Preprint Vs.2

The role of biases in on-line learning of 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

(January 24, 1998)

Abstract

The influence of biases on the learning dynamics of a two-layer neural network, a normalized soft-committee machine, is studied for on-line gradient descent learning. Within a statistical mechanics framework, numerical studies show that the inclusion of adjustable biases dramatically alters the learning dynamics found previously. The symmetric phase which has often been predominant in the original model all but disappears for a non-degenerate bias task. The extended model furthermore exhibits a much richer dynamical behavior, e.g., attractive suboptimal symmetric phases even for realizable cases and noiseless data.

PACS number(s): 87.10.+e, 05.2.0-y, 02.50.-r, 02.30.Hq

I. INTRODUCTION

The theoretical understanding of the learning dynamics of multilayer feedforward perceptrons (MLPs) has attracted widespread interest due to their universal approximation ability [1] and their subsequent paramount use in practical applications. Until recently progress has been hampered by the inability to perform the necessary (quenched) average over the training set in order to study their performance independent of the particularities of an individual training set. A method to overcome this problem has been introduced recently in [2]. It studies *on-line* learning in two-layer networks with an arbitrary number of hidden unit, allowing insight into the learning behavior of neural network models whose complexity is of the same order as those used in real world applications.

The on-line learning paradigm, whereby the network parameters are updated serially after the presentation of each single example, allows to avoid the difficulties of averaging over a whole (finite) training set necessary for the more commonly studied *batch* learning algorithm, where all examples are used simultaneously to update the network parameters. The network model studied in particular, the soft-committee machine [3], consists of a single hidden layer with adjustable input-hidden, but fixed hidden-output weights. The average learning dynamics of these networks are calculated in the thermodynamic limit of infinite input dimensions and in a student-teacher scenario, where a *student* network is presented with training examples $(\boldsymbol{\xi}^\mu, \zeta^\mu)$. The input vectors $\boldsymbol{\xi}^\mu$ are Gaussian random variables and the outputs ζ^μ are labeled by a *teacher* network of the same architecture but possibly with a different number of hidden units. Although the framework allows in principle for any on-line learning algorithm to update the student parameter; gradient descent on the squared example error is studied here.

The above learning scenario is already quite similar to the problems faced in the real world, but the approach still suffers from several drawbacks. First, the analysis of the mean learning dynamics relies on the thermodynamic limit of infinite input dimension — a problem which has been addressed in [4], where finite size effects have been studied and it

was shown that the thermodynamic limit is relevant in most cases. Second, examples are not resampled, describing a scenario with an unrealistically large training set compared to most real cases, where training examples are scarce and therefore repeatedly cycled over. This problems has so far proved evasive, although the issue has been considered at least for the linear perceptron [5]. Third, the hidden-output weights are kept fixed, a constraint which has been relaxed in [6,7], where it has been shown that the learning dynamics are usually dominated by the input-hidden weights. Fourth, the biases of the hidden units are fixed to zero, a constraint which is actually more severe than fixing the hidden-output weights. One can show [8] that soft-committee machines are universal approximators provided one allows for adjustable biases in the hidden layer.

In this paper, we address the fourth limitation by studying the model of a normalized soft-committee machine with dynamic biases following the framework set out in [2]. In Sect. II the model is defined and the calculation of the differential equations governing the training evolution is derived. In Sect. III numerical studies of a few typical learning scenarios are presented to show the qualitative difference in the dynamics to the model with fixed biases, most notably the emergence of attractive suboptimal network configurations. These and their dependence on the teacher task, the influence of weight and bias initialization, and the choice of the learning rates for weights and biases will be studied in Sect. IV. We will also set our results in context to previous works on weight initialization which devised heuristic rules. In Sect. V the optimal learning rates are calculated analytically for arbitrary network size and a range of teacher tasks for the convergence phase, where the student network is close to the optimal solution. In Sect. VI we will outline possible extensions of this framework 0and in particular briefly assess the impact of unrealizable teacher rules. This is followed by a summary and discussion of the main results in Sect. VII.

II. DYNAMICAL EQUATIONS

The student network considered is a normalized soft-committee machine of K hidden units with adjustable biases. Each hidden unit i consists of a bias θ_i and a weight vector $\boldsymbol{W}_i$ which is connected to the N-dimensional inputs $\boldsymbol{\xi}$. All hidden units are connected to a 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., [3]); however, this makes the necessary scaling of the learning rates more explicit and leads to more elegant results for optimal learning rates. The implemented mapping is therefore

$$f_{\boldsymbol{w}}(\boldsymbol{\xi}) = \frac{\gamma}{\sqrt{K}} \sum_{i=1}^{K} g\left(x_i - \theta_i\right), \tag{1}$$

where $x_i = \boldsymbol{W}_i \cdot \boldsymbol{\xi}/\sqrt{N}$ is the student activation and $g()$ is a sigmoidal transfer function. Note, although the biases add only K degrees of freedom to the network, their influence on the hidden unit response is still of the same order as the complete weight vector.

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 $\boldsymbol{B}_n$ and biases ϱ_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 and the outputs are labeled by the teacher according to

$$\zeta^\mu = \frac{\gamma}{\sqrt{M}} \sum_{n=1}^{M} g\left(y_n^\mu - \varrho_n\right), \tag{2}$$

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.

In on-line learning the student parameters $\mathbf{\Omega}$, i.e., all weights and biases, are modified to reduce the error the student makes on a presented single example $(\boldsymbol{\xi}^\mu, \zeta^\mu)$

$$\epsilon(\mathbf{\Omega}, \boldsymbol{\xi}^\mu) = \frac{1}{2}[\zeta^\mu - f_{\boldsymbol{w}}(\boldsymbol{\xi}^\mu)]^2. \tag{3}$$

Gradient descent on the error (3), in this scenario commonly identified with *back-propagation* [9,10], results in updates of the student parameters

$$\boldsymbol{W}_i^{\mu+1} - \boldsymbol{W}_i^\mu = \eta_{\boldsymbol{w}} \delta_i^\mu \frac{\boldsymbol{\xi}^\mu}{\sqrt{N}}, \tag{4a}$$

$$\theta_i^{\mu+1} - \theta_i^\mu = -\frac{\eta_\theta}{N} \delta_i^\mu, \tag{4b}$$

with

$$\delta_i^\mu = \delta^\mu g'(x_i^\mu) = [\zeta^\mu - f_{\boldsymbol{w}}(\boldsymbol{\xi}^\mu)]\, g'(x_i^\mu), \tag{4c}$$

where g' is the derivative of the activation function g. The two learning rates, $\eta_{\boldsymbol{w}}$ for the weights and η_θ for the biases (which has been rescaled explicitly by $1/N$), have to be set by the user to ensure both fast training and convergence to a minimum of the generalization error.

The above Markovian stochastic dynamics (4) are hard to solve generally since this necessitates solving a master equation for the time evolution of the weight and bias probability distributions. Usually approximations such as small learning rates must be employed [11] to make any progress.

However, one is ultimately interested mainly in the typical performance of the student network on a randomly selected input example given by the *generalization error*

$$\epsilon_{\mathrm{g}}(\mathbf{\Omega}) = \langle \epsilon(\mathbf{\Omega}, \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}}\,. \tag{5}$$

Since the dependence of the inputs enter only through the student and teacher activations $\boldsymbol{x} = (x_1, \ldots, x_K)$ and $\boldsymbol{y} = (y_1, \ldots, y_M)$, the probability of $\boldsymbol{\xi}$ can be rewritten in terms of a joint probability distribution in the activations. The resulting distribution is Gaussian with 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:

$$\langle x_i\, x_j\rangle_{\boldsymbol{\xi}} = \tfrac{\sigma^2}{N}\boldsymbol{W}_i\cdot\boldsymbol{W}_j \equiv Q_{ij}, \tag{6a}$$

$$\langle x_i\, y_n\rangle_{\boldsymbol{\xi}} = \tfrac{\sigma^2}{N}\boldsymbol{W}_i\cdot\boldsymbol{B}_n \equiv R_{in}, \tag{6b}$$

$$\langle y_n\, y_m\rangle_{\boldsymbol{\xi}} = \tfrac{\sigma^2}{N}\boldsymbol{B}_n\cdot\boldsymbol{B}_m \equiv T_{nm}. \tag{6c}$$

Since also the weights solely enter through the activations, the generalization error must be a function of these order parameters and the biases θ_i and ϱ_n only. This provides the motivation for replacing the difference equations (4) for the weights $\boldsymbol{W}_i$ by difference equations for Q_{ij} and R_{in}, which replace the $\boldsymbol{W}_i$ as dynamical variables, whereas the T_{nm} are fixed and given by the task.

In the thermodynamic limit ($N\to\infty$), the dynamical order parameters Q_{ij} and R_{in} become self-averaging with respect to the randomness in the training data, i.e., their probability distributions become δ-functions at their mean value, and it is sufficient to study their mean evolution by averaging over the input distribution or rather the joint Gaussian distribution of the activations.

Although it is known that self-averaging holds for overlap-type order parameter dynamics, this is not entirely self-evident for the bias dynamics and one anticipates that the updates of the biases have to be of $O(1/N)$, i.e., the bias learning rate needs to be scaled by $1/N$. This has been confirmed by extensive simulations for a number of finite system sizes N which conclusively show that the bias dynamics are also self-averaging and their variances exhibit a $1/N$ scaling behavior. For the details of the simulations we refer the reader to Sect. III. In the case of adjustable hidden-output weights, a rigorous proof (which can be extended to apply to biases) for self-averaging for $O(1/N)$ updates is given in [7].

If one further interprets the normalized example number $\alpha=\mu/N$ as a continuous time variable, the difference equations can be conveniently rewritten as first-order coupled differential equations

$$\frac{dQ_{ij}}{d\alpha} = \eta_{\boldsymbol{w}}\,\langle\delta_i x_j+\delta_j x_i\rangle_{\boldsymbol{\xi}} + \eta_{\boldsymbol{w}}^2\,\langle\delta_i\delta_j\rangle_{\boldsymbol{\xi}}\,, \tag{7a}$$

$$\frac{dR_{in}}{d\alpha} = \eta_{\boldsymbol{w}}\,\langle\delta_i y_n\rangle_{\boldsymbol{\xi}}\,, \tag{7b}$$

$$\frac{d\theta_i}{d\alpha} = -\eta_\theta \left\langle \delta_i \right\rangle_{\boldsymbol{\xi}} . \tag{7c}$$

The scaling of the bias learning rate with $1/N$ may suggest that the dynamics of the biases and the weights are mismatched in this framework for at least some of the learning stages, leading to an optimal learning rate for the biases at infinity. This effect has already been observed in the case of adaptive hidden-output weights [7].

For dynamics on different time scales or different order of learning rates, it is natural to apply the method of adiabatic elimination [12] to the fast variables, here the hidden-output weights or biases. In this approximation, it is assumed that the fast variables driven by the large learning rates are forced to relax to an attractive fixed point of the their dynamics assuming the slow variables, i.e., input-hidden weight order parameters, to be constant. This method has already been employed successfully for adaptive hidden-output weights [7], where it has been shown also that the ensuing dynamics for the order parameters are again self-averaging. One can further show [13], that adiabatic elimination for the hidden-output weights is not only locally optimal by minimizing the generalization error with respect to the hidden-output weights instantly but also globally optimal. In the case of adiabatic elimination of the bias dynamics, neither can be shown since the equilibrium values of the biases are calculated from a set of nonlinear equations, whereas the equilibrium of the hidden-output weights is given by a set of linear equations. Furthermore, the solution of the nonlinear set of equations does not necessarily need to be unique, a problem which can be removed by demanding that the bias dynamics should relax dynamically to an attractive solution from their previous equilibrium values. A detailed treatment would therefore go beyond the scope of this paper although we will present some results derived by this approximation where deemed appropriate.

Most integrations in Eqs. (7) can be performed analytically for the choice of the error function $g_\nu(x) = \mathrm{erf}(\nu x/\sqrt{2})$ as the sigmoidal transfer function, but for single Gaussian integrals remaining for $\eta_{\boldsymbol{w}}^2$-terms and the generalization error. For the exact form of the dynamical equations and the generalization error the reader is referred to Appendix A. We

only mention in passing that the variance of the input distribution σ^2 merely rescales the weight order parameters and the weight learning rates by σ^2. The sigmoidal gain ν rescales the weight order parameters and weight learning rate by ν^2 and the biases and bias learning rate by ν. The output gain γ rescales all learning rates by γ^2. In the following these parameters are therefore set to one without loss of generality.

Before we will present some typical results for the training evolution by numerically integrating the differential equations (7), we would like to classify the huge variety of learning scenarios in this framework to some distinct generic tasks. In the original model with fixed biases [2], it has been found useful to classify a learning scenario according to the isotropy of its teacher weight vectors. Tasks with very similar norms of the hidden unit weight vectors exhibit a much longer training time than tasks with strongly graded norms, which can especially be attributed to the problem of symmetry breaking in the space of the student hidden units. This may somewhat be attributed to the identical output distributions of the individual teacher hidden units with the same norm. Only the differences in the initial student-teacher overlaps R_{in} introduced by the random initial conditions, allow the student hidden units to distinguish between the teacher hidden units in this case. For graded teacher lengths, the hidden unit output distributions still have zero mean but differ in the variance and higher cumulants. In this case, asymmetric initialization of the student-student overlaps Q_{ij} is sufficient to break student node symmetry.

The extra degrees of freedom introduced by the biases should have similar symmetry breaking effects. For simplicity, assume for the moment that the teacher weight vectors are isotropic. In the case that all teacher biases are degenerate ($\varrho_n = \varrho$), the identical hidden unit output distributions are shifted, with means

$$\langle g(y_n - \varrho_n) \rangle_{\boldsymbol{\xi}} = -g\left(\frac{\varrho_n}{\sqrt{1+T_{nn}}}\right). \tag{8}$$

Again, one finds that only asymmetric initial conditions of the student-teacher overlaps R_{in} can break the symmetry. If, however, the teacher biases are non-degenerate, the teacher hidden unit output distributions are all different, e.g., have shifted means. In this case,

asymmetric initial values of the student biases are sufficient to break the student hidden-unit symmetry. We will later see, that this symmetry breaking effect is stronger than that introduced by graded teacher lengths. For graded teachers, the only obvious choice for "degenerate" teacher biases is $\varrho_n = 0$. For non-zero teacher biases, the mean of the output distribution will shift according to Eq. (8). The choice $\varrho_n = \varrho$ leads to student hidden unit symmetry breaking even for identical initial weight vectors as long as the initial student biases are not identical as well; clearly a sign of "non-degenerate" biases when compared to isotropic teacher weights. Two other possible scaling ansätze for "degenerate" teacher biases in the case of graded teacher lengths are

$$\hat{\varrho} = \frac{\varrho}{\sqrt{1+T}}, \tag{9a}$$

$$\check{\varrho} = \frac{\varrho}{\sqrt{T}}, \tag{9b}$$

where $\hat{\varrho}$ restores identical means of the individual teacher hidden unit output distributions, whereas $\check{\varrho}$ restores identical distances of the decision hyperplane (in the following termed *abscissa*) of the sigmoidal transfer function to the origin. Neither of these ansätze (or any other ansatz inspired by numerical results) seems to restore "degenerate" teacher biases perfectly, reflecting the fact that it is impossible to preserve output distribution symmetries for non-zero means, due to the skewed distributions induced by the nonlinearity. However, once the teacher lengths and one teacher bias is fixed, one can numerically always find a set of teacher biases which exhibit at least a very slow learning progress. Unfortunately, we have not been able to find a consistent ansatz that can predict these correctly, although they are in many cases close to the values given by the ansatz (9a). In general, we have found this ansatz more useful in most cases and we will therefore term $\hat{\varrho}$ the *effective bias*.

Summarizing the above argument, it makes sense to classify teacher tasks according to the following two criteria:

- Degree of isotropy in the teacher norms. Isotropic teacher tasks are defined by similar weight vector lenghts ($T_{nm} = T\delta_{nm}$), whereas graded teachers tasks feature norms with different values. These are referred to as $\mathcal{T}^i$ and $\mathcal{T}^g$, respectively.

- Degree of degeneracy in the student biases. For isotropic teacher weights, degenerate teachers tasks are defined by similar biases ($\varrho_n = \varrho$), whereas non-degenerate teachers tasks exhibit biases with distinct values. These tasks are referred to as $\mathcal{T}_d$ and $\mathcal{T}_n$, respectively.

For graded teacher weights, degenerate biases as such are only given for $\varrho_n = 0$, although one can also find sets of non-zero biases numerically which are approximately "degenerate."

III. TYPICAL EVOLUTION OF THE DYNAMICAL EQUATIONS

The differential equations can only be solved accurately in moderate times for smaller student networks ($K \leq 5$) but any teacher size M due to the required numerical integrations. For small learning rates, where $\eta_{\boldsymbol{w}}^2$-terms can be neglected, the differential equations can be solved for any K. For the remainder of the paper, we would like to focus on the influence of different bias scenarios and the influence of the learning rates. We therefore restrict ourselves otherwise mainly to small realizable networks ($K = M$ with $K = 2, 3$) and uncorrelated isotropic teacher weight vectors of arbitrary length ($T_{nm} = T\delta_{nm}$).

The dynamical evolution of the overlaps Q_{ij}, R_{in} and the biases θ_i follows from integrating the equations of motion (7) from initial conditions determined by the (random) initialization of the student weights $\boldsymbol{W}_i$ and biases θ_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}, the biases θ_i 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 $[0, 1]$, $[-1, 1]$, and $[-10^{-12}, 10^{-12}]$ intervals, respectively. We find that the results of the numerical integration are sensitive to these random initial values which has not been the case to this extent for fixed biases. To study the effect of different weight initialization, we have fixed the initial values of the student-student overlaps Q_{ij} and biases θ_i for some of the numerical examples, as these can be manipulated freely in any learning scenario. The

initial student-teacher overlaps R_{in} are always randomized as suggested above.

In our first example (Fig. 1), we demonstrate the potential influence of the adjustable biases in the learning dynamics of the soft-committee machine model, by comparing two typical realizable learning tasks ($K = M = 2$) with isotropic teacher weight vectors $\mathcal{T}^i$ ($T_{nm} = \delta_{nm}$). The student parameters denoted by $*$ represent a learning scenario in the original model, where both student and teacher lack biases, i.e., $\theta_i = 0$ and $\varrho_n = 0$. The other scenarios feature student networks from the extended model, i.e., with adjustable biases. They are trained by an isotropic teacher task with small non-degenerate biases ($\varrho_{1,2} = \mp 0.1$). For both scenarios, the learning rate and the initial conditions were judiciously chosen to be $\eta_0 = 2.0$, $Q_{11} = 0.1$, $Q_{22} = 0.2$, $\hat{R}_{in} = \hat{Q}_{12} = U[-10^{-12}, 10^{-12}]$ with $\theta_1 = 0.0$ and $\theta_2 = 0.5$ for the student with adjustable biases.

In both cases, the student weight vectors [Fig. 1(a)] are drawn quickly from their initial values into a suboptimal symmetric phase, characterized by the lack of specialization of the student hidden units on a particular teacher hidden unit, as can be depicted from the similar values of R_{in} in Fig. 1(b). This symmetry is broken almost immediately in the learning scenario with adjustable student biases and non-degenerate teacher biases. The student converges quickly to the optimal solution, characterized by the evolution of the overlap matrices $\mathbf{Q}$, $\mathbf{R}$ and biases $\boldsymbol{\theta}$ [see Fig. 1(c)] to their optimal values $\mathbf{T}$ and $\boldsymbol{\varrho}$ (up to the permutation symmetry due to the arbitrary labeling of the student nodes). Likewise, the generalization error ϵ_g decays to zero in Fig. 1(d). The student with fixed biases is trapped for most of its training time in the symmetric phase before it converges eventually.

Before analyzing the differences between the original soft-committee and the extended model further, we would like to briefly assess the influence of finite input dimension N on the dynamics, especially in order to confirm that the dynamic variables are self-averaging. In Fig. 1 we therefore also compare the theoretical evolution of the overlaps, the biases and the generalization error with the simulation results for input dimensions $N = 10 \ldots 500$, for the above student and teacher scenario with adjustable biases. The initialization for the simulations are identical to the theory for the student norms and biases, but the overlaps

were scaled appropriately with input dimension ($\hat{R}_{in} = \hat{Q}_{12} = U[-N^{-1/2}, N^{-1/2}]$).

