Letter

Asymptotic generalization errors in the online learning of random feature models

Roman Worschech [1,2] and Bernd Rosenow[2]

[1]*Max Planck Institute for Mathematics in the Sciences, D-04103, Leipzig, Germany*

[2]*Institut für Theoretische Physik, Universität Leipzig, Brüderstrasse 16, 04103 Leipzig, Germany*

(Received 12 September 2023; accepted 29 April 2024; published 3 June 2024)

Deep neural networks are widely used prediction algorithms whose performance often improves as the number of weights increases, leading to over-parametrization. We consider a two-layered neural network whose first layer is frozen while the last layer is trainable, known as the random feature model. We study over-parametrization in the context of a student-teacher framework by deriving a set of differential equations for the learning dynamics. For any finite ratio of hidden layer size and input dimension, the student cannot generalize perfectly, and we compute the non-zero asymptotic generalization error. Only when the student's hidden layer size is exponentially larger than the input dimension, an approach to perfect generalization is possible.

DOI: 10.1103/PhysRevResearch.6.L022049

Deep neural networks are very versatile, with applications extending beyond typical areas like image classification and speech recognition into the domain of physics [1–5]. A captivating observation is that the performance of these networks often improves with an increasing number of weights. This results in networks with many free parameters, which notably surpass the available training data, thus unveiling intriguing properties [6–10]. When weights are initialized independently with zero mean and variance inversely proportional to the network width, the output in the infinite width limit is described by a Gaussian process, primarily defined by its covariance matrix [11–15].

The ultrawide limit, where the network size tends towards infinity, has emerged as a suitable starting point to study the intricacies of weight dynamics during training and their implication on the prediction performance. If the network output is scaled by the inverse square root of its width, then for training with gradient flow the network is in the neural tangent kernel (NTK) limit [16,17], known colloquially as the "lazy training" regime. A generalized NTK description is possible for training with noisy stochastic gradient descent with weight decay [18,19]. In the NTK limit, the weights remain close to their initial values throughout the learning, rendering weight dynamics akin to a linearized network [20–22]. This phenomenon offers a promising avenue for the exploration of generalization capabilities and convergence characteristics of highly over-parametrized neural networks.

This Letter aims to analytically describe the behavior of strongly lazy two-layer neural networks. Here, we keep the first layer weights constant, effectively mimicking their lazy nature, while focusing on training the weights from the hidden layer to the output. Such an architecture is known as random feature model [23]. Using statistical mechanics, we study the learning dynamics of highly over-parametrized random feature models, endowed with large hidden and input layers. Training occurs via one-pass stochastic gradient descent, which defines the system dynamics. We contextualize over-parametrization through the student-teacher framework, where the student learns from the teacher's outputs [24]. In this framework, the number of student hidden nodes, denoted as K, surpasses the number of teacher hidden nodes, M. Given the many degrees of freedom in these networks, statistical mechanics becomes essential in deriving macroscopic observables from the interplay of network weights [25–29]. Historically, this approach has proven fruitful across diverse network architectures [19,30–42].

Earlier studies have highlighted that the random feature model struggles to accurately predict the output of a rectified linear unit (ReLU) perceptron unless the hidden layer size K grows exponentially with the input dimension N [43]. Here, we report similar behavior for an error function activation through direct calculations of the asymptotic generalization error as a function of K/N. In Ref. [44], the generalization error for random feature models was analyzed in detail, relating it to the projection of the target function onto the span of kernel eigenfunctions, which remains unexplored in the context of a finite K/N ratio. In this work, we present an explicit formula that shows how the generalization error depends on K/N specifically for error function and ReLU activation. Importantly, we establish that the student's learning behavior is mainly determined by its initial weight configurations, achieving optimal performance only when the random features perfectly match the teacher's features. Other pertinent research indicates that when K scales linearly with N, the student is restricted to learning solely a linear approximation of the teacher [45,46]. Such findings emphasize a discernible performance disparity between the random feature

model and fully trained over-parametrized neural networks. Furthermore, we present a framework using ordinary differential equations to track the learning path of the student under infinitely large input and hidden layers, employing either a ReLU or error activation function. Our analysis reveals that the generalization error as a function of the number of training examples saturates at a certain plateau value as long as K/N remains finite. We determine the dependence of these plateaus on K/N for different activation functions, eliminating the need for error bounds.

For our setup, the student and teacher are presented input samples $\boldsymbol{\xi}^\mu \in \mathbb{R}^N$ once and in a sequential order, where each component is generated by the normal distribution $\xi_i^\mu \in \mathcal{N}(0, 1)$ with $\mu \in \{1, ..., p\}$. The student has K hidden neurons and the connection between the input layer with the ith hidden node is expressed in terms of the student vectors $\boldsymbol{J}_i \in \mathbb{R}^N$. The outputs of the hidden nodes are modulated by a continuous activation function g. The overall output of the network is a linear combination of the outputs of the hidden units

$$\sigma(\boldsymbol{J}, \boldsymbol{\xi}) = \sum_{i=1}^{K} c_i g(x_i), \tag{1}$$

with $x_i = \frac{\boldsymbol{J}_i \cdot \boldsymbol{\xi}}{\sqrt{N}}$ and c_i being the hidden-to-output weights. Note that the rescaling factor $\frac{1}{\sqrt{N}}$ guarantees preactivations of $\mathcal{O}(1)$. The teacher has M hidden neurons characterized by the teacher vectors $\boldsymbol{B}_n \in \mathbb{R}^N$ and provides the output $\zeta(\boldsymbol{B}, \boldsymbol{\xi}) = \frac{1}{\sqrt{M}} \sum_{n=1}^{M} g(y_n)$ with $y_n = \frac{\boldsymbol{B}_n \cdot \boldsymbol{\xi}}{\sqrt{N}}$. However, as shown in the Supplemental Material [47], it turns out that increasing the teacher size M does not influence the learning process since scaling of the teacher output with $1/\sqrt{M}$ leads to consistent statistical characteristics of the generalization error. We therefore consider a teacher perceptron with $M = 1$.

