Letter

Soft mode in the dynamics of over-realizable online learning for soft committee machines

Frederieke Richert,[1,*] Roman Worschech,[2,1,*] and Bernd Rosenow[1]
[1]*Institut für Theoretische Physik, Universität Leipzig, Brüderstrasse 16, 04103 Leipzig, Germany*
[2]*Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany*

(Received 8 July 2021; revised 8 March 2022; accepted 4 April 2022; published 17 May 2022)

Overparametrized deep neural networks trained by stochastic gradient descent are successful in performing many tasks of practical relevance. One aspect of overparametrization is the possibility that the student network has a larger expressivity than the data generating process. In the context of a student-teacher scenario, this corresponds to the so-called over-realizable case, where the student network has a larger number of hidden units than the teacher. For online learning of a two-layer soft committee machine in the over-realizable case, we present evidence that the approach to perfect learning occurs in a power-law fashion rather than exponentially as in the realizable case. All student nodes learn and replicate one of the teacher nodes if teacher and student outputs are suitably rescaled and if the numbers of student and teacher hidden units are commensurate.

DOI: 10.1103/PhysRevE.105.L052302

The research field of deep learning has recently attracted considerable attention due to significant progress in performing tasks relevant to many different applications [1–5]. Neural networks are learning machines inspired by the structure of the human brain [6], which have been studied with methods from statistical mechanics [5,7–11], starting with simpler versions such as the perceptron [12–14] and also including two-layer networks [15–29]. Often, learning is studied in the framework of the student-teacher scenario, in which a student has to learn the connection vectors according to which a teacher classifies input patterns [10].

One of the surprising properties of multilayer neural networks is their ability to generalize well even in the overparametrized regime, when the number of model parameters exceeds the number of training examples [30]. Recently an understanding has started to emerge that generalization beyond the training data set can be successful even for strongly overparametrized networks due to implicit regularization in a gradient descent based learning process [31–33]. Bounds on the generalization error have been found to depend on the size of the training data set according to a power law [34–38], in contrast to the exponential decrease of the generalization error in a student-teacher scenario for soft committee machines in the realizable scenario [19,24], where the number of hidden nodes is the same for student and teacher.

Motivated by the fact that in the overparametrized regime the student network may have a larger expressivity than the process which generates the training data, we study the evolution of the generalization error in the over-realizable case of learning in a student-teacher setup, where the student network has a larger number of hidden units than the teacher network. The soft committee machines studied in this Letter are two-layer neural networks that incorporate a smooth activation function, and have a trainable first layer and a fixed second layer of weights. For online learning (i.e., one-pass stochastic gradient descent) of a soft committee machine, where each training example is presented to the student only once, it was found that additional student nodes (beyond the number of teacher nodes) do not learn at all, i.e., the weight vectors of the additional nodes decay to zero [19,24]. Only in a fully trained two-layer network has learning of all student nodes been observed [21,26–28]. Here, we present a rescaling of the output of soft committee machines such that all student nodes learn in the asymptotic limit of a large number of training data. Then, the approach to perfect learning is strikingly different as compared to the realizable case with an exponentially fast convergence to zero generalization error: convergence is of power-law type in the over-realizable case due to the presence of soft modes, which we demonstrate both numerically and analytically. In addition, for the case of a noisy teacher we present numerical evidence that the generalization error is smaller in the over-realizable case than in the realizable one (similar to the case of the fully trained two-layer network studied in [21]).

In our setup, both the student and the teacher network receive inputs $\boldsymbol{\xi}^{\mu} \in \mathbb{R}^N$ at time steps $\mu = 1, \ldots, p$ in the input layer, where the components are independent normally distributed $\xi_i^{\mu} \sim \mathcal{N}(0, 1)$. These inputs are processed in the hidden layer using a nonlinear mapping. The student network has K hidden nodes, with the kth node being characterized by the student vector $\boldsymbol{J}_k^{\mu} \in \mathbb{R}^N$ at time step μ, with a normally distributed initialization $\boldsymbol{J}_{ih}^0 \sim \mathcal{N}(0, \chi^2)$. The teacher network is structured similarly to the student vector, having M hidden nodes, with a teacher vector $\boldsymbol{B}_m \in \mathbb{R}^N$ associated with node m, normalized according to $\boldsymbol{B}_i \cdot \boldsymbol{B}_j = \delta_{ij}$. A linear combination

[*]These authors contributed equally to this work.

