Explicit loss asymptotics in the gradient descent training of neural networks

Maksim Velikanov
Skolkovo Institute of Science and Technology
maksim.velikanov@skoltech.ru

Dmitry Yarotsky
Skolkovo Institute of Science and Technology
d.yarotsky@skoltech.ru

Abstract

Current theoretical results on optimization trajectories of neural networks trained by gradient descent typically have the form of rigorous but potentially loose bounds on the loss values. In the present work we take a different approach and show that the learning trajectory of a wide network in a lazy training regime can be characterized by an explicit asymptotic at large training times. Specifically, the leading term in the asymptotic expansion of the loss behaves as a power law $L(t) \sim Ct^{-\xi}$ with exponent ξ expressed only through the data dimension, the smoothness of the activation function, and the class of function being approximated. Our results are based on spectral analysis of the integral operator representing the linearized evolution of a large network trained on the expected loss. Importantly, the techniques we employ do not require a specific form of the data distribution, for example Gaussian, thus making our findings sufficiently universal.

1 Introduction

A major challenge in the research of neural networks is the quantitative theoretical description of their optimization by gradient descent. At present, many aspects of network training seem to be understood rather well on a qualitative level, or admit convincing heuristic explanations, but we seem to lack tools for making reasonably accurate quantitative predictions, even for relatively simple models and data. In this sense, the theory of neural networks compares unfavorably to physics, which is also an application-driven field but with an apparently much more successful penetration of theoretical methods. The main difficulty here is probably the complex structure of the data and models, which are hard to describe in terms of convenient and simple mathematical abstractions.

In recent years, a significant progress in the theoretical analysis of gradient descent of neural networks has been associated with the limit of large networks, which can be studied using various methods from partial differential equations [26, 31], kernel methods [20, 21], spin glass theory [12], random matrix theory [29], dynamical systems [30], and other mathematical fields.

In the present work, we consider a setting of large networks and large, smoothly distributed data sets that allows us to obtain explicit leading terms in the long-term evolution of the loss under gradient descent. We are inspired by the spectral theory of singular integral operators [5], which we apply to the linearized evolution of the network. While this linearized evolution has been widely studied recently, most related research seems to focus on theoretical convergence guarantees and upper bounds for the loss values [3, 27], or on a highly symmetric problems admitting explicit solution [35, 38]. In contrast, we focus on explicit loss evolution formulas, which we find as power laws

$$L(t) \sim Ct^{-\xi}. \tag{1}$$

We argue that the exponents ξ here exhibit some form of universality, in that they are essentially determined by the input dimension d and by the smoothness classes of the activation function and the target function. In particular, we find that in the case of ReLU networks approximating an indicator