We choose the student and teacher vectors from the uniform distribution over the N sphere and initialize the output weights of the student by the normal distribution with $c_i \in \mathcal{N}(0, \frac{1}{K})$. In order to express the similarity of the student and teacher vectors, we introduce the correlation matrices $R_{in} = \frac{\boldsymbol{J}_i \cdot \boldsymbol{B}_n}{N}$, $Q_{ij} = \frac{\boldsymbol{J}_i \cdot \boldsymbol{J}_j}{N}$, and set $\frac{\boldsymbol{B}_n \cdot \boldsymbol{B}_m}{N} = \delta_{nm}$. For the case $K < N$, one refers to these parameters as order parameters and makes the transition from a microscopic to a macroscopic description of the system. However, in our setup such an interpretation is no longer true as we are mainly interested in the relation $K > N$. In this case, the number of potential order parameters is of order $\mathcal{O}(K^2)$ and surpasses the number of degrees of freedom in the system, which is of order $\mathcal{O}(KN)$. This contradicts the purpose of order parameters, which is to reduce the system's complexity.

The performance of the student with respect to the teacher is measured by the loss function $\epsilon = \frac{1}{2}[\zeta - \sigma]^2$ known as the mean-squared error. As the distribution of the input patterns is accessible, we can take the expectation value of the loss function and define the generalization error $\epsilon_g = \langle \epsilon(c_i, \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}}$ depending on the correlation matrices (cf. Supplemental Material). This makes it possible to analyze the typical error of the student evaluated on unseen test data. During the learning process, we update the student weights c_i via stochastic gradient descent after each representation of a specific input example

$$c_i^{\mu+1} - c_i^\mu = -\frac{\eta}{K} \nabla_{c_i} \epsilon\left(c_i^\mu, \xi^\mu\right), \tag{2}$$

with η denoting the learning rate which controls the step size in the weight space, and μ being a discrete time index for the input pattern in step μ. Commonly, one would rescale the learning rate by $\frac{1}{N}$ in order to study the dynamics of the learning process [28,32,33,48]. Since in our case the first layer is fixed and the K output weights are trained, we scale the learning rate by $\frac{1}{K}$ in Eq. (2) in order to guarantee small fluctuations for a large hidden layer size [49]. In the ultrawide limit, the parameters N, K, and p all tend to infinity, but with a finite ratio $\frac{p}{K} = \alpha$. We then find a Langevin equation for the student weights

$$\frac{d\boldsymbol{c}}{d\alpha} = -\eta \nabla \epsilon_g + \frac{\eta}{\sqrt{K}} \boldsymbol{\gamma} \ , \tag{3}$$

where $\boldsymbol{\gamma}$ is a random vector with $\langle \boldsymbol{\gamma} \rangle = 0$, $\langle \gamma_i(\alpha) \gamma_j(\alpha') \rangle = \Sigma_{ij} \delta(\alpha - \alpha')$ and covariance matrix $\boldsymbol{\Sigma} = \langle (\nabla \epsilon - \nabla \epsilon_g)(\nabla \epsilon - \nabla \epsilon_g)^T \rangle$. In order to make further assertions about the large-K limit, on the right-hand side of Eq. (3), we need to take a closer look at the variance of the trajectory. As the system size increases, one can replace the above stochastic Langevin equation by its mean trajectory leading to a deterministic differential equation if the fluctuations get negligible [50]. As a measure for the fluctuations, we consider the relative variance of the stochastic trajectory and find for its scaling

$$\frac{\left\langle \left(\frac{dc}{d\alpha}\right)^2 \right\rangle - \left\langle \frac{dc}{d\alpha} \right\rangle^2}{\left\langle \frac{dc}{d\alpha} \right\rangle^2} \propto \frac{N}{K}, \tag{4}$$

directly related to the fluctuations of the loss function as shown in the Supplemental Material. Thus, for small ratios $\frac{N}{K}$, the relative variance of the loss function becomes small and we approximate the stochastic Langevin equation with its mean. In statistical mechanics, such a scaling relation for the variance of a system's property is known as a self-averaging character. A rigorous treatment for the relationship between the Langevin equation and the stochastic gradient descent can be found in Ref. [49].