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Open access publication funded by the Max Planck Society.

of the outputs of hidden units, obtained using a nonlinear activation function [in our case $g(x) = \mathrm{erf}(x/\sqrt{2})$ due to its analytic properties] yields the student output

$$\sigma(\boldsymbol{J}^\mu, \boldsymbol{\xi}^\mu) = \frac{\sqrt{M}}{K} \sum_{k=1}^{K} g\big(\boldsymbol{J}_k^\mu \cdot \boldsymbol{\xi}^\mu\big). \tag{1}$$

Importantly, the normalization factor $\sqrt{M}/K$ is chosen such that it matches the teacher output $\zeta(\boldsymbol{B}, \boldsymbol{\xi}^\mu) = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} g(\boldsymbol{B}_m \cdot \boldsymbol{\xi}^\mu)$, while at the same time ensuring that the over-realizability is attended to by the additional factor of M/K as compared to the teacher normalization. Learning of the student is achieved by minimizing the loss function $\epsilon(\boldsymbol{J}^\mu, \boldsymbol{\xi}^\mu) = \frac{1}{2}[\sigma(\boldsymbol{J}^\mu, \boldsymbol{\xi}^\mu) - \zeta(\boldsymbol{B}, \boldsymbol{\xi}^\mu)]^2$ in each time step μ, using stochastic gradient descent [17,18]. In this procedure the ith student vector is updated from time μ to time $\mu + 1$ via the gradient of the loss function with respect to this particular student. The update rule for the ith student vector with learning rate η is

$$\boldsymbol{J}_i^{\mu+1} - \boldsymbol{J}_i^\mu = \frac{\eta}{N} \delta_i^\mu \boldsymbol{\xi}^\mu, \tag{2}$$

$$\delta_i^\mu = \frac{1}{K} g'\big(\boldsymbol{J}_i^\mu \cdot \boldsymbol{\xi}^\mu\big) \left[\sum_{m=1}^{M} g(\boldsymbol{B}_m \cdot \boldsymbol{\xi}^\mu) - \frac{M}{K} \sum_{k=1}^{K} g\big(\boldsymbol{J}_k^\mu \cdot \boldsymbol{\xi}^\mu\big) \right]. \tag{3}$$

In order to analyze the learning behavior of the student, we need to define the generalization error ϵ_g, which is an average over all possible input vectors $\boldsymbol{\xi}$.

$$\epsilon_g(\boldsymbol{J}^\mu) = \big\langle \tfrac{1}{2}[\sigma(\boldsymbol{J}^\mu, \boldsymbol{\xi}^\mu) - \zeta(\boldsymbol{B}, \boldsymbol{\xi}^\mu)]^2 \big\rangle_{\{\boldsymbol{\xi}\}}. \tag{4}$$

In order to compute this average in a statistical mechanics approach, we introduce order parameter matrices $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{T}$ with elements $R_{in}^\mu = \boldsymbol{J}_i^\mu \cdot \boldsymbol{B}_n$, $Q_{ik}^\mu = \boldsymbol{J}_i^\mu \cdot \boldsymbol{J}_k^\mu$, $T_{mn} = \boldsymbol{B}_m \cdot \boldsymbol{B}_n = \delta_{mn}$, with $i, k = 1, \ldots, K$ and $m, n = 1, \ldots, M$. They are the covariances of a multivariate Gaussian distribution, over which the averaging simplifies considerably [10,24]. The generalization error thus becomes a function of these order parameters $\epsilon_g(\boldsymbol{R}^\mu, \boldsymbol{Q}^\mu)$. As we are interested in the typical behavior of the dynamics of our network, we take the thermodynamic limit $N \to \infty$, $p \to \infty$, where $\alpha = p/N$ stays finite [19–21]. Thus, we obtain a time evolution in the continuous time variable α for the student vector, and as a consequence, the dynamics of $\boldsymbol{R}(\alpha)$ and $\boldsymbol{Q}(\alpha)$ is given by [19]

$$\frac{d}{d\alpha}\boldsymbol{R} = \eta \boldsymbol{F}(\boldsymbol{R}, \boldsymbol{Q}), \tag{5a}$$

