The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization

Ben Adlam [*][†][1] **Jeffrey Pennington** [*][1]

Abstract

Modern deep learning models employ considerably more parameters than required to fit the training data. Whereas conventional statistical wisdom suggests such models should drastically overfit, in practice these models generalize remarkably well. An emerging paradigm for describing this unexpected behavior is in terms of a *double descent* curve, in which increasing a model's capacity causes its test error to first decrease, then increase to a maximum near the interpolation threshold, and then decrease again in the overparameterized regime. Recent efforts to explain this phenomenon theoretically have focused on simple settings, such as linear regression or kernel regression with unstructured random features, which we argue are too coarse to reveal important nuances of actual neural networks. We provide a precise high-dimensional asymptotic analysis of generalization under kernel regression with the Neural Tangent Kernel, which characterizes the behavior of wide neural networks optimized with gradient descent. Our results reveal that the test error has non-monotonic behavior deep in the overparameterized regime and can even exhibit additional peaks and descents when the number of parameters scales quadratically with the dataset size.

1. Introduction

Machine learning models based on deep neural networks have achieved widespread success across a variety of domains, often playing integral roles in products and services people depend on. As users rely on these systems in increasingly important scenarios, it becomes paramount to establish a rigorous understanding for when the models might work, and, crucially, when they might not. Unfortunately, the current theoretical understanding of deep learning is modest at best, as large gaps persist between theory and observation and many basic questions remain unanswered.

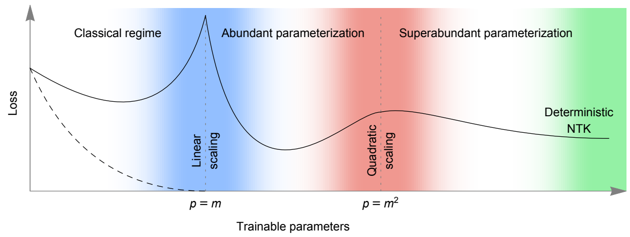

Figure 1. An illustration of multi-scale generalization phenomena for neural networks and related kernel methods. The classical U-shaped under- and over-fitting curve is shown on the far left. After a peak near the interpolation threshold, when the number of parameters p equals the number of samples m, the test loss decreases again, a phenomenon known as *double descent*. On the far right is the limit when $p \to \infty$, which is described by the Neural Tangent Kernel. In this work, we identify a new scale of interest in between these two regimes, namely when p is quadratic in m, and show that it exhibits complex non-monotonic behavior, suggesting that double descent does not provide a complete picture. Putting these observations together we define three regimes separated by two transitional phases: (i) the *classical regime* of underparameterization when $p < m$, (ii) the *abundant parameterization* regime when $m < p < m^2$, and (iii) the *superabundant parameterization* regime when $p > m^2$. The transitional phases between them are of particular interest as they produce non-monotonic behavior.

One of the most conspicuous such gaps is the unexpectedly good generalization performance of large, heavily-overparameterized models. These models can be so expressive that they can perfectly fit the training data (even when the labels are replace by pure noise), but still manage to generalize well on real data (Zhang et al., 2016). An emerging paradigm for describing this behavior is in terms of a double descent curve (Belkin et al., 2019a), in which increasing a model's capacity causes its test error to first decrease, then increase to a maximum near the interpolation threshold (where the number of parameters equals the number of samples), and then decrease again in the overparameterized regime.

[*] Equal contribution [†] Work done as a member of the Google AI Residency Program. [1] Google Brain. Correspondence to: Jeffrey Pennington <jpennin@google.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

There are of course more elaborate measures of a model's capacity than a naive parameter count. Recent empirical and theoretical work studying the correlation of these capacity measures with generalization has found mixed results, with many measures having the opposite relationship with generalization that theory would predict (Neyshabur et al., 2017). Other work has questioned whether it is possible in principle for uniform convergence results to explain the generalization performance of neural networks (Nagarajan & Kolter, 2019).

Our approach is quite different. We consider the algorithm's asymptotic performance on a specific data distribution, leveraging the large system size to get precise theoretical results. In particular, we examine the high-dimensional asymptotics of kernel ridge regression with respect to the Neural Tangent Kernel (NTK) (Jacot et al., 2018) and conclude that double descent does not always provide an accurate or complete picture of generalization performance. Instead, we identify complex non-monotonic behavior in the test error as the number of parameters varies across multiple scales and find that it can exhibit additional peaks and descents when the number of parameters scales quadratically with the dataset size.

Our theoretical analysis focuses on the NTK of a single-layer fully-connected model when the samples are drawn independently from a Gaussian distribution and the targets are generated by a wide teacher neural network. We provide an exact analytical characterization of the generalization error in the high-dimensional limit in which the number of samples m, the number of features n_0, and the number of hidden units n_1 tend to infinity with fixed ratios $\phi := n_0/m$ and $\psi := n_0/n_1$. By adjusting these ratios, we reveal the intricate ways in which the generalization error depends on the dataset size and the effective model capacity.

We investigate various limits of our results, including the behavior when the NTK degenerates into the kernel with respect to only the first-layer or only the second-layer weights. The latter corresponds to the standard setting of random feature ridge regression, which was recently analyzed in (Mei & Montanari, 2019). In this case, the total number of parameters p is equal to the width n_1, i.e. $p = n_1 = (\phi/\psi)m$, so that p is linear in the dataset size. In contrast, for the full kernel, the number of parameters is $p = (n_0+1)n_1 = (\phi^2/\psi)m^2 + (\phi/\psi)m$, *i.e.* it is quadratic in the dataset size. By studying these two kernels, we derive insight into the generalization performance in the vicinities of linear and quadratic overparameterization, and by piecing these two perspectives together, we infer the existence of multi-scale phenomena, which sometimes can include triple descent. See Fig. 1 for an illustration and Fig. 4 for empirical confirmation of this behavior.

1.1. Our Contributions

1. We derive exact high-dimensional asymptotic expressions for the test error of NTK ridge regression.
2. We prove that the test error can exhibit non-monotonic behavior deep in the overparameterized regime.
3. We investigate the origins of this non-monotonicity and attribute them to the kernel with respect to the second-layer weights.
4. We provide empirical evidence that triple descent can indeed occur for finite-sized networks trained with gradient descent.
5. We find exceptionally fast learning curves in the noiseless case, with $E_{\text{test}} \sim m^{-2}$.

1.2. Related Work

A recent line of work studying the behavior of interpolating models was initiated by the intriguing experimental results of (Zhang et al., 2016; Belkin et al., 2018b), which showed that deep neural networks and kernel methods can generalize well even in the interpolation regime. A number of theoretical results have since established this behavior in certain settings, such as interpolating nearest neighbor schemes (Belkin et al., 2018a) and kernel regression (Belkin et al., 2019c; Liang et al., 2020b).