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

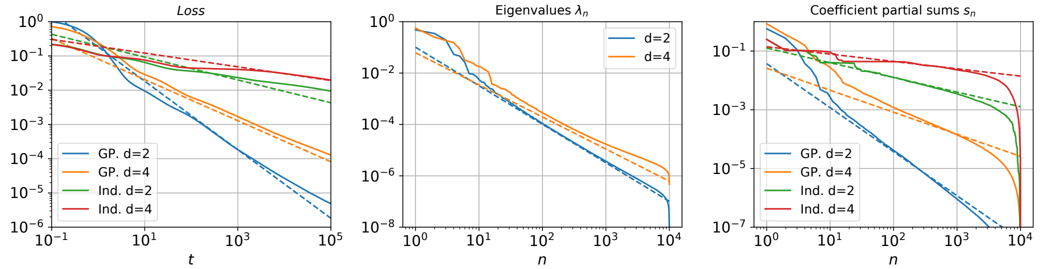


Figure 1: The loss trajectories and spectral properties of the neural tangent kernels of shallow networks in the NTK regime. The target function (i.e., the initial displacement between the network output and the approximated function) is either generated by a Gaussian process (*GP*) modeled by a larger network of the same architecture, or is an indicator function of a d-dimensional ball (*Ind*). The data distributions μ are modeled as mixtures of 8 Gaussian distributions with random centers, and the data dimension is either $d = 2$ or $d = 4$. The **solid** lines show the numerically obtained values, while the **dashed** lines show the respective theoretical power-law asymptotics. The dataset size is $M = 10^4$ (see Section A (SM) for further details of experiments).
Left: Loss evolution for a shallow network with width $N = 3000$. The scaling exponent giving the slope of the theoretical asymptotic is $\xi = \frac{\beta}{d+\alpha} = \frac{3}{d+1}$ for GP and $\xi = \frac{1}{d+\alpha}$ for Ind (see Section 5.2). **Center:** Distribution of the infinite network NTK eigenvalues λ_n. The theoretical scaling exponent is $\nu = 1 + \frac{1}{d}$ (see Section 5.1). **Right:** Distributions of the coefficient partial sums s_n (see Eq. (7)). The theoretical scaling exponent is $\kappa = \frac{\beta}{d} = \frac{3}{d}$ for GP and $\kappa = \frac{1}{d}$ for Ind (see Section 5.2).

function of some region in the d-dimensional space (a classification problem target), the natural value of the exponent is $\xi = \frac{1}{d+1}$. On the other hand, in the case of target functions generated by a randomly initialized wide ReLU network, the exponent is $\xi = \frac{3}{d+1}$. Our approach also allows us to obtain explicit expressions for the coefficient C in these cases.

The power law (1) is established using similar power laws (but with different exponents and coefficients) for the eigenvalues of the evolution operator and for the coefficients in the expansion of the target function over corresponding eigenvectors. These power laws are indeed confirmed by our experiments (see Figure 1).

Our main scenario is approximation by shallow ReLU network in the NTK regime, but we also briefly consider several modifications of this scenario, namely the activation functions $(x_+)^q$ with $q > 0$, approximation by a deep network in the NTK regime, and approximation in the mean field regime.

2 Related work

The approximation of linearized network evolution and its applications were studied in many works, see in particular [21, 10, 23, 22, 19]. The role of the eigenvalues and eigenvectors of the NTK in the linearized network was emphasized in [3], where the GD dynamic of the finite network was linked to the dynamic of its infinite width counterpart, determined by spectral properties of the corresponding NTK. In subsequent works, the NTK spectrum was central for description of network training [27] and generalization [7, 4]. These papers use the assumption of power law NTK spectrum, but justify it empirically or for highly symmetric problems.

Because of the importance of the NTK spectrum, a number of works focused on its description in different settings. We first mention directions and settings that are different from ours. The case of very deep networks was studied in [37, 18, 17, 2]. This analysis is relying on convergence of the NTK to simple fixed points in the limit of infinite depth. [17] also studied the choice of activation function, in particular its smoothness. Another line of research [1, 13] uses techniques from Random Matrix Theory to analyze the setting where the dataset size M goes to infinity together with data dimension d and layer widths n_l.

In this work we consider the setting of fixed data dimension and effectively infinite network width and dataset size. In this case [35] showed that the network evolution can be described by a deterministic integral operator that is easier to analyze than a large but finite matrix. Also, in that paper and in the papers [38, 8, 27], the integral operator was explicitly diagonalized in the special case of uniform distribution on a sphere.

A power law upper bound for the training loss was obtained in [27], but with an exponent ξ smaller than ours. The paper [4] describes a different power law, relating the test loss at the end of training to the dataset size and the network width. This result also relies on a power law assumption for the NTK spectrum.

Another related line of research is the case of univariate functions. The gradient descent evolution of univariate shallow networks has been analytically studied in the papers [39, 36].

