lambda_{1,2}-A_2)$$
$$+c_{41}D_1+c_{23}C_2]\}. \quad \text{(B14g)}$$

Comparing the eigenvectors (B14) with the expansion of the generalization error (B10), one finds that the modes $\boldsymbol{v}_{3,4}$ are orthogonal to the first-order terms in the generalization error and therefore cannot contribute to their decay. These modes are therefore only relevant for second-order terms in the generalization error with a decay rate of $2\lambda_{3,4}$. As discussed in Sec. IV B, the fastest convergence is given by Eq. (15). This is achieved either for η_r^{opt}, where $2\lambda_3=\lambda_1$, or for η_m^{opt}, which is defined by the minimum of λ_1. The critical (maximal) learning rates are defined by the zeros of the determinant in η

$$A_1B_1=C_1D_1, \quad \text{(B15a)}$$

$$A_2B_2=C_2D_2, \quad \text{(B15b)}$$

where only one nonzero learning rate solution exists in Eq. (B15b), coinciding with $\lambda_1=0$.

Unfortunately, it is in general infeasible to optimize the eigenvalues with respect to the learning parameters η and β analytically for arbitrary K and T. However, one can make some progress in certain limits of K and T, which we will investigate below.

a. Large-K limit

The dominant terms for a large number of hidden units for all relevant quantities can be extracted by an asymptotic series expansion under the self-consistent ansatz $\eta=O(1)$ and $\beta=O(1)$. For the two relevant eigenvalues one makes the ansatz $\lambda_i=O(K^{-1})$ and finds to leading order

$$\lambda_1(\beta)=-\frac{4}{\pi}\frac{\eta}{K}\frac{\pi\chi_1-\eta\chi_2}{\mathcal{E}_1\mathcal{E}_2\mathcal{E}_3(\pi\mathcal{E}_1-\eta)}, \quad \text{(B16a)}$$

$$\lambda_3(\beta)=-\frac{2}{\pi}\frac{\eta}{K}(\mathcal{E}_3^{-3}-\mathcal{E}_1^{-3}), \quad \text{(B16b)}$$

with the auxiliary variables

$$\chi_1=\mathcal{E}_1\mathcal{E}_2(\mathcal{E}_1-\mathcal{E}_3), \quad \text{(B16c)}$$

$$\chi_2=\mathcal{E}_1\mathcal{E}_2-\mathcal{E}_3[\sqrt{1+2\beta T}(1+T)+\sqrt{1+2T}(1+\beta T)-\mathcal{E}_1^2], \quad \text{(B16d)}$$

$$\mathcal{E}_1=\sqrt{(1+T)(1+\beta T)}, \quad \text{(B16e)}$$

$$\mathcal{E}_2=\sqrt{(1+2T)(1+2\beta T)}, \quad \text{(B16f)}$$

$$\mathcal{E}_3=\sqrt{1+(1+\beta)T}. \tag{B16g}$$

These define two critical learning rates

$$\eta^0_{\rm crit}(\beta)=\pi\frac{\chi_1}{\chi_2}, \tag{B17a}$$

$$\eta^\infty_{\rm crit}(\beta)=\pi\mathcal{E}_1>\eta^0_{\rm crit}, \tag{B17b}$$

where λ_1 is identical to zero ($\eta^0_{\rm crit}$) and diverges ($\eta^\infty_{\rm crit}$), respectively. Solving Eq. (B15b), one finds $\eta_{\rm max}=\eta^0_{\rm crit}$, as expected. It is important to realize that Eq. (B16a) is only a valid expansion for λ_1 for $\eta<\eta^\infty_{\rm crit}$, beyond which the ansatz $\lambda_1=O(K^{-1})$ breaks down as $\lambda_1=O(1)$. In fact, the order of the two eigenvalues λ_1 and λ_2 changes at $\eta^\infty_{\rm crit}$ and Eq. (B16a) is the correct asymptotic expansion of λ_2 for $\eta>\eta^\infty_{\rm crit}$. This change in the order of eigenvalues can be seen quite well in Fig. 6(a), as the natural continuation for λ_1 for large η follows the curve representing λ_2 and vice versa. As mentioned above, one has to calculate, in general, both $\eta_r^{\rm opt}$ and $\eta_m^{\rm opt}$ by solving $2\lambda_3=\lambda_1$ and $d\lambda_1/d\eta=0$, respectively. Due to the breakdown of the ansatz for λ_1 above $\eta^\infty_{\rm crit}$, solutions with $\eta^{\rm opt}>\eta^\infty_{\rm crit}$ are spurious.