These observations, coupled with classical notions of the bias-variance tradeoff, have given rise to the double descent paradigm for understanding how test error depends on model complexity. These ideas were first discussed in (Belkin et al., 2019a), and empirical evidence was obtained in (Advani & Saxe, 2017; Geiger et al., 2020) and recently in (Nakkiran et al., 2019). Precise theoretical predictions soon confirmed this picture for linear regression in various scenarios (Belkin et al., 2019b; Hastie et al., 2019; Mitra, 2019).

Linear models struggle to capture all of the phenomena relevant to double descent because the parameter count is tied to the number of features. Recent work found multiple descents in the test loss for minimum-norm interpolants in Reproducing Kernel Hilbert Spaces (Liang et al., 2020a), but it similarly requires changing the data distribution to vary model capacity. A precise analysis of a nonlinear system for a fixed data generating process is the most direct way to draw insight into double descent. A recent preprint (Mei & Montanari, 2019) shares this view and adopts a similar analysis to ours, but focuses entirely on the standard case of unstructured random features. Such a setup can indeed model double descent, and certainly bears relevance to certain wide neural networks in which only the top-layer weights are optimized (Neal, 1996; Rahimi & Recht, 2008; Lee et al., 2018; de G. Matthews et al., 2018; Lee et al., 2019), but its connection to neural networks trained with gradient descent remains less clear.

Gradient-based training of wide neural networks initialized in the standard way was recently shown to correspond to kernel gradient descent with respect to the Neural Tangent Kernel (Jacot et al., 2018). This result has spawned renewed interest in kernel methods and their connection to deep learning; a woefully incomplete list of papers in this direction includes Lee et al. (2019); Chizat et al. (2019); Du et al. (2019; 2018); Arora et al. (2019); Xiao et al. (2019).

To connect these research directions, our analysis requires tools and recent results from random matrix theory and free probability. A central challenge stems from the fact that many of the matrices in question have nonlinear dependencies between the elements, which arises from the nonlinear feature matrix $F = \sigma(WX)$. This challenge was overcome in (Pennington & Worah, 2017), which computed the spectrum of F, and in (Pennington & Worah, 2018), which examined the spectrum of the Fisher information matrix; see also (Louart et al., 2018). We also utilize the results of (Adlam et al., 2019; Péché et al., 2019), which established a linear signal plus noise model for F that shares the same bulk statistics. This linearized model allows us to write the test error as the trace of a rational function of the underlying random matrices. The methods we use to compute such quantities rely on so-called *linear pencils* that represent the rational function in terms of the inverse of a larger block matrix (Helton et al., 2018), and on operator-valued free probability for computing the trace of the latter (Far et al., 2006).

2. Preliminaries

In this section, we introduce our theoretical setting and some of the tools required to state our results.

2.1. Problem Setup and Notation

We consider the task of learning an unknown function from m independent samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^{n_0} \times \mathbb{R}$, $i \le m$, where the datapoints are standard Gaussian, $\mathbf{x}_i \sim \mathcal{N}(0, I_{n_0})$, and the labels are generated by a wide[1] single-hidden-layer neural network:

$$y_i | \mathbf{x}_i, \Omega, \omega \sim \omega \sigma_{\mathrm{T}}(\Omega \mathbf{x}_i / \sqrt{n_0}) / \sqrt{n_{\mathrm{T}}} + \varepsilon_i. \tag{1}$$

The teacher's activation function σ_{T} is applied coordinatewise, and its parameters $\Omega \in \mathbb{R}^{n_{\mathrm{T}} \times n_0}$ and $\omega \in \mathbb{R}^{1 \times n_{\mathrm{T}}}$ are matrices whose entries are independently sampled once for all data from $\mathcal{N}(0, 1)$. We also allow for independent label noise, $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Let $\hat{y}(\mathbf{x})$ denote the model's predictive function. We consider squared error, so the test loss is,

$$\mathbb{E}(y - \hat{y})^2 = \mathbb{E}_{\mathbf{x},\varepsilon} (\omega \sigma_{\mathrm{T}}(\Omega \mathbf{x} / \sqrt{n_0}) / \sqrt{n_{\mathrm{T}}} + \varepsilon - \hat{y}(\mathbf{x}))^2, \tag{2}$$

where the expectation is over an iid test point $(\mathbf{x}, y)$ conditional on the training set, the teacher parameters, and any randomness in the learning algorithm producing $\hat{y}$, such as the random parameters defining the random features. Note that the test loss is a random variable; however, in the high-dimensional asymptotics we consider here, it concentrates about its mean.

2.2. Neural Tangent Kernel Regression

We consider predictive functions $\hat{y}$ defined by approximate (*i.e.* random feature) kernel ridge regression using the Neural Tangent Kernel (NTK) of a single-hidden-layer neural network. The NTK can be considered a kernel K that is approximated by random features corresponding to the Jacobian J of the network's output with respect to its parameters, *i.e.* $K(\mathbf{x}_1, \mathbf{x}_2) = J(\mathbf{x}_1) J(\mathbf{x}_2)^\top$. As the width of the network becomes very large (compared to all other relevant scales in the system), the approximate NTK converges to a constant kernel determined by the network's initial parameters and describes the trajectory of the network's output under gradient descent. In particular,

$$N_t(\mathbf{x}) = N_0(\mathbf{x}) + (Y - N_0(X)) K^{-1} (I - e^{-\eta t K}) K_{\mathbf{x}}\,, \tag{3}$$

where $N_t(\mathbf{x})$ is the output of the network at time t, $K := K(\gamma) = K(X, X) + \gamma I_m$, $K_{\mathbf{x}} := K(X, \mathbf{x})$, η is the learning rate, and γ is a ridge regularization constant[2]. For this work, we are interested in the $t \to \infty$ limit of (3), which defines the predictive function,

$$\hat{y}(\mathbf{x}) := N_\infty(\mathbf{x}) = N_0(\mathbf{x}) + (Y - N_0(X)) K^{-1} K_{\mathbf{x}}\,. \tag{4}$$

We remark that if the width is not asymptotically larger than the dataset size, the validity of (3) can break down and (4) may not accurately describe the late-time predictions of the neural network. While this potential discrepancy is an interesting topic, we defer an in-depth analysis to future work (but see Fig. 4) for an empirical analysis of gradient descent). Instead, we regard (4) as the definition of our predictive function and focus on kernel regression with the NTK. We believe this setup is interesting its own right; for example, recent work has demonstrated its effectiveness as a kernel method on complex image datasets (Li et al., 2019) and found it to be competitive with neural networks in small data regimes.