However, from Fig. 1 one sees that the generalization error is self-averaging already for $K/N = \mathcal{O}(1)$. This can be explained by the fact that ϵ_g is a weighted sum over terms that individually fluctuate, due to their dependency on c_i, Q_{ij}, and R_i. The self-averaging nature of the generalization error is due to the scaling of its variance with the inverse system size, combined with a non-zero expectation value as shown in the Supplemental Material. Hence, we obtain for the mean evolution of weights

$$\left\langle \frac{d\boldsymbol{c}}{d\alpha} \right\rangle = -\eta[\tilde{\boldsymbol{Q}}\boldsymbol{c} - \tilde{\boldsymbol{R}}], \tag{5}$$

with $\tilde{Q}_{ij} = \langle g(x_i) g(x_j) \rangle_\xi$ and $\tilde{R}_i = \langle g(x_i) g(y) \rangle_\xi$ depending on the choice of the activation function. Thus, we obtain a set of deterministic differential equations by Eq. (5) characterizing the dynamical behavior of the learning process valid for arbitrary ratios $\frac{K}{N}$. The fixed point of Eq. (5) determines the asymptotic solution for the weights $\boldsymbol{c}^* = \tilde{\boldsymbol{Q}}^{-1} \tilde{\boldsymbol{R}}$, and allows

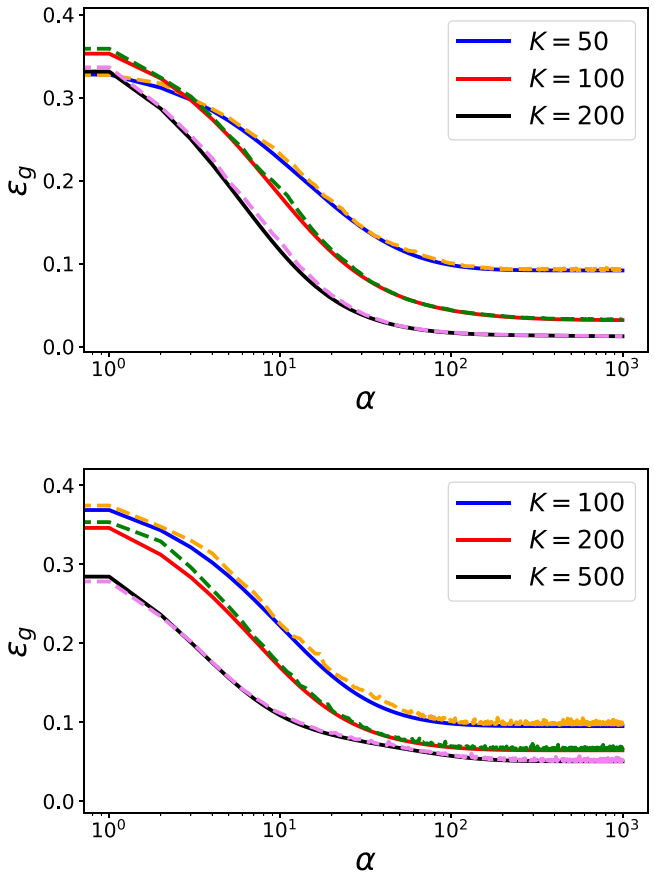

FIG. 1. Generalization error as a function of α for an error function activation $g(x) = \mathrm{erf}(\frac{x}{\sqrt{2}})$ (top) and ReLU activation (bottom) for $N = 100$. The numerical solutions of the differential equations (solid lines) fit well to the simulations. We use for the simulations $\eta_{\mathrm{simulation}} = \frac{0.1}{K}$ (dashed lines) corresponding to the rescaling of the learning rate in Eq. (2), and for the solutions of the differential equations given by Eq. (5) $\eta = 0.1$. For both methods, we set the same values for $\boldsymbol{R}$, $\boldsymbol{Q}$ and use the same initial values of $\boldsymbol{c}_0 \in \mathcal{N}(0, \frac{1}{K})$.

obtaining the generalization error as

$$\epsilon_g^* = \frac{\langle \zeta(\boldsymbol{B}, \boldsymbol{\xi})^2 \rangle_\xi}{2} - \frac{1}{2} \tilde{\boldsymbol{R}}^T \tilde{\boldsymbol{Q}}^{-1} \tilde{\boldsymbol{R}} \,. \tag{6}$$

As one can show, the asymptotic solution corresponds to the minimal generalization error, implying that the error will be minimized after a long training period. Figure 1 compares the generalization error of the random feature model for simulations according to the update rule given by Eq. (2) with the solutions of Eq. (5). We find excellent agreement between the two.

In order to analyze how the asymptotic generalization error depends on the student size K, we consider a finite hidden-to-input layer size ratio $K = \beta N$ with $\beta \in (1, \infty)$. In the ultrawide limit, we can linearize $\tilde{\boldsymbol{Q}}$ and $\tilde{\boldsymbol{R}}$ to first order in R_i and Q_{ij} since large overlaps are small for a finite β. This assumption is based on the curse of dimensionality where we consider K random N-dimensional student vectors leading to small overlaps of $O(1/\sqrt{N})$. Furthermore, the generalization error depends on the distribution of the student and teacher vectors and we therefore analyze the dynamics of the typical asymptotic generalization by taking the expectation value of Eq. (6) in this linearized regime.