For GD the eigenvalues and the critical learning rates simplify to

$$\lambda_1(1)=-\frac{4}{\pi}\frac{\eta}{K}[(1+T)-\sqrt{1+2T}]\times\frac{\pi\sqrt{1+2T}-\eta}{(1+2T)[\pi(1+T)-\eta]}, \tag{B18a}$$

$$\lambda_3(1)=-\frac{2}{\pi}\frac{\eta}{K}[(1+2T)^{-3/2}-(1+T)^{-3}], \tag{B18b}$$

$$\eta^0_{\rm crit}(1)=\pi\sqrt{1+2T}, \tag{B18c}$$

$$\eta^\infty_{\rm crit}(1)=\pi(1+T), \tag{B18d}$$

resulting in the two candidates for the optimal learning rate taking the form

$$\eta_r^{\rm opt}(1)=\frac{\eta^\infty_{\rm crit}T[2(1+T)^3-(2+T)(1+2T)^{3/2}]}{(1+T)^4(\sqrt{1+2T}-2)+(1+2T)^{3/2}}, \tag{B19a}$$

$$\eta_m^{\rm opt}(1)=\eta^\infty_{\rm crit}-\pi\sqrt{1+T}[(1+T)-\sqrt{1+2T}]^{1/2}. \tag{B19b}$$

To decide on the correct learning rate for given T, one has to evaluate whether $\eta_r^{\rm opt}(1)<\eta^\infty_{\rm crit}(1)$ and then calculate the convergence rates for the two learning rates. We find that $\eta^{\rm opt}(1)=\eta_r^{\rm opt}(1)$ for $T>T^{\rm crit}$ and $\eta^{\rm opt}(1)=\eta_m^{\rm opt}(1)$ for $T<T^{\rm crit}$, where $T^{\rm crit}=1.2780$ is defined by $\eta_r^{\rm opt}(1)=\eta_m^{\rm opt}(1)$.

When optimizing β, one always finds that the fastest convergence is achieved for $2\lambda_3=\lambda_1$ and the optimal learning rate is determined by

$$\eta^{\rm opt}(\beta)=\pi\mathcal{E}_2T\{\mathcal{E}_1^4(1+\beta)+\mathcal{E}_1\mathcal{E}_3^3[1+\beta(1+T)]\}\times\{\mathcal{E}_1^3\mathcal{E}_2(1+\beta)T-\mathcal{E}_3^3[\sqrt{1+2T}(1+\beta T)\mathcal{E}_1^2+\sqrt{1+2\beta T}(1+T)\mathcal{E}_1^2-\mathcal{E}_1^4-\mathcal{E}_2]\}^{-1}. \tag{B20}$$

The optimal convergence rate, which is just given as $2\lambda_3$ at $\eta^{\rm opt}$, however, cannot be further optimized analytically with respect to β and this optimization has to be done numerically. The results for $\beta^{\rm opt}$ and all other interesting quantities in this limit can be seen in Figs. 7 and 8.

To make further progress in the $K\to\infty$ limit, one can look at the limits $T\to\infty$ and $T\to0$. These results turn out to be equivalent, to leading order in K and T, to results where both T and K go to their limits simultaneously, i.e., taking the limit $K\to\infty$ with $T=T_\infty K$ and $T=T_0/K$, respectively. T_0 and T_∞ are prefactors controlling the significance between T and K. Below, we therefore used the more general expansion in both variables for higher-order terms. Unfortunately, this was infeasible for higher-order terms for optimized ABP in the small-T limit, where we present the results obtained by taking the large-K limit first.

(i) *Small T limit* ($T=T_0/K$). For GD the leading terms of the relevant quantities in this limit are

$$\eta_{\rm max}=\pi\left[1+T-\frac{1}{2}T^2+\frac{1}{2}\frac{T^2}{K}(TK-4)\right], \tag{B21a}$$

$$\eta^{\rm opt}=\pi\left[1+\frac{1}{2}(2-\sqrt{2})T-\frac{\sqrt{2}}{4}\frac{T}{K}\right], \tag{B21b}$$

$$\lambda^{\rm opt}=-2\frac{T^2}{K}\left[1-(2+\sqrt{2})T+\frac{19+12\sqrt{2}}{4}T^2+\frac{\sqrt{2}}{2}\frac{T}{K}\right], \tag{B21c}$$

with $TK=T_0=O(1)$. The optimization for ABP yields, for $K\to\infty$ preceding $T\to0$,

$$\beta^{\rm opt}=\frac{2}{T}+\frac{3}{10}\frac{5^{3/4}\sqrt{6}(\sqrt{5}-1)}{\sqrt{T}}, \tag{B22a}$$