In this work, we restrict our study to the NTK of single-hidden-layer fully-connected networks. In particular, consider a network of with width n_1 and pointwise activation function σ, defined by,

$$N_0(\mathbf{x}) = W_2 \sigma(W_1 \mathbf{x} / \sqrt{n_0}) / \sqrt{n_1}\,, \tag{5}$$

[1] We assume the width $n_{\mathrm{T}} \to \infty$, but the rate is not important.

[2] These overloaded definitions of K can be distinguished by the number of arguments and should be clear from context.

for initial weight matrices $W_1 \in \mathbb{R}^{n_1 \times n_0}$ and $W_2 \in \mathbb{R}^{1 \times n_1}$ with iid entries $[W_1]_{ij} \sim \mathcal{N}(0,1)$[3] and $[W_2]_i \sim \mathcal{N}(0, \sigma_{W_2}^2)$.

We collect our assumptions on the activation functions below, in Assumption 1. Their main purpose is to ensure that certain moments and derivatives exist almost surely, but for simplicity we state somewhat stronger conditions than are actually required for our analysis. To simplify the already cumbersome algebraic manipulations, we assume that σ has zero Gaussian mean. We emphasize that this condition is not essential and our techniques easily generalize to all commonly used activation functions.

Assumption 1. *The activation functions $\sigma, \sigma_{\mathrm{T}} : \mathbb{R} \to \mathbb{R}$ are assumed to be differentiable almost everywhere. We assume $|\sigma(x)|, |\sigma'(x)|, |\sigma_{\mathrm{T}}(x)| = \mathcal{O}(\exp(Cx))$ for some positive constant C, which implies all the Gaussian moments of σ, σ', and σ_{T} exist, and we assume $\mathbb{E}\sigma(Z) = 0$ for $Z \sim \mathcal{N}(0,1)$.*

The Jacobian of (5) with respect to the parameters naturally decomposes into the Jacobian with respect to W_1 and W_2, *i.e.* $J(\mathbf{x}) = [\partial N_0(\mathbf{x})/\partial W_1, \partial N_0(\mathbf{x})/\partial W_2] = [J_1(\mathbf{x}), J_2(\mathbf{x})]$. Therefore the kernel K also decomposes this way, and we can write.

$$K(\mathbf{x}_1, \mathbf{x}_2) = J_1(\mathbf{x}_1) J_1(\mathbf{x}_2)^\top + J_2(\mathbf{x}_1) J_2(\mathbf{x}_2)^\top \tag{6}$$

$$=: K_1(\mathbf{x}_1, \mathbf{x}_2) + K_2(\mathbf{x}_1, \mathbf{x}_2) \tag{7}$$

A simple calculation yields the per-layer constituent kernels,

$$K_1(X, X) = \frac{X^\top X}{n_0} \odot \frac{(F')^\top \operatorname{diag}(W_2)^2 F'}{n_1} \tag{8}$$

$$K_2(X, X) = \frac{1}{n_1} F^\top F\,, \tag{9}$$

where we have introduced the abbreviations $F = \sigma(W_1 X/\sqrt{n_0})$ and $F' = \sigma'(W_1 X/\sqrt{n_0})$. Notice that when $\sigma_{W_2}^2 \to 0$, $K = K_2$, *i.e.* the NTK degenerates into the standard random features kernel. However, the predictive function (4) contains an offset $N_0(\mathbf{x})$ which would typically be set to zero in standard random feature kernel regression because it simply increases the variance of test predictions. Removing this variance component has an analogous operation in neural network training: either the function value at initialization can be subtracted throughout training, or a symmetrization trick can be used in which two copies of the NN are initialized identically, and their normalized difference $N \equiv \left(N^{(a)} - N^{(b)}\right)/\sqrt{2}$ is trained with gradient descent. Either method preserves the kernel K while enforcing $N_0 \equiv 0$. We call this setup *centering*, and present results with and without it.

Finally, we note that ridge regularization in the kernel perspective corresponds to using L2 regularization of the neural network's weights toward their initial values.

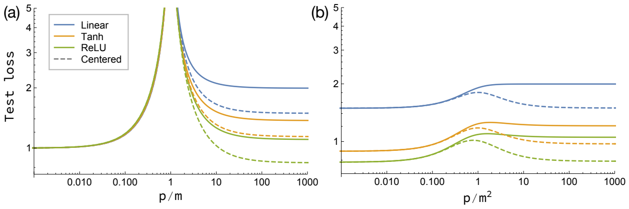

Figure 2. Theoretical results for the test error with and without centering for different activation functions with $\phi = 2, \gamma = 10^{-3}$, and SNR $= 1$ for (a) the second-layer kernel K_2 and (b) the full NTK K as the number of parameters p is varied by changing the network width. Nonmonotonic behavior is clearly visible at the linear scaling transition ($p = m$) and the quadratic scaling transition ($p = m^2$). Here, ReLU denotes the zero-mean function $\sigma(x) = \max(x, 0) - 1/\sqrt{2\pi}$.

3. Three Regimes of Parameterization

In this section, we outline an argument based on the structure of the NTK as to why one should expect the test error to exhibit non-trivial phenomena at two different scales of overparameterization. From the expressions for the test error (2) and the predictive function (4), it is evident that the behavior of the test error is determined by the spectral properties of the NTK. Although the fine details of the relationship can only be revealed by the explicit calculation, we can nevertheless make some basic high-level observations based on the coarser structure of the kernel.

The number of trainable parameters p relative to the dataset size m controls the amount of parameterization or complexity of a model. In our setting of a single-hidden-layer fully-connected neural network, $p = n_1(n_0 + 1)$, and for a fixed dataset, we can adjust the ratio p/m by varying the hidden-layer width n_1.

The simplest way to see that there should be two scales comes from examining the two terms in the kernel separately. Because $K_1 = J_1 J_1^T$ and $J_1 \in \mathbb{R}^{m \times n_0 n_1}$, the first-layer kernel has rank at most $\min\{n_0 n_1, m\}$, which suggests nontrivial transitional behavior when $p = \Theta(m)$. Similarly, the rank of K_2 is at most $\min\{n_1, m\}$, which suggests a second interesting scale when $n_1 = \Theta(m)$, or equivalently, when $p = \Theta(m^2)$ if $n_0 = \Theta(n_1)$. Our explicit calculations confirm that interesting phenomena indeed occur at these scales, as can be seen in Fig. 2.