$$\frac{d}{d\alpha}\boldsymbol{Q} = \eta \boldsymbol{G}(\boldsymbol{R}, \boldsymbol{Q}) + \eta^2 \boldsymbol{H}(\boldsymbol{R}, \boldsymbol{Q}), \tag{5b}$$

where the detailed expressions for $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{H}$ are given in our Supplemental Material [39]. The graph of the generalization error of soft committee machines has a characteristic shape [19], as can be seen in Fig. 1. The first structure of interest is the plateau of the generalization error, which corresponds to a plateau in the dynamics of $\boldsymbol{R}$ and $\boldsymbol{Q}$ [17,24]. For this part of the time evolution the analytic ansatz $R_{in} \equiv R$, $Q_{ik} \equiv Q$ can be made, because all student vectors are found to have the same overlap with each other and with all the

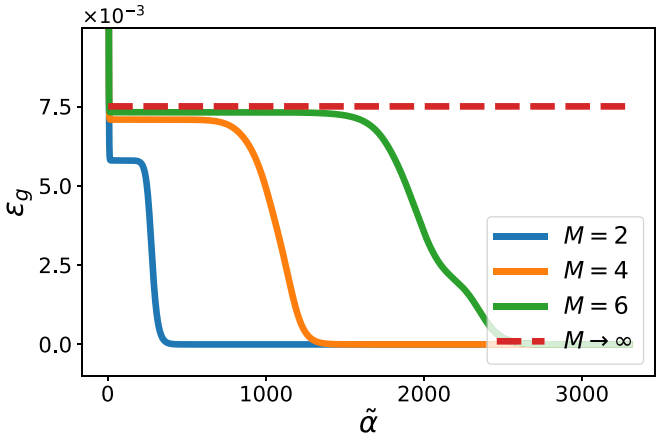


FIG. 1. Dependence of the generalization error in the realizable scenario, $K = M$, on the normalized number of examples $\tilde{\alpha} = \alpha\eta$ ("time"), computed to $O(\eta)$. Different colors indicate different widths M of the teacher. The height of the plateau converges monotonously with M. The initialization of the ordinary differential equations is given by $\chi^2 = 10^{-4}$, where χ^2 is the variance of initial student vector elements.

teacher vectors. In the small η limit, i.e., neglecting η^2 terms, the equations for R and Q in this symmetric regime are

$$\frac{dR}{d\alpha} = \frac{2\eta}{\pi} \frac{1}{1+Q} \frac{1}{K} \left\{ \frac{1 + Q - MR^2}{\sqrt{2(1+Q) - R^2}} - \frac{MR}{\sqrt{1+2Q}} \right\}, \tag{6a}$$

$$\frac{dQ}{d\alpha} = \frac{4\eta}{\pi} \frac{1}{1+Q} \frac{1}{K} \left\{ \frac{MR}{\sqrt{2(1+Q) - R^2}} - \frac{MQ}{\sqrt{1+2Q}} \right\}. \tag{6b}$$

These equations show that with a proper normalization of student and teacher output, we achieve an effective scaling of the learning rate with $1/K$, which leads to interesting further results. The fixed points of (6) are $R^* = [M(2M-1)]^{-1/2}$, $Q^* = (2M-1)^{-1}$, the same as in the realizable case [24]. But the length and height of the plateau in ϵ_g are interesting: The escape time of the symmetric plateau is proportional to the learning rate [22], leading to a prolongation of the plateau by a factor of K in the over-realizable case. The height of the generalization error in the plateau region is given by

$$\epsilon_g^* = \frac{1}{6} - \frac{1}{\pi} M \arcsin\left(\frac{1}{2M}\right), \tag{7}$$

which converges to $\lim_{M\to\infty} \epsilon_g^* = (1/6 - 1/2\pi) \approx 0.0075$ in the limit of an infinitely wide teacher network, independent of the number of student hidden nodes K. These analytical results agree well with numerical simulations: Fig. 1 shows the increasing plateau height for increasing dimensions M of the teacher, which converges to the analytically derived value. In addition, the value of ϵ_g^* does not depend on the dimension of the student K, as can be seen in Fig. 2, where one can also observe that the plateau length does depend on K in a linear fashion. We also investigated the correlation between plateau length and the initialization of $\boldsymbol{R}$. It turns out that the plateau length is inversely proportional to the logarithm of the variance taken for the initialization of the student-teacher overlaps (see also [22]). Further numerical details can be found in the Supplemental Material [39].