First, we use an error function activation $g(x) = \mathrm{erf}(\frac{x}{\sqrt{2}})$ and discuss the asymptotic generalization error $\epsilon_{\mathrm{erf}}^* = \frac{1}{2\pi}(\frac{\pi}{3} - \boldsymbol{R}^T \boldsymbol{S}^{-1} \boldsymbol{R})$ in the linearized regime, where $S_{ii} = \frac{\pi}{3}$ and $S_{ij} = Q_{ij}$. Thus, we take the expectation value of Eq. (6) in the linearized regime and obtain for its random part

$$\langle \boldsymbol{R}^T \boldsymbol{S}^{-1} \boldsymbol{R} \rangle_{\boldsymbol{J},\boldsymbol{B}} = \frac{1}{N} \sum_i^K \left\langle \frac{\lambda_i}{\frac{\pi}{3} - 1 + \lambda_i} \right\rangle_\lambda , \tag{7}$$

with λ_i as the ith eigenvalue of $\boldsymbol{Q}$. Thereby, we have exploited $\langle B_a B_b \rangle = \delta_{ab}$. Since $\boldsymbol{J}$ is a $K \times N$ random matrix whose entries have zero mean and bounded variance, the eigenvalues of the correlation matrix for $N \to \infty$ are distributed according to the Marčenko-Pastur distribution [51]. For ratios $\frac{K}{N} > 1$, the Marčenko-Pastur distribution consists of two parts: the first $K - N$ eigenvalues are zero and just N eigenvalues contribute to the sum in Eq. (7). After evaluating the expectation value over the eigenvalue distribution for the remaining sum, we obtain

$$\lim_{\substack{K,N\to\infty \\ \frac{N}{K}=\mathrm{const.}}} \frac{1}{N} \sum_i^K \left\langle \frac{\lambda_i}{\frac{\pi}{3} - 1 + \lambda_i} \right\rangle = \frac{1}{2} \left[\frac{K}{N} + \frac{\pi}{3} - \sqrt{\left(\frac{\pi - 3}{3} + \left(\frac{K}{N} + 1 \right) \right)^2 - \frac{4K}{N}} \right]. \tag{8}$$

For large ratios, we find that $\lim_{\frac{K}{N}\to\infty} \langle \boldsymbol{R}^T \boldsymbol{S}^{-1} \boldsymbol{R} \rangle = 1$ leading to a limiting value of the asymptotic generalization error in the linearized regime $\lim_{\frac{N}{K}\to 0} \lim_{K,N\to\infty} \langle \epsilon_{g_{\mathrm{erf}}}^* \rangle = \frac{1}{2\pi}(\frac{\pi}{3} - 1) \approx 0.007512$.

Second, we evaluate the asymptotic generalization error Eq. (6) for the ReLU activation function in the linearized regime and obtain $\epsilon_{\mathrm{ReLU}}^* = \frac{1}{4} - \frac{1}{2} \hat{\boldsymbol{R}}^T \hat{\boldsymbol{Q}}^{-1} \hat{\boldsymbol{R}}$ with $\hat{Q}_{ii} = \frac{1}{2}$, $\hat{Q}_{ij} = \frac{Q_{ij}}{4} + \frac{1}{2\pi}$ and $\hat{R}_i = \frac{1}{4} R_i + \frac{1}{2\pi}$. Similar to the case of the error function, we take the expectation value of its random part

$$\begin{aligned} \langle \hat{\boldsymbol{R}}^T \hat{\boldsymbol{Q}}^{-1} \hat{\boldsymbol{R}} \rangle = {} & \frac{1}{16} \langle \boldsymbol{R}^T \hat{\boldsymbol{Q}}^{-1} \boldsymbol{R} \rangle \\ & + \frac{1}{8\pi} [\langle \boldsymbol{R}^T \boldsymbol{T} \rangle + \langle \boldsymbol{T}^T \boldsymbol{R} \rangle] \\ & + \frac{1}{4\pi^2} \sum_{ij} (\hat{\boldsymbol{Q}}^{-1})_{ij}, \end{aligned} \tag{9}$$

where $\boldsymbol{T}^T$ is a vector containing the sum of the columns of $\hat{\boldsymbol{Q}}^{-1}$, i.e. $T_i = \sum_j (\hat{\boldsymbol{Q}}^{-1})_{ij}$. In the Supplemental Material, we use estimates for large systems and find

$$\lim_{\substack{K,N\to\infty \\ \frac{N}{K}=\mathrm{const.}}} \langle \hat{\boldsymbol{R}}^T \hat{\boldsymbol{Q}}^{-1} \hat{\boldsymbol{R}} \rangle = \frac{1}{2\pi} + \frac{1}{8} \left[\frac{K}{N} + \frac{2(\pi - 1)}{\pi} - \sqrt{\left(\frac{\pi - 2}{\pi} + \left(\frac{K}{N} + 1 \right) \right)^2 - \frac{4K}{N}} \right]. \tag{10}$$