Since the learning trajectory for finite N is stochastic, there is a probability for a student node permutation in the specialization process leading to multimodal probability distributions of the dynamic variables. To be able to calculate meaningful mean trajectories and variances, student nodes were therefore relabeled a posteriori. However, this permutation probability decreases in the simulations with $1/N^3$, leading to a well defined deterministic behavior in the thermodynamic limit, i.e., the probability distributions of the dynamic variables become asymptotically unimodal. The resulting mean trajectories of the dynamic variables are shown for two input dimensions ($N = 10, 100$) in Figs. 1(a)–1(c), where some of the order parameters (Q_{22}, R_{22}, and R_{21}) were omitted as they have very similar values to others (Q_{11}, R_{11}, and R_{12}) due to the symmetry in the learned task. The size of the symbols is only a guide to the eye, but is generally much larger than the standard deviation in the mean. Even for the smallest input dimension of $N = 10$, the agreement of the simulations with the theoretical predictions is qualitatively good but the trajectories exhibit a systematic shift to smaller α values. For $N = 100$ the finite size effects on the mean trajectory are already very small. For comparison, the simulated value of the generalization error in Fig. 1(d) for larger input dimensions ($N = 200, 500$) are already virtually indistinguishable from the theoretical predictions. In general, one finds that the deviations of the mean from their thermodynamic predictions and the variances of the dynamical fluctuations scale with $1/N$ as expected [4].

One of the most striking differences between the soft-committee machine with and without biases is the length of the symmetric phase for non-degenerate teacher biases. In the model with fixed biases, the symmetric phase dominates the overall training time for the isotropic teacher scenarios in back-propagation training even for an optimized learning rate [14,15], as the training time grows linearly with more than K^2 in the symmetric phase and only with K in the convergence phase. For small learning rates the trapping time is furthermore linearly extended with η_0. The influence of the initial conditions is only log-

arithmic through the differences in the initial student-teacher overlaps R_{in} [16] which are typically of $O(1/\sqrt{N})$ and cannot be influenced in real scenarios without *a priori* knowledge. The initialization of the biases, however, can be controlled by the user and its influence on the learning dynamics is shown in Figs. 1(c) and 1(d) for the biases and the generalization error, respectively. For initially identical biases ($\theta_1 = \theta_2 = 0$), the evolution of the order parameters and hence the generalization error is almost indistinguishable from the fixed biases case. A breaking of this symmetry leads to a decrease of the symmetric phase linear in $\log(|\theta_1 - \theta_2|)$ until it has all but disappeared. The dynamics are again slowed down for very large initialization of the biases [see Fig. 1(d)], where the biases have to be modified significantly before reaching their optimal values.

The influence of bias dynamics in the case of degenerate teacher biases is demonstrated in Fig. 2; here we show the evolution of the overlaps, the biases and the generalization error from random initial conditions for $K = 3$ and a common learning rate ($\eta_0 = \eta_\theta = \eta_{\boldsymbol{w}} = 2$) for a realizable task ($M = 3$) with isotropic weight vectors ($T_{nm} = \delta_{nm}$) and degenerate but non-zero biases ($\varrho_n = 1$). As before the student-student overlaps [Fig. 2(a)] are quickly drawn into a symmetric subspace, characterized by similar overlaps R_{in} [Fig. 2(b)] between each student node and all teacher nodes. The student biases [Fig. 2(c)] take values which are symmetrically grouped around the true degenerate teacher biases. The breaking of the symmetry occurs in two stages. First, the third hidden unit, whose single student bias is located closest to the true bias value, begins to specialize on the third teacher unit. The other two student units decorrelate from the third and its associated teacher unit, but remain strongly correlated with each other and the two other teacher units. The two biases keep their symmetry around the true teacher bias value. These symmetries are eventually also broken and the student finally converges to the optimal solution. Although the evolution is therefore still characterized by three learning stages, transient to the symmetric phase, breaking of the symmetry and final convergence, similar to the evolution of the model with fixed biases, the extra degrees of freedom introduced by the biases enrich the dynamical evolution considerably.

To contrast the training behavior in this very symmetric task $\mathcal{T}_d^i$ with the three other generic tasks that exhibit less symmetry, we introduce small deviations from the original symmetry by choosing $T_{nm} = (1 + 0.1n)\delta_{nm}$ instead of $T_{nm} = \delta_{nm}$ for teacher overlaps and/or $\varrho_n = 0.8 + 0.1n$ instead of $\varrho_n = 1$ for the biases. These deviations have a dramatic effect on the evolution of the generalization error in Fig. 2(d). The task $\mathcal{T}_d^i$ has by far the slowest training behavior, with the sequential specialization process already described above for the order parameters. This is followed by the approximation to the task $\mathcal{T}_d^g$ which also features a sequential breaking of the symmetry but on a much shorter time scale. The fastest training times are exhibited for tasks $\mathcal{T}_n^g$ and $\mathcal{T}_n^i$ with no measurable speed up for the graded task, suggesting that non-degenerate biases affect the breaking of node symmetry more significantly than graded weight vectors. The strong symmetry breaking effect of the biases is arguably due to a steep minimum in the generalization error surface along the direction of the biases caused by the shift of the means of the individual hidden unit output distributions. This picture can be confirmed by the fact that the trajectories of the biases do not cross, i.e., the rank ordering according to the value of the bias is preserved at all times, whereas the ordering according to the norms is not. We have found this to be true for a range of other learning scenarios studied, including larger networks and more strongly graded teachers, provided that the biases were not initialized highly symmetrically. This seems to promote initialization schemes where the biases of the student hidden units are spread evenly across the input domain as has been suggested previously on a heuristic basis [17].

For the cases of degenerate teacher biases, the grouping of student biases found above is typical for all cases studied. For an even number of degenerate teacher biases, the student units combine in pairs. Each pair is characterized by its two biases having the same distance to the true teacher bias value with opposite sign and by its weight vectors being highly correlated. For an odd degeneracy, as above, the behavior is similar but for a single remaining student bias which is stabilized around the true teacher bias value. The breaking of the symmetries in these cases can take a lot longer than for fixed biases and can be extremely

complicated. It is often broken in stages as in the example given above, but can also occur simultaneously. We also find a strong influence of the training outcome on the initial conditions and the learning rate chosen, in some cases not all symmetries are broken and the student remains trapped in a suboptimal configuration, i.e., some of the symmetric fixed points are attractive.

To illustrate this point, the dynamics of the student biases θ_i are shown in Fig. 3 for $K = M = 2$, $\eta_0 = 1$ and random initial conditions, and an isotropic teacher with degenerate biases ($\varrho_n = 0$). The student was initialized identically for the different runs (i.e., the same seed was used for the random number generator), but for a change in the range of the random initialization of the biases ($U[-b, b]$). We find that the student progress is inversely related to the magnitude of the bias initialization until a critical value of b is reached, where the student fails to converge at all. It remains in a suboptimal phase characterized by biases of the same large magnitude but opposite sign and highly positively correlated weight vectors which have identical overlap with all respective teacher vectors. This behavior may be explained by the fact that the generalization error decreases with increasing magnitude of the symmetric bias arrangement in the symmetric phase, suggesting the possibility of a local minimum in the generalization error surface. This may cause the dynamic competition between the specialization process of the student hidden units and the increase in magnitude of the biases observed in Fig. 3, where the basin of attraction is determined by the initial conditions and the learning rates. Fastest convergence for this scenario is achieved for $b = 0$ and a reasonable bias initialization strategy seems therefore almost opposite to the above case of non-degenerate teacher biases.

In order to devise an initialization strategy which can cope well with all learning scenarios, we explore the influence of the initial conditions and the learning rate on the learning process more systematically in the following section.

IV. ATTRACTIVE FIXED POINTS

Although attractive symmetric fixed points have been found also for the soft-committee machine model with fixed biases [16], these needed careful preparation of the initial conditions and were restricted to over-realizable cases. In the case of adaptive biases, one finds a multitude of attractive sub-optimal fixed points for realizable cases with, in some cases, large basins of attraction. They exist not only in cases where both teacher weight vectors are isotropic and the biases degenerate but also for graded teachers and non-degenerate biases, although in these cases, the basins of attraction tend to shrink with increasing task asymmetry. In real world problems, the problem of poor local minima and the influence of the initial conditions on these is well known for back-propagation training. One can find numerous examples in the literature (e.g., [18,19]) which produce training error dynamics that look very similar to the evolution of the generalization error found in this work.

Subsequently, many algorithms (see e.g., [20] and references therein) have been proposed that aim to find good initial conditions. However, we are aware only of two [17,18] which do not rely on information extracted from an *a priori* known training set and are therefore the only ones applicable in the framework studied. Below, we will therefore try to gain a qualitative understanding of how the initial conditions and the learning rates can be chosen to avoid becoming attracted to suboptimal network solutions. Our findings are then compared to the heuristicly based suggestions in [17,18].

Due to the quadratic increase in the number of dynamic variables with the system size K, we restrict ourselves to the the smallest network size $K = 2$, although we have verified the validity of the drawn conclusions for larger networks. In particular, three elements which influence the size of the basin of attraction for given initial conditions were investigated: the task asymmetry (in terms of the teacher lengths and biases), the initial conditions and the learning rates.

Since the initialization space and hence the basins of attraction are still of high dimensionality, we have restricted ourselves to one-dimensional slices in one of the biases, θ_2,

parameterized by a further variable. The remaining variables of the student were chosen to be $\eta_\theta = \eta_{\boldsymbol{w}} = 2.0$, $Q_{11} = 0.1$, $Q_{22} = 0.2$, $\theta_1 = 0.0$, and $\hat{R}_{in} = \hat{Q}_{12} = U[-10^{-12}, 10^{-12}]$ (with a fixed random seed). The teacher task was usually chosen to be of the form $\mathcal{T}_n^i$ with $T_{nm} = \delta_{nm}$ and $\varrho_n = 0$, if not otherwise stated. The convergence time $\alpha_{\rm C}$ was defined as the example number at which the generalization error has decayed to a small value, here judiciously chosen to be 10^{-8} requiring the student to have broken the symmetries in weight space successfully. The convergence time diverges in the case that the student is attracted to a suboptimal fixed point.

A. Task asymmetry

In Figs. 4–6 we compare the influence of the initialization of θ_2 on the convergence time $\alpha_{\rm C}$ and the resulting basin of attraction for three different teacher tasks of the form $\mathcal{T}_d^i$, $\mathcal{T}_n^i$ and $\mathcal{T}_d^g$, where some sort of asymmetry was applied gradually to the original teacher task ($T_{nm} = \delta_{nm}$ and $\varrho_n = 0$).

In the case of degenerate teacher biases $\mathcal{T}_d^i$ (Fig. 4) for which the biases were chosen to be $\varrho_n = \varrho$, the convergence time diverges beyond some critical absolute values $\theta_c^{\pm}$ of θ_2 and the basin of attraction to the optimal solution is restricted to $\theta_{\rm crit}^+ < \theta_2 < \theta_{\rm crit}^-$. For small ϱ this basin is symmetric ($\theta_{\rm crit}^+ = \theta_{\rm crit}^-$) and almost constant in size, whereas for large ϱ, the basin is skewed and increases in size. The fastest convergence is always achieved for $\theta_2 = \theta_1 = 0$, i.e., when the teacher task degeneracy is reflected in the bias initialization. This effect becomes increasingly more pronounced for larger teacher bias values ϱ, which also generally show shorter convergence times. This effect may be explained by the fact that for small ϱ most examples are drawn from the region where the sigmoidal transfer function is linear, making the symmetry breaking process more difficult.

This behavior is to be contrasted to the case of non-degenerate teacher bias tasks $\mathcal{T}_n^i$ characterized by $\varrho_n = \pm\varrho$ shown in Fig. 5. Here, one finds that the basin of attraction to the optimal solution already increases substantially for very small values of ϱ, although we

still find that the student is drawn into a suboptimal solution for large enough initial θ_2. However, above a certain value in the teacher bias asymmetry $\varrho_{\text{crit}} \approx 0.174$, the suboptimal solution ceases to be an attractive fixed point, although the dynamics can still be slowed down considerably due to the influence of the symmetric fixed point. Above ϱ_{crit} and very large initial values θ_2, one finds that the convergence time increases exponentionally with θ_2, arguably due to the fact that the student hidden unit is initially highly saturated and the gradient decreases exponentionally.

We further find that the basin of attraction is always perfectly symmetric, unlike in the degenerate case since the hidden unit symmetry is broken by the biases and not the weights. This also explains the sharp peak in the convergence time for initial values around $\theta_2 = 0$ with

$$\alpha_{\text{C}}(\theta_2') - \alpha_{\text{C}}(\theta_2) \propto \log\left(\frac{|\theta_2|}{|\theta_2'|}\right) \tag{10}$$

for small initial values θ_2' and θ_2, as already shown in Fig. 1(d). Eq. (10) holds exactly in the limit $\theta_2 \to 0$ only for $R_{in} = 0$, in which case the the convergence time diverges as only the biases can break the symmetry. Otherwise, the convergence time is affected by the specialization process triggered by the asymmetric initial conditions in R_{in}. This is also true for the other laws [Eqs. (11) and (12)] found below.

Similarly, the shortest possible convergence time decreases initially with increasing task asymmetry according to

$$\alpha_{\text{C}}^{\text{opt}}(\varrho') - \alpha_{\text{C}}^{\text{opt}}(\varrho) \propto \log\left(\frac{\varrho}{\varrho'}\right), \tag{11}$$

and the minimum becomes sharper in terms of θ_2 for large ϱ. This minimum defines the optimal initial value $\theta_2^{\text{opt}}(\varrho)$, which increases as expected with increasing ϱ, but is always considerably larger than ϱ. This effect is especially remarkable when taking the initial student norm into account, comparing the actual effective bias or alternatively the abscissa of the hidden units (i.e., $\varrho/\sqrt{1+T}$ and $\theta_2/\sqrt{1+Q_{22}}$ or $\varrho/\sqrt{T}$ and $\theta_2/\sqrt{Q_{22}}$).

The graded teacher task $\mathcal{T}_d^g$ also speeds up the breaking of hidden unit symmetry as shown in Fig. 6 and reduces the optimal convergence time $\alpha_{\text{C}}^{\text{opt}}$ substantially. The difference

in convergence time due to a small task asymmetry is given in terms of the teacher length difference $\delta T = T_{22} - T_{11}$ by

$$\alpha_{\rm C}^{\rm opt}(\delta T') - \alpha_{\rm C}^{\rm opt}(\delta T) \propto \log\left(\frac{\delta T}{\delta T'}\right). \tag{12}$$

The total reduction in $\alpha_{\rm C}$ for a given asymmetry is smaller when compared to $\mathcal{T}_n^i$. This confirms the observation made in Sect. III that the biases have a stronger symmetry breaking effect than the weights. This is also mirrored in the basin of attraction increase, which is not as substantial as in the case of asymmetric biases, and the critical bias $\theta_{\rm crit}$ follows approximately $\theta_{\rm crit}(\delta T) - \theta_{\rm crit}(0) \propto \delta T^{0.141(3)}$.

We have found qualitatively similar results for larger networks, where the basin of attraction to the optimal solution also grows with the teacher task asymmetry. However, one also finds that the range of initial conditions attracted to the optimal solution shrinks with network size for a given teacher task asymmetry (e.g., $\varrho_n - \varrho_{n-1} = 0.1$) and the number of suboptimal attractive fixed points grows significantly. We have found this to be true especially where the asymmetry is purely in the weight vectors.

B. The initial conditions

Since the largest basin of attraction to the suboptimal fixed point is found for learning scenarios with degenerate teacher biases, we will investigate the influence of the other initial conditions and the learning rates for the task $T_{nm} = \delta_{nm}$ and $\varrho_n = 0$.

In Fig. 7 it is shown that the influence of the initialization of the first bias θ_1 consists almost exclusively of a linear shift in the range of initial θ_2 values that lead to convergence of the training. In particular, we find that the results become invariant under the transformation $\theta_2' = \theta_2 - 0.9745(9) \times \theta_1$, i.e., the basin of attraction depends almost solely on the difference $\theta_2 - \theta_1$. This is somewhat surprising since one may have assumed that the basin of attraction should depend on the individual abscissas or the effective biases of the student.

In Fig. 8 the basin of attraction for different initial student lengths is shown. All the initial student-student overlaps were magnified from their original values [21] by factors M

given in the legend. The influence of the student lengths is clearly twofold. First, the basin of attraction in θ_2 grows approximately with $0.068(5) + 0.331(6) \times M^{0.445(8)}$, making the training process less sensitive to the initial bias values. However, this growth translates into a decrease of the critical abscissa since Q grows with M, which could be interpreted as another sign that the raw initial values are the crucial parameters and not the abscissas. Second, the optimal convergence time is slowed down slightly for increasing M and one finds approximately $\alpha_{\mathrm{C}}^{\mathrm{opt}} = 643(1) + 12(1) \times M^{0.34(3)}$.

Similarly in Fig. 9, we assess the influence of finite size effects on the basin of attraction through the typical initial normalized student-teacher overlaps $\hat{r} = O(1/\sqrt{N})$ (ignoring other stochastic finite size effects). As predicted in [16], the optimal convergence time is reduced linearly in $\log(\hat{r})$ [$\alpha_{\mathrm{C}} = 187.70(7) - 16.923(4) \times \log(\hat{r})$]. More relevant for the purpose of this work is the increase in the basin of attraction to the optimal solution with the critical initial bias $\theta_{\mathrm{crit}} = 0.370(1) + 0.507(5) \times \hat{r}^{0.103(1)}$.

The results found for $K = 2$ again carry over qualitatively to larger networks with the decrease in the basin of attraction with network size as already mentioned in Sect. IV A. Especially interesting in this respect is, that even for $K = 2$, the maximal initial abscissas that guarantee convergence for the case of degenerate teacher biases are generally smaller than the size of the input domain, a tendency which becomes more emphasized for larger networks. These results therefore contradict heuristics presented in [17], where it has been suggested to spread the abscissas across the input domain. In [17], it also has been assumed implicitly that the abscissas are the relevant quantities, whereas our work indicates that the raw bias values are more important in determining the basin of attraction.

C. The learning rates

Beside the initial conditions and the teacher task to be learned, the learning rates used also strongly influence the learning process. In Fig. 10 the convergence time as a function of θ_2 is shown for a range of common learning rates η_0. For convenience, the convergence time

has been normalized with $1/\eta_0$. One finds that the convergence time diverges for all learning rates, above a critical initial value of θ_2. For increasing learning rates, this transition first becomes sharper and occurs at smaller θ_2 until the learning rate is reached that provides the fastest convergence to the optimal solution for small θ_2, beyond which the basin of attraction widens again.

The increase of the basin of attraction has been postulated in [18], however, the functional relationship given ($\eta_0 < Q_{ii} + \theta_i^2$) cannot be supported by our findings. It is not only quantitatively incorrect, it also fails to predict a finite boundary for an infinitesimal small learning rate. This work further does not account for interaction between the hidden units and the different roles of weights and biases in determining the basin of attraction (see Sect. IV B).

In Figs. 11 and 12 it is shown that it can be beneficial to separate the weight and bias learning rates. In Fig. 11 the normalized convergence time $\widehat{\alpha_C}(\theta_2)$ is plotted for fixed bias learning rate ($\eta_\theta = 2$) but allowing for variations in the weight learning rate $\eta_{\boldsymbol{w}}$. One can readily see that the basin of attraction increases when the weight and bias learning rates are well separated. This advantage, however, is relative as a very small weight learning rate increases the convergence time linearly.

Similarly in Fig. 12, the convergence time $\alpha_C(\theta_2)$ is shown for fixed weight learning rate ($\eta_{\boldsymbol{w}} = 2$) but variable bias learning rate η_θ. Again, the basin of attraction is clearly enlarged when separating the time scale for the training of biases and weights. Whereas training is slowed down for small bias learning rates, this is not the case for large η_θ where the basin of attraction increases to very large values. It is therefore more reasonable to achieve the desirable separation of the learning rates by choosing a large bias learning rate. In fact, a maximal bias learning rate does not exist in this scenario, suggesting a possible different scaling. It further poses the question whether in this case the basin of attraction encompasses the whole space of initial conditions.

Unfortunately, a closer inspection using larger networks and other learning tasks reveals several limitations of large bias learning rates and adiabatic elimination. First of all, the use

of adiabatic elimination for very small α leads to extremely large initial equilibrium values of opposing signs for the biases, effectively cancelling the outputs of pairs of hidden units. This effect can be attributed to the initial lack of information about the teacher, reflected by the inherently small values of the student-teacher overlaps R_{in} favoring the hidden units to be switched off effectively Consequently, the progress of the student weights is inhibited to such an extent that training does not converge in finite time for all practical purposes [22]. Similarly, very large but finite bias learning rates also slow down the training time due to the biases blowing up in the very early stages of learning. It is therefore necessary to restrict the bias learning rate for very small α, i.e., for the initial transient, to a finite value. It is unclear, whether this is also a problem for finite size systems where adiabatic elimination corresponds to a bias learning rate of $O(1)$ instead of $O(1/N)$.

Even when adiabatic elimination or a very large bias learning rate are only triggered once training has reached the stable symmetric plateau, their usefulness in terms of basin of attraction enlargement is, in general, not pronounced for larger networks. In fact, using large bias learning rates can actually decrease the basin of attraction to the optimal network parameters especially in degenerate bias tasks with isotropic weight vectors, e.g., training with a bias learning rate above $\eta_\theta = 3$ in the learning scenario of Fig. 2 converges to a suboptimal fixed point.