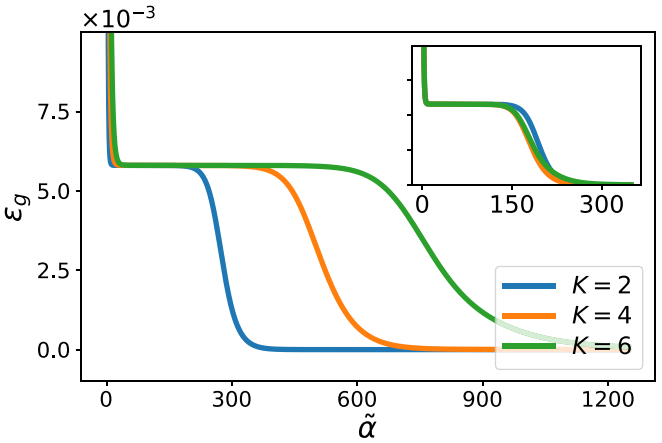


FIG. 2. Evolution of the generalization error in realizable and over-realizable scenarios, $K \geqslant M$, as a function of $\tilde{\alpha} = \alpha\eta$, computed to $O(\eta)$ for $M = 2$. The plateau height is independent of K, while the plateau length is proportional to K. The inset shows that a rescaling of the learning rate with K leads to a collapse of the curves with different K. The initialization of the ordinary differential equations is given by $\chi^2 = 10^{-4}$.

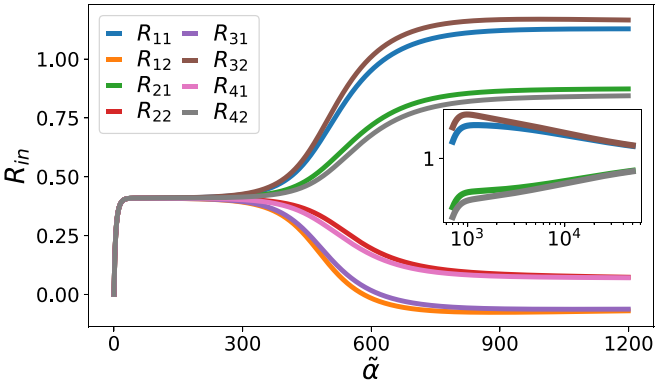


FIG. 3. Evolution of the student-teacher overlap R_{in} as a function of $\tilde{\alpha} = \alpha\eta$, computed to $O(\eta)$ for $M = 2$, $K = 4$. Students 1 and 2 imitate teacher 1 after the specialization transition, while students 3 and 4 imitate teacher 2, indicated by the R_{in} being close to one. The remaining overlaps tend to zero. The initialization of the ordinary differential equations is given by $\chi^2 = 10^{-4}$. The inset shows the asymptotic course of the upper of student-teacher overlaps displayed by a logarithmic x axis.

In the symmetric regime, all student vectors are found to behave similarly, in the sense that they have the same overlap with a given teacher vector. An exit from this region is achieved via the so-called specialization transition [17,19,24]. As the name indicates, the student vectors now start to imitate one particular teacher vector each. We focus on the over-realizable scenario ($K > M$), such that there are more student vectors than teacher vectors. For $Z := K/M \in \mathbb{N}$, our model, due to its particular normalization, allows all student vectors to learn, differing from the previous finding that $K - M$ student vectors are redundant and reduce their length to zero [24]. This behavior is reflected in the asymptotic regime of the generalization error, which we examine both numerically and analytically. Numerically, we find an algebraic convergence of the asymptotic evolution of both the order parameters $\boldsymbol{R}$, $\boldsymbol{Q}$ and the generalization error ϵ_g (see Figs. 3 and 4, respectively). Figure 3 shows the evolution of the overlaps R_{in} between student vector i and teacher vector n. One can observe that there is a plateau regime, where all the overlaps are equal. After the specialization transition the overlaps of the first two student vectors with the first teacher vector and the overlaps of the third and fourth student vectors with the second teacher vector are in the vicinity of one, while the remaining overlaps slowly approach zero (for more details and larger values of M and K see [39]). In principle, any student vector can imitate any given teacher vector, but due to the permutation symmetry between nodes we relabel the student vectors such that the first two imitate teacher vector one, and the remaining approach teacher vector two. Interestingly, the overlaps of two student vectors imitating the same teacher vector add up to one, which will be exploited in our analytical solution. Figure 4 shows the generalization error in the asymptotic regime for different scenarios. From the inset it can be seen that the exponential decay of the realizable case [17,24] is replaced by a slower, power-law rate of convergence in the over-realizable case, with $\epsilon_g \propto 1/\alpha^2$. We find the same behavior for $M = 3$ and $K = 6, 9, 12$ (see [39]). Interestingly, for noncommensurate values of M and K, the generalization error does not seem to approach zero but instead saturates at a finite value (see [39]).