These two scales partition the degree of parameterization into three regimes. We consider the *classical* regime to be when $p \lesssim m$ because classical generalization theory tends to hold and the U-shaped test error curve is observed. The transition around $p = \Theta(m)$ manifests as a sharp rise in the test loss near the interpolation threshold, followed by a quick descent as p increases further, as can be seen in Fig. 2(a). We call this the *linear scaling* transition. After this, we enter a regime we call *abundant parameterization*

[3] Any non-zero $\sigma_{W_1}^2$ can be absorbed into a redefinition of σ.

when $m \lesssim p \lesssim m^2$. In this regime, the test error tends to decrease until p nears the vicinity of m^2, where it can sometimes increase again, producing a second U-shaped curve. When $p = \Theta(m^2)$, another transition is observed, which we call the *quadratic scaling* transition, which can be seen in Fig. 2(b). On the other side of this transition, $p \gtrsim m^2$, a regime we call *superabundant parameterization*. See Fig 1 for an illustration of this general picture.

While the classical regime has been long studied, and the superabundant regime has generated considerable recent interest due to the NTK, our main aim in delineating the above regimes is to highlight the existence of the intermediate scale containing complex phenomenology. For this reason, we focus our theoretical analysis on the novel scaling regime in which $p = \Theta(m^2)$. In particular, as mentioned in Sec. 1, we consider the high-dimensional asymptotics in which $n_0, n_1, m \to \infty$ with $\phi := n_0/m$ and $\psi := n_0/n_1$ held constant.

4. Overview of Techniques

In this section, we provide a high-level overview of the analytical tools and mathematical results we use to compute the generalization error. To begin with, let us first describe the main technical challenges in computing explicit asymptotic limits of (2).

The first challenge, which is evident upon inspecting (8), is that the kernel contains a Hadamard product of random matrices, for which concrete results in the random matrix literature are few and far between. We address this problem in Sec. 4.1.

The second challenge, which is apparent by inspecting (9), is that the kernel depends on random matrices with nonlinear dependencies between the entries. We describe how to circumvent this difficulty in Sec. 4.2.

Finally, by expanding the square in (2) and substituting (4), we find terms that are constant, linear, and quadratic in K^{-1}. Some of the random matrices that appear inside the matrix inverses (*e.g.* X, and W_1) also appear outside of them as multiplicative factors, a situation that prevents the straightforward application of many standard proof techniques in random matrix theory. We describe how to overcome this challenge in Sec. 4.3.

4.1. Simplification of First-Layer Kernel

A straightforward central limiting argument shows that in the asymptotic limit the entries of $W_1 X/\sqrt{n_0}$ are marginally Gaussian with mean zero and unit variance. As such, the first and second moments of the entries in the matrix $F' = \sigma'(W_1 X/\sqrt{n_0})$ are equal to

$$\sqrt{\zeta} := \mathbb{E}_{z\sim\mathcal{N}(0,1)}\sigma'(z)\,, \quad \eta' := \mathbb{E}_{z\sim\mathcal{N}(0,1)}\sigma'(z)^2\,. \quad (10)$$

It follows that we can split K_1 into two terms,

$$\frac{X^\top X}{n_0} \odot \frac{\left(\bar{F}'\right)^\top \operatorname{diag}(W_2)^2 \bar{F}'}{n_1} + \sigma_{W_2}^2 \zeta \frac{X^\top X}{n_0}, \quad (11)$$

where $\bar{F}'$ is a centered version of F'. Focusing on the first term, because $n_0 n_1 = \phi^2/\psi m^2$, the random fluctuations in the off-diagonal elements are $\mathcal{O}(1/m)$, which are too small to contribute to the spectrum or moments of an $m \times m$ matrix whose diagonal entries are order one. In fact, the diagonal entries are simply proportional to the variance of the entries of F', namely $(\eta' - \zeta)$. Putting this together, we can eliminate the Hadamard product entirely and write,

$$K_1 \cong \sigma_{W_2}^2(\eta' - \zeta) I_m + \frac{\sigma_{W_2}^2 \zeta}{n_0} X^\top X\,, \quad (12)$$

where the $\cong$ notation means the two matrices share the same bulk statistics asymptotically. We make this argument precise in Sec. S1.

4.2. Linearization 1: Gaussian Equivalents

The test error (2) involves large random matrices with nonlinear dependencies, which are not immediately amenable to standard methods of analysis in random matrix theory. The main culprit is the random feature matrix $F = \sigma(W_1 X/\sqrt{n_0})$, but $f := \sigma(W_1 \mathbf{x}/\sqrt{n_0})$, $Y = \omega\sigma_{\mathrm{T}}(\Omega X/\sqrt{n_0})/\sqrt{n_{\mathrm{T}}} + \mathcal{E}$, and $y := \omega\sigma_{\mathrm{T}}(\Omega \mathbf{x}/\sqrt{n_0})/\sqrt{n_{\mathrm{T}}}$ all suffer from the same issue.

The solution is to replace each of these matrices with an equivalent matrix without nonlinear dependencies, but chosen to maintain the same first- and second-order moments for all of the terms that appear in the test error (2). This approach was described for F in (Adlam et al., 2019) (see also (Péché et al., 2019)). The upshot is that the test error is asymptotically invariant to the following substitutions,

$$F \to F^{\mathrm{lin}} := \sqrt{\frac{\zeta}{n_0}} W_1 X + \sqrt{\eta - \zeta}\,\Theta_F \quad (13)$$

$$Y \to Y^{\mathrm{lin}} := \sqrt{\frac{\zeta_{\mathrm{T}}}{n_{\mathrm{T}} n_0}} \omega \Omega X + \sqrt{\frac{\eta_{\mathrm{T}} - \zeta_{\mathrm{T}}}{n_{\mathrm{T}}}} \omega \Theta_Y + \mathcal{E} \quad (14)$$

$$f \to f^{\mathrm{lin}} := \sqrt{\frac{\zeta}{n_0}} W_1 \mathbf{x} + \sqrt{\eta - \zeta}\,\theta_f \quad (15)$$

$$y \to y^{\mathrm{lin}} := \sqrt{\frac{\zeta_{\mathrm{T}}}{n_{\mathrm{T}} n_0}} \omega \Omega \mathbf{x} + \sqrt{\frac{\eta_{\mathrm{T}} - \zeta_{\mathrm{T}}}{n_{\mathrm{T}}}} \omega \theta_y\,. \quad (16)$$

The new objects Θ_F, Θ_Y, θ_f, and θ_y are matrices of the appropriate shapes with iid standard Gaussian entries. The constants $\eta, \zeta, \eta_{\mathrm{T}}$, and ζ_{T} are chosen so that the mixed moments up to second order are the same for the original and linearized versions. In particular,

$$\zeta := \left[\mathbb{E}_{z\sim\mathcal{N}(0,1)}\sigma'(z)\right]^2, \quad \eta := \mathbb{E}_{z\sim\mathcal{N}(0,1)}\sigma(z)^2\,, \quad (17)$$

$$\zeta_{\mathrm{T}} := [\mathbb{E}_{z\sim\mathcal{N}(0,1)}\sigma_{\mathrm{T}}'(z)]^2\,, \quad \eta_{\mathrm{T}} := \mathbb{E}_{z\sim\mathcal{N}(0,1)}\sigma_{\mathrm{T}}(z)^2\,. \tag{18}$$