However, once all hidden unit symmetries have been broken, adiabatic elimination or a very large bias learning rate can be employed in all circumstances and generally results in slightly faster training when compared to using a finite learning rate. This will be investigated analytically in more detail in the following section.

V. ANALYSIS OF THE CONVERGENCE PHASE

For the soft-committee machine model with fixed zero biases, realizable learning scenario ($K = M$), and isotropic teachers ($T_{nm} = T\delta_{nm}$), the order parameter space could be very well characterized throughout the learning process by similar diagonal and off-diagonal elements

of the overlap matrices $\mathbf{Q}$ and $\mathbf{R}$, simplifying the linear analysis around the symmetric and zero generalization error fixed points [14] considerably since the number of dynamic variables could be reduced to four.

For the model with dynamic biases this dimensionality reduction for the equivalent teacher task with isotropic weights and degenerate biases is in general not a good approximation as can be clearly seen in Fig. 2. However, if the student biases are initialized quite symmetrically, we find the ansatz

$$Q_{ij} = Q\delta_{ij} + C(1 - \delta_{ij}), \tag{13a}$$

$$R_{in} = R\delta_{in} + S(1 - \delta_{in}), \tag{13b}$$

$$\theta_i = \theta \tag{13c}$$

to be justified for the student-student overlaps, (apart from a relabeling of the student nodes) student-teacher overlaps, and the student biases in the convergence phase.

The reduction of the number of order parameters from $O(K^2)$ to just five allows us to analyze the learning dynamics in the convergence phase as a function of the network size K, the length of the teacher hidden units T, the size of the teacher biases ϱ, and the user adjustable learning rates η_0 and η_θ.

A. The eigenvalue spectrum

In order to predict the optimal learning rates for the convergence phase, we linearize the equations of motion (A4) in $\{R, Q, C, S, \theta\}$ around the zero generalization error fixed point $R^* = Q^* = T$, $S^* = C^* = 0$ and $\theta^* = \varrho$ (see Appendix B). The matrix $\mathbf{M}$ of the resulting system of five coupled linear differential equations in $r = T - R$, $q = T - Q$, $s = S$, $c = C$ and $\vartheta = \varrho - \theta$ has two sets of eigenvalues.

Two eigenvalues ($\lambda_{1,2}$) are the solutions to a quadratic equation (B3) consisting of the same matrix elements of $\mathbf{M}$ as in the fixed bias case and are therefore independent of the bias learning rate η_θ. These eigenvalues are nonlinear in the learning rate $\eta_{\boldsymbol{w}}$ and λ_1 becomes

positive for large enough $\eta_{\boldsymbol{w}}$. The other three eigenvalues ($\lambda_{3,4,5}$) are the solution to a cubic equation (B4). These eigenvalues depend on both learning rates and are negative for all values of $\eta_{\boldsymbol{w}}$ and η_θ. These eigenvalues are minimized with respect to η_θ in the limit $\eta_\theta \to \infty$, i.e., the optimal bias learning rate in the convergence phase is at infinity (for a more detailed discussion see Appendix B). Below, we will therefore restrict ourselves to the study of two learning rate parameterizations: a common learning rate $\eta_0 = \eta_{\boldsymbol{w}} = \eta_\theta$ or the weight learning rate $\eta_{\boldsymbol{w}}$ with the bias learning rate η_θ eliminated by taking the limit $\eta_\theta \to \infty$. We will adopt the convention to use a generic learning rate η and eigenvalues λ whenever a statement is applicable for both parameterizations, whereas parameterization dependent symbols denoted by superscripts or subscripts are used otherwise.

The behavior of the eigenvalues described above is graphically illustrated for both learning rate parameterizations in Fig. 13(a) for $K = 5$, $T = 1$, and $\varrho = 1$. Within these parameterizations, the eigenvalues $\lambda_{3,4,(5)}$ are linear in η, whereas $\lambda_{1,2}$ have higher orders in η. $\lambda_{1,2}$ are identical for both parameterizations since they are functions of $\eta_{\boldsymbol{w}}$ only, whereas the slopes of $\lambda_{3,4}$ are clearly minimized for the parameterization $\eta_\theta \to \infty$ ($\lambda_5^{\boldsymbol{w}}$ is omitted since $\lambda_5 \to -\infty$ for $\eta_\theta \to \infty$). One can further distinguish between two slow modes associated with eigenvalues λ_1 and λ_3 and three fast modes associated with eigenvalues λ_2 and $\lambda_{4,5}$, 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 is more closely illustrated in Fig. 13(b) in the same learning scenario. 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 $\eta_{\max}$ as $\lambda_1(\eta_{\max}) = 0$. For all $\eta < \eta_{\max}$ the generalization error decays exponentionally to $\epsilon_{\mathrm{g}}^* = 0$.

B. The optimal dynamics

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, c, and ϑ (B8). Numerically, we find that the eigenvector associated with the linear eigenvalue λ_3 is orthogonal to the first-order terms in the generalization error and can therefore not contribute to their decay, but controls only the decay of second-order term with $2\lambda_3$.

The learning rate η^{opt} which provides the fastest asymptotic decay rate λ^{opt} of the generalization error is therefore given by the condition

$$\lambda^{\mathrm{opt}} = \left| \min_{\eta} \left[\max \left(\lambda_1, 2\lambda_3 \right) \right] \right| . \tag{14}$$

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. In Fig. 13(b) one finds that for this particular case the fastest decay is achieved at the minimum of λ_1 for $\eta_\theta \to \infty$ but at the root of $\lambda_1 - 2\lambda_3$ for $\eta_\theta = \eta_{\boldsymbol{w}}$.

Unfortunately, the calculation of λ^{opt} (and η_0 or $\eta_{\boldsymbol{w}}$) via Eq. (14) and the determination of the kind of optimum is analytically infeasible for general K, T and ϱ. However, for some special cases further analytical progress can be made: $K \to \infty$, $T \to \infty$ and $T \to 0$. For the T limits, it is necessary to adopt a scaling for the teacher bias ϱ, and we have used both natural scaling ansätze (see Eq. (9) in Sect. II). These analytic limits are studied in detail in Appendices B 1–B 5 and the main results will be referred to in the discussion of the appropriate figures and are summarized in Table I.

1. The critical teacher length T^{crit}

We find that in the small-T limit, the optimum is always given by the minimum of λ_1 and both learning rate parameterizations are identical, whereas for the large-T limit, the root solution ($\lambda_1 = 2\lambda_3$) applies resulting in a faster decay for $\eta_\theta \to \infty$. For finite T there exists

a $T^{\text{crit}}(K, \varrho)$, which depends on the kind of learning rate parameterization and divides these two solution regimes. The functional dependence of T_0^{crit} and $T_{\boldsymbol{w}}^{\text{crit}}$ is graphically illustrated in Fig. 14 as a function of ϱ for a range of K values including the $K \to \infty$ limit, where it is implicitly assumed that $\exp \varrho^2 \ll K$.

In Fig. 14(a) T_0^{crit} decreases monotonically with ϱ. The $K \to \infty$ limit exhibits a finite limit ($T_0^{\text{crit}} \approx 0.21$) for $\varrho \to \infty$, but acquires a power-law decay $T_0^{\text{crit}} \propto \varrho^{-2}$ for all finite K [see inset of Fig. 14(a)]. For $T > T_0^{\text{crit}}(K, 0) \approx 1.278$, the root solution applies for all ϱ due to monotonously decreasing T_0^{crit}, whereas for all other T values the solution type changes from the minimum to the root above a T and K dependent value of ϱ. The dependence of T_0^{crit} on K is relatively weak and varies with ϱ. For small ϱ ($\varrho \lessapprox 0.45$), T_0^{crit} increases with K, whereas for medium ϱ ($0.45 \lessapprox \varrho \lessapprox 1.64$), T_0^{crit} decreases with K. Above $\varrho \gtrapprox 1.64$, T_0^{crit} increases again with K and reaches the qualitatively different solution for finite and infinite K.

On the other hand, $T_{\boldsymbol{w}}^{\text{crit}}$ does not behave monotonically in ϱ (with the exception of $K = 2$) as shown in Fig. 14(b). It also decreases initially like T_0^{crit} up to $\varrho \approx 1.3$, but then increases up to a maximum whose height and position increases in K, before it falls towards the asymptotic value of $T_{\boldsymbol{w}}^{\text{crit}}(K, \infty) = 1/2$ for all finite K. We again find a qualitatively different behavior for $K \to \infty$ as $T_{\boldsymbol{w}}^{\text{crit}}$ grows unabatedly with ϱ. Depending on the value of K and T, the type of solution can therefore change up to three times for increasing ϱ. Similar to T_0^{crit}, we also find that the $T_{\boldsymbol{w}}^{\text{crit}}$ grows with K initially ($\varrho \lessapprox 0.52$), then decreases ($0.52 \lessapprox \varrho \lessapprox 1.97$) and then increases again.

It is also clear from the graphs and from the fact that $\lambda_3^{\boldsymbol{w}} \leq \lambda_3^0$, that $T_{\boldsymbol{w}}^{\text{crit}}$ must be greater than T_0^{crit} for all K and ϱ besides $\varrho = 0$ where $T_{\boldsymbol{w}}^{\text{crit}} = T_0^{\text{crit}}$. We can therefore divide the optimal convergence behavior for all K, T, and ϱ into three regimes:

1. $T \leq T_0^{\text{crit}}(K, \varrho) \leq T_{\boldsymbol{w}}^{\text{crit}}(K, \varrho)$: The minimum of λ_1 defines the optimum and both learning rate parameterizations behave identically ($\lambda_{\boldsymbol{w}}^{\text{opt}} = \lambda_0^{\text{opt}}$ and $\eta_{\boldsymbol{w}}^{\text{opt}} = \eta_0^{\text{opt}}$).

2. $T_0^{\text{crit}}(K, \varrho) < T < T_{\boldsymbol{w}}^{\text{crit}}(K, \varrho)$: The optimal solution is different for both parameteriza-

tion. The minimum of λ_1 is still optimal for $\eta_\theta \to \infty$, but $\lambda_1 - 2\lambda_3 = 0$ defines the optimum for $\eta_\theta = \eta_{\boldsymbol{w}}$. The optimal convergence rates and learning rates are different with $\lambda_{\boldsymbol{w}}^{\text{opt}} > \lambda_0^{\text{opt}}$ and $\eta_{\boldsymbol{w}}^{\text{opt}} < \eta_0^{\text{opt}}$.

3. $T_0^{\text{crit}}(K, \varrho) \leq T_{\boldsymbol{w}}^{\text{crit}}(K, \varrho) < T$: Although the optimal solution is now the root of $\lambda_1 - 2\lambda_3$ for both parameterizations, we still find $\lambda_{\boldsymbol{w}}^{\text{opt}} \geq \lambda_0^{\text{opt}}$ and $\eta_{\boldsymbol{w}}^{\text{opt}} \leq \eta_0^{\text{opt}}$ since $\lambda_3^{\boldsymbol{w}} \leq \lambda_3^0$.

Since the 3-dimensional parameter space is difficult to visualise, we study the optimal convergence exemplary for two slices.

2. Optimal dynamics in K-ϱ space

In Fig. 15 we show the convergence behavior of the parameterization $\eta_\theta = \eta_{\boldsymbol{w}} = \eta_0$ [Figs. 15(a)–15(c)] in comparison to $\eta_\theta \to \infty$ [Figs. 15(d)–15(f)] as a function of K for $T = 1$ and a range of ϱ values. In Fig. 15(a) one can see that the optimal learning rate η_0^{opt} is hardly K dependent for small ϱ (beside the inherent rescaling with $1/K$ implied by the normalization of the soft-committee machine), but increases proportionally to K for large ϱ before it eventually levels off at a ϱ dependent value. The $K \to \infty$ analysis suggests a scaling of the optimal learning rate with $\log \eta_0^{\text{opt}} \propto \varrho^2$ since the maximal learning rate scales in this fashion. This is mirrored in the behavior of the optimal convergence rate in Fig. 15(b) (for graphical purposes multiplied by K) which exhibits the expected $1/K$ behavior for small ϱ. For large ϱ, however, the increase in $\eta_0^{\text{opt}} \propto K$ for small K causes λ_0^{opt} to be constant until η_0^{opt} levels off, when λ_0^{opt} reverts back to the $1/K$ decay. We further note that the absolute value of the convergence rate λ_0^{opt} initially increases for small ϱ for all values of K, which is a T-dependent effect we will study in more detail below. In Fig. 15(c) we further show the normalized difference between the maximal and optimal learning rate defined as

$$\Delta\eta_{\max}^{\text{opt}} = \frac{\eta_{\max} - \eta^{\text{opt}}}{\eta^{\text{opt}}}.$$

We find that $\Delta\eta_0{}_{\max}^{\text{opt}}$ initially increases with ϱ for all K, which is again a feature dependent on T, before it decreases monotonically, reflecting a steeper and more skewed curve for λ_1.

To compare the two learning rate parameterizations, the ratio of the optimal learning rates $\eta_{\boldsymbol{w}}^{\text{opt}}$ and η_0^{opt} shown in Fig. 15(d) shows that for small ϱ the ratio is identical since $T = 1 < T_0^{\text{crit}} < T_{\boldsymbol{w}}^{\text{crit}}$. For increasing ϱ the ratio falls below 1 since η_0^{opt} is now determined by the root of $2\lambda_3 - \lambda_1$ ($T_0^{\text{crit}} < T < T_{\boldsymbol{w}}^{\text{crit}}$). Increasing ϱ even further, one finds that also $\eta_{\boldsymbol{w}}^{\text{opt}}$ is determined initially by the root solution ($T_0^{\text{crit}} < T_{\boldsymbol{w}}^{\text{crit}} < T$). For larger K one finds kinks in the curves when the ratio approaches $1/2$. A ratio of $1/2$ suggests for an assumed quadratic eigenvalue λ_1, that η_0^{opt} is close to the maximal learning rate $\eta_{\max}$, whereas $\eta_{\boldsymbol{w}}^{\text{opt}}$ is close to the minimum located at $\eta_{\max}/2$. The kinks therefore coincide with a change to $T_0^{\text{crit}} < T < T_{\boldsymbol{w}}^{\text{crit}}$ above a value of K dependent on ϱ [e.g., for $\varrho = 6$ the kink is at $K \approx 100$, which coincides with $T^{\text{crit}}(100, 6) \approx 1$ as can be seen in Fig. 14(b)]. For even larger ϱ this solution change is pushed out to larger values of K.

The ratio of the optimal convergence rates $\lambda_{\boldsymbol{w}}^{\text{opt}}$ and λ_0^{opt} shown in Fig. 15(e) reflects above observations. For small ϱ the minimum of λ_1 is optimal and the ratio is 1. Even for larger T values, where the root solutions apply for $\varrho = 0$, ratios very close to 1 are observed for small ϱ. For larger ϱ, however, the root solutions apply either for both learning rate parameterizations or at least for $\eta_\theta = \eta_{\boldsymbol{w}}$ and the widening gap between λ_1 for the two learning rate parameterizations leads to ratios above 1 increasing with ϱ. The benefit achievable is, however, limited eventually for large K when the optimal convergence of the $\eta_\theta \to \infty$ parameterization reverts back to the minimum of λ_1.

This behavior holds similarly for the ratio of the normalized separation of maximal and optimal learning rates $\Delta\eta_{\boldsymbol{w}}{}_{\max}^{\text{opt}}$ and $\Delta\eta_0{}_{\max}^{\text{opt}}$ [Fig. 15(f)]. The widening gap between λ_1 increases the ratio significantly above 1, once η_0^{opt} is given by the root solution. The nonmonotonic behavior for some of the lines in Fig. 15(f) can be explained by the change in the degree of skewness of λ_1 away from a parabolic form when the minimum solution applies for $\eta_{\boldsymbol{w}}^{\text{opt}}$.

3. Optimal dynamics in ϱ-T space

When considering the optimal dynamics as a function of ϱ and T, two natural scaling ansätze for the bias ϱ present themselves (see discussion in Sect. II), which become especially relevant in the limits $T \to \infty$ and $T \to 0$. The first ansatz ($\varrho = \hat{\varrho}\sqrt{1+T}$), here termed effective bias, fixes the mean hidden unit output independent of T, the other ansatz ($\varrho = \check{\varrho}\sqrt{T}$), here termed abscissa, keeps the distance of the decision hyperplane to the origin constant. For large $T \gg 1$, both ansätze become identical to leading orders. For small T, however, there are significant differences. In this section we have adopted $\hat{\varrho}$ as the preferred variable since it results in the more universal behavior for finite T, but we will discuss their differences in detail in Sect. V C.

In Fig. 16 the influence of different teacher length values T is studied, where the convergence behavior of the parameterization $\eta_\theta \to \infty$ [Figs. 16(a)–16(c)] is shown as a function of $\hat{\varrho}$ for $K = 10^2$ and a range of T values (including theoretical predictions from asymptotic analyses when useful). Fig. 16(a) shows that the optimal learning rate increases exponentionally in $\hat{\varrho}^2$. For small $\hat{\varrho}$, the prefactor of the exponentional increase approaches 1/2 for large T, whereas it approaches 1 for small T, in agreement with the prediction from the $K \to \infty$ and $T \to 0$ analyses [included in Fig. 16(a)]. For larger $\hat{\varrho}$, however, one finds a prominent change in the slope of the $\eta_{\boldsymbol{w}}^{\text{opt}}$ curves, where the position of the transition and its significance is dependent on T. For very small but finite T this transition is beyond the range of the graph and the change in the slope becomes less significant. The limiting behavior is in agreement with the $T \to 0$ analysis [included in Fig. 16(a)]. For finite T, $\eta_{\boldsymbol{w}}^{\text{opt}}$ still increases exponentionally in $\hat{\varrho}^2$ after the transition, but the constant prefactor in the exponent is altered and decreases for large T. The limiting behavior is in agreement with the findings of the $T \to \infty$ analysis for finite K in Appendix B 5, which predicts a finite limit of $\eta_{\boldsymbol{w}}^{\text{opt}}$ for large $\hat{\varrho}$ also shown in Fig. 16(a).

The dependence of the optimal convergence eigenvalue $\lambda_{\boldsymbol{w}}^{\text{opt}}$ shown in Fig. 16(b) is similarly intriguing. One finds that the convergence rate increases initially with $\hat{\varrho}$ up to max-

imum, whose position shifts to larger $\hat{\varrho}$ values for decreasing T and becomes flatter for increasing T. Beyond the maximum, $\lambda_{\boldsymbol{w}}^{\text{opt}}$ decreases exponentionally in $\hat{\varrho}^2$, with the prefactor in the exponentional increasing with T, but saturating at $1/2$ as predicted from the $T \to \infty$ analysis. The small T expansion predicts the steep initial increase in $\lambda_{\boldsymbol{w}}^{\text{opt}}$ correctly, as the order of the optimal convergence rate for non-zero $\hat{\varrho}$ is not $O(T^2/K)$ as for zero $\hat{\varrho}$ but $O(T/K)$. The expansion is a good approximation for small finite T and small $\hat{\varrho}$ but breaks down for larger $\hat{\varrho}$, where the optimal convergence rate $\lambda_{\boldsymbol{w}}^{\text{opt}}$ reaches a almost T independent maximum of $O(1/K)$ and can also not account for the eventual exponentional decrease of $\lambda_{\boldsymbol{w}}^{\text{opt}}$ with $\hat{\varrho}$ beyond the maximum. This failure is caused by the implicit assumption $\hat{\varrho}^2 \ll -\log T$ in the $T \to 0$ limit which shifts the maximum in $\lambda_{\boldsymbol{w}}^{\text{opt}}$ to $\hat{\varrho} = \infty$. For larger network sizes K not shown here, one finds that the position of the maximum shifts to larger $\hat{\varrho}$ and becomes flatter. This effect leads to the shift of the maximum to $\hat{\varrho} = \infty$ in the $K \to \infty$ expansion.

The behavior of the normalized separation $\Delta\eta_{\boldsymbol{w}}{}_{\max}^{\text{opt}}$ in Fig. 16(c) reflects the kind of solution present. For small $T < T_{\boldsymbol{w}}^{\text{crit}}$, the minimum of λ_1 is optimal and $\Delta\eta_{\boldsymbol{w}}{}_{\max}^{\text{opt}}$ increases monotonically towards 1, i.e., λ_1 becomes parabolic for large $\hat{\varrho}$. For $T = 1$, we find the same behavior for small $\hat{\varrho}$, but find a prominent kink at $\hat{\varrho} \approx 4.25$ [i.e., $\varrho \approx 6$, see Fig. 14(b)], which coincides with $T^{\text{crit}} = 1$. For $\hat{\varrho} > 4.25$, $T^{\text{crit}} < 1$ and $\Delta\eta_{\boldsymbol{w}}{}_{\max}^{\text{opt}}$ falls to a constant below 1. For larger T, the behavior is similar but smoother in comparison to $T = 1$, reflecting the fact that although the optimal solution is always given by the root, its distance to the minimum changes with $\hat{\varrho}$ as $T_{\boldsymbol{w}}^{\text{crit}}$ rises and falls.