Inspired by the numerical results for the asymptotic evolution of our dynamical system, we investigate the case $M = 2$, $K = 4$ using the ansatz $\boldsymbol{J}_1 = (1 + \lambda_1)\mathbf{B}_1 + \lambda_2\mathbf{B}_2$, $\boldsymbol{J}_2 = (1 + \lambda_3)\mathbf{B}_1 + \lambda_4\mathbf{B}_2$, with $\{\lambda_j\}_{j=1}^4$ being small parameters. Thus, the first two student vectors both try to imitate the first teacher vector, while the third and fourth student vectors attempt to imitate the second teacher vector, described by an analogous ansatz. We thus obtain a four-parameter ansatz for the

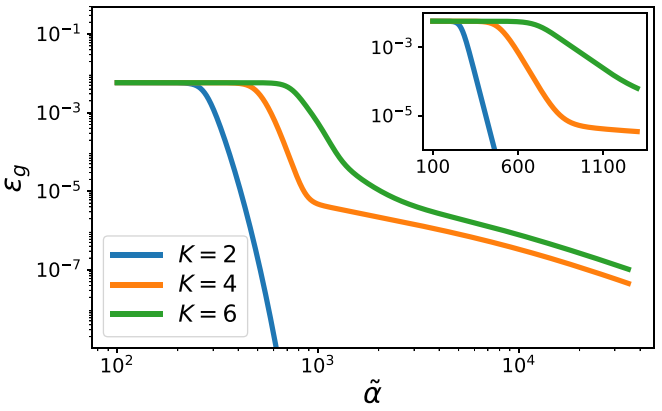


FIG. 4. Double logarithmic plot of the asymptotic generalization error as a function of $\tilde{\alpha} = \alpha\eta$, computed for $M = 2$ to $O(\eta)$, in both the realizable $K = M$ scenario and the over-realizable scenario $K \geqslant M$. The inset shows the evolution of the generalization error on a single logarithmic scale during the early stage of the specialization transition. While the generalization error approaches perfect learning with an exponential decay in the realizable case, one finds an exponentially decaying transient followed by a power law $\epsilon_g \propto \tilde{\alpha}^{-2}$ for $K > 2$. The initialization of the ordinary differential equations is given by $\chi^2 = 10^{-4}$.

asymptotic regime, with $\boldsymbol{R}$ and $\boldsymbol{Q}$ being of block matrix structure:

$$\boldsymbol{R} = \begin{pmatrix} 1+\lambda_1 & \lambda_2 \\ 1+\lambda_3 & \lambda_4 \\ \lambda_2 & 1+\lambda_1 \\ \lambda_4 & 1+\lambda_3 \end{pmatrix}, \quad \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \\ \boldsymbol{Q}_2 & \boldsymbol{Q}_1 \end{pmatrix}. \tag{8}$$

Due to the linear dependence of $\boldsymbol{R}$ on the parameters λ_j, we can express (5a) as dynamical equations in terms of a vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$, yielding $\frac{d}{d\alpha}\boldsymbol{\lambda} = \eta \boldsymbol{F}(\boldsymbol{\lambda})$. In order to gain insight into the dynamics of the system, we linearize $\eta \boldsymbol{F}(\boldsymbol{\lambda})$, obtaining $\frac{d}{d\alpha}\boldsymbol{\lambda} = \boldsymbol{L}\boldsymbol{\lambda}$. We find that the matrix $\boldsymbol{L}$ has eigenvalues $\{0, 0, l_+, l_-\}$, with $l_\pm = \frac{(\pm\sqrt{129-8\sqrt{3}})\eta}{72\pi} < 0$. In contrast to the realizable case studied earlier [17,24], where exponential convergence was found, we obtain a doubly degenerate eigenvalue zero, hinting at a slower mode of convergence. The two eigenvectors corresponding to these modes are found to be $v_1 = (0, -1, 0, 1)^T$ and $v_2 = (-1, 0, 1, 0)^T$.