The statement that the test error only depends on Y^{lin} is consistent with the observations made in (Ghorbani et al., 2019; Mei & Montanari, 2019) that in the high-dimensional regime where $n_0 = \Theta(m)$, only linear functions of the data can be learned. Indeed, Y^{lin} is equivalent to a linear teacher plus noise with signal-to-noise ratio given by,

$$\text{SNR} = \frac{\zeta_{\mathrm{T}}}{\eta_{\mathrm{T}} - \zeta_{\mathrm{T}} + \sigma_\varepsilon^2}\,. \tag{19}$$

We often make this equivalence to a linear teacher explicit by setting $\sigma_{\mathrm{T}}(x) = x$, which implies $\eta_{\mathrm{T}} = \zeta_{\mathrm{T}} = 1$. Doing so also removes the noise from the test label, but since this noise merely contributes an additive shift to the test loss, removing it does not change any of our conclusions.

4.3. Linearization 2: Linear Pencil

Next we turn our attention to the actual computation of the asymptotic test loss. Expanding the test error (2) we have[4],

$$\begin{aligned} E_{\text{test}} &:= \mathbb{E}_{(\mathbf{x},y)}(y - \hat{y}(\mathbf{x}))^2 \qquad (20)\\ &= \mathbb{E}_{(\mathbf{x},\varepsilon)}\Big[\operatorname{tr}(y^\top y) - 2\operatorname{tr}(K_{\mathbf{x}}^\top K^{-1} Y^\top y)\\ &\qquad + \operatorname{tr}(K_{\mathbf{x}}^\top K^{-1} Y^\top Y K^{-1} K_{\mathbf{x}})\Big]\,. \qquad (21)\end{aligned}$$

The simplification (12) gives,

$$K = \sigma_{W_2}^2\left[(\eta' - \zeta)I_m + \frac{\zeta X^\top X}{n_0}\right] + \frac{F^\top F}{n_1} + \gamma I_m \tag{22}$$

$$K_{\mathbf{x}} = \frac{\sigma_{W_2}^2\zeta}{n_0}X^\top \mathbf{x} + \frac{1}{n_1}F^\top f\,, \tag{23}$$

which, when applied to (21) together with the substitutions (13)-(16), expresses the test error directly in terms of the iid Gaussian random matrices $W_1, X, \Theta_F, \Omega, \Theta_Y, \mathcal{E}, \theta_f, \theta_y$ and $\mathbf{x}$. The expectations over $\mathbf{x}$ and $\mathcal{E}$ are trivial because these variables do not appear inside the matrix inverse K^{-1}. Moreover, asymptotically the traces concentrate around their means with respect to $\Omega, \Theta_Y, \theta_f$ and θ_y, which we can also compute easily for the same reason. Therefore, the test error can be written as,

$$E_{\text{test}} = a_0 + \sum_i b_i \operatorname{tr}(B_i K^{-1}) + \sum_i c_i \operatorname{tr}(C_i K^{-1} D_i K^{-1}) \tag{24}$$

where B_i, C_i, D_i are monomials in $\{W_1, X, \Theta_F\}$ and their transposes, and $a_0, b_i, c_i \in \mathbb{R}$.

Eqn. (24) is a rational function of the noncommutative random variables W_1, X, and Θ_F. A useful result from noncommutative algebra guarantees that such a rational function can be *linearized* in the sense that it can be expressed in terms of the inverse of a matrix whose entries are linear in the noncommutative variables. This representation is often called a linear pencil, and is not unique; see *e.g.* (Helton et al., 2018) for details.

To illustrate this concept, consider the simple case of K^{-1}. After applying the substitutions (13)-(16) to (22), a linear pencil is given by

$$\begin{bmatrix} [\gamma + \sigma_{W_2}^2(\eta' - \zeta)]I & \frac{\sigma_{W_2}^2\zeta}{n_0}X^\top & \frac{\sqrt{\eta-\zeta}}{n_0}\Theta_F^\top & \frac{\sqrt{\zeta}}{\sqrt{n_0}n_1}X^\top \\ -X & I & 0 & 0 \\ -\sqrt{\eta-\zeta}\Theta_F & -\frac{\sqrt{\zeta}}{\sqrt{n_0}}W_1 & I & 0 \\ 0 & 0 & -W_1^\top & I \end{bmatrix}^{-1}_{11}\,,$$

which can be checked by an explicit computation of the block matrix inverse. After obtaining a linear pencil for each of the terms in (24), the only task that remains is computing the trace. Since each linear pencil is a block matrix whose blocks are iid Gaussian random matrices, its trace can be evaluated using the techniques described in (Far et al., 2006) or through the general formalism of operator-valued free probability. We refer the reader to the book (Mingo & Speicher, 2017) for more details on these topics.

5. Asymptotic Training and Test Error

The calculations described in the previous section are presented in the Supplementary Materials. Here we present the main results.

Proposition 1. *As $n_0, n_1, m \to \infty$ with $\phi = n_0/m$ and $\psi = n_0/n_1$ fixed, the traces $\tau_1(z) := \frac{1}{m}\mathbb{E}\operatorname{tr}(K(z)^{-1})$ and $\tau_2(z) := \frac{1}{m}\mathbb{E}\operatorname{tr}(\frac{1}{n_0}X^\top X K(z)^{-1})$ are given by the unique solutions to the coupled polynomial equations,*

$$\begin{aligned} &\phi\left(\zeta\tau_2\tau_1 + \phi(\tau_2 - \tau_1)\right) + \zeta\tau_1\tau_2\psi\left(z\tau_1 - 1\right)\\ &\quad = -\zeta\tau_1\tau_2\sigma_{W_2}^2\left(\zeta\left(\tau_2 - \tau_1\right)\psi + \tau_1\psi\eta' + \phi\right)\\ &\zeta\tau_1^2\tau_2\left(\eta' - \eta\right)\sigma_{W_2}^2 + \zeta\tau_1\tau_2\left(z\tau_1 - 1\right)\\ &\quad = \left(\tau_2 - \tau_1\right)\phi\left(\zeta\left(\tau_2 - \tau_1\right) + \eta\tau_1\right)\,, \end{aligned} \tag{25}$$

such that $\tau_1, \tau_2 \in \mathbb{C}^+$ for $z \in \mathbb{C}^+$.