The results for the parameterization $\eta_\theta = \eta_{\boldsymbol{w}}$ are quite similar to $\eta_\theta \to \infty$ and to enhance the differences we show the ratios of the relevant quantaties in Figs. 16(d)–16(f). For the optimal learning rate η_0^{opt}, we also find the change in the exponentional behavior. For large enough $T > T_0^{\text{crit}}$, the ratio of the $\eta_{\boldsymbol{w}}^{\text{opt}}/\eta_0^{\text{opt}}$ falls below 1 [see Fig. 16(d)] and approaches a constant limit for large $\hat{\varrho}$. For medium T (e.g., $T = 1$), the difference is most pronounced, reflecting the many changes in the type of solutions due to the variability of $T_{\boldsymbol{w}}^{\text{crit}}$ and T_0^{crit}. For small $\hat{\varrho}$, the minimum solution of λ_1 is optimal for both learning rate parameterizations.

In the range of $0.40 \lessapprox \hat{\varrho} \lessapprox 4.25$ (i.e., $0.55 \lessapprox \varrho \lessapprox 6$), $T_0^{\rm crit} < T < T_{\boldsymbol{w}}^{\rm crit}$ and the ratio drops significantly [23] towards 1/2 until also $T_{\boldsymbol{w}}^{\rm crit} < T$ and the ratio rises again towards the asymptotic behavior.

The improvement by using a large bias learning rate is reflected in the ratio $\lambda_{\boldsymbol{w}}^{\rm opt}/\lambda_0^{\rm opt}$ [Fig. 16(e)], which increases monotonically with $\hat{\varrho}$, for T or $\hat{\varrho}$ large enough so that $T > T_0^{\rm crit}$. In the $T > T_{\boldsymbol{w}}^{\rm crit}$ region, the ratio $\lambda_{\boldsymbol{w}}^{\rm opt}/\lambda_0^{\rm opt}$ increases with $a_0 + a_2\hat{\varrho}^2$, where a_0 and a_2 are T dependent constants which approach $a_0 = 1$ and $a_2 = 1$ for large T as predicted by the $T \to \infty$ analysis. Using large η_θ is similarly benefical in the same region of T and $\hat{\varrho}$ with respect to the separation of maximal and optimal learning rates as depicted in Fig. 16(f). For larger T, we find the same regression behavior of the ratio $\Delta\eta_{\boldsymbol{w}}{}^{\rm opt}_{\rm max}/\Delta\eta_0{}^{\rm opt}_{\rm max}$ with $b_0 + b_2\hat{\varrho}^2$, where b_0 and b_2 are again T dependent constants with the asymptotic limit $1+\hat{\varrho}^2$ for $T \to \infty$. In the curve for $T = 1$, one observes several swerves and a kink due to $T_0^{\rm crit}$ or $T_{\boldsymbol{w}}^{\rm crit}$ crossing $T = 1$.

C. The impact of adaptive biases

In comparison to the analysis of the convergence phase for zero-fixed biases [14], the extension to variable non-zero biases, has revealed several insights. For small T, where the training for the zero-bias case is slowed down by a factor $1/T^2$, arguably due to the nearly linear network output making the distinction between different units difficult, one finds that the scaling assumption for the bias has a dramatic impact. This can be understood qualitatively by considering the network output distribution which can be calculated in closed form in the $T \to 0$ limit.

For finite abscissa (using the scaling $\varrho = \check{\varrho}\sqrt{T}$), the hidden unit output distribution is Gaussian with mean $\mu = -\sqrt{2K/\pi}\check{\varrho}\sqrt{T}$ and standard deviation $\sigma = \sqrt{2/\pi}\sqrt{T}$. The probability of a positive (and hence negative) output remains constant for $T \to 0$ and is equal to $H(\check{\varrho}\sqrt{K})$, where $H(x) = \int_x^\infty dx/\sqrt{2\pi}\exp(-x^2/2)$, i.e., even for small T the output of the hidden unit will have some probability of being both negative and positive, but the

mean goes to zero. For this scaling, one finds a slight improvement in the convergence rate for non-zero bias by a factor $1 + 2\breve{\varrho}^2$, suggesting that breaking the symmetry of the network output distribution around zero is beneficial, but a more significant improvement is not possible since the hidden unit outputs are mainly in the linear regime where the student cannot discriminate efficiently between the teacher hidden units and the convergence rate still decays with T^2.

For finite effective bias (using the scaling $\varrho = \hat{\varrho}\sqrt{1+T}$), the network output distribution is also Gaussian for small T, but with mean $\mu = -\sqrt{K}g(\hat{\varrho})$ and standard deviation $\sigma = \sqrt{2/\pi}\exp(-\hat{\varrho}^2/2)\sqrt{T}$. The probability of an output of opposite sign to the mean output vanishes for $T \to 0$. The single hidden unit output is concentrated in the nonlinear region of the sigmoidal activation function and one could argue that most information about a teacher parameters can be extracted by the student in this region as long as the hidden units are not too saturated, leading to the improvement in the convergence rate by $O(\hat{\varrho}^2/T)$.

One could further speculate, that the increase of the optimal learning rate matching the suppression of the gradient is facilitated by the exponentional decrease of the network output variance with $\hat{\varrho}$. For finite T and larger $\hat{\varrho}$, the results for $T \to 0$ expansion become inaccurate for $\hat{\varrho}^2 \not\ll -\log T$ and one finds that the optimal learning rate growth cannot be sustained, leading to the eventual exponentional decay of the convergence eigenvalue with $\hat{\varrho}^2$ as observed for finite K. Due to the T dependence of this breakdown, one even finds the anomaly that training can be momentarily improved when decreasing T slightly [see Fig. 16(b)].

The unsustainability of the optimal learning rate growth is epitomized in the $T \to \infty$ limit, where the optimal learning rate stays constant for all $\hat{\varrho}$. However, if the $K \to \infty$ limit is taken simultaneously with $T \to \infty$, the convergence rate either remains constant for $\eta_\theta \to \infty$ or decays algebraically with $(1+\hat{\varrho})^2$ for $\eta_\theta = \eta_{\boldsymbol{w}}$. Similar behavior is also found for finite T and large K for small enough $\hat{\varrho}$.

The underlying reasons of this difference can be explained most easily for the infinite T case, where the hidden unit output becomes binary and the subsequent network output

probability distribution is binomial, as teacher hidden units are uncorrelated. The probability of a single hidden output to be $+1$ parameterizes the binomial distribution and is $1/2[1-g(\hat{\varrho})]$, i.e., $1/2$ for $\hat{\varrho}=0$ and decays exponentionally fast for large $\hat{\varrho}$ ($\propto \mathrm{e}^{-\hat{\varrho}^2}$). The corresponding mean and standard deviation are $\mu=-\sqrt{K}g(\hat{\varrho})$ and $\sigma=\sqrt{1-g^2(\hat{\varrho})}$, respectively. Since both student and teacher network are highly correlated, the error signal should be at most $O(1/K)$, i.e., at most two hidden units disagree, leading to a possible increase of the learning rate with K. For large effective bias $\hat{\varrho}$, this event becomes exponentionally unlikely and the error signal is identically zero most of the time. The learning rate, however, cannot be increased accordingly since this would lead to an exponentionally large update step size in an error event. The convergence rate has therefore to decay exponentionally. For $K\to\infty$, the binomial output distribution becomes Gaussian with the above mean and variance, leading to smooth network outputs and error signals. Here, the learning rate can be increased exponentionally, which may be linked to the exponentional decrease of the output variance for large $\hat{\varrho}$ combined with the implicit assumption that $\hat{\varrho}^2 \ll \log K$. This behavior carries over qualitatively to finite T and K for $\hat{\varrho}^2$ small enough, and can explain the initial matching increase of the optimal learning rate and the extension of the region of almost constant convergence rate for larger K.

VI. TOWARDS MORE REALISTIC SCENARIOS

The scope of this work has so far been restricted in several ways. One obvious restriction has been the fixed hidden-output weights. Although soft-committee machine with biases are universal approximators [8], in practice it is advantageous to use adjustable hidden-output weights. This extension is straightforward in terms of feasibility, but adds a further dimension to the space of parameters to be investigated. We expect our results to be at least qualitatively correct, but we cannot rule out that the dynamics become even richer with more suboptimal fixed points. Unfortunately, the works to date which have allowed for adjustable hidden-output units [6,7] have not discussed the issue of hidden unit symmetry

breaking.

We have furthermore restricted ourselves to realizable scenarios, where the student network can learn to imitate the teacher network perfectly. In real learning scenarios, one expects both structural unrealizability, due to a mismatch between the function space of the student and the task, as well as unrealizability due to corrupted training data. Both types of unrealizability can be incorporated in this framework, by studying $K \neq M$ and by allowing for noise on the teacher weights and/or outputs, respectively. Both have been addressed already for the soft-committee machine without biases [2,24,25].

Here we will briefly assess the effects arising due to the introduction of adjustable biases in the case of structural unrealizability. In Fig. 17 the evolution of the training is shown for $K = 3$ and $M = 4$, i.e., when the target function is more complicated than the mapping the student can achieve. The teacher overlaps are $T_{nm} = \delta_{nm}(n+1)/2$ for graded and $T_{nm} = \delta_{nm}$ for isotropic teachers. The teacher biases are $\varrho_n = (2n-5)/5\sqrt{1+T_{nn}}$ for non-degenerate and $\varrho_n = 0$ for degenerate teachers. The common learning rate is always $\eta_0 = 2$ and the weight initialization is $Q_{ii} = (18+n)/100$, $\theta_i = (n-2)/100$, and random overlaps as outlined in Sect. III. The initialization was chosen quite symmetrically to make differences between the tasks more pronounced and to ensure convergence to a fixed point with the lowest generalization error for the most symmetric task $\mathcal{T}_d^i$.

The main focus will be on the $\mathcal{T}_n^i$ since for this task the effect of non-degenerate teacher biases can be separated from the effect of graded teacher norms. In Figs. 17(a)–17(c) the evolution of the overlaps Q_{ij}, R_{in} and the biases θ_i is shown. The student is initially drawn into a symmetric phase with similar values for student lengths Q_{ii} and correlations Q_{ij} [Fig. 17(a)]. This is mirrored by similar student-teacher overlaps R_{in} shown in Fig. 17(b), signalling the lack of significant specialization with a specific teacher node. The specialization is driven by the student biases depicted in Fig. 17(c), whose symmetry is broken first and whose trajectories do not cross, although they were initialized quite symmetrically. Since the student network does not have enough resources to model the teacher task adequately, it chooses to dedicate two units (1 and 3) to specialise primarily on the teacher hidden units

(1 and 4) with the largest absolute bias value; which is reflected by large R_{11} and R_{34} values and the proximity of the student biases θ_1 and θ_3 to the corresponding teacher biases ϱ_1 and ϱ_4. This seems sensible since these two units have on average the largest (absolute) output. The last student unit 2 specializes almost equally on the two remaining teacher units 2 and 3 (large R_{22}, R_{23} and θ_2 lies between ϱ_2 ϱ_3). The remaining student-teacher overlaps fall roughly into two groups: The student units (1,4) which are highly specialized on one unit acquire a relatively large overlap with the remaining teacher units (2,3) for which no dedicated student unit exists, whereas they retain only small correlations of either positive or negative sign with those teacher units, which are already modelled almost entirely by another student unit. The size of the individual student-teacher overlaps is also highly correlated with the proximity of the associated student and teacher biases (e.g., $R_{23} > R_{13} > R_{33}$ for fixed teacher unit or $R_{34} > R_{33} > R_{32} > R_{31}$ for fixed student unit). One further notices that the student biases are positioned to ensure that the means of the student and teacher network output distributions (which is just the sum of the means of the individual hidden unit output distributions in a network) are very similar. Matching the mean of the teacher output distribution is obviously a necessary but not sufficient condition for achieving a small generalization error.

Obviously, the specialization process described above is dependent on the teacher task presented. For graded teacher tasks, the larger teacher hidden unit weight vectors lead to a larger variance of their output distributions (and ultimately the output distribution of the whole network). The student hidden unit have therefore to compromise between primarily modelling large variance by specialising on teacher units with large weight norms and large mean by specialising on teacher units with large (effective) biases. We still find that the student biases are positioned to ensure that the mean output is approximately identical, but the student also accounts for larger variances. For degenerate biases, one finds that the dynamics and the optimal attractive fixed point are very similar to the fixed bias case for both graded and isotropic teacher, with the student biases taking values close to the degenerate (effective) teacher bias position [26].

In Fig. 17(d) the dynamics of the four different generic tasks are compared by following the evolution of the generalization error. As for realizable learning scenarios, one finds that the specialization process for the task $\mathcal{T}_d^i$ is by far the slowest due to the slow breaking of the symmetries. For the task $\mathcal{T}_d^g$ one finds more than one plateau in the generalization error [see inset of Fig. 17(d)] characteristic of the sequential symmetry breaking for graded teacher lengths. The fastest training is exhibited by the tasks with non-degenerate biases $\mathcal{T}^n$, with a slight speed-up for graded teacher lengths $\mathcal{T}_g^n$. Unlike in realizable scenarios, the dynamics approach a non-zero asymptotic generalization error, which is smallest for the task $\mathcal{T}_d^i$ with most symmetries. For the tasks presented here, the breaking of the bias degeneracy results in a smaller increase of the generalization error than the breaking of length isotropy. This feature, however, depends on the particular choice of teacher norms and biases.

Similar to the realizable case, we also find that the dynamics are sensitive to the initial conditions, especially for tasks with many symmetries such as $\mathcal{T}_d^i$, and the asymptotic network configuration can vary significantly in their generalization error. For the $\mathcal{T}_d^i$, the basin of attraction to the optimal solution described above is quite small and requires highly symmetric initial bias values. Otherwise the bias dynamics show the grouping around the true teacher bias value similar to the realizable case with the notable difference, that the bias values seem to diverge instead of converging to (suboptimal) fixed values.

For non-degenerate biases, one also finds a multitude of stable network configuration depending on the initial conditions, which all feature quite similar generalization error. For the task $\mathcal{T}_n^i$ for example, a different set of initial conditions (changing only the norms $Q_{ii} = (1+n)/10$) leads to student unit 2 specializing primarily on teacher unit 3 instead of specializing almost equally on teacher units 2 and 3 and results in a slightly smaller generalization error. We find that the evolution of the dynamics to solutions with similar asymptotic generalization error are qualitatively similar, but one does not find a dominant basin of attraction to a particular solution as in the case of fixed biases. A more detailed investigation is therefore beyond the scope of this paper and will be reported elsewhere.

Finally we would like to point out, that in the case of student-teacher mismatch $K \neq M$,

the difference between the normalized and unnormalized committee machine are substantial and the results are therefore quite different. For $K > M$, the unnormalized soft-committee machine is overrealizable and the excess nodes can be pruned away to achieve perfect generalization. This is obviously not possible for the normalized soft-committee machine due to the different normalization factor, and the task becomes unrealizable with a finite asymptotic generalization error. For $K < M$, the normalisation of the committee machine leads to generally lower asymptotic values of the order parameters with a resulting generalization error which is always lower than for the unnormalized case. This seems due to the normalization keeping the variance of the network output distribution of constant order (for uncorrelated teacher weight vectors) irrespective of the number of hidden units, whereas the order of the output variance is mismatched ($\sqrt{K}$ and $\sqrt{M}$) in the unnormalized model.

VII. SUMMARY AND DISCUSSION

This research has been motivated by recent progress in the theoretical study of on-line learning in realistic two-layer neural network models — the soft-committee machine, trained with back-propagation [2]. The studies so far have excluded biases to the hidden layers, a constraint which has been removed in this paper. Such a network is in principle a universal approximator [8], although within the framework at issue the model can only be studied in a limit where the approximation proof does not necessarily hold as it may require the number of hidden units to scale with N. Nevertheless, the dynamics of the extended model turn out to be very rich and more complex than the original model, although we had to restrict ourselves for computational reasons to small networks.

For non-degenerate teacher biases, one finds that the symmetry in the student hidden unit space can be broken almost immediately by the biases, provided the student biases were initialized asymmetrically, speeding up the learning process considerably in comparison to the fixed bias model where the training process can easily be dominated by the symmetric phase characterized by a lack of hidden unit specialization. These results suggest that

student biases should in practice be initially spread evenly across the input domain if there is no *a priori* knowledge of the target function. For degenerate teacher biases, however, especially in combination with similar teacher lengths, such a scheme can be extremely counterproductive as asymmetric initial student biases severely prolong the training and can in many cases even trap the learning process permanently in attractive fixed points. Although attractive suboptimal fixed points were also found in the original soft-committee machine model [16], these seem to have been restricted to over-realizable cases and the associated basins of attraction have been very small.

Unlike in the fixed bias case, the initial conditions, Q_{ij} and θ_i, which can manipulated in real scenarios, influence the training time considerably and can even cause complete training failure. To gain a qualitative understanding of the influence of the initial conditions, the basins of attraction to the optimal solution were therefore studied exhaustively for $K = M = 2$. One finds that attractive suboptimal fixed points exist for many training scenarios, including graded teachers and even non-degenerate teacher biases. The range of initial conditions attracted to these suboptimal network configurations diminishes with increasing asymmetry of the task, especially for non-degenerate teacher biases, where the attractive fixed point vanishes eventually. In the task with the smallest basin of attraction, isotropic teacher weight vectors and degenerate teacher biases, which was studied in great detail, one finds several unexpected results. First, the basin of attraction is mainly dependent on the difference in the initial student biases, rather than their individual abscissas or the resulting mean. Second, the basin of attraction, with respect to the student biases θ_i, grows with increasing student norms, but the corresponding abscissa ($\hat{\theta} = \theta/\sqrt{Q}$) decreases. Third, the basin of attraction is enlarged by larger initial student-teacher overlaps and training should therefore be less prone to failure for smaller input dimension.

Additionally, the influence of the learning rates on the basin of attraction was studied for the same isotropic and degenerate task. For a common learning rate for biases and weights, the basin of attraction shrinks to a minimum in the region of fastest convergence, i.e., for the overall optimal learning rate. The basin of attraction increases especially for small learning

rate but always remains finite. More effective in increasing the basin of attraction seems, however, the separation of the bias and weight learning rate. Whereas one must necessarily pay dearly for stability with a decrease in convergence speed when employing a small bias or weight learning rate, a large bias learning rate does not compromise training efficiency.

Although most of the results found for $K = 2$ also carry over qualitatively to larger networks, the size of the basin of attraction shrinks considerably with network size, which may partly be contributed to the substantial increase of the number of attractive suboptimal fixed points with different internal symmetries. In particular, we have found that the use of a large bias learning rate or the adiabatic elimination of the biases can actually decrease the basin of attraction for larger networks and degenerate biases.

Unlike preliminary results [8] which seemed to support the heuristic suggestion in an earlier work [17] to spread the abscissas across the input domain in order to speed up training; our more extensive work, clearly suggest that such an initialization scheme may in general not be advisable. Our results show that in terms of the initialization, the difference in the threshold values and not the individual abscissas are the more relevant variables. Furthermore, such a scheme will most likely fail to convergence to the optimal solution when some of the biases are degenerate, although one can only speculate how common these tasks are encountered in practice.

Other previous work [18], which relates the basin of attraction of the weight initialization with the learning rate, seems also to be partially contradicted by our findings. Although the basin of attraction does grow with decreasing learning rate, as found in [18], the functional relationship given for convergence in this work ($\eta_0 < Q_{ii} + \theta_i^2$) fails to predict a finite boundary for an infinitesimal learning rate. Furthermore, the treatment of the biases as just another weight parameter suggests a growing basin of attraction with both increasing weights and biases, whereas we find that biases actually have the reverse effect. The work also neglects the strong interaction between the hidden units, e.g., the importance of the difference in initial thresholds or the shrinking of the basin of attraction for larger networks.

An initialization procedure which provides both stability and fast convergence speed for

all tasks, seems therefore difficult to realize due to the inherently different requirements for tasks with degenerate and non-degenerate biases. The probably most successful approach is to opt for a combined approach of medium spread of the biases, large initial weights, a reasonable separation of weight and bias learning rate. This must be combined with a criteria which restarts network biases for hidden units trapped in an attractive suboptimal fixed point. Since for most attractive fixed points found, the student hidden units are not highly saturated, i.e., the absolute values of their mean output is reasonably less than 1, it is not sufficient to just select saturated units with large effective bias. This criteria must therefore account for the actual bias values in combination with correlations between the student hidden unit weight vectors. For persistently large correlation between a pair of weight vectors and very similar lengths, the biases could for example be reset to their mean value. If such a strategy works in all situations remains to be shown, which goes beyond the scope of this paper. Possible difficulties are likely to be unrealizable scenarios, where persistent correlation are caused by a lack of student resources and a successful algorithm would have to be able to distinguish between the two. Its usefulness would then have to be further tested in finite size systems and real world problems. However, as already mentioned, in cases where the training set is known in advance, many algorithms are available that aim to infer good initial conditions from the training data (see e.g., [20] and references therein).