3 Asymptotic evolution of the loss function

We consider a linearized training of a neural network by gradient descent. Such linear approximations arise naturally in various "lazy training" scenarios [10]. Consider the standard quadratic loss function $L(\mathbf{W}) = \frac{1}{2}\int_{\mathbb{R}^d} |\widetilde{f}(\mathbf{W},\mathbf{x})-f(\mathbf{x})|^2\mu(\mathbf{x})d\mathbf{x}$, where $\mathbf{x}$ is the d-dimensional input, f is the approximated function, $\widetilde{f}$ is the network, $\mathbf{W}$ are the network weights, and μ is the data distribution on which the network is trained. Gradient descent can be written as the differential equation $\frac{d}{dt}\mathbf{W} = -\nabla_{\mathbf{W}} L(\mathbf{W})$. We assume that the weight vector $\mathbf{W}$ is close to a global minimum $\mathbf{W}_*$ where $L(\mathbf{W}_*) = 0$ and $\widetilde{f}(\mathbf{W}_*, \mathbf{x}) = f(\mathbf{x})$ μ-a.e., and that the evolution equation can be linearized at $\mathbf{W} = \mathbf{W}_*$. It is convenient to write this linearized equation in terms of the difference $\delta f(\mathbf{x}) = \widetilde{f}(\mathbf{W},\mathbf{x}) - f(\mathbf{x})$ between the current output and the target. The corresponding linear equation is $\frac{d}{dt}\delta f = -\mathcal{A}\delta f$, where $\mathcal{A}$ can be written as the integral operator $\mathcal{A}\delta f(\mathbf{x}) = \int_{\mathbb{R}^d}\Theta(\mathbf{x},\mathbf{x}')\mu(\mathbf{x}')\delta f(\mathbf{x}')d\mathbf{x}'$ with the NTK (neural tangent kernel)

$$\Theta(\mathbf{x},\mathbf{x}') = \nabla_{\mathbf{W}}\widetilde{f}(\mathbf{W}_*,\mathbf{x})^T\nabla_{\mathbf{W}}\widetilde{f}(\mathbf{W}_*,\mathbf{x}') \tag{2}$$

The evolution operator $\mathcal{A}$ is a symmetric non-negative definite operator with respect to the scalar product $\langle f_1, f_2\rangle_\mu = \int f_1(\mathbf{x})f_2(\mathbf{x})\mu(\mathbf{x})d\mathbf{x}$. By multiplying functions f by $\mu^{1/2}$, the operator $\mathcal{A}$ can be brought to the form $\widetilde{\mathcal{A}} = \mu^{1/2}\mathcal{A}\mu^{-1/2}$ with a symmetric kernel,

$$\widetilde{\mathcal{A}}\delta f(\mathbf{x}) = \int_{\mathbb{R}^d}\mu^{1/2}(\mathbf{x})\Theta(\mathbf{x},\mathbf{x}')\mu^{1/2}(\mathbf{x}')\delta f(\mathbf{x}')d\mathbf{x}'. \tag{3}$$

In this form $\widetilde{\mathcal{A}}$ is symmetric w.r.t. the usual scalar product $\langle f_1, f_2\rangle = \int f_1(\mathbf{x})f_2(\mathbf{x})d\mathbf{x}$. Observe that the loss at time t can be written as $L(t) = \frac{1}{2}\|g_t\|^2$, where $g_t = e^{-t\widetilde{\mathcal{A}}}g$, the norm $\|\cdot\|$ corresponds to the scalar product $\langle\cdot,\cdot\rangle$, and the function g is given by

$$g(\mathbf{x}) = \mu^{1/2}(\mathbf{x})(\widetilde{f}(\mathbf{w}(t=0),\mathbf{x}) - f(\mathbf{x})). \tag{4}$$

We can now describe the evolution of the loss by diagonalizing the operator $\widetilde{\mathcal{A}}$. We will be interested in the scenario where μ is a smooth function that is compactly supported or falls off at infinity sufficiently fast. (In particular, in the context of a finite training set this means that this set is large enough to be legitimately approximated by μ.) In this case, for typical kernels Θ, the operator $\widetilde{\mathcal{A}}$ will have a discrete spectrum with eigenvalues converging to 0. Let λ_n denote the eigenvalues of $\widetilde{\mathcal{A}}$ sorted in decreasing order, and let c_n be the respective coefficients in the expansion of the initial error g (given by Eq. (4)) over the normalized orthogonal eigenvectors. Then, the loss evolves by

$$L(t) = \frac{1}{2}\sum_{n=0}^{\infty} e^{-2\lambda_n t}|c_n|^2. \tag{5}$$

To compute the asymptotic of $L(t)$ at large times t, we need to know the distribution of the eigenvalues λ_n and the coefficients c_n at large n. The key assumption of our work (verified later for certain scenarios) is that these distributions have power law forms. Specifically, regarding the eigenvalues λ_n we assume that