In order to make further progress, we perform a coordinate transformation using the transpose of the eigenvector matrix. Exploiting the symmetry of $\boldsymbol{L}$, we introduce the transformation $\boldsymbol{\Lambda} = \boldsymbol{O}\boldsymbol{\lambda}$, which diagonalizes $\boldsymbol{L}$. The dynamical system in the new coordinates $\{\Lambda_j\}_{j=1}^4$ is then given by

$$\frac{d}{d\alpha}\boldsymbol{\Lambda} = \boldsymbol{O}F(\boldsymbol{O}^{-1}\boldsymbol{\Lambda}). \tag{9}$$

This is still a coupled system of ordinary differential equations (ODEs), which we expanded to third order in the parameters Λ_j. The system of equations we solved is of the form

$$\frac{d}{d\alpha}\begin{pmatrix}\Lambda_1 \\ \Lambda_2\end{pmatrix} = \boldsymbol{F}_1\left(\Lambda_1^3, \Lambda_1^2\Lambda_2, \Lambda_1\Lambda_2^2, \Lambda_2^3\right) + \boldsymbol{F}_2(\Lambda_1\Lambda_3, \Lambda_1\Lambda_4, \Lambda_2\Lambda_3, \Lambda_2\Lambda_4), \tag{10a}$$

$$\frac{d}{d\alpha}\begin{pmatrix}\Lambda_3 \\ \Lambda_4\end{pmatrix} = \begin{pmatrix} l_+\Lambda_3 \\ l_-\Lambda_4 \end{pmatrix} + \boldsymbol{F}_3\left(\Lambda_1^2, \Lambda_1\Lambda_2, \Lambda_2^2\right). \tag{10b}$$

As a first guess for solving these equations, we assume that the exponential convergence of the fast modes Λ_3, Λ_4 is dominant, such that we can set $\Lambda_3 = 0 = \Lambda_4$. Then, (10a) transforms into an equation of the form $\frac{d}{d\alpha}f = \hat{c}f^3$ for Λ_1, Λ_2, solved by $f = \frac{\hat{c}}{\sqrt{\alpha}}$. Inserting this ansatz for Λ_1, Λ_2 into (10b), we obtain the differential equation $\frac{d}{d\alpha}\Lambda_{3/4} = l_\pm\Lambda_{3/4} + \frac{\hat{c}_{3/4}}{\alpha}$, which is solved by $\Lambda_{3/4} = [\hat{c}_{3/4}\mathrm{Ei}(-l_\pm\alpha) + \tilde{c}_{3/4}]e^{l_\pm\alpha}$, where Ei is the standard exponential integral function. Studying the asymptotics of this function, one finds a convergence to zero as $\Lambda_{3/4} \propto \frac{1}{\alpha}$. To realize a self-consistent solution of (10), we at last introduce also the coupling term $\boldsymbol{F}_2$, which has only the effect of changing the constants in the solutions for Λ_1, Λ_2. Thus, we have found an asymptotic solution $\Lambda_{1/2} = \frac{\hat{c}_{1/2}}{\sqrt{\alpha}}$, $\Lambda_{3/4} = \frac{\hat{c}_{3/4}}{\alpha}$.

To obtain the solution in our original parameters λ_j we perform the coordinate backtransformation, yielding

$$\boldsymbol{\lambda} = \begin{pmatrix} \frac{c_1}{\alpha} + \frac{c_2}{\sqrt{\alpha}} \\ \frac{c_3}{\alpha} + \frac{c_4}{\sqrt{\alpha}} \\ \frac{c_1}{\alpha} - \frac{c_2}{\sqrt{\alpha}} \\ \frac{c_3}{\alpha} - \frac{c_4}{\sqrt{\alpha}} \end{pmatrix}. \tag{11}$$