Unlike for the entire training process and general learning scenarios, where we had to restrict ourselves to small networks, the dynamics can be studied and optimized for all network sizes for the isotropic degenerate teacher task in the convergence phase, where hidden unit symmetry is already broken successfully and the student approaches the optimal solution. Since this type of task is not only the slowest in terms of overall training time, but also in the convergence phase itself, the results should give us a bound on the performance of other tasks.

One finds that optimal convergence is achieved for an infinite bias learning rate, suggesting that an $O(1)$ rather than an $O(1/N)$ bias learning rate is appropriate for finite systems once hidden unit symmetry is broken and that the input-hidden weights dominate

the learning behavior in this phase. The dependence of the optimal (weight) learning rate has been studied as a function of the number of hidden units K and the teacher length T with special emphasis on the influence of non-zero (effective) bias $\hat{\varrho}$, which provides the most useful scaling of the bias in the convergence phase. We have restricted ourselves also to two learning rate parameterizations for the biases: $\eta_\theta = \eta_{\boldsymbol{w}}$ and $\eta_\theta \to \infty$. One finds that both for small T or small $\hat{\varrho}$, there is either no or little difference between the two parameterizations. The advantage of an increased bias learning rate grows, however, for large enough T approximately proportionally to $\hat{\varrho}^2$.

The influence of the value of the effective teacher biases $\hat{\varrho}$ manifests itself for both parameterizations in the initially surprising effect that for most T values the learning performance actually improves for small non-zero bias. This can be explained by postulating that most information on the parameters of an individual hidden unit can be obtained in the region where the sigmoid is already nonlinear but not quite saturated. In this region one finds an exponentional increase in the optimal learning rate matching the suppression of the gradient. This increase, however, cannot be sustained for larger $\hat{\varrho}$ and leads to an eventual exponentional decay of the convergence speed in $\hat{\varrho}^2$ for any finite K. This exponentional decay is delayed to larger $\hat{\varrho}$ values for small teacher length T and large network size K, which may be attributed to the increasing smoothness of the error signals allowing for a larger learning rate. This fact is epitomized in the $T \to 0$ and $K \to \infty$ limits, where the convergence rate does increase unabatedly or decreases at most algebraically in $\hat{\varrho}$, respectively.

The choice of the learning rate is therefore important in both the symmetric phase, where it can help to avoid attractive fixed point as well as in the convergence phase, where the optimal value varies significantly in the relevant region of input space, making it difficult to choose good learning rates in practice. The problem of training is also exacerbated by the difficulty of student parameter initialization without *a priori* knowledge about the learning task present, which can change the basin of attraction to the optimal solution considerably.

Future research effort should therefore be aimed at devising more sophisticated on-line learning algorithms, which are able to infer information about the teacher task and the

progress made in training by monitoring the student parameters and subsequently adjust the learning rates accordingly or restart hidden units trapped in suboptimal fixed points. The introduction of individual learning rates for each hidden unit, already shown to be beneficial for the fixed biased model [15], seems a further direction worthwhile to pursue. Since the learning dynamics have shown to change significantly with the introduction of adjustable biases for realizable scenarios, it appears to be of obvious interest to investigate the influence of unrealizability more systematic than could be achieved within the scope of this paper.

ACKNOWLEDGMENTS

A.H.L.W. would like to acknowledge gratefully financial support by the EPSRC, a scholarship of 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 (3) 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

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

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 the sigmoidal transfer function g is here taken to be the error function $g_\nu(u) = \mathrm{erf}(\nu u/\sqrt{2})$. We denote with I_d, J_d averages over d variables with one and two g terms, respectively. Unlike in the case of fixed zero-biases, only integrals involving a single g terms can be calculated analytically, whereas general Gaussian integrals involving g^2 terms of shifted arguments have no known analytical solution. However, these integrals can be simplified considerably to make a numerical integration feasible. There are several possible representations, e.g., the Kendall series expansion, but we have chosen one which consists of a single Gaussian integral of two error functions. We have found that this form has the advantage that the summation over units and the integration can be interchanged, greatly improving numerical accuracy for fixed computational cost.

In this form the integral $J_2()$ is given by

$$J_2(1,2) = \int Dt\; g_\nu\Big(\sqrt{C_{11}}\,t - \vartheta_1\Big) \times g_\nu\left(\frac{C_{12}t - \vartheta_2}{\sqrt{C_{11}\psi_2 - \nu^2 C_{12}^2}}\right), \tag{A3}$$

where

$$\psi_i = 1 + \nu^2 C_{ii}, \quad \text{and} \quad Dt = \frac{dt}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)$$

is the Gaussian measure, with any integral without explicit limits is from $-\infty$ to $+\infty$. The dependence of the integral on the sigmoidal gain ν can be absorbed by redefining

$$\tilde{\vartheta}_i = \nu\vartheta_i, \quad \text{and} \quad \tilde{C}_{ij} = \nu^2 C_{ij},$$

a rescaling which 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 $\mathbf{Q}$, $\mathbf{R}$, and $\boldsymbol{\theta}$ are calculated similarly and take the form

$$\frac{dQ_{ij}}{d\alpha} = \frac{\eta_{\boldsymbol{w}}\gamma^2}{K}\left\{\sqrt{\frac{K}{M}}\sum_{m=1}^{M} I_3(i,j,m) + I_3(j,i,m) - \sum_{k=1}^{K} I_3(i,j,k) + I_3(j,i,k)\right\}$$
$$+ \left(\frac{\eta\gamma^2}{K}\right)^2\left\{\frac{K}{M}\sum_{n,m=1}^{M} J_4(i,j,n,m) - 2\sqrt{\frac{K}{M}}\sum_{k,n=1}^{K,M} J_4(i,j,k,n) + \sum_{k,l=1}^{K} J_4(i,j,k,l)\right\}, \quad \text{(A4a)}$$

$$\frac{dR_{in}}{d\alpha} = \frac{\eta_{\boldsymbol{w}}\gamma^2}{K}\left\{\sqrt{\frac{K}{M}}\sum_{m=1}^{M} I_3(i,n,m) - \sum_{k=1}^{K} I_3(i,n,k)\right\}, \quad \text{(A4b)}$$

$$\frac{d\theta_i}{d\alpha} = -\frac{\eta_\theta\gamma^2}{K}\left\{\sqrt{\frac{K}{M}}\sum_{m=1}^{M} I_2(i,n) - \sum_{k=1}^{K} I_2(i,k)\right\}, \quad \text{(A4c)}$$

where the two integrals $I_2(1,2) = \langle g'(u_1)g(u_2)\rangle$ and $I_3(1,2,3) = \langle g'(u_1)u_2\, g(u_3)\rangle$ can be evaluated analytically, whereas $J_4(1,2,3,4) = \langle g'(u_1)g'(u_2)g(u_3)g(u_4)\rangle$ can be simplified to a form similar to $J_2()$ and one finds

$$I_2(1,2) = \nu\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\psi_1}}\exp\left(-\frac{1}{2}\frac{\tilde{\vartheta}_1^2}{\psi_1}\right) g_1(\Theta_{12}), \quad \text{(A5a)}$$

$$I_3(1,2,3) = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\psi_1}}\exp\left(-\frac{1}{2}\frac{\tilde{\vartheta}_1^2}{\psi_1}\right)\left[\frac{\tilde{C}_{13}\tilde{\vartheta}_1}{\psi_1} g_1(\Theta_{12})\right.$$
$$\left. + \sqrt{\frac{2}{\pi}}\frac{\Psi_{12}\Gamma_{13}}{\sqrt{\Psi_{13}}\sqrt{\psi_1}}\exp\left(-\frac{1}{2}\Theta_{12}^2\right)\right], \quad \text{(A5b)}$$

$$J_4(1,2,3,4) = \nu^2\exp\left(-\frac{1}{2}\frac{\psi_2\tilde{\vartheta}_1^2 - 2\tilde{C}_{12}\tilde{\vartheta}_1\tilde{\vartheta}_2 + \psi_1\tilde{\vartheta}_2^2}{\Psi_{12}}\right)$$
$$\times\frac{2}{\pi}\frac{1}{\sqrt{\Psi_{12}}}\int Dt\; g_1\left(\sqrt{\tilde{C}'_{33}}\, t - \tilde{\vartheta}'_3\right) \quad \text{(A5c)}$$
$$\times g_1\left(\frac{\tilde{C}'_{34}t - \tilde{\vartheta}'_4}{\sqrt{\tilde{C}'_{33}\psi'_4 - \tilde{C}'^2_{34}}}\right),$$

where we conveniently define

$$\Psi_{ij} = \psi_i\psi_j - \tilde{C}_{ij}^2,$$

$$\Theta_{ij} = \frac{\tilde{C}_{ij}\tilde{\vartheta}_i - \psi_i\tilde{\vartheta}_j}{\sqrt{\psi_i\Psi_{ij}}},$$
$$\Gamma_{1i} = \frac{\psi_1\tilde{C}_{2i} - \tilde{C}_{12}\tilde{C}_{1i}}{\Psi_{12}},$$
$$\Gamma_{2i} = \frac{\psi_2\tilde{C}_{1i} - \tilde{C}_{12}\tilde{C}_{2i}}{\Psi_{12}},$$

and the primed variables

$$\tilde{C}'_{ij} = \tilde{C}_{ij} - (\tilde{C}_{1i}\Gamma_{2j} + \tilde{C}_{2i}\Gamma_{1j}),$$
$$\tilde{\vartheta}'_i = \tilde{\vartheta}_i - (\tilde{\vartheta}_1\Gamma_{2i} + \tilde{\vartheta}_2\Gamma_{1i}),$$

with the obvious extensions, e.g., $\psi'_i = 1+\tilde{C}'_{ii}$. Again, one infers the elements of the reduced covariance matrix using the unit labeling convention and the appropriate dimensionality reduction.

As mentioned above the gain ν rescales all order parameters and the biases explicitly and furthermore leads to an implicit rescaling of both learning rates by ν^2 in the differential equations (A4). The learning rates are further rescaled by the linear output gain by γ^2. The total rescaling of any bias and the bias learning rate η_0 therefore is

$$\tilde{\vartheta} = \nu\vartheta, \quad \text{and} \quad \tilde{\eta}_\theta = \frac{\nu^2\gamma^2}{(K)}\eta_\theta. \tag{A6a}$$

For the weight order parameters and their learning rate $\eta_{\boldsymbol{w}}$, the input variance σ^2 can also be absorbed to give

$$\tilde{C} = \nu^2\sigma^2 C \quad \text{and} \quad \tilde{\eta}_{\boldsymbol{w}} = \frac{\nu^2\gamma^2\sigma^2}{(K)}\eta_{\boldsymbol{w}}. \tag{A6b}$$

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

APPENDIX B: THE REDUCED EQUATIONS CONVERGENCE DYNAMICS

For a realizable isotropic teacher scenario characterized by $K = M$, $T_{nm} = T\delta_{nm}$, and degenerate biases $\varrho_n = \varrho$, the number of free parameters can be reduced with the ansatz

(13), to just five variables R, S, Q, C, and θ, which gives an accurate description for the dynamics when the student biases were not initialized too unsymmetrically.

In the convergence phase one can expand the differential equations (A4) in a Taylor series to first order around the zero generalization error fixed point, $Q^* = R^* = T$, $C^* = S^* = 0$, and $\theta^* = \varrho$,

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

where $p_i = P_i - P_i^*$ and P_i are generic order parameters (we use the ordering $P_1 = R$, $P_2 = Q$, $P_3 = S$, $P_4 = C$, and $P_5 = \theta$ following the convention of earlier work [2]), and the eigenvalues and eigenvectors of the Jacobian matrix $\mathbf{M}$ of first derivatives determine the solution of the linearized differential equation.

The elements of the Jacobian matrix are explicitly given by

$$m_{11} = -\frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\frac{\exp\left(-\frac{\varrho^2}{1+2T}\right)}{(1+2T)^{3/2}}\left[(1+3T) - \frac{2T\varrho^2}{1+2T}\right], \tag{B1a}$$

$$m_{12} = \frac{1}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\left\{3\frac{[(1+2T)-2\varrho^2]T}{(1+2T)^{5/2}}\exp\left(-\frac{\varrho^2}{1+2T}\right) - (K-1)\frac{T\varrho^2}{(1+T)^3}\exp\left(-\frac{\varrho^2}{1+T}\right)\right\}, \tag{B1b}$$

$$m_{13} = \frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\frac{K-1}{(1+T)^2}\mathrm{e}^{-\frac{\varrho^2}{1+T}}\left[(1+2T) - \frac{T\varrho^2}{1+T}\right], \tag{B1c}$$

$$m_{14} = -\frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}(K-1)\mathrm{e}^{-\frac{\varrho^2}{1+T}}\frac{[(1+T)-\varrho^2]T}{(1+T)^3}, \tag{B1d}$$

$$m_{15} = \frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\varrho T\left[2\frac{\mathrm{e}^{-\frac{\varrho^2}{1+2T}}}{(1+2T)^{3/2}} + \frac{K-1}{(1+T)^2}\mathrm{e}^{-\frac{\varrho^2}{1+T}}\right], \tag{B1e}$$

$$m_{21} = \frac{4}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\left\{\mathrm{e}^{-\frac{\varrho^2}{1+2T}}\frac{(1+T)(1+2T)+2T\varrho^2}{(1+2T)^{5/2}}\right. \tag{B1f}$$

$$\left. - \frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\left[\frac{1}{\sqrt{1+4T}}\mathrm{e}^{-\frac{2\varrho^2}{1+4T}} + \frac{K-1}{1+2T}\mathrm{e}^{-\frac{2\varrho^2}{1+2T}}\right]\right\},$$

$$m_{23} = -\frac{4}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}(K-1)\exp\left(-\frac{\varrho^2}{1+T}\right) \times \left[\frac{(1+T)+T\varrho^2}{(1+T)^3} - m'_{23}\right], \tag{B1g}$$

$$m'_{23} = \frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\left[\frac{2}{\sqrt{(1+T)(1+3T)}}\mathrm{e}^{-\frac{\varrho^2}{(1+T)(1+3T)}} + \frac{(K-2)}{(1+T)\sqrt{1+2T}}\mathrm{e}^{-\frac{\varrho^2}{1+2T}}\right], \tag{B1h}$$

$$m_{31} = \frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\frac{1}{1+T}\exp\left(-\frac{\varrho^2}{1+T}\right), \tag{B1i}$$

$$m_{32} = -\frac{1}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\frac{(1+T)-\varrho^2}{(1+T)^3}T\exp\left(-\frac{\varrho^2}{1+T}\right), \tag{B1j}$$

$$m_{33} = -\frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\left[\frac{(K-2)(1+T)^2 - T\varrho^2}{(1+T)^3}\exp\left(-\frac{\varrho^2}{1+T}\right) + \frac{1}{\sqrt{1+2T}}\exp\left(-\frac{\varrho^2}{1+2T}\right)\right], \tag{B1k}$$

$$m_{34} = -\frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\frac{T\varrho^2}{(1+T)^3}\exp\left(-\frac{\varrho^2}{1+T}\right), \tag{B1l}$$

$$m_{35} = -\frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\frac{T\varrho}{(1+T)^2}\exp\left(-\frac{\varrho^2}{1+T}\right), \tag{B1m}$$

$$m_{41} = -\frac{4}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\exp\left(-\frac{\varrho^2}{1+T}\right)\left[\frac{1}{1+T} - m'_{23}\right], \tag{B1n}$$

$$m_{43} = \frac{4}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\left\{\frac{(K-2)(1+T)^2 + T\varrho^2}{(1+T)^3}\exp\left(-\frac{\varrho^2}{1+T}\right) + \frac{1}{\sqrt{1+2T}}\exp\left(-\frac{\varrho^2}{1+2T}\right) - \frac{2}{\pi}\frac{\eta_{\boldsymbol{w}}}{K}\left[\frac{2}{1+2T}\mathrm{e}^{-\frac{2\varrho^2}{1+2T}} + (K-2)\mathrm{e}^{-\frac{\varrho^2}{1+T}}\times\left(\frac{4}{\sqrt{1+2T}}\mathrm{e}^{-\frac{\varrho^2}{1+2T}} + \frac{K-3}{1+T}\mathrm{e}^{-\frac{\varrho^2}{1+T}}\right)\right]\right\}, \tag{B1o}$$

$$m_{51} = -\frac{2}{\pi}\frac{\eta_\theta}{K}\frac{\varrho}{(1+2T)^{3/2}}\exp\left(-\frac{\varrho^2}{1+2T}\right), \tag{B1p}$$

$$m_{52} = -\frac{1}{\pi}\frac{\eta_\theta}{K}\varrho\left[\frac{\mathrm{e}^{-\frac{\varrho^2}{1+2T}}}{(1+2T)^{3/2}} + \frac{K-1}{(1+T)^2}\mathrm{e}^{-\frac{\varrho^2}{1+T}}\right], \tag{B1q}$$

$$m_{53} = \frac{2}{\pi}\frac{\eta_\theta}{K}(K-1)\frac{\varrho}{(1+T)^2}\exp\left(-\frac{\varrho^2}{1+T}\right), \tag{B1r}$$

$$m_{55} = -\frac{2}{\pi}\frac{\eta_\theta}{K}\left[\frac{\exp\left(-\frac{\varrho^2}{1+2T}\right)}{\sqrt{1+2T}} + \frac{K-1}{1+T}\mathrm{e}^{-\frac{\varrho^2}{1+T}}\right]. \tag{B1s}$$

The remaining elements can be deduced by the matrix relations

$$m_{11} - \frac{1}{2}m_{21} = m_{22} - 2m_{12},$$

$$
\begin{aligned}
m_{33} - \frac{1}{2} m_{43} &= m_{44} - 2m_{34}, \\
m_{13} - \frac{1}{2} m_{23} &= m_{24} - 2m_{14}, \\
m_{31} - \frac{1}{2} m_{41} &= m_{42} - 2m_{32}, \\
m_{25} &= 2m_{15}, \\
m_{45} &= 2m_{45}, \\
m_{54} &= -m_{53}.
\end{aligned}
\tag{B2}
$$

The characteristic polynomial of such a Jacobian matrix, whose zeros define the eigenvalues, separates in a quadratic and a cubic equation. The two eigenvalues given by the quadratic equation correspond to those of the 4×4-matrix with fixed biases and are given by

$$
\lambda_{1,2} = \frac{1}{2} \left[A_1 + B_1 \pm \sqrt{(A_1 - B_1)^2 + 4 C_1 D_1} \right], \tag{B3a}
$$

with

$$
\begin{aligned}
A_1 &= m_{11} - \frac{1}{2} m_{21} \qquad & B_1 &= m_{44} - 2m_{34}, \\
C_1 &= m_{31} - \frac{1}{2} m_{41} \qquad & D_1 &= m_{24} - 2m_{14}.
\end{aligned}
\tag{B3b}
$$

These eigenvalues are nonlinear in the learning rate $\eta_{\boldsymbol{w}}$.

The remaining eigenvalues are given by the solutions to the cubic equation

$$
\lambda^3 + a_2 \lambda^2 + a_1 \lambda + a_0 = 0, \tag{B4a}
$$

with coefficients

$$
\begin{aligned}
a_2 &= -(m_{55} + A_2 + B_2), \\
a_1 &= m_{55}(A_2 + B_2) + (A_2 B_2 - C_2 D_2) \\
&\quad - E_2 m_{15} - m_{54} m_{35}, \\
a_0 &= -m_{55}(A_2 B_2 - C_2 D_2) + m_{54}(m_{35} A_2 - m_{15} D_2) \\
&\quad + E_2 (m_{15} B_2 - m_{35} C_2),
\end{aligned}
\tag{B4b}
$$

where

$$A_2 = m_{11} + 2m_{12}, \qquad B_2 = m_{44} + \frac{1}{2}m_{43},$$
$$C_2 = m_{31} + 2m_{32}, \qquad D_2 = m_{24} + \frac{1}{2}m_{23}, \tag{B4c}$$
$$E_2 = m_{51} + 2m_{52}.$$

These eigenvalues are negative for all values of $\eta_{\boldsymbol{w}}$ and η_θ. For $\eta_{\boldsymbol{w}} = \eta_\theta = \eta_0$, these eigenvalues are also linear in η_0.