Theorem 1. *Let $\gamma = Re(z)$ and let τ_1 and τ_2 be defined as in Proposition 1 with $Im(z) \to 0^+$. Then the asymptotic training error $E_{train} = \frac{1}{m}\mathbb{E}\|Y - \hat{y}(X)\|_F^2$ is given by,*

$$\begin{aligned} E_{train} = &-\gamma^2(\sigma_\varepsilon^2\tau_1' + \tau_2') + \nu\sigma_{W_2}^2\gamma^2(\tau_1 + \gamma\tau_1')\\ &+ \nu\sigma_{W_2}^4\gamma^2\left((\eta' - \zeta)\tau_1' + \zeta\tau_2'\right)\,, \end{aligned} \tag{26}$$

and the asymptotic test error $E_{test} = \mathbb{E}(y - \hat{y}(\mathbf{x}))^2$ is given by

$$E_{test} = (\gamma\tau_1)^{-2}E_{train} - \sigma_\varepsilon^2\,. \tag{27}$$

[4] For simplicity, we discuss the centered setting with $N_0 = 0$, which captures all of the technical complexities.

Remark 1. *The subtraction of σ_ε^2 in eqn. (27) is because we have assumed that there is no label noise on the test points. Had we included the same label noise on both the training and test distributions, that term would be absent.*

Remark 2. *When $\nu = 0$, the quantity $(\gamma\tau_1)^{-2}E_{train}$ on the right hand side of eqn. (27) is precisely the generalized cross-validation (GCV) metric of (Golub et al., 1979). Theorem 1 shows that the GCV gives the exact asymptotic test error for the problem studied here.*

6. Test Error in Limiting Cases

While the explicit formulas in preceding section provide an exact characterization of the asymptotic training and test loss, they do not readily admit clear interpretations. On the other hand, eqn. (25) and therefore the expressions for E_{test} simplify considerably under several natural limits, which we examine in this section.

6.1. Large Width Limit

Here we examine the test error in the superabundant regime in which the width n_1 is larger than any constant times the dataset size m, which can be obtained by letting $\psi \to 0$ and $\psi/\phi \to 0$. In this setting we find,

$$\begin{aligned} E_{\text{test}}|_{\psi=0} = &\frac{1}{2\phi\chi_0}\left(\chi_0(\phi-1)+\xi\phi(1+\phi)+\rho(1-3\phi)\right) \\ &+\frac{\nu\sigma_{W_2}^2}{2\phi\chi_0}\left((\eta\phi+\zeta)(\rho+\xi\phi)-4\zeta\rho\phi\right) \\ &+\frac{\nu\sigma_{W_2}^2}{2\phi}(\eta\phi-\zeta)+\frac{\phi\xi+\rho-\chi_0}{2\chi_0\text{SNR}}\,, \end{aligned} \tag{28}$$

where $\nu = 0$ with centering and $\nu = 1$ without it and $\rho := \zeta(1+\sigma_{W_2}^2)$, $\xi := \gamma+\eta+\sigma_{W_2}^2\eta'$, and

$$\chi_0 := \sqrt{(\rho+\xi\phi)^2-4\phi\rho^2}\,. \tag{29}$$

The learning curve is remarkably steep with centering. To see this, we expand the result as $m \to \infty$, *i.e.* as $\phi \to 0$,

$$E_{\text{test}}|_{\psi=0} = \begin{cases} \frac{\phi}{\text{SNR}}+\mathcal{O}(\phi^2) & \text{SNR} < \infty \\ (1-\frac{\xi}{\rho})^2\phi^2+\mathcal{O}(\phi^3) & \text{SNR} = \infty \end{cases}. \tag{30}$$

Interestingly, we see that when the network is super abundantly parameterized, we obtain very fast learning curves: for finite SNR, $E_{\text{test}} \sim m^{-1}$, and in the noiseless case $E_{\text{test}} \sim m^{-2}$. See Fig 3(b).

6.2. Small Width Limit

Here we consider the limit in which the width n_1 is smaller than any constant times the dataset size m or the number of features n_0, which can be obtained by letting $\psi \to \infty$ with ϕ held constant. In this setting we find,

$$\begin{aligned} E_{\text{test}}|_{\psi\to\infty} = &\frac{1}{2\phi\chi_1}\left(\chi_1(\phi-1)+\xi_1\phi(1+\phi)+\zeta(1-3\phi)\right) \\ &+\frac{1}{2\chi_1\text{SNR}}\left(\phi\xi_1+\zeta-\chi_1\right)\,, \end{aligned} \tag{31}$$

where $\xi_1 := \eta'+\gamma/\sigma_{W_2}^2$, and

$$\chi_1 := \sqrt{(\zeta+\xi_1\phi)^2-4\phi\zeta^2}\,. \tag{32}$$

The small width limit characterizes one boundary of the abundant parameterization regime and as such provides an upper bound on the test loss in that regime. Therefore, a sufficient condition for the global minimum to occur at intermediate widths is $E_{\text{test}}|_{\psi\to\infty} < E_{\text{test}}|_{\psi=0}$. By comparing eqn. (28) to eqn. (31), precise though unenlightening constraints on the parameters can be derived for satisfying this condition. One such configuration is illustrated in Fig. 4(b).

6.3. Large Dataset Limit

Here we consider the limit in which the dataset m is larger than any constant times the width n_1, which can be obtained by letting $\phi \to 0$ with $\phi/\psi \to 0$. In this setting we find,

$$E_{\text{test}}|_{\phi\to0} = \begin{cases} \frac{1+\psi}{\text{SNR}}(\frac{\phi}{\psi})+\mathcal{O}(\frac{\phi}{\psi})^2 & \text{SNR} < \infty \\ \frac{\tau^2(\nu\zeta^2\sigma_{W_2}^4+\kappa)}{(\eta-\zeta)\zeta^2\sigma_{W_2}^4}(\frac{\phi}{\psi})^2+\mathcal{O}(\frac{\phi}{\psi})^3 & \text{SNR} = \infty \end{cases},$$

where $\nu = 0$ with centering and $\nu = 1$ with without it and,

$$\tau := \gamma+\sigma_{W_2}^2(\eta'-\zeta)\,, \quad \kappa := \zeta\psi+(\eta-\zeta)\psi^2\,. \tag{33}$$

Here again we observe very steep learning curves, similar to the large width limit above.

6.4. Ridgeless Limit: First-Layer Kernel

Here we examine the ridgeless limit $\gamma \to 0$ of the first-layer kernel K_1. We find that the result can be obtained through a degeneration of (28),

$$E_{\text{test}}^{K_1}|_{\gamma=0} = \lim_{\sigma_{W_2}\to\infty} E_{\text{test}}|_{\psi=0} \tag{34}$$

$$\begin{aligned} = &\frac{1}{2\phi\bar\chi}\left(\bar\chi(\phi-1)+\eta'\phi(1+\phi)+\zeta(1-3\phi)\right) \\ &+\frac{1}{2\bar\chi\text{SNR}}\left(\phi\eta'+\zeta-\bar\chi\right)\,, \end{aligned} \tag{35}$$

where, $\bar\chi := \sqrt{(\zeta+\eta'\phi)^2-4\phi\zeta^2}$ and we have specialized to the centered case $\nu = 0$. The expansion as $m \to \infty$ also looks similar to (30) and can be obtained from that equation by substituting $\xi/\rho \to \eta'/\zeta$.