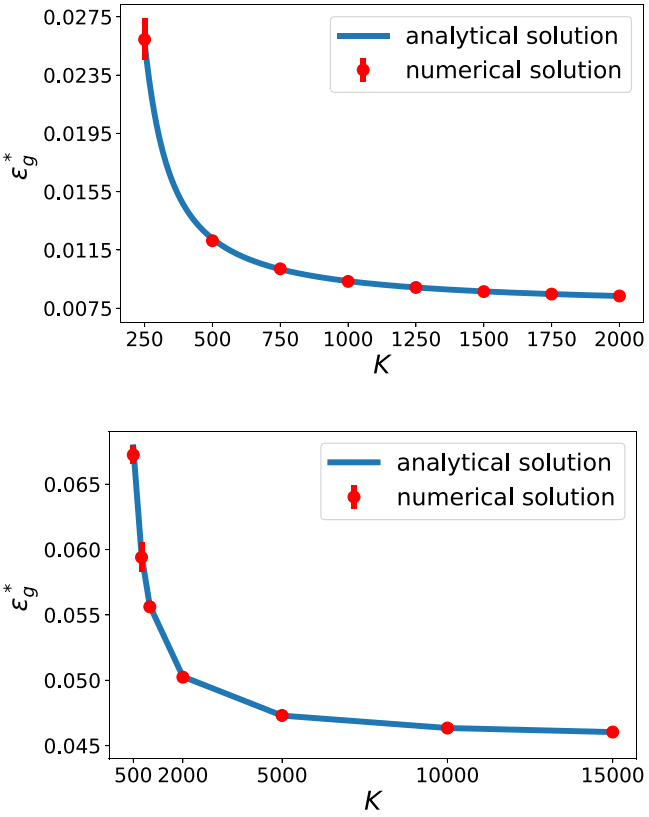

FIG. 2. Asymptotic generalization error ϵ_g^* as a function of K for $M = 1$ and $N = 200$ for an error function as the activation (top) and ReLU activation (bottom). The numerical solution of Eq. (6) is obtained from ten initializations of random matrices $\boldsymbol{B}$ and $\boldsymbol{Q}$ under the linearized setup. The error bars show the standard deviation of these averages. The analytical solution is based on Equation (8) for the error function and on Eq. (10) for the ReLU activation.

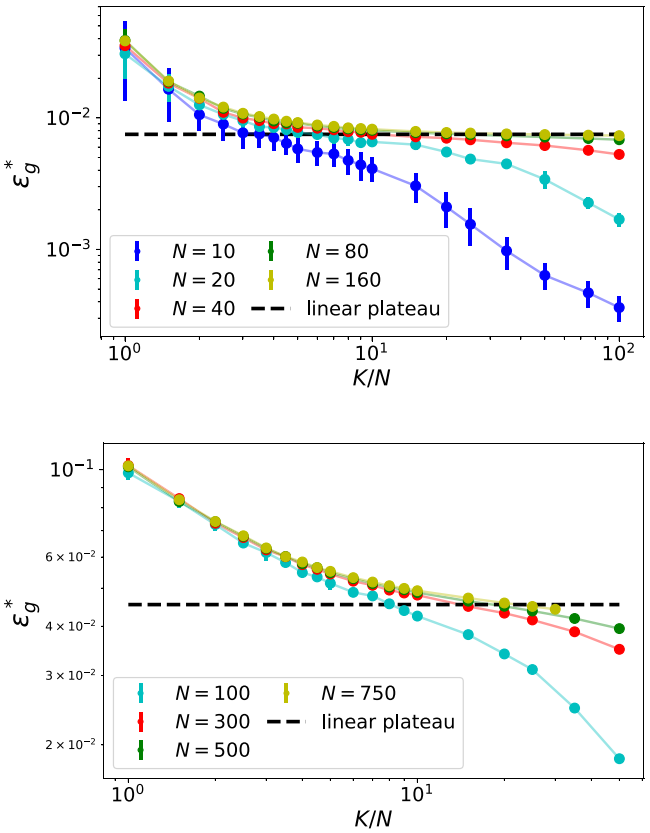

FIG. 3. Asymptotic generalization error as a function of $\frac{K}{N}$ for an error function activation $g(x) = \mathrm{erf}(\frac{x}{\sqrt{2}})$ (top) and ReLU (bottom). Here, we evaluate Eq. (6) numerically and compute the expectation value of the generalization error for a given $\frac{K}{N}$. The error bars show the standard deviation of the average over 100 initializations for $N = 10$ and 10 initializations for $N > 10$.

For the corresponding asymptotic generalization error for large ratios $\frac{K}{N}$, we obtain $\lim_{\frac{N}{K}\to 0} \lim_{K,N\to\infty} \langle \epsilon_g^* \rangle_{\mathrm{ReLU}} = \frac{1}{4}(\frac{1}{2} - \frac{1}{\pi}) \approx 0.0454$. Figure 2 shows the asymptotic generalization error evaluated using Eqs. (8) and (10) together with the corresponding numerical solution of Eq. (6) in the linearized setting.

Figure 3 displays the numerical solution of Eq. (6) for both activation functions not restricted to the linearized setting and shows how the linearized plateau is reached as a function of the ratio $\frac{K}{N}$ for different input dimensions N. All curves are monotonically decreasing with $\frac{K}{N}$. For small N, the generalization error stays near the plateau value only for a limited range of $\frac{K}{N}$, while for larger N the generalization error is close to the plateau value even for large values of K/N.