This can be confirmed by finding the zeros of the determinant in the two learning rates $\eta_{\boldsymbol{w}}$ and η_θ, which correspond to an eigenvalue becoming zero and therefore define critical (maximal) learning rates. For the equations for the determinant roots

$$a_2 = 0, \tag{B5a}$$
$$A_2 B_2 - C_2 D_2 = 0, \tag{B5b}$$

we obtain only one non-trivial, i.e., non-zero, solution for the weight learning rate $\eta_{\boldsymbol{w}}$, coinciding with $\lambda_1 = 0$, and in particular no non-trivial solution for the bias learning rate η_θ. This and numerical solutions suggest that the optimal bias learning rate is located at infinity.

This can be explicitly shown for the special case $\varrho = 0$, where the eigenvalue spectrum separates further. A closer inspection of the matrix elements reveals that all m_{5i} and m_{i5} for $i \neq 5$ become zero and the eigenvalues take the form

$$\lambda_{3,4} = \frac{1}{2}\left[A_2 + B_2 \pm \sqrt{(A_2 - B_2)^2 + 4C_2 D_2}\right], \tag{B6a}$$
$$\lambda_5 = m_{55}, \tag{B6b}$$

recovering the convergence dynamics of the weight order parameters in the isotropic case with fixed biases studied previously [14], but for an extra eigenvalue describing the decay of the student biases to their optimal value. Since only this eigenvalue depends (linearly) on η_θ, the optimal bias learning rate is at infinity.

To make progress in the general case of non-zero teacher bias, we restrict our study to two possible parameterizations $\eta_0 = \eta_\theta = \eta_{\boldsymbol{w}}$ and a finite weight learning rate $\eta_{\boldsymbol{w}}$ with $\eta_\theta \to \infty$. In the following, we use the convention that the (weight) learning rate will be denoted by η for the generic case or when a result is valid for both parameterizations.

For large η_θ, we expand the characteristic polynomial (B4a) asymptotically with the two ansätze $\lambda = O(\eta_\theta)$ and $\lambda = O(1)$. One finds that the characteristic polynomial separates as expected into

$$\begin{aligned}\lambda_{3,4} = \frac{1}{2}(A_2 + B_2) - \frac{E_2 m_{15} + m_{54} m_{35}}{2 m_{55}} \\ \pm \frac{1}{2} \Bigg[(A_2 - B_2)^2 + 4 C_2 D_2 \\ + \left(\frac{E_2 m_{15} + m_{54} m_{35}}{m_{55}} \right)^2 \\ - 2 \frac{(A_2 - B_2)(E_2 m_{15} - m_{54} m_{35})}{m_{55}} \\ - 4 \frac{E_2 m_{35} C_2 + m_{54} m_{15} D_2}{m_{55}} \Bigg]^{1/2} ,\end{aligned} \tag{B7a}$$

$$\lambda_5 = m_{55} , \tag{B7b}$$

which is similar to the zero bias case, but with corrections to the eigenvalues $\lambda_{3,4}$ due to the finite biases. However, these eigenvalues become independent of the value of η_θ.

In order to study the optimal value of the learning rate η, which gives the fastest decay to zero generalization error, one has to assess which mode, i.e., eigenvalue and associated eigenvector, contributes to its decay. We therefore expand the generalization error (A2) to second order in $\{q, r, s, c, \vartheta\}$

$$\begin{aligned}\epsilon_{\mathrm{g}} = \frac{\mathrm{e}^{-\frac{\varrho^2}{1+2T}}}{\pi \sqrt{1+2T}} \Bigg[(2r - q) - \frac{1}{4}(2r - q)^2 \frac{T(1+2T) + \varrho^2}{(1+2T)^2} \\ + q(r - q) \frac{(1+2T) - 2\varrho^2}{(1+2T)^2} + \vartheta \varrho \frac{2r - 3q}{1+2T} + \vartheta^2 \Bigg] \\ - \frac{K-1}{\pi(1+T)} \mathrm{e}^{-\frac{\varrho^2}{1+T}} \Bigg[(2s - c) + q(s - c) \frac{(1+T) - \varrho^2}{(1+T)^2} \\ + \frac{1}{4}(4s^2 - 2c^2 - q^2) \frac{\varrho^2}{(1+T)^2}\end{aligned}$$

$$+ \vartheta(2s - 2c + q)\frac{\varrho}{1+T} - \vartheta^2\Bigg] . \tag{B8}$$

Unfortunately, we were unable to find analytical solutions to the eigenvectors. Numerical solutions, however, show that the eigenvectors associated with the eigenvalues $\lambda_{3,4,5}$ are orthogonal to the first-order terms in the generalization error and thus cannot contribute to their decay. These modes are therefore only relevant for second-order terms in the generalization error with a decay rate of $2\lambda_{3,4,5}$. As discussed in Sect. V, the fastest convergence is given by Eq. (14). This is usually 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.

It is in general infeasible to optimize the eigenvalues with respect to the learning parameter η ($\eta_{\boldsymbol{w}}$ or η_0) analytically for arbitrary K, T and ϱ. However, one can make some progress in certain limits of K, T, and ϱ which we will investigate below.

1. Large-K limit

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

$$\begin{aligned}\lambda_1 = -\frac{4}{\pi}\frac{\eta_{\boldsymbol{w}}\mathcal{E}_1}{K}\frac{(1+T) - \sqrt{1+2T}\mathcal{E}_2}{1+2T} \\ \times \frac{\pi\sqrt{1+2T} - \eta_{\boldsymbol{w}}\mathcal{E}_1}{\pi(1+T) - \eta_{\boldsymbol{w}}\mathcal{E}_1\mathcal{E}_2},\end{aligned} \tag{B9a}$$

$$\begin{aligned}\lambda_3 = -\frac{2}{\pi}\frac{1}{K}\frac{\eta_{\boldsymbol{w}}\eta_\theta\mathcal{E}_1}{\eta_\theta(1+T)^2 + \eta_{\boldsymbol{w}}T\varrho^2}\Bigg[\frac{(1+T)^2}{(1+2T)^{3/2}} \\ + \frac{T\varrho^2}{(1+2T)^{5/2}} - \frac{\mathcal{E}_2}{1+T}\Bigg],\end{aligned} \tag{B9b}$$

with the auxiliary variables

$$\mathcal{E}_1 = \exp\left(-\frac{\varrho^2}{1+2T}\right), \tag{B9c}$$

$$\mathcal{E}_2 = \exp\left(-\frac{T\varrho^2}{(1+T)(1+2T)}\right). \tag{B9d}$$

These define two critical learning rates

$$\eta_{\boldsymbol{w}}^{\max} = \pi \frac{\sqrt{1+2T}}{\mathcal{E}_1}, \tag{B10a}$$

$$\eta_{\boldsymbol{w}}^{\mathrm{crit}} = \pi \frac{1+T}{\mathcal{E}_1 \mathcal{E}_2} > \eta_{\boldsymbol{w}}^{\max}, \tag{B10b}$$

where λ_1 is identical to zero ($\eta_{\boldsymbol{w}}^{\max}$) [corresponding to the maximal learning rate that can also be obtained by solving Eq. (B5b)] and diverges ($\eta_{\boldsymbol{w}}^{\mathrm{crit}}$), respectively. Inspecting Eqs. (B9) and (B10) suggests that the natural rescaling for the learning rates for non-zero teacher bias in this limit is

$$\hat{\eta}_{\boldsymbol{w}} = \eta_{\boldsymbol{w}} \mathcal{E}_1 \quad \text{and} \quad \hat{\eta}_{\theta} = \eta_{\theta} \mathcal{E}_1. \tag{B11}$$

We further mention in passing, that Eq. (B9a) is only a valid expansion of λ_1 for $\eta_{\boldsymbol{w}} < \eta_{\boldsymbol{w}}^{\mathrm{crit}}$, beyond which the ansatz $\lambda_1 = O(K^{-1})$ breaks down, a fact that becomes important when optimizing the dynamics with respect to the learning rate.

For both of parameterizations ($\eta_0 = \eta_{\boldsymbol{w}} = \eta_{\theta}$ and $\eta_{\boldsymbol{w}}$ with $\eta_{\theta} = \eta_{\theta}^{\mathrm{opt}} \to \infty$) this optimization is performed by calculating both η_r^{opt} and η_m^{opt}, i.e., solving $2\lambda_3 = \lambda_1$ and $d\lambda_1/d\eta = 0$, respectively. Since λ_1 is only a function of $\eta_{\boldsymbol{w}}$, η_m^{opt} is identical for both parameterizations, whereas η_r^{opt} is in general different. The candidates for the optimal learning rate take the form

$$\begin{aligned}
\eta_{0,r}^{\mathrm{opt}} = {} & \pi(1+T)T\mathcal{E}_1^{-1} \left\{ 2(1+T)^2 \left[(1+T)(1+2T) + 2T\varrho^2 \right] - (1+2T)^{5/2}(2+T+\varrho^2)\mathcal{E}_2 \right\} \\
& \times \left\{ (1+2T)^{3/2}(1+T)^2 \left[(1+T)^2 + T\varrho^2 \right] \right. \\
& \left. - 2(1+T)\left[2(1+2T)(1+T)^3 + T(1+2T+2T^2)\varrho^2 \right] \mathcal{E}_2 + (1+2T)^{5/2}\mathcal{E}_2^2 \right\}^{-1},
\end{aligned} \tag{B12a}$$

$$\eta_{\boldsymbol{w},r}^{\mathrm{opt}} = \frac{\pi(1+T)T}{\mathcal{E}_1} \frac{(1+T)\left[2(1+T)^2(1+2T) - \varrho^2 \right] - (1+2T)^{5/2}(2+T)\mathcal{E}_2}{(1+2T)^{3/2}(1+T)^4 - (1+T)\left[2(1+2T)(1+T)^3 + T\varrho^2 \right]\mathcal{E}_2 + (1+2T)^{5/2}\mathcal{E}_2^2}, \tag{B12b}$$

$$\eta_m^{\mathrm{opt}} = \eta_{\mathrm{crit}} - \pi \frac{\sqrt{1+T}}{\mathcal{E}_1 \mathcal{E}_2} \left[(1+T) - \sqrt{1+2T}\mathcal{E}_2 \right]^{1/2}. \tag{B12c}$$

To decide on the correct optimal learning rate η^{opt}, one has to evaluate whether $\eta_{r,m}^{\mathrm{opt}} < \eta_{\mathrm{crit}}$ since the solution is otherwise spurious due to the breakdown of the ansatz for λ_1 above η_{crit}. For the remaining valid candidates the optimal convergence rate is calculated. In general,

one finds for given T and ϱ that $\eta^{\text{opt}} = \eta_r^{\text{opt}}$ for $T > T^{\text{crit}}(\varrho)$ and $\eta^{\text{opt}} = \eta_m^{\text{opt}}$ for $T < T^{\text{crit}}(\varrho)$, where $T^{\text{crit}}(\varrho)$ is defined by $\eta_r^{\text{opt}} = \eta_m^{\text{opt}}$.

To make further progress in the $K \to \infty$ limit, one can look at several limits for T and ϱ. For the limits $T \to \infty$ and $T \to 0$, one has to consider scaling ansätze for the biases with T which ensure that the biases remain meaningful. As discussed in Sect. II and subsequently Sect. V C, one can adopt two possible interpretations of the influence of the biases which are identical to leading orders for $T \to \infty$ but qualitatively different for $T \to 0$. The *effective bias* ($\varrho = \hat{\varrho}\sqrt{1+T}$) keeps the mean hidden unit output constant for all T. The *abscissa* ($\varrho = \check{\varrho}\sqrt{T}$) keeps the distance of the decision hyperplane (or root) constant.

There are some further subtleties when studying various limits. The results for first taking the $K \to \infty$ limit and than the large-T limit 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 $T = T_\infty K$, where T_∞ controls the significance between T and K. However, there is a significant difference to the case where the $T \to \infty$ limit is taken first, which will also be studied below. For small T on the other hand, the limits $K \to \infty$ and $T \to 0$ are interchangeable to third order. Below, we therefore only use those expansions which give us the more general solutions.

2. Small-T limit

In this limit, the slowest mode is associated with λ_1 and the optimal learning rate is determined by η_m^{opt} which is identical for both learning rate parameterizations and the leading terms of the interesting quantities are

$$\eta_{\max} = \pi \mathrm{e}^{\hat{\varrho}^2} \left[1 + \left(1 - \frac{K+4}{K} \hat{\varrho}^2 \right) T \right], \tag{B13a}$$

$$\eta^{\text{opt}} = \eta_{\max} - \pi \mathrm{e}^{\hat{\varrho}^2} \left[\sqrt{\frac{(K-1)\hat{\varrho}^2}{K}} \sqrt{T} - \frac{K-2}{K} \hat{\varrho}^2 T \right], \tag{B13b}$$

$$\lambda^{\rm opt} = -4\frac{T}{K}\left\{\hat{\varrho}^2 - 2\sqrt{\frac{K-1}{K}}\hat{\varrho}^2\sqrt{T} + \frac{1}{2}\left[1 - 4\hat{\varrho}^2 + 5\frac{K-4}{K}\hat{\varrho}^4\right]T\right\}. \tag{B13c}$$

The result for the model without biases can be recovered to leading order by simply setting $\hat{\varrho} = 0$. This shows that learning speed is improved by a factor of T for non-zero (finite) bias since the two leading terms of $\lambda^{\rm opt}$ vanish for $\hat{\varrho} = 0$. In this limit, the effective bias $\hat{\varrho}$ dominates the dynamics. It is obvious, that this expansions suffers from two drawbacks. First, the limit of zero bias cannot be taken adequately for higher orders (this is especially obvious for higher-order terms in $\eta^{\rm opt}$, which have not been included here for brevity, where $\hat{\varrho}$ appear in the denominator). Second, the expansion predicts a unabated increase of the optimal convergence rate $\lambda^{\rm opt}$ with $\hat{\varrho}$, which is not the case for any finite T, where $\lambda^{\rm opt}$ levels off and eventually decays exponentionally. This is due to the implicit assumption in the $T \to 0$ expansion that $\hat{\varrho}^2 \ll -\log T$, i.e., the small T terms always dominate the solution over exponentional terms in $\hat{\varrho}$. Below, we will address the first of the inadequacies, by analyzing the $T \to 0$ limit, with the scaling $\check{\varrho} = \varrho/\sqrt{T}$, i.e., ϱ vanishes with T.

3. Small-T limit and $\check{\varrho}$

As in the small-T limit with ϱ finite, the slowest mode is associated with λ_1 and both parameterizations are identical. In particular, one finds

$$\eta_{\rm max} = \pi\left[1 + (1+\check{\varrho}^2)T + \frac{\check{\varrho}^4}{2}T^2 - \frac{K+4}{2K}(1+2\check{\varrho}^2)T^2\right], \tag{B14a}$$

$$\eta^{\rm opt} = \eta_{\rm max} - \pi\sqrt{\frac{K-1}{2K}}\sqrt{1+2\check{\varrho}^2}\,T, \tag{B14b}$$

$$\lambda^{\rm opt} = -2\frac{T^2}{K}\left\{(1+2\check{\varrho}^2) - 2\left[(1+3\check{\varrho}^2) + \sqrt{\frac{K-1}{2K}}(1+2\check{\varrho}^2)^{3/2}\right]T\right\}. \tag{B14c}$$

In this case, the results for the model without biases are recovered for all orders for $\check{\varrho} = 0$. One can still see, that the learning is improved for non-zero biases, but for this scaling only by a factor of $1 + 2\check{\varrho}^2$ and not by $O(T)$. This expansion holds only for $\check{\varrho}^2 \ll T$ due to the algebraic expansion of all exponentional terms.

4. Large-T and -K limit ($T = T_\infty K$):

For large T, the two scaling ansätze for ϱ are equivalent and the eigenvalue λ_3 has the smallest order. The optimal solution is therefore given by the solution of η_r^{opt} and the leading terms of the relevant quantities become

$$\eta_{\max} = \pi\sqrt{2}\sqrt{T}\mathrm{e}^{\hat{\varrho}^2/2}\left[1 - \frac{\sqrt{T}}{K}\mathrm{e}^{\hat{\varrho}^2/2} + \frac{1 + 4T_\infty + 4T_\infty^2\mathrm{e}^{\hat{\varrho}^2} - \hat{\varrho}^2}{4T}\right], \tag{B15a}$$

$$\eta_0^{\text{opt}} = \eta_{\max} - \frac{\pi\sqrt{2}}{2\sqrt{T}(1+\hat{\varrho}^2)}\mathrm{e}^{\hat{\varrho}^2/2}, \tag{B15b}$$

$$\eta_{\boldsymbol{w}}^{\text{opt}} = \eta_{\max} - \frac{\pi\sqrt{2}}{2\sqrt{T}}\mathrm{e}^{\hat{\varrho}^2/2}, \tag{B15c}$$

$$\lambda_0^{\text{opt}} = -\frac{2}{KT(1+\hat{\varrho}^2)}\left[1 - \frac{\sqrt{T}\mathrm{e}^{\hat{\varrho}^2/2}}{K} + \frac{T_\infty^2\mathrm{e}^{\hat{\varrho}^2} + T_\infty}{T} + \frac{\hat{\varrho}^4 + 4\hat{\varrho}^2 - 2}{2T(1+\hat{\varrho}^2)}\right], \tag{B15d}$$

$$\lambda_{\boldsymbol{w}}^{\text{opt}} = -\frac{2}{KT}\left[1 - \frac{\sqrt{T}\mathrm{e}^{\hat{\varrho}^2/2}}{K} + \frac{2T_\infty^2\mathrm{e}^{\hat{\varrho}^2} + 2T_\infty + \hat{\varrho}^2 - 2}{2T}\right]. \tag{B15e}$$

The comparison for zero biases ($\hat{\varrho} = 0$) reveals that in this limit, the existence of biases slows down the training process to leading order only in the case where $\eta_{\boldsymbol{\theta}} = \eta_{\boldsymbol{w}}$. Furthermore, this decrease is surprisingly only algebraic in $\hat{\varrho}$. This can be explained by the exponential growth of the optimal learning rates matching the gradient decrease due to the saturation of the error function for large $\hat{\varrho}$. Again, this solution is only a good approximation for finite

K and T as long as $\hat{\varrho}^2 \ll \log K$ or $\hat{\varrho}^2 \ll \log T$.

5. Large-T limit

Unlike for small T, the learning behavior changes qualitatively in the $T \to \infty$ limit for K finite, as indicated by numerical solutions. Again λ_3 controls the convergence and one finds to leading orders

$$\eta_{\max} = \pi\sqrt{2}K\left[1 - \frac{K-1}{\sqrt{T}}\mathrm{e}^{-\hat{\varrho}^2/2}\right], \tag{B16a}$$

$$\eta_0^{\mathrm{opt}} = \eta_{\max} - \frac{\pi\sqrt{2}K}{2(1+\hat{\varrho}^2)T}, \tag{B16b}$$

$$\lambda_0^{\mathrm{opt}} = -\frac{2\exp(-\hat{\varrho}^2/2)}{(1+\hat{\varrho}^2)T^{3/2}}\left[1 - \frac{K-1}{\sqrt{T}}\mathrm{e}^{-\hat{\varrho}^2/2}\right], \tag{B16c}$$

$$\eta_{\boldsymbol{w}}^{\mathrm{opt}} = \eta_{\max} - \frac{\pi\sqrt{2}K}{2T}, \tag{B16d}$$

$$\lambda_{\boldsymbol{w}}^{\mathrm{opt}} = -\frac{2}{T^{3/2}}\mathrm{e}^{-\hat{\varrho}^2/2}\left[1 - \frac{K-1}{\sqrt{T}}\mathrm{e}^{-\hat{\varrho}^2/2}\right]. \tag{B16e}$$

In this case, the optimal learning rate is independent of $\hat{\varrho}$ to leading order in T. The exponentionally decreasing gradient therefore directly affects the optimal convergence rate.

REFERENCES

[1] C. Cybenko, Math. Control Signals Syst. **2**, 303 (1989).

[2] D. Saad and S. A. Solla, Phys. Rev. E **52**, 4225 (1995).

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

[4] D. Barber, D. Saad, and P. Sollich, Europhys. Lett. **34**, 151 (1996).

[5] P. Sollich and D. Barber, in *Advances in Neural Information Processing Systems*, edited by M. C. Mozer, M. I. Jordan and T. Petsche (MIT Press, Cambridge, MA, 1997) Vol. 9 p. 274; P. Sollich and D. Barber, Europhys. Lett. **38**, 477 (1997).

[6] P. Riegler and M. Biehl, J. Phys. A **28**, L507 (1995).

[7] P. Riegler, Ph.D. thesis, University of Würzburg, 1997.

[8] A. H. L. West, D. Saad, and I. T. Nabney, in *Advances in Neural Information Processing Systems*, edited by M. C. Mozer, M. I. Jordan and T. Petsche (MIT Press, Cambridge, MA, 1997) Vol. 9 p. 288.

[9] P. J. Werbos, Ph.D. thesis, Harvard University, 1974 (unpublished).

[10] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature* **323**, 533 (1986); in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, edited by D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (MIT Press, Cambridge MA, 1986), Vol. 1, p. 318.