6.5. Ridgeless Limit: Second-Layer Kernel

Here we examine the ridgeless limit $\gamma \to 0$ when the kernel is due to the second-layer weights only, *i.e.* K_2. This limit

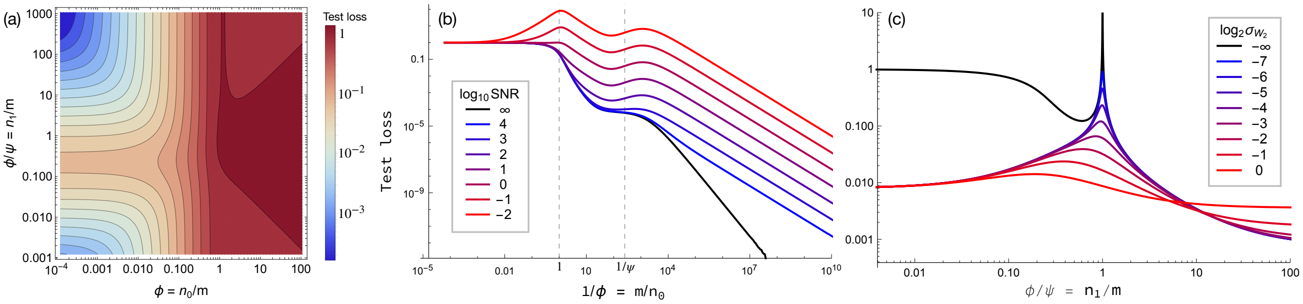

Figure 3. Test error for NTK regression with $\sigma = \tanh$ under various scenarios. (a) Contour plot of the error as a function of $\phi = n_0/m$ and $\phi/\psi = n_1/m$ for $\gamma = 0$ and $\mathrm{SNR} = 1$. The non-monotonic behavior is evident not just in the width n_1, but also in the number of features n_0. (b) Learning curves for the NTK for different signal-to-noise ratios. With no noise (black curve), the error decreases quadratically in the dataset size m, otherwise it decreases linearly. Dashed lines indicate $m = n_0$ and $m = n_1$, where humps emerge for low SNR. (c) Test error as a function of width for various values of σ_{W_2}, which controls the relative contribution of K_1 and K_2. As σ_{W_2} decreases (red to blue), the kernel becomes more like K_2 and the small hump at the quadratic transition increases in size until it resembles the large spike at the linear transition, suggesting that K_2 is responsible for the non-monotonicity in the overparameterized regime.

can be obtained by letting $\sigma_{W_2} \to 0$. In this setting, the result can be expressed as,

$$E_{\text{test}}^{K_2}|_{\gamma=0} = \frac{\phi}{\mathrm{SNR}}\frac{1}{|\phi-\psi|} + \frac{2\omega\zeta-\beta}{2\zeta|\phi-\psi|} + \delta_{\phi>\psi}\left(\frac{\beta-2\chi}{2\chi\mathrm{SNR}} - \frac{\beta(\eta-\zeta)}{2\zeta\chi}\right), \tag{36}$$

where $\omega := \max\{\phi, \psi\}$, $\beta := \zeta + \omega\eta - \chi$, and

$$\chi = \sqrt{(\zeta + 4\omega\eta)^2 - 4\omega\zeta^2}\,, \tag{37}$$

and we have again specialized to the centered case $\nu = 0$. This expression is in agreement with the result presented in (Mei & Montanari, 2019).

When the system is far in the regime of abundant parameterization, namely $p = n_1 \gg m$ (or $\psi/\phi \to 0$), we can examine the large dataset behavior by first sending $\psi \to 0$ and then expanding as $\phi \to 0$. The result is described by (30) by substituting $\xi/\rho \to \eta/\zeta$.

7. Quadratic Overparameterization

In this section, we investigate the implications of our theoretical results about the generalization performance of NTK regression in the quadratic scaling limit $n_0, n_1, m \to \infty$ with $\phi = n_0/m$ and $\psi = n_0/n_1$ held constant. Our high-level observation is that there is complex non-monotonic behavior in this regime as these ratios are varied, and that this behavior can depend on the signal-to-noise ratio and the initial parameter variance $\sigma_{W_2}^2$ in intricate ways. We highlight a few examples in Fig. 3.

In Fig. 3(a), we plot the test error as a function of ϕ and ϕ/ψ, which reveals the behavior of jointly varying the number of features n_0 and the number of hidden units n_1. As expected from Fig. 2(b), for fixed ϕ the test error has a hump near $n_1 = m$. Perhaps unexpectedly, for large n_1, the test loss exhibits non-monotonic dependence on n_0, with a spike near $n_0 = m$. Notice that for small n_1, this non-monotonicity disappears. It is clear that the test error depends in a complex way on both variables, underscoring the richness of the quadratically-overparameterized regime.

Fig. 3(b) shows learning curves for fixed ψ and various values of the SNR. For small enough SNR, there are visible bumps in the vicinity of $m = n_0$ and $m = n_1$ that reveal the existence of regimes in which more training data actually hurts test performance. Note that $n_0 = \Theta(n_1)$ so these two humps are separated by a constant factor, so the presence of two humps in this figure is not evidence of multi-scale behavior, though it surely reflects the complex behavior at the quadratic scale.

It is natural to wonder about the origins of this complex behavior. Can it be attributed to a particular component of the kernel K? We investigate this question in Fig. 3(c), which shows how the test error changes as the relative contributions of the per-layer kernels K_1 and K_2 are varied. By decreasing σ_{W_2}, the contribution of K_1 decreases and the kernel becomes more like K_2, and the small hump at the quadratic transition increases in size until it resembles the large spike at the linear transition (*c.f.* Fig. 2), suggesting that K_2 is in fact responsible for the non-monotonicity in the quadratically-overparameterized regime.