A more detailed numerical investigation of Eq. (6) for $K, N \to \infty$ while maintaining a fixed $\frac{N}{K}$ ratio, reveals that the leading finite-N correction to the asymptotic generalization has the form

$$\epsilon_g^* = \epsilon_g^\infty + \frac{b}{N} \quad . \tag{11}$$

Here, ϵ_g^∞ signifies the asymptotic value of the generalization error as N approaches infinity, and b is a regression parameter. Furthermore, if we plot ϵ_g^∞ against $\frac{N}{K}$, we obtain again a linear dependence as shown in Fig. 4, in agreement with our analytical solution Eq. (8). These linear dependencies are validated through finite size scaling for different $\frac{N}{K}$ ratios as presented in the Supplemental Material. After extracting these asymptotic values ϵ_g^∞, we study the limit where $\frac{N}{K}$ approaches zero for large K, N and extrapolate ϵ_g^∞ for $\frac{N}{K} \to 0$. Our numerical analysis yields $\lim_{\frac{N}{K}\to 0} \epsilon_{\mathrm{erf}}^\infty = 0.007512 \pm 2 \times 10^{-6}$ and $\lim_{\frac{N}{K}\to 0} \epsilon_{\mathrm{ReLU}}^\infty = 0.0453 \pm 2 \times 10^{-4}$ for the error function and ReLU activation, respectively. Thus, our numerical results evaluated outside the linearized regime for $\frac{K}{N} \to \infty$ and then $\frac{N}{K} \to 0$ are in excellent agreement with our analytical predictions found within the linearized regime for $\frac{N}{K} \to 0$.

Figure 3 shows that for small N the asymptotic generalization error decreases below the plateau value towards zero as a function of K/N. The reason for this behavior lies in the probability distribution of the overlaps R_i. As K increases with fixed N, the probability of selecting a student vector closely aligned with the teacher vector grows and the remaining problem is to determine the scaling of K with N to obtain a small asymptotic generalization error. Moreover, the more student vectors are approximately in the direction of the teacher vector, the lower the generalization error, making a linearization of $\tilde{\boldsymbol{R}}$ and $\tilde{\boldsymbol{Q}}$ infeasible. If even one student vector

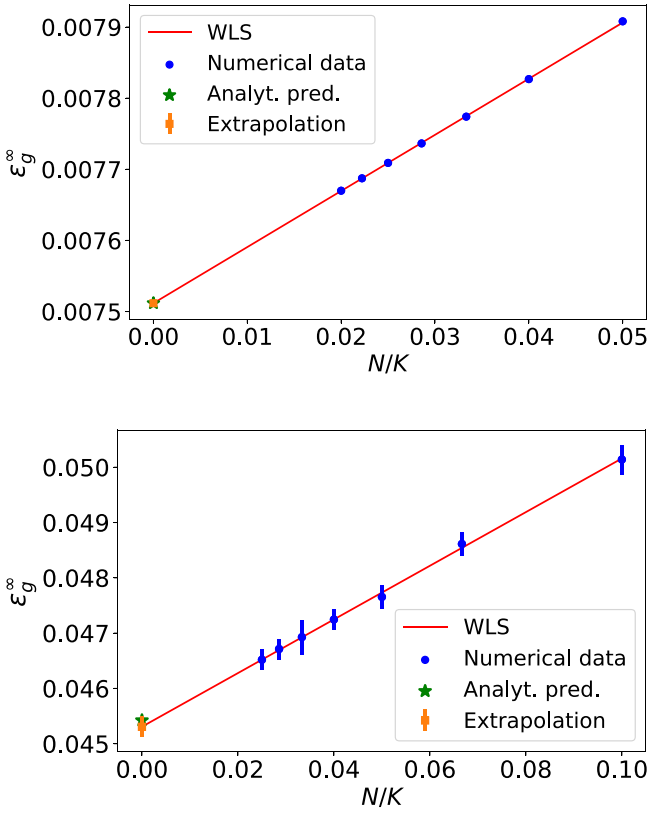

FIG. 4. Extrapolated asymptotic generalization error ϵ_g^∞ as a function of $\frac{N}{K}$ for error function (top) and ReLU activation (bottom). For each $\frac{N}{K}$ ratio, we use at least 100 random matrix initializations and utilize Eq. (11) for a numerical evaluation, determining ϵ_g^∞. Error bars in the graph represent the uncertainties calculated through error propagation in linear regression for $N \to \infty$. For the error function the error bars are smaller than the symbols. The red curve shows a weighted least squares (WLS) on ϵ_g^∞ in order to perform a second extrapolation for $\frac{N}{K} \to 0$ (orange) and compare it with the analytical prediction (green).

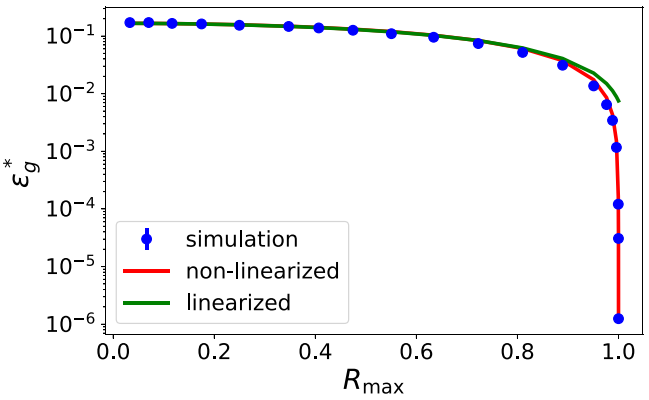

FIG. 5. Asymptotic generalization error as a function of $R_{\max} = \max_{1\leqslant i\leqslant K} R \in \{R_i\}$ for $K = 7, M = 1, N = 5$ and $\eta = \frac{0.1}{K}$ for an error function as the activation $g(x) = \mathrm{erf}(\frac{x}{\sqrt{2}})$. We initialize the student and set the teacher in such a way that the first component of $\boldsymbol{R}$ is the largest one and the others are small and of similar size. The blue curve shows the corresponding simulation. The orange and green curves show the solution of Eq. (6) for a perceptron and a linearized perceptron, respectively. For the simulation, we averaged the generalization error over a predefined interval to get its asymptotic value. The error bars are smaller than the symbol sizes.

has a high overlap, the generalization error already decreases towards zero as $R_i \to 1$, cf. Fig. 5.