$$\eta_{\rm max}=\pi\sqrt{3}\left[1+\frac{5^{3/4}\sqrt{6}(\sqrt{5}-1)}{20}\sqrt{T}\right], \tag{B22b}$$

$$\eta^{\rm opt}=\pi\sqrt{3}\left[1-\frac{1519\sqrt{5}-3315}{300(3-\sqrt{5})}T\right], \tag{B22c}$$

$$\lambda^{\rm opt}=-\frac{4}{3}\frac{T}{K}\left[1-\frac{5^{3/4}\sqrt{6}(3-\sqrt{5})}{5(\sqrt{5}-1)}\sqrt{T}\right]. \tag{B22d}$$

In this limit ABP yields in leading order a factor of $2/3T^{-1}$ in reduction of training time due to the increase of $\beta^{\rm opt}\propto T^{-1}$. Furthermore, the decrease in the normalized gap between $\eta_{\rm max}$ and $\eta^{\rm opt}$ is slowed down proportional to $1/\sqrt{T}$.

(ii) *Large-T limit* ($T=T_\infty K$). For GD the leading terms of the relevant quantities in this limit are

$$\eta_{\max}=\pi\sqrt{2}\sqrt{T}\left[1-\frac{\sqrt{T}}{K}+\frac{(1+2T_\infty)^2}{4T}\right], \tag{B23a}$$

$$\eta^{\mathrm{opt}}=\eta_{\max}-\frac{\pi\sqrt{2}}{2\sqrt{T}}, \tag{B23b}$$

$$\lambda^{\mathrm{opt}}=-\frac{2}{KT}\left[1-\frac{\sqrt{T}}{K}+\frac{T_\infty^2+T_\infty-1}{T}\right], \tag{B23c}$$

whereas the optimization for ABP gives

$$\beta^{\mathrm{opt}}=\frac{1}{3}-\frac{1}{18}\frac{3\sqrt{2}T_\infty+8\sqrt{6}-12-2\sqrt{3}}{\sqrt{T}}, \tag{B24a}$$

$$\eta_{\max}=\pi\sqrt{T}-\frac{\pi}{16}[11\sqrt{2}T_\infty+20+14\sqrt{3}-8\sqrt{2}(2+\sqrt{3})], \tag{B24b}$$

$$\eta^{\mathrm{opt}}=\eta_{\max}-\frac{3}{4}\frac{\pi}{\sqrt{T}}, \tag{B24c}$$

$$\lambda^{\mathrm{opt}}=-\frac{3}{2}\frac{\sqrt{3}}{KT}\left[1-\frac{T_\infty-(2-\sqrt{2})(\sqrt{3}-\sqrt{2})}{\sqrt{2}\sqrt{T}}\right]. \tag{B24d}$$

In this limit ABP yields only a constant factor of $3\sqrt{3}/4\approx 1.2990$ in reduction of training time and an increase in the learning rate gap by a factor 3/2. This should be contrasted to the increase in training time for both algorithms by a factor T and a decrease in the normalized learning rate gap of T^{-1}. Two logical further extensions are to look at the limits $T\to 0$ and $T\to\infty$ for K finite, especially as the numerical solutions indicate [see Fig. 7(b)] that there are qualitative changes in the learning behavior at least for $T\to\infty$.

b. Small-T limit

For small T, where the network becomes nearly linear, one should only expect minor changes to the limits studied previously since the network behaves smoothly. In particular, we find for GD

$$\eta_{\max}=\pi\left[1+T-\frac{K+4}{2K}T^2\right], \tag{B25a}$$

$$\eta^{\mathrm{opt}}=\pi\left[1+\left(1-\sqrt{\frac{K-1}{2K}}\right)T(1+T)\right], \tag{B25b}$$

$$\lambda^{\mathrm{opt}}=-2\frac{T^2}{K}\left[1-2\left(1+\sqrt{\frac{K-1}{2K}}\right)T\right]. \tag{B25c}$$

For ABP only the leading term is feasible to calculate, resulting in

$$\beta^{\mathrm{opt}}=\frac{2}{T}, \tag{B26a}$$

$$\eta_{\max}=\pi\sqrt{3}\frac{5K}{5(K-1)+3\sqrt{5}}, \tag{B26b}$$

$$\eta^{\mathrm{opt}}=\eta_{\max}, \tag{B26c}$$

$$\lambda^{\mathrm{opt}}=-\frac{4}{3}\frac{5T}{5(K-1)+3\sqrt{5}}, \tag{B26d}$$

which explains the very weak influence of K on the previous results (besides the natural rescaling of λ^{opt} with K^{-1}).