8. Empirical Validation

Our theoretical results establish the existence of nontrivial behavior of the test loss at $p = m$ for the second-layer kernel K_2 and at $p = m^2$ for the full kernel K. While these results are strongly suggestive of multi-scale behavior, they do not prove this behavior exists for a single kernel, nor do they guarantee it will be revealed for finite-size systems, let

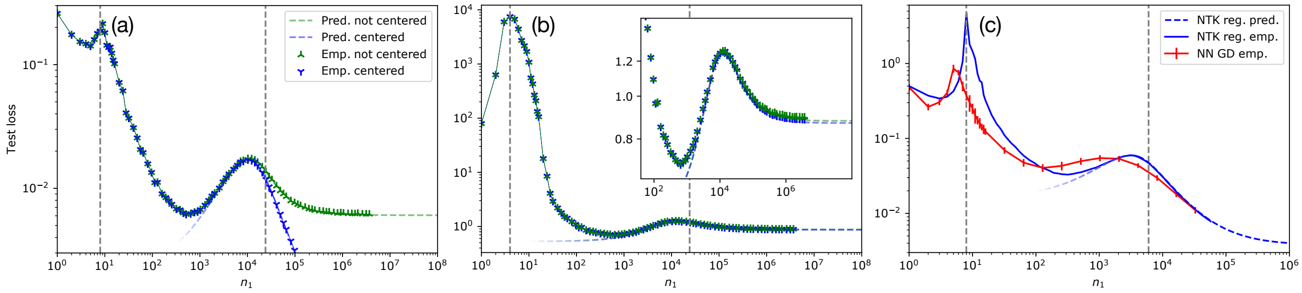

Figure 4. Empirical validation of multi-scale phenomena, triple descent, and the linear and quadratic scaling transitions for kernel regression (a,b,c) and gradient descent (c). All cases show a peak near the linear parameterization transition (first dashed vertical line), as well as a bump near the quadratic transition (second dashed vertical line). Theoretical predictions (dashed blue) agree with kernel regression in their regime of validity (quadratic parameterization). While the global minimum is often at $n_1 = \infty$, it need not be as illustrated in (b). The NTK does not perfectly describe gradient dynamics in high dimensions, so the deviations between the red (GD) and blue (kernel regression) curves in (c) are expected. (a) Mean of five trials with $m = 24000$, $n_0 = 3000$, $\sigma^2_{W_2} = 1/8$, $\sigma^2_\varepsilon = 0$, $\gamma = 10^{-6}$, and $\sigma = \mathrm{erf}$. (b) Mean of five trials with $m = 24000$, $n_0 = 6000$, $\sigma^2_{W_2} = 1/8$, $\sigma^2_\varepsilon = 4$, and $\sigma = c(\mathrm{erf}(6(x+1) + \mathrm{erf}(6(x-1))$ with c chosen so $\zeta = 1/4$. (c) Mean and standard deviation of 20 trials with $m = 6000$, $n_0 = 750$, $\sigma^2_{W_2} = 1/8$, $\sigma^2_\varepsilon = 0$, and $\sigma = \mathrm{ReLU}$.

alone for models trained with gradient descent. Here we provide positive empirical evidence on all counts.

Fig. 4 demonstrates multi-scale phenomena, triple descent, and the linear and quadratic scaling transitions for random feature NTK regression and gradient descent for finite-dimensional systems. The simulations all show a peak near the linear parameterization transition, as well as a bump near the quadratic transition. The asymptotic theoretical predictions agree well with kernel regression in their regime of validity, which is when n_1 is near m. While we found that the global minimum of the test error is often at $p = \infty$, there are some configurations for which the optimal p lies between m and m^2, as illustrated in Fig. 4(b).

Fig. 4(a) clearly shows triple descent for NTK regression and a marked difference in loss with and without centering, suggesting that this source of variance may often dominate the error for large n_1.

Fig. 4(c) confirms the existence of triple descent for a single-layer neural network trained with gradient descent. The noticeable difference between kernel regression and the actual neural network is to be expected because the NTK can change during the course of training when the width is not significantly larger than the dataset size. Indeed, the deviation diminishes for large n_1. In any case, the qualitative behavior is similar across all scales, providing support for the validity of our framework beyond pure kernel methods.

9. Conclusion

In this work, we provided a precise description of the high-dimensional asymptotic generalization performance of kernel regression with the Neural Tangent Kernel of a single-hidden-layer neural network. Our results revealed that the test error has complex non-monotonic behavior deep in the overparameterized regime, indicating that double descent does not always provide an accurate or complete picture of generalization performance. Instead, we argued that the test error may exhibit additional peaks and descents as the number of parameters varies across multiple scales, and we provided empirical evidence of this behavior for kernel ridge regression and for neural networks trained with gradient descent. We conjecture that similar multi-scale phenomena may exist for broader classes of architectures and datasets, but we leave that investigation for future work.

References

Adlam, B., Levinson, J., and Pennington, J. A random matrix perspective on mixtures of nonlinearities for deep learning. *arXiv preprint arXiv:1912.00827*, 2019.

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pp. 8139–8148, 2019.

Belkin, M., Hsu, D. J., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in neural information processing systems*, pp. 2300–2311, 2018a.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549, 2018b.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019b.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611–1619, 2019c.

Benigni, L. and Péché, S. Eigenvalue distribution of nonlinear models of random matrices. *arXiv preprint arXiv:1904.03090*, 2019.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pp. 2933–2943, 2019.

de G. Matthews, A. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1-nGgWC-.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/du19c.html.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

El Karoui, N. et al. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

Erdos, L. The matrix dyson equation and its applications for random matrices. *arXiv preprint arXiv:1903.10060*, 2019.

Far, R. R., Oraby, T., Bryc, W., and Speicher, R. Spectra of large block matrices. *arXiv preprint cs/0610045*, 2006.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., dAscoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Golub, G. H., Heath, M., and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Helton, J. W., Mai, T., and Speicher, R. Applications of realizations (aka linearizations) to free probability. *Journal of Functional Analysis*, 274(1):1–79, 2018.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

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

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pp. 8570–8581, 2019.

Li, Z., Wang, R., Yu, D., Du, S. S., Hu, W., Salakhutdinov, R., and Arora, S. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

Liang, T., Rakhlin, A., and Zhai, X. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. *arXiv preprint arXiv:1908.10292 [cs, math, stat]*, 2020a.

Liang, T., Rakhlin, A., et al. Just interpolate: Kernel ridgeless regression can generalize. *Annals of Statistics*, 48(3): 1329–1347, 2020b.

Louart, C., Liao, Z., Couillet, R., et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Mingo, J. A. and Speicher, R. *Free probability and random matrices*, volume 35. Springer, 2017.

Mitra, P. P. Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*, 2019.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 11611–11622, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019.

Neal, R. M. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Péché, S. et al. A note on the pennington-worah distribution. *Electronic Communications in Probability*, 24, 2019.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, pp. 2637–2646, 2017.

Pennington, J. and Worah, P. The spectrum of the fisher information matrix of a single-hidden-layer neural network. In *Advances in Neural Information Processing Systems*, pp. 5410–5419, 2018.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8 (1-2):1–230, 2015.

Xiao, L., Pennington, J., and Schoenholz, S. S. Disentangling trainability and generalization in deep learning. *arXiv preprint arXiv:1912.13053*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.