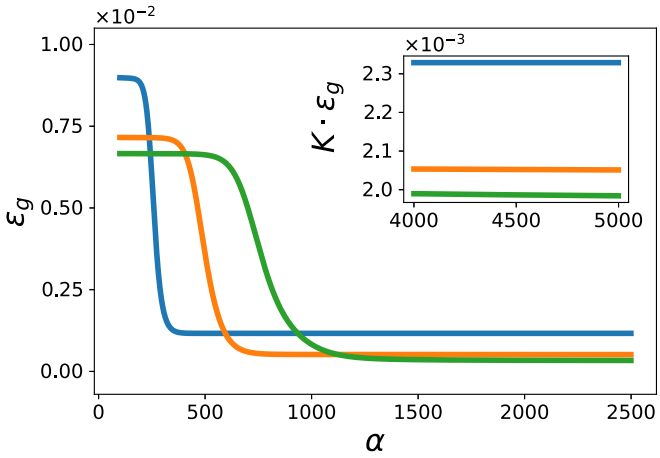


FIG. 5. Evolution of the generalization error in the case of a noisy teacher $M = 2$ and noise variance $\sigma_\gamma^2 = 0.01$ for $K \in \{2, 4, 6\}$ and $\eta = 1$. The asymptotic generalization error decreases with increasing K. The inset shows that the rescaled asymptotic generalization error $K\epsilon_g$ is only weakly dependent on K. The initialization of the ODEs is given by $\chi^2 = 10^{-4}$.

This result enables us to compare the analytical findings for the asymptotic behavior of $\boldsymbol{R}$ directly to the numerical findings, with excellent agreement. Inserting our solution into the $\boldsymbol{R}$ and $\boldsymbol{Q}$ matrices and determining the generalization error, we find $\epsilon_g \propto 1/\alpha^2$, in agreement with the numerical results discussed earlier.

So far, we have focused on the learning dynamics in the over-realizable regime as compared to the realizable one. However, in the case of a noisy teacher it is interesting to also study how the asymptotic generalization error depends on the degree of overparametrization. Specifically, we consider a noisy teacher with output $\zeta_\gamma(\boldsymbol{B}, \boldsymbol{\xi}) = \zeta(\boldsymbol{B}, \boldsymbol{\xi}) + \gamma$, where random noise $\gamma \sim \mathcal{N}(0, \sigma_\gamma^2)$ is added to the output of the teacher network. The dependence of the asymptotic generalization error on K is shown in Fig. 5. It is consistent with the scaling $\epsilon_g^\infty \propto \frac{1}{K}$ found in [21] for the case in which both layers of the network were fully trained. We thus demonstrate that learning of the second layer is not required for an improved asymptotic generalization error if a suitable normalization is applied to student and teacher outputs.

In conclusion, we have provided evidence that using an appropriate normalization of teacher and student outputs in a two-layer soft committee machine, all student nodes can learn in the over-realizable regime for K being an integer multiple of M. The generalization error stays finite in the limit of a large number of hidden units, and the value of the symmetric plateau is independent of the number of student nodes. Following the specialization transition, groups of students will imitate one teacher node, with a power-law approach to perfect learning due to the presence of soft modes. In the case of a noisy teacher, the asymptotic generalization error is lower in the over-realizable regime as compared to the realizable case.

We would like to thank M. Biehl for an interesting discussion. This work was supported by the IMPRS MiS Leipzig.

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, 2016).

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, in Advances in Neural Information Processing Systems, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Red Hook, NY, 2012), Vol. 25, p. 1097.

[4] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, Mastering the game of Go without human knowledge, Nature (London) **550**, 354 (2017).

[5] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborova, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[6] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (John Wiley and Sons, Inc., New York, 1949).

[7] John A. Hertz, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading MA, 1991).

[8] T. L. H. Watkin, A. Rau, and M. Biehl, The statistical mechanics of learning a rule, Rev. Mod. Phys. **65**, 499 (1993).

[9] D. Saad, *On-Line Learning in Neural Networks* (Cambridge University Press, New York, 1998).

[10] A. Engel and C. van den Broeck, *The Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, UK, 2001).

[11] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annu. Rev. Condens. Matter Phys. **11**, 501 (2020).

[12] E. Gardner, The space of interactions in neural network models, J. Phys. A: Math. Gen. **21**, 257 (1988).