Therefore, we ask for the probability of finding at least one large overlap $\max\{R_i\} > R^*$ after the initialization of K student vectors for a given N and threshold R^*. We use the relation

$$P(\max\{R_i\} > R^*; N, K) = 1 - F(R^*; N)^K, \qquad (12)$$

where $F(R^*; N) = \Pr(R \leqslant R^*; N) = \int_0^{R^*} dR_i \rho(R_i)$ is the cumulative probability to find $R \leqslant R^*$ after randomly drawing a student vector, obtained by integrating over a density function $\rho(R)$. Since the student and teacher vectors are drawn from a uniform distribution over the N sphere, the shifted overlaps $\frac{R_i+1}{2}$ are generated by the beta distribution. For the error function activation, we obtain the density $\rho(y) = |\cos(y)|\beta(\frac{2\sin(y)+1}{2}; a, b)$, where $\beta(t; a, b) = \frac{1}{B(a,b)} t^{a-1}(1-t)^{b-1}$ is the beta density function with normalization constant $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$ and $a = b = \frac{N-1}{2}$. We can now perform the integration for $R^* > R$, thus obtaining the probability for large overlaps between student and teacher vectors

$$\begin{aligned} F(R^*; N) &= 1 - \int_{\arcsin(\frac{R^*}{2})}^{\frac{\pi}{6}} \rho(y)\, dy \\ &= \frac{B\left(\frac{R^*+1}{2}, a, a\right)}{B(a, a)} \\ &= I_z(a, a)\,. \end{aligned} \qquad (13)$$

Here, $B(z, \alpha, \beta) = \int_0^z t^{\alpha-1}(1-t)^{\beta-1}dt$ is the incomplete beta function with $z = \frac{R^*+1}{2}$ for the case above and $I_z(a, b)$ is the regularized incomplete beta function. Therefore, our cumulative distribution function is related to the Binomial cumulative distribution function via the regularized incomplete beta function $1 - F(R^*; N) = F_{\text{Binomial}}(\frac{N-1}{2}; N-2, \frac{R^*+1}{2})$. For the case $R^* \lesssim 1$, one can estimate the tail of the Binomial distribution function by the Chernoff bound $F_{\text{Binomial}}(\frac{N-1}{2}; N-2, \frac{R^*+1}{2}) \geqslant \frac{1}{\sqrt{2N-4}} \exp[-(N-2)D(\frac{N-3}{2N-4} \| \frac{R*+1}{2})]$ with the Kullback-Leibler divergence $D(P\|Q) = P\ln(\frac{P}{Q}) + (1-P)\log(\frac{1-P}{1-Q})$ [52].

Furthermore, we demand $P(\max\{R_i\} > R^*; N, K) > P^*$, where P^* is a probability threshold or confidence and insert this condition in Eq. (12) which yields $K = \frac{\ln(1-P^*)}{\ln(F(R^*;N))}$. Finally, we obtain

$$K > \sqrt{2N-4}\; e^{\frac{N}{2}\ln\left(\frac{1}{1-R^{*2}}\right)} \,|\ln(1-P^*)|. \qquad (14)$$

Therefore, the student size K has to increase exponentially fast with the input-layer size N for a fixed R^* and P^* leading to exponential long training times if one wants to keep a small generalization error below the threshold of the linearized regime. This result is consistent with the conclusions in [43].

In conclusion, we have studied the learning dynamics of the random feature model trained by the stochastic gradient

descent embedded in the student-teacher framework. We obtained asymptotic solutions of the generalization error out of a set of coupled differential equations describing the weight dynamics. For a regime with a finite ratio of hidden layer width and input dimension, we computed the asymptotic generalization error and found that it stays finite for two choices of activation functions. In the second part of this work, we found by a simple ansatz that the generalization error can become arbitrarily small under an exponential increase of the student size in relation to the input dimension.

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, MA, 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. Burges, L. Bottou, and K. Weinberger (Curran Associates, Inc., Glasgow, Scotland, 2012), Vol. 25.

[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. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[6] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM **64**, 107 (2021).

[7] M. Theunissen, M. Davel, and E. Barnard, Benign interpolation of noise in deep learning, South African Comput. J. **32**, 12 (2020).

[8] A. Canziani, A. Paszke, and E. Culurciello, An analysis of deep neural network models for practical applications, arXiv:1605.07678 (2017).

[9] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, The role of over-parametrization in generalization of neural networks, in *International Conference on Learning Representations* (Curran Associates, Inc., Glasgow, Scotland, 2023).

[10] P. Bartlett, The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network, IEEE Trans. Inf. Theory **44**, 525 (1998).

[11] R. M. Neal, *Priors for Infinite Networks* (Springer, New York, 1996), p. 29.

[12] J. Lee, J. Sohl-Dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri, Deep neural networks as gaussian processes, in *International Conference on Learning Representations* (2018).