c. Large-T limit

Unlike for small T, we find significant changes in the learning behavior of both algorithms in the large-T limit. For GD one finds for the leading orders

$$\eta_{\max}=\pi\sqrt{2}K\left[1-\frac{K-1}{\sqrt{T}}\right], \tag{B27a}$$

$$\eta^{\mathrm{opt}}=\eta_{\max}-\frac{\pi\sqrt{2}K}{2T}, \tag{B27b}$$

$$\lambda^{\mathrm{opt}}=-\frac{2}{T^{3/2}}\left[1-\frac{K-1}{\sqrt{T}}\right]. \tag{B27c}$$

For ABP the numerical solutions suggest the self-consistent ansatz $\beta^{\mathrm{opt}}\propto T^{-1/3}$ and the leading terms are

$$\beta^{\mathrm{opt}}=\frac{1}{6}\left[\frac{12(K-1)^2}{T}\right]^{1/3}-\frac{5K+19}{54}\left[\frac{18(K-1)}{T^2}\right]^{1/3}, \tag{B28a}$$

$$\eta_{\max}=\pi K\left\{\sqrt{2}-\left[\frac{3\sqrt{2}(K-1)^2}{T}\right]^{1/3}-\frac{3K+1}{18}\left[\frac{36\sqrt{2}(K-1)}{T^2}\right]^{1/3}\right\}, \tag{B28b}$$

$$\eta^{\mathrm{opt}}=\eta_{\max}-\frac{\pi\sqrt{2}K}{T}, \tag{B28c}$$

$$\lambda^{\mathrm{opt}}=-\frac{1}{T^{3/2}}\left\{4\sqrt{2}-6\left[\frac{3\sqrt{2}(K-1)^2}{T}\right]^{1/3}+\frac{37K+11}{12}\left[\frac{36\sqrt{2}(K-1)}{T^2}\right]^{1/3}\right\}. \tag{B28d}$$

In this limit ABP yields a larger constant factor of $2\sqrt{2}\approx 2.828$ in reduction of training time and an increase in the learning rate gap by a factor 2, which is somewhat better than for the infinite-K case.

[1] C. Cybenko, Math. Control Signals Syst. **2**, 303 (1989).
[2] D. Saad and S. A. Solla, Phys. Rev. E **52**, 4225 (1995).
[3] T. K. Leen and G. B. Orr, in *Advances in Neural Information Processing Systems*, edited by J. D. Cowan, G. Tesauro, and J. Alspector (Kaufmann, San Francisco, 1994), Vol. 6, p. 477.
[4] A. Prügel-Bennett (unpublished).
[5] S. Amari, in *Advances in Neural Information Processing Systems*, edited by M. C. Mozer, M. I. Jordan, and T. Petsche (MIT Press, Cambridge, 1997), Vol. 9, p. 127.
[6] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992); M. Copelli and N. Caticha, *ibid.* **28**, 1615 (1995); M. Copelli, O. Kinouchi, and N. Caticha, Phys. Rev. E **53**, 6341 (1996).
[7] M. Biehl and P. Riegler, Europhys. Lett. **28**, 525 (1994).
[8] J. Kim and H. Sompolinsky, Phys. Rev. Lett. **76**, 3021 (1996).
[9] M. Biehl and H. Schwarze, J. Phys. A **28**, 643 (1995).
[10] D. Saad and M. Rattray, in Proceedings of the Minerva Workshop on Mesoscopics, Fractals and Neural Networks, Eilat, 1997 (unpublished).
[11] D. Barber, D. Saad, and P. Sollich, Europhys. Lett. **34**, 151 (1996).
[12] A. H. L. West (unpublished).
[13] A. H. L. West, D. Saad, and I. T. Nabney, in *Advances in Neural Information Processing Systems* (Ref. [5]), p. 288.
[14] A. H. L. West and D. Saad, in *Advances in Neural Information Processing Systems*, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (MIT Press, Cambridge, MA, 1996), Vol. 8, p. 323.
[15] The term *adaptive* in ABP therefore refers to the adjustability of β and to the subsequent deformation of the search space. It does not imply an ability of the algorithm to tune its adjustable parameters on-line.
[16] Although we find numerically exponents for I smaller than -1 for larger K, it remains unclear if these hold even for $TK \ll 1$. However, for $K \to \infty$, $T \to 0$, and $TK = \text{const}$, the power law seems to approach 3/2.
[17] This result only seems to differ from the result in [2] due to different scaling for η.
[18] P. Riegler, Ph.D. thesis, University of Würzburg, 1997 (unpublished).