[11] T. Heskes, J. Phys. A **27**, 5145 (1994).

[12] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer Verlag, Heidelberg, 1983), Chap. 6.4, p. 195 ff.

[13] M. Rattray and D. Saad (unpublished).

[14] A. H. L. West and D. Saad, Phys. Rev. E **56** (1997) (to be published).

[15] D. Saad and M. Rattray, in *Proceedings of the Minerva Workshop on Mesoscopics, Fractals and Neural Networks, Eilat, 1997* (to be published).

[16] M. Biehl, P. Riegler, and C. Wöhler, J. Phys. A **29**, 4769 (1996).

[17] D. Nguyen and B. Widrow, in *IJCNN International Conference on Neural Networks*, (IEEE, Piscataway, New Jersey, 1990) Vol. 1-3, Ch. 430, p. C21.

[18] Y. K. Kim and J. B. Ra, in *IEEE International Joint Conference on Neural Networks*, (IEEE, Piscataway, New Jersey, 1991) Vol. 1-3, Ch. 444, p. 2396.

[19] A. van Ooyen and B. Nienhuis, Neural Networks **5**, 465 (1992).

[20] M. Lehtokangas, J. Saarinen, K. Kaski, and P. Huuhtanen, Neural Computation **7**, 982 (1995).

[21] The increase in Q_{ii} leads to a rescaling of the overlaps R_{in} since the normalized overlaps $\hat{R}_{in}$ were randomly fixed. Note also, that similar results are obtained when increasing the initial student lengths individually.

[22] For adiabatic elimination of the hidden-output weights one finds similarly that the outputs of the student hidden units are suppressed initially by an equilibrium of the output weights close to 0 [13]. However, this does not inhibit the progress of the student as in the case of the biases.

[23] Note that for $T = 1$, $T_{\boldsymbol{w}}^{\rm crit}$ also falls briefly below T in the range $0.80 \lessapprox \hat{\varrho} \lessapprox 1.10$ ($1.15 \lessapprox \varrho \lessapprox 1.50$) and $T_0^{\rm crit} < T_{\boldsymbol{w}}^{\rm crit} < T$. The ratio of the learning rates still drops due to the widenin cap between $\lambda_3^{\boldsymbol{w}}$ and λ_3^0 for increasing $\hat{\varrho}$.

[24] D. Saad and S. A. Solla, in *Advances in Neural Information Processing Systems*, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (MIT Press, Cambridge, U.S.A., 1996) Vol. 8, p. 302.

[25] D. Saad and S. A. Solla, in *Advances in Neural Information Processing Systems*, edited

by M. C. Mozer, M. I. Jordan and T. Petsche (MIT Press, Cambridge MA, 1997) Vol. 9, p. 260.

[26] For zero degenerate teacher biases, the student biases converge exactly to zero, whereas for non-zero "degenerate" teacher biases one finds the most self-consistent results for *effective* biases, i.e., degenerate effective teacher biases lead to approximately degenerate effective student biases.

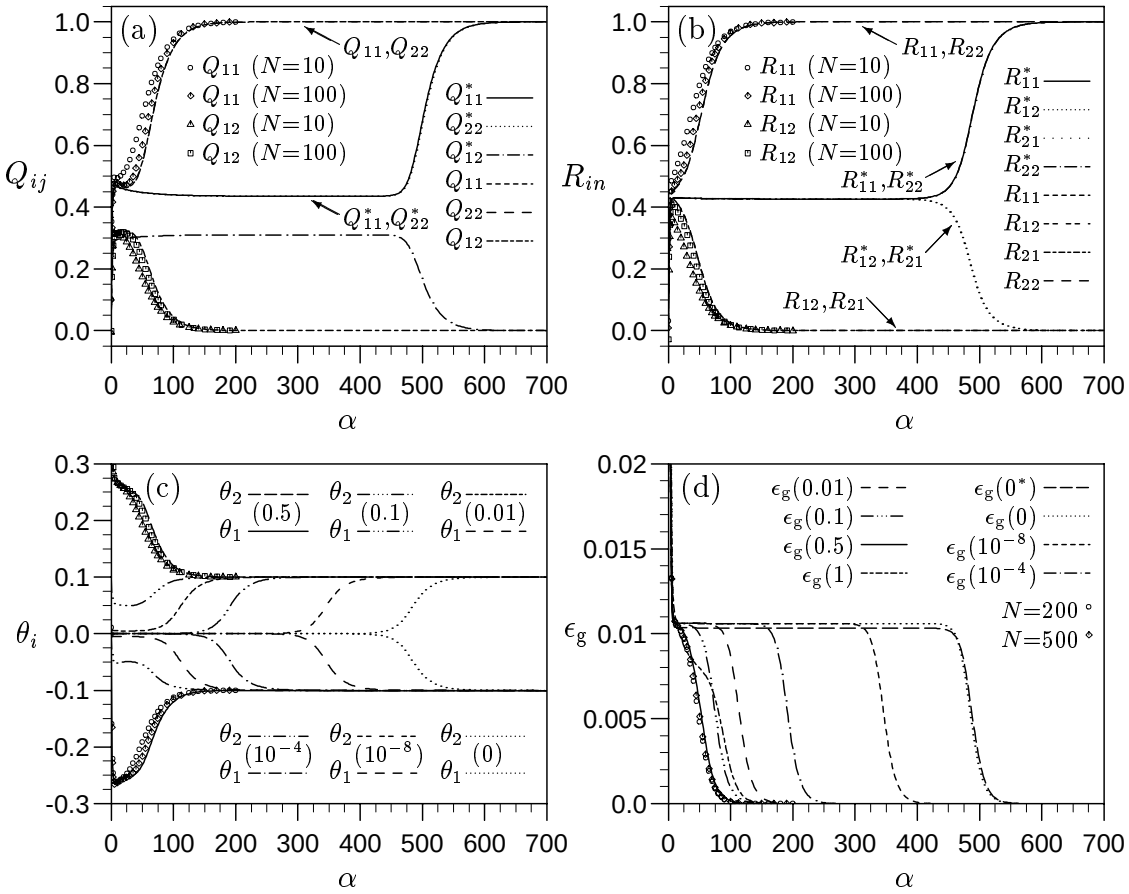

FIG. 1. The dynamical evolution of (a) the student-student overlaps Q_{ij} and (b) the student-teacher overlaps R_{in} as a function of the normalized example number α is compared for two student-teacher scenarios. One student network (denoted by $*$) has fixed zero biases and is trained using examples generated by a bias-less teacher network. Other student networks have adjustable biases and are learning to imitate a teacher task with non-zero biases. The influence of the symmetry in the initialization of the biases on the dynamics is shown for (c) the student biases θ_i and (d) the generalization error ϵ_g. The initial value of $\theta_1 = 0$ is kept for all runs, but θ_2 varies and is given in brackets in the legends. Finite size simulations for input dimensions $N = 10 \ldots 500$ show that the dynamical variables are self-averaging. For all order parameters and the biases the mean trajectories for $N = 10$ and $N = 100$ are shown for the relevant order parameters (see the legends, for biases: θ_1 [$N = 10$ ($\circ$), $N = 100$ ($\diamond$)]; θ_2 [$N = 10$ ($\triangle$), $N = 100$ ($\square$)]). For the generalization error we show the results for $N = 200$ and $N = 500$ for comparison.

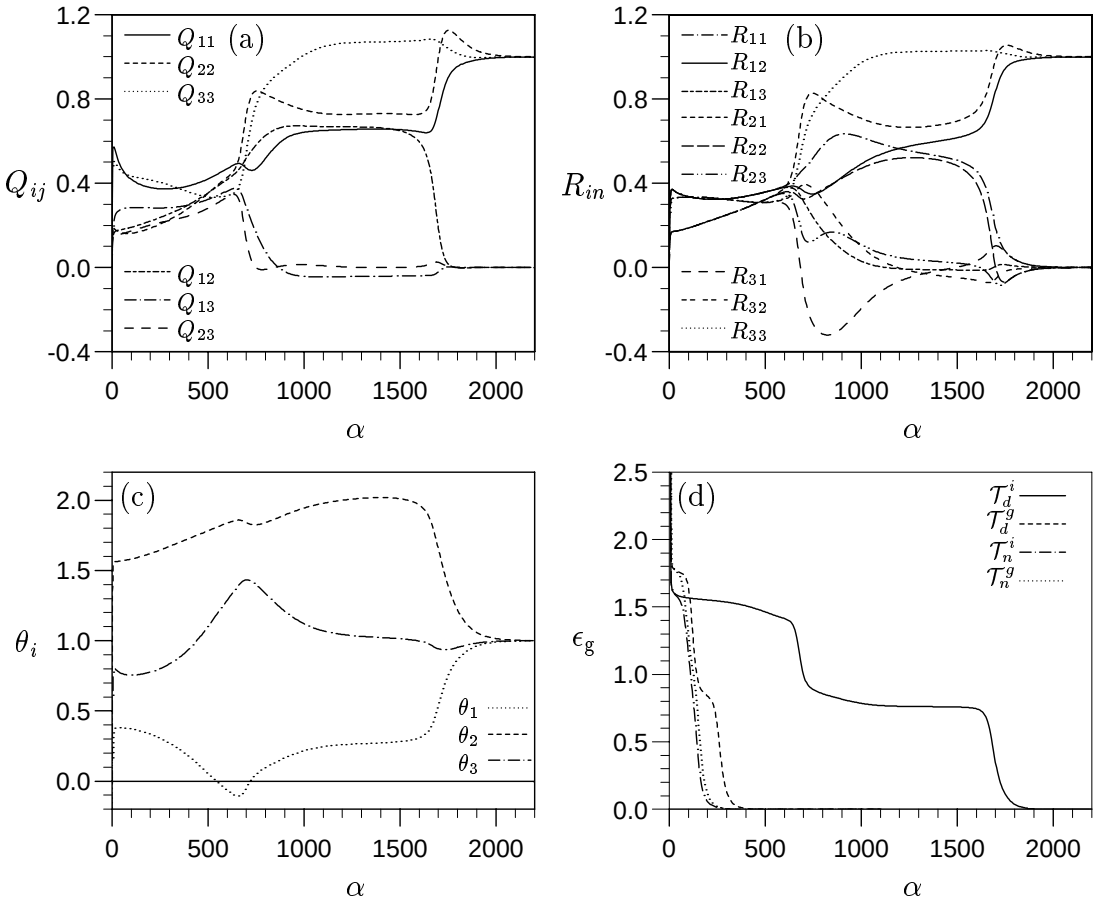

FIG. 2. The dynamical evolution of the student-student overlaps Q_{ij} (a), the student-teacher overlaps R_{in} (b), the student biases θ_i (c) and the generalization error ϵ_g (d) as a function of the normalized example number α is shown for a realizable scenario $K = M = 3$ and $\eta_0 = \eta_\theta = \eta_{\boldsymbol{w}} = 2$. The teacher tasks $\mathcal{T}_d^i$ large degree of symmetry ($T_{nm} = \delta_{nm}$ and $\varrho_n = 1$) is responsible for the very slow specialization process that takes place in two identifiable stages. Training time is shortened considerably when the teacher vector isotropy or bias degeneracy is broken.

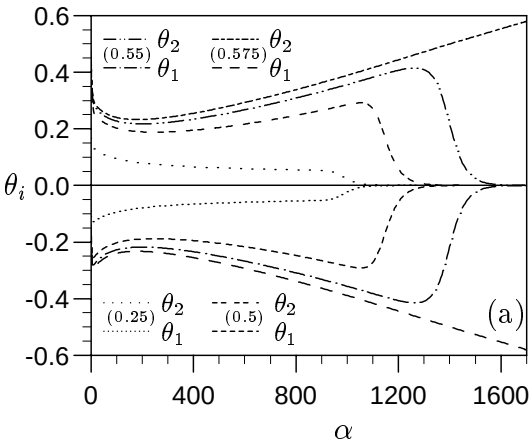

FIG. 3. The dynamical evolution of the biases θ_i for a student imitating an isotropic teacher with zero biases reveals symmetric dynamics for θ_1 and θ_2. The student was randomly initialized identically for the different runs, but for a change in the range of the random initialization of the biases ($U[-b, b]$), with the value of b given in the legend. Above a critical value of b the student remains trapped in a suboptimal phase.

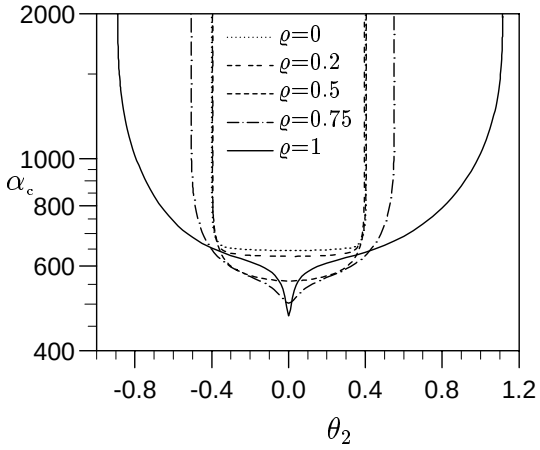

FIG. 4. The convergence time $\alpha_c(\theta_2)$ (see the text) is shown for several values of the common teacher bias for the degenerate teacher bias task $\mathcal{T}_d^i$ ($\varrho_n = \varrho$). α_c diverges for large enough initial magnitude of θ_2 for all values of ϱ (see the legend). For increasing ϱ the basin of attraction to the optimal solution becomes asymmetric and larger.

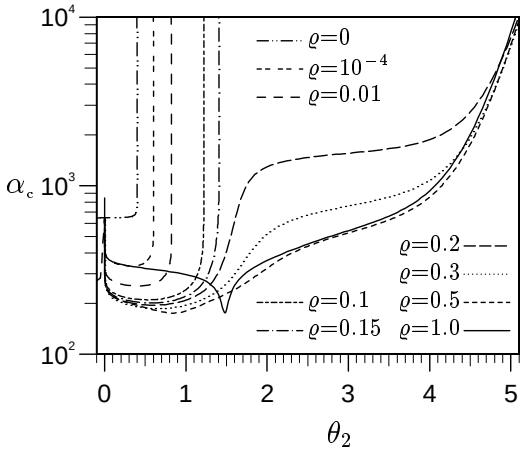

FIG. 5. The convergence time $\alpha_{\rm C}(\theta_2)$ is shown in terms of the asymmetry in the teacher biases $\mathcal{T}_n^i$ ($\varrho_n = \pm\varrho$). These tasks also exhibit an attractive suboptimal fixed point for small ϱ, but with a smaller basin of attraction. Above a critical value the suboptimal fixed point becomes unstable although it still can influence the learning process considerably. For very large initial values θ_2 (and large enough ϱ), the learning process is slowed down exponentionally, but the student is still able to converge to the optimal solution eventually.

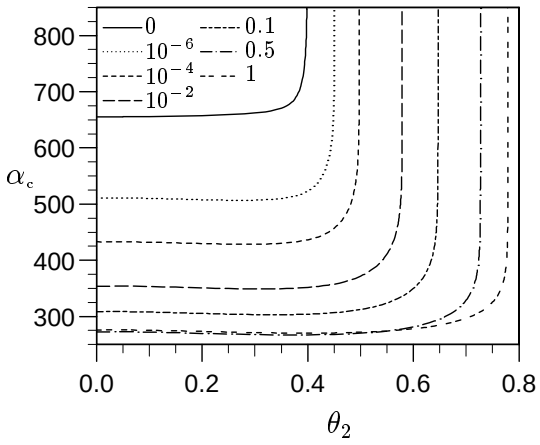

FIG. 6. The convergence time $\alpha_{\rm C}$ is shown as a function of difference in the teacher lengths $\delta T = T_{22} - T_{11}$ (see the legend). $\alpha_{\rm C}$ is also reduced as for the asymmetric bias case (Fig. 5), but the basin of attraction does not grow as significantly for the tasks $\mathcal{T}_d^g$.

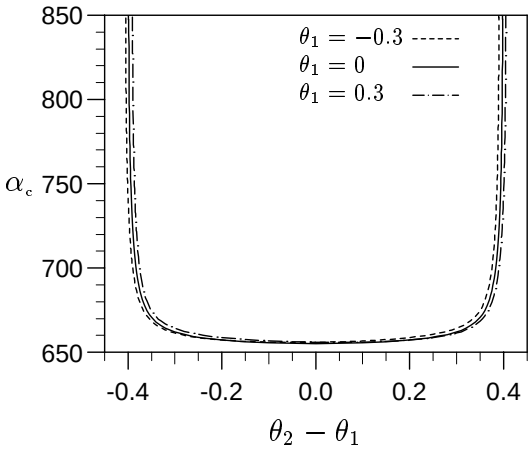

FIG. 7. The basin of attraction for initial θ_2 shown for several values of θ_1 depends almost solely on the difference $\theta_2 - \theta_1$.

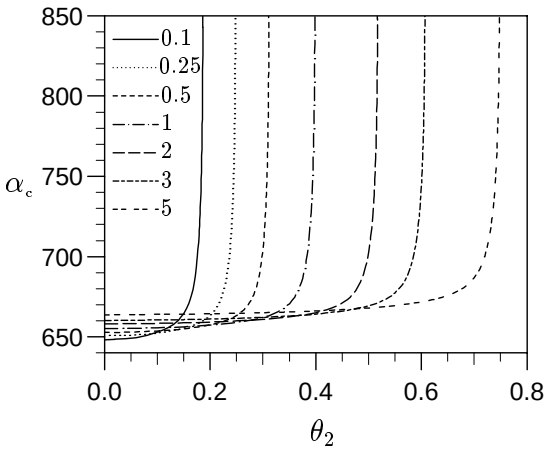

FIG. 8. The basin of attraction for initial θ_2 shown for several magnification factors M of the initial student-student overlaps Q_{ij} (see the legend) increases with the size of these initial values.

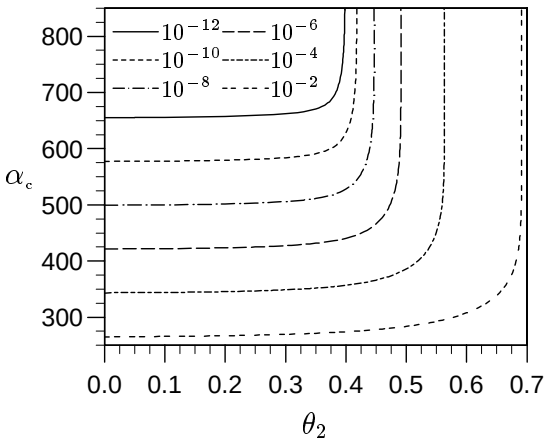

FIG. 9. Although the basin of attraction for initial θ_2 grows with the range of initial student-teacher overlaps $\hat{r}$ (for values see the legend) the dynamics still get trapped in a suboptimal configuration for large enough θ_2. Since $\hat{r} \sim 1/\sqrt{N}$, this gives some indication how finite size systems may behave.

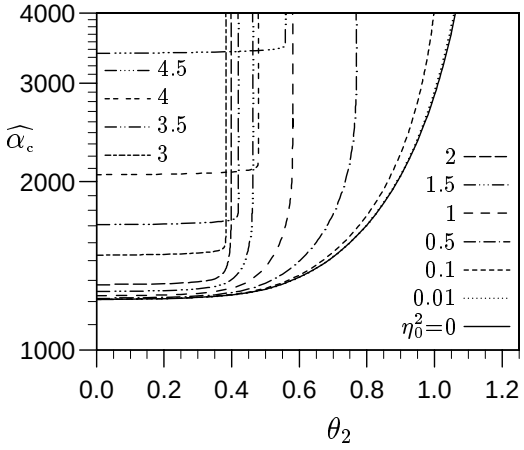

FIG. 10. The normalized convergence time $\widehat{\alpha_c} \equiv \eta_0 \alpha_c$ is shown as a function of the initialization of θ_2 for various learning rates η_0 (see the legend, $\eta_0^2 = 0$ represents the dynamics neglecting η_0^2 terms.).

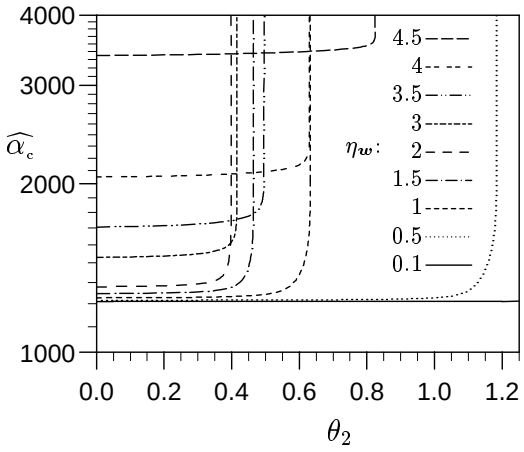

FIG. 11. The normalized convergence time as a function of θ_2 is shown for various weight learning rates $\eta_{\boldsymbol{w}}$ (see the legend) with the bias learning rate fixed at $\eta_\theta = 2$. For very small weight learning rate the basin of attraction increases quickly (for $\eta_{\boldsymbol{w}} = 0.1$ the training diverges for $\theta_{\rm crit} = 5.415$).