[13] R. Novak, L. Xiao, Y. Bahri, J. Lee, G. Yang, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, Bayesian deep convolutional networks with many channels are gaussian processes, in *International Conference on Learning Representations* (2019).

[14] A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, Gaussian process behaviour in wide deep neural networks, in *International Conference on Learning Representations* (2018).

[15] C. Williams, Computing with infinite networks, in *Advances in Neural Information Processing Systems*, edited by M. Mozer, M. Jordan, and T. Petsche (MIT Press, Cambridge, MA, 1996), Vol. 9.

[16] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Glasgow, Scotland, 2018), Vol. 31.

[17] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, On exact computation with an infinitely wide neural net, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Glasgow, Scotland, 2019), Vol. 32.

[18] Z. Chen, Y. Cao, Q. Gu, and T. Zhang, A generalized neural tangent kernel analysis for two-layer neural networks, in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., Glasgow, Scotland, 2021).

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

[20] 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*, edited by K. Chaudhuri and R. Salakhutdinov, Vol. 97 of Proceedings of Machine Learning Research (Curran Associates, Inc., Glasgow, Scotland, 2019), pp. 322–332.

[21] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, in *Advances in Neural Information Processing Systems*, edited by, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Glasgow, Scotland, 2019), Vol. 32.

[22] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Glasgow, Scotland, 2019), Vol. 32.

[23] A. Rahimi and B. Recht, Random features for large-scale kernel machines, in *Advances in Neural Information Processing Systems*, edited by, J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran Associates, Inc., Glasgow, Scotland, 2007), Vol. 20.

[24] E. Gardner and B. Derrida, Three unfinished works on the optimal storage capacity of networks, J. Phys. A **22**, 1983 (1989).

[25] J. Hertz, A. Krough, and R. Palmer, *Introduction To The Theory Of Neural Computation* (West View Press, Boulder, Colorado, 1991), Vol. 44, p. 12.
[26] T. L. H. Watkin, A. Rau, and M. Biehl, The statistical mechanics of learning a rule, Rev. Mod. Phys. **65**, 499 (1993).
[27] D. Saad, On-line learning in neural networks, J. Am. Stat. Assoc. **95**, 452 (2000).
[28] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).
[29] 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).
[30] E. Gardner, The space of interactions in neural network models, J. Phys. A **21**, 257 (1988).
[31] H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Phys. Rev. A **45**, 6056 (1992).
[32] P. Riegler and M. Biehl, On-line backpropagation in two-layered neural networks, J. Phys. A **28**, L507 (1995).
[33] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Glasgow, Scotland, 2019), Vol. 32.
[34] M. Biehl and H. Schwarze, Learning by on-line gradient descent, J. Phys. A. **28**, 643 (1995).
[35] D. Saad and S. A. Solla, On-line learning in soft committee machines, Phys. Rev. E **52**, 4225 (1995).
[36] D. Saad and S. A. Solla, Exact solution for on-line learning in multilayer neural networks, Phys. Rev. Lett. **74**, 4337 (1995).
[37] M. Biehl, P. Riegler, and C. Wöhler, Transient dynamics of on-line learning in two-layered neural networks, J. Phys. A **29**, 4769 (1996).
[38] M. Biehl, E. Schlösser, and M. Ahr, Phase transitions in soft-committee machines, Europhys. Lett. **44**, 261 (1998).
[39] E. Oostwal, M. Straat, and M. Biehl, Hidden unit specialization in layered neural networks: Relu vs. sigmoidal activation, Physica A **564**, 125517 (2021).
[40] F. Richert, R. Worschech, and B. Rosenow, Soft mode in the dynamics of over-realizable online learning for soft committee machines, Phys. Rev. E **105**, L052302 (2022).
[41] G. Naveh, O. Ben David, H. Sompolinsky, and Z. Ringel, Predicting the outputs of finite deep neural networks trained with noisy gradients, Phys. Rev. E **104**, 064301 (2021).
[42] Q. Li and H. Sompolinsky, Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization, Phys. Rev. X **11**, 031059 (2021).
[43] G. Yehudai and O. Shamir, *On the Power and Limitations of Random Features for Understanding Neural Networks* (Curran Associates Inc., Glasgow, Scotland, 2019).
[44] S. Mei, T. Misiakiewicz, and A. Montanari, Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration, Appl. Comput. Harmon. Anal. **59**, 3 (2022), Special Issue on Harmonic Analysis and Machine Learning.
[45] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, Linearized two-layers neural networks in high dimension, Ann. Statist. **49**, 1029 (2021).
[46] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, Limitations of lazy training of two-layers neural network, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Glasgow, Scotland, 2019), Vol. 32.
[47] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.6.L022049 for the setup of differential equations, the derivation of the Langevin equation and the the asymptotic generalization error. The analytical results are futher supported by numerical investigations.
[48] G. Reents and R. Urbanczik, Self-averaging and on-line learning, Phys. Rev. Lett. **80**, 5445 (1998).
[49] G. Rotskoff and E. Vanden-Eijnden, Trainability and accuracy of artificial neural networks: An interacting particle system approach, Commun. Pure Appl. Math. **75**, 1889 (2022).
[50] N. V. Kampen, *Stochastic Processes in Physics and Chemistry* (North Holland, Amsterdam, 2007).
[51] V. A. Marčenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR-Sbornik **1**, 457 (1967).
[52] R. B. Ash, *Information Theory* (Dover, New York, 1965).