[13] G. Gyorgyi and N. Tishby, *Statistical Theory of Learning a Rule*, in Workshop on Neural Networks and Spin Glasses, edited by K. Theumann and W. K. Koeberle (World Scientific, Singapore, 1990), pp. 3–36.

[14] M. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Phys. Rev. A **45**, 6056 (1992).

[15] H. Schwarze and J. Hertz, Generalization in fully connected committee machines, Europhys. Lett. **21**, 785 (1993).

[16] M. Opper, Learning and Generalization in a Two-Layer Neural Network: The Role of the Vapnik-Chervonenkis Dimension, Phys. Rev. Lett. **72**, 2113 (1994).

[17] M. Biehl and H. Schwarze, Learning by online gradient descent, J. Phys. A: Math. Gen. **28**, 643 (1995).

[18] P. Riegler and M. Biehl, On-line backpropagation in two-layered neural networks, J. Phys. A: Math. Gen. **28**, L507 (1995).

[19] D. Saad and S. A. Solla, Exact Solution for On-Line Learning in Multilayer Neural Networks, Phys. Rev. Lett. **74**, 4337 (1995).

[20] G. Reents and R. Urbanczik, Self-Averaging and On-Line Learning, Phys. Rev. Lett. **80**, 5445 (1998).

[21] S. Goldt, M. Advani, A. M. Sace, F. Krzakala, and L. Zdeborová, Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), p. 6981.

[22] M. Biehl, P. Riegler, and C. Wöhler, Transient dynamics of online learning in two-layered neural networks, J. Phys. A: Math. Gen. **29**, 4769 (1996).

[23] M. Biehl, E. Schlösser, and M. Ahr, Phase transitions in soft-committee machines, Europhys. Lett. **44**, 261 (1998).

[24] D. Saad and S. A. Solla, On-line learning in soft committee machines, Phys. Rev. E **52**, 4225 (1995).

[25] Y. Yoshida and M. Okada, Data-dependence of plateau phenomenon in learning with neural network—Statistical mechanical analysis, J. Stat. Mech. (2020) 124013.

[26] S. Goldt, M. Mezard, F. Krzakala, and L. Zdeborova, Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model, Phys. Rev. X **10**, 041044 (2020).

[27] S. Mei, A. Montanari, and P. Nguyen, A mean field view of the landscape of two-layer neural networks, Proc. Natl. Acad. Sci. USA **115**, E7665 (2018).

[28] L. Chizat and F. Bach, *On the Global Convergence of Gradient Descent for Over-Parameterized Models Using Optimal Transport*, in Advances in Neural Information Processing Systems (Curran Associates, Inc., 2018), Vol. 31, pp. 3040–3050.

[29] E. Oostwal, M. Straat, and M. Biehl, Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation, Physica A **564**, 125517 (2021).

[30] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, in Proceedings of International Conference on Learning Representations (Association for Computing Machinery, 2017).

[31] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: Convergence and Generalization in Neural Networks*, in NIPS '18: Proceedings of the 32nd International Conference on Neural Information Processing Systems (Curran Associates, Inc., 2018), p. 8580.

[32] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang, On exact computation with an infinitely wide neural net, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), Vol. 32.

[33] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off, Proc. Natl. Acad. Sci. USA **116**, 15849 (2019).

[34] Y. Cao and Q. Gu, *Generalization Error Bounds of Gradient Descent for Learning Over-Parameterized Deep ReLU Networks*, in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI Press, Palo Alto, California USA, 2020), Vol. 34, p. 3349.

[35] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, *Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks*, in Proceedings of the 36th International Conference on Machine Learning (PMLR) (Proceedings of Machine Learning Research (PMLR), 2019), Vol. 97, p. 322.

[36] Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang, A Generalized Neural Tangent Kernel Analysis for Two-Layer Neural Networks, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2020), Vol 33, pp. 13363–13373.

[37] O. Cohen, O. Malka, and Z. Ringel, Learning curves for overparametrized deep neural networks: A field theory perspective, Phys. Rev. Research **3**, 023034 (2021).

[38] G. Naveh, O. B. David, H. Sompolinsky, and Z. Ringel, Predicting the outputs of finite deep neural networks trained with noisy gradients, Phys. Rev. E **104**, 064301 (2021).

[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.105.L052302 for more details on the differential equations and further numerical results.