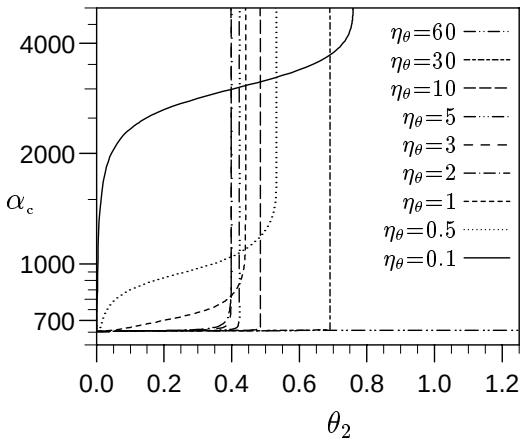

FIG. 12. The convergence time $\alpha_{\rm C}(\theta_2)$ is plotted for various bias learning rates η_θ (see the legend) with the weight learning rate fixed at $\eta_{\boldsymbol{w}} = 2$. For very large bias learning rate the basin of attraction extends to very large values, e.g., to $\theta_{\rm crit} = 5.735$ for $\eta_\theta = 60$, although the training is still eventually slowed down exponentionally for very large initial values of θ_2.

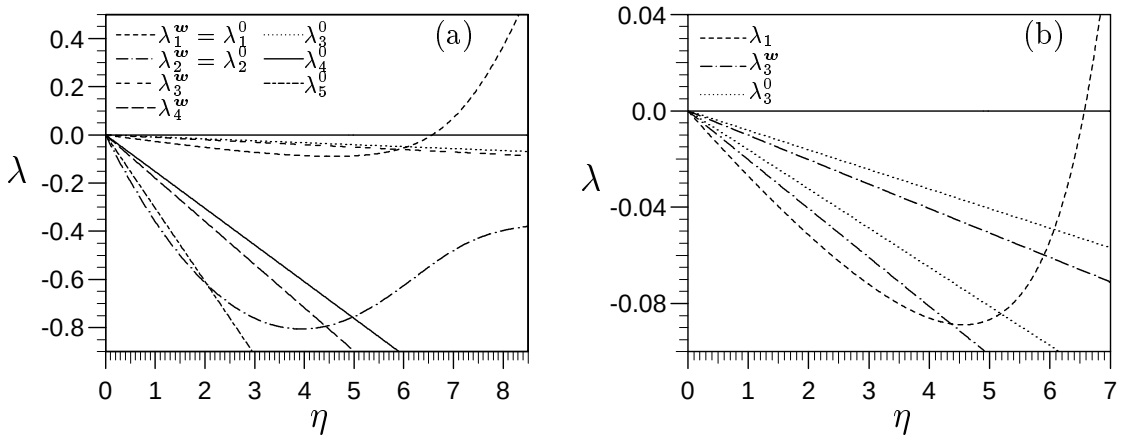

FIG. 13. (a) The eigenvalues λ_i^0 and $\lambda_i^{\boldsymbol{w}}$ are shown as a function of the applicable learning rate η for $K = 5$, $T = 1$ and $\varrho = 1$ for the cases $\eta_\theta = \eta_{\boldsymbol{w}} = \eta_0$ and $\eta_\theta \to \infty$, respectively. (b) The two relevant eigenvalues (see the text) λ_1 and λ_3 are magnified for the same scenario. For comparison we plot $2\lambda_3$ and find that the optimal learning rate $\eta^{\rm opt}$ is given by the minimum of λ_1 for $\eta_\theta \to \infty$ but by the root of $\lambda_1 - 2\lambda_3$ for $\eta_\theta = \eta_{\boldsymbol{w}}$.

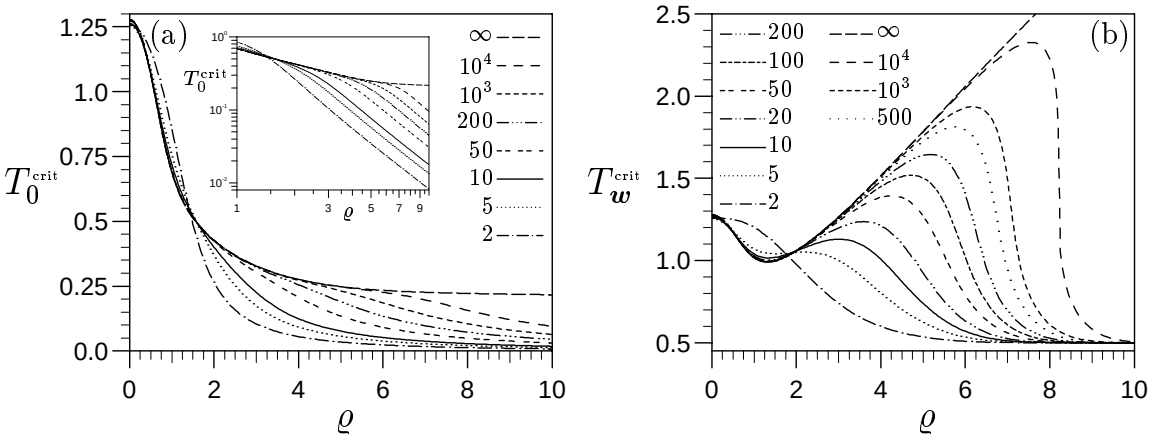

FIG. 14. The critical teacher lengths $T_0^{\rm crit}$ (a) for $\eta_\theta = \eta_{\boldsymbol{w}}$ and $T_{\boldsymbol{w}}^{\rm crit}$ (b) for $\eta_\theta \to \infty$ as a function of $\hat{\varrho}$ for several K values given in the legend (∞ represents the $K \to \infty$ limit). $T^{\rm crit}$ defines the transition between the optimal convergence given by the minimum of λ_1 and by the root of $\lambda_1 - 2\lambda_3$. Notice that for given T, the solution type can change for increasing $\hat{\varrho}$ at most once for $\eta_\theta = \eta_{\boldsymbol{w}}$, whereas it can change up to three times for $\eta_\theta \to \infty$. The inset in (a) shows the power-law decay of $T_0^{\rm crit} \propto \hat{\varrho}^{-2}$.

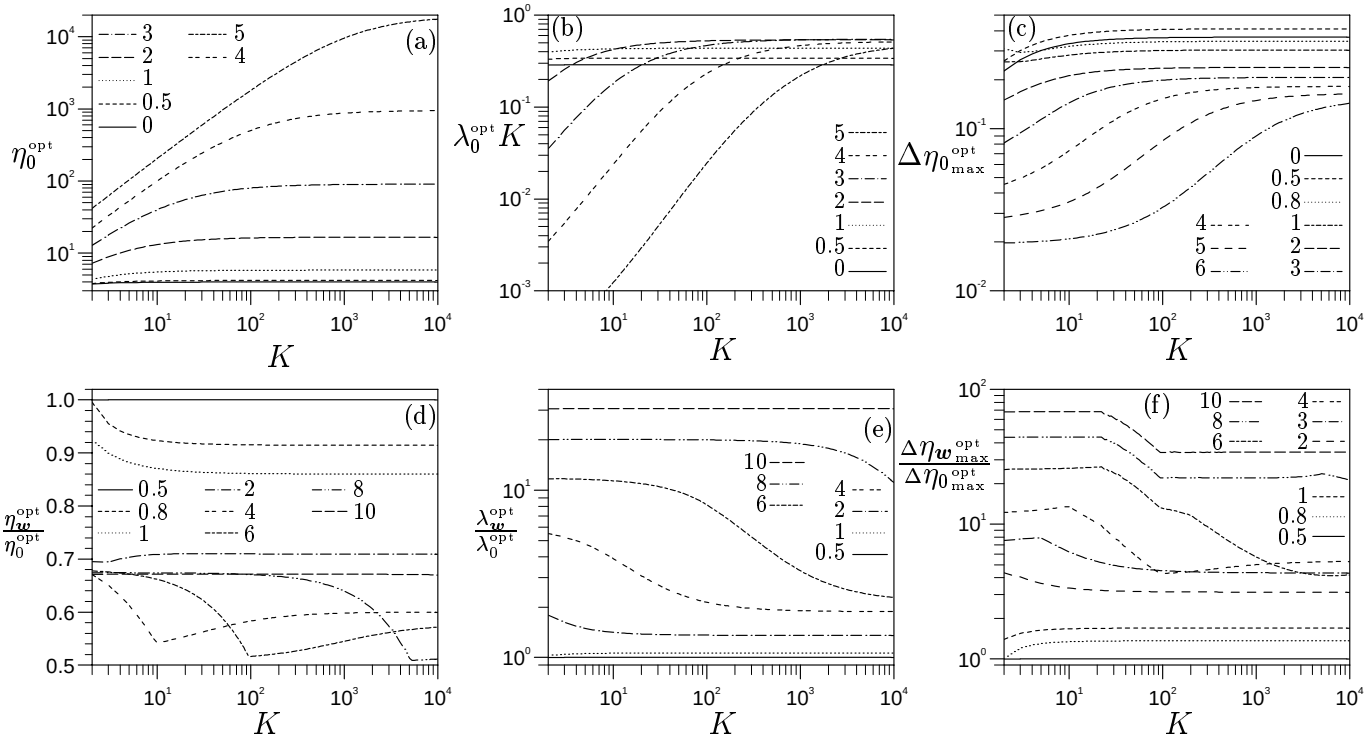

FIG. 15. The convergence scenario as a function of K for $T = 1$ and various ϱ given in the legends. (a) Optimal learning rate $\eta_0^{\rm opt}$ for $\eta_\theta = \eta_0$. (b) Optimal convergence rate $\lambda_0^{\rm opt}$, multiplied by K for convenience. (c) The normalized difference between the optimal and maximal learning rates $\Delta\eta_{0\,\rm max}^{\rm opt}$. Ratio of the optimal learning rates $\eta_{\boldsymbol{w}}^{\rm opt}$ and $\eta_0^{\rm opt}$ (d), the optimal convergence rates $\lambda_{\boldsymbol{w}}^{\rm opt}$ and $\lambda_0^{\rm opt}$ (e), and the normalized differences $\Delta\eta_{\boldsymbol{w}\,\rm max}^{\rm opt}$ and $\Delta\eta_{0\,\rm max}^{\rm opt}$ (f).

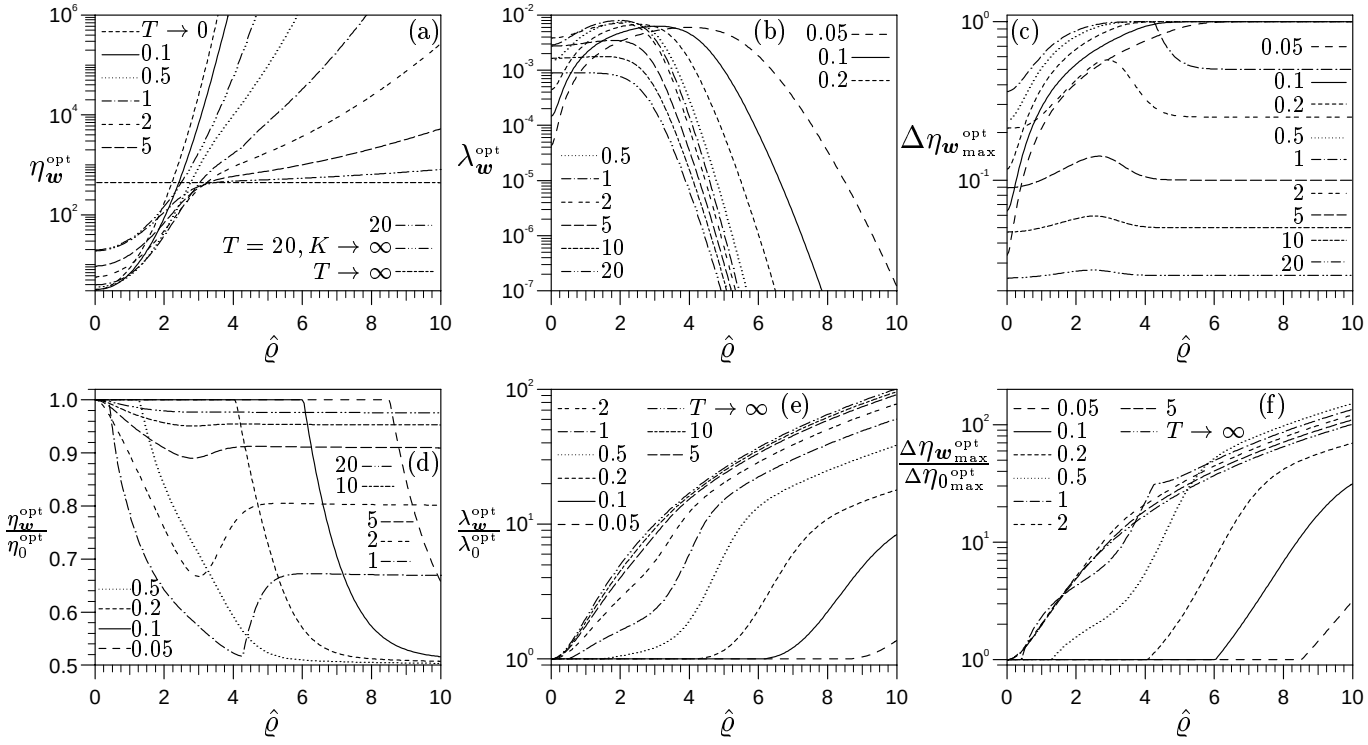

FIG. 16. The convergence scenario as a function of $\hat{\varrho}$ for $K = 10^2$ and various T given in the legends including predictions from expansions for $T \to 0$, $T \to \infty$. (a) Optimal learning rate $\eta_{\boldsymbol{w}}^{\rm opt}$ for $\eta_\theta = \infty$. (b) Optimal convergence rate $\lambda_{\boldsymbol{w}}^{\rm opt}$. (c) The normalized difference between the optimal and maximal learning rates $\Delta\eta_{\boldsymbol{w}\,\rm max}^{\rm opt}$. Ratio of the optimal learning rates $\eta_{\boldsymbol{w}}^{\rm opt}$ and $\eta_0^{\rm opt}$ (d), the optimal convergence rates $\lambda_{\boldsymbol{w}}^{\rm opt}$ and $\lambda_0^{\rm opt}$ (e), and the normalized differences $\Delta\eta_{\boldsymbol{w}\,\rm max}^{\rm opt}$ and $\Delta\eta_{0\,\rm max}^{\rm opt}$ (f).

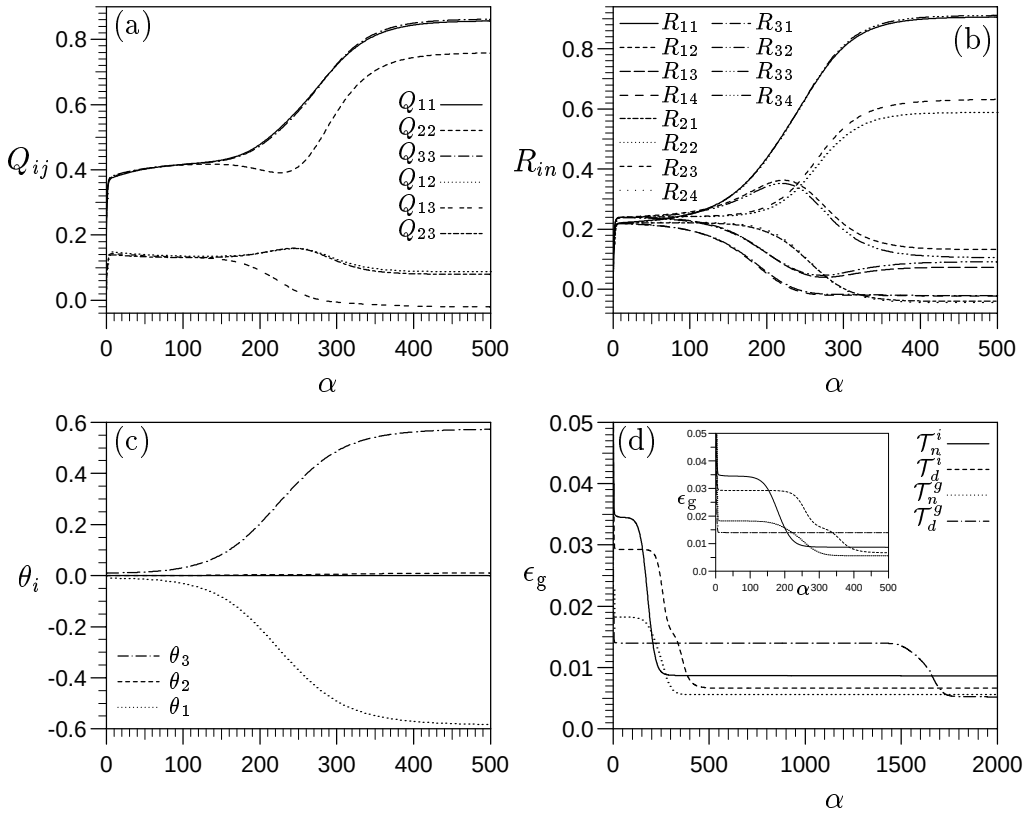

FIG. 17. A typical training dynamics is shown as a function of α for an unrealizable case $K = 3$ and $M = 4$. The teacher task are of the form: $T_{nm} = \delta_{nm}(n+1)/2$ for graded and $T_{nm} = \delta_{nm}$ for isotropic teachers; $\varrho_n = (2n-5)/5\sqrt{1+T_{nn}}$ for non-degenerate and $\varrho_n = 0$ for degenerate teacher biases. The common learning is always $\eta_0 = 2$. The evolution of the student-student overlaps Q_{ij} (a), the student-teacher overlaps R_{in} (b), and the student biases θ_i (c) are shown for $\mathcal{T}_n^i$. The generalization error ϵ_g (d) is shown for all tasks, with the inset magnifying the escape out of the symmetric phase for the students learning the less symmetric tasks.

TABLES

TABLE I. For $T \to 0$ and $T \to \infty$ the optimized dynamics in the convergence phase show power-law behavior in leading order (for more detail including higher-order terms consult Appendix B) for both learning rate parameterizations $\eta_\theta = \eta_{\boldsymbol{w}}$ and $\eta_\theta \to \infty$. The table shows the power laws and the $\hat{\varrho} = \varrho/\sqrt{1+T}$ dependence of the optimal learning parameters $\eta_{\boldsymbol{w}}^{\text{opt}}$ and η_0^{opt}, their respective optimal convergence eigenvalue $\lambda_{\boldsymbol{w}}^{\text{opt}}$ and λ_0^{opt} and the normalized difference between maximal and optimal learning rate $\Delta\eta_{\text{max}}^{\text{opt}} = (\eta_{\text{max}} - \eta^{\text{opt}})/\eta^{\text{opt}}$. Note that for the $T \to 0$ limit both learning rate parameterizations are identical. In this limit, an alternative scaling for the biases ($\check{\varrho} = \varrho/\sqrt{T}$) has been investigated as well.

	$T \to 0$		$T \to \infty$ (K finite)		$T \to \infty$ [$TK^{-1} = O(1)$]	
	$\eta_\theta \geq \eta_{\boldsymbol{w}}$ ($\check{\varrho}$)	$\eta_\theta \geq \eta_{\boldsymbol{w}}$ ($\hat{\varrho}$)	$\eta_\theta = \eta_{\boldsymbol{w}}$	$\eta_\theta \to \infty$	$\eta_\theta = \eta_{\boldsymbol{w}}$	$\eta_\theta \to \infty$
η^{opt}	π	$\pi \mathrm{e}^{\hat{\varrho}^2}$	$\pi\sqrt{2}K$	$\pi\sqrt{2}K$	$T^{1/2}\mathrm{e}^{\hat{\varrho}^2/2}$	$T^{1/2}\mathrm{e}^{\hat{\varrho}^2/2}$
$\Delta\eta_{\text{max}}^{\text{opt}}$	$T\sqrt{1+2\check{\varrho}^2}$	$\sqrt{\hat{\varrho}^2 T}$	$[T(1+\hat{\varrho}^2)]^{-1}$	T^{-1}	$[T(1+\hat{\varrho}^2)]^{-1}$	T^{-1}
λ^{opt}	$T^2K^{-1}(1+2\check{\varrho}^2)$	$TK^{-1}\hat{\varrho}^2$	$T^{-3/2}(1+\hat{\varrho}^2)^{-1}\mathrm{e}^{-\hat{\varrho}^2/2}$	$T^{-3/2}\mathrm{e}^{-\hat{\varrho}^2/2}$	$[TK(1+\hat{\varrho}^2)]^{-1}$	$(TK)^{-1}$