$$\lambda_n \sim \Lambda n^{-\nu} \tag{6}$$

with some coefficient Λ and exponent ν. Regarding the coefficients c_n, we assume that they also have a power law distribution on a large length scale in n, but possibly deviate from this law locally (e.g., due to oscillations). For this reason, we describe their large n behavior by the partial sums

$$s_n \equiv \sum_{k\geq n} |c_k|^2. \tag{7}$$

We then assume that, for some coefficient K and exponent κ,

$$s_n \sim Kn^{-\kappa}. \tag{8}$$

Under assumption of the power laws (6) and (8), it is easy to check (see SM, Section B) that the loss also has a power law asymptotic (1) with the constant C and exponent ξ expressible through the constants Λ, K and exponents ν, κ :

$$L(t) \sim \tfrac{K}{2}\Gamma\left(\tfrac{\kappa}{\nu}+1\right)(2\Lambda t)^{-\frac{\kappa}{\nu}}\,, \tag{9}$$

where $\Gamma(z)$ is the Gamma function.

In Figure 1 we illustrate this approach to the long-term loss evolution with several examples of target functions having different smoothness and dimension and, as a result, exhibiting different exponents.

In the remainder of the paper we show that the power laws (6) and (8), and hence the large-t asymptotic (9) of the loss, are indeed valid for some natural network training scenarios. The asymptotic (6) of the eigenvalues is primarily determined by the singularities of the kernel Θ. These singularities can be explicitly described for shallow neural network with piecewise smooth activations such as ReLU. Then, the power law (6) can be derived from general results on integral operators with singular kernels. In particular, in the case of ReLU we find that $\nu = 1 + \frac{1}{d}$.

The asymptotic (8) of the coefficients is more subtle, as it depends significantly on the class of the initial error function g (which in turn depends, by Eq. (4), on the target function f and the initial approximation $\widetilde{f}$). We derive this power law for one natural class of discontinuous functions g, and for functions g that are realizations of a Gaussian process of a particular "roughness".

Moreover, for these two classes, we also find an explicit form of the coefficient C appearing in loss asymptotic (1). This requires us, however, to modify the above derivation of the loss asymptotic (9) by what can be called "integrated localization". Roughly speaking, in the large-t limit we can think of the eigenvectors of $\widetilde{\mathcal{A}}$ as infinitesimally localized in $\mathbb{R}^d$. We then apply the above derivation of Eq. (9) not to the full set of eigenvectors, but separately to each infinitesimal sub-domain, and then integrate the results. See details in Sections 5.2 and D (SM).

The subsequent exposition is structured as follows. In Section 4 we provide the background on the NTK kernel and spectral properties of singular integral operators. Then, in Section 5, we derive the loss asymptotic (1) for our main setting – the NTK training with a shallow ReLU network. After that, in Section 6 we consider various modifications of this setting: other activation functions (Section 6.1), deep networks (Section 6.2), and training in the mean field regime (Section 6.3).

4 Background

4.1 Infinitely wide networks

Lazy training scenarios discussed in section 3 naturally arise for the networks in the limit of infinite width. However, there are several ways to scale parameters of the network with width, which lead to different operating regimes of infinitely wide networks [15, 16].

NTK regime. The first option we consider is the NTK regime [20], for which the NTK Θ defined in (2) is deterministic and constant during training. This simplification immediately leads to a linear dynamic in the space of network outputs. We consider feed-forward networks parametrized as

$$\begin{cases} z_j^1 = \sum_{i=1}^d \sigma_w w_{ij}^1 x_i + \sigma_b b_j^1 \\ z_j^l = \sum_{i=1}^{n_{l-1}} \frac{\sigma_w}{\sqrt{n_{l-1}}} w_{ij}^l x_i^{l-1} + \sigma_b b_j^l, \quad l>1 \\ x_j^l = \phi(z_j^l) \end{cases} \tag{10}$$

Here n^l is the width of layer l, x_i is a network input and z_j^L is the network output. We also consider the last layer without bias term b_j^L and having width $n_L = 1$ (scalar output). Trainable parameters $w_{ij}^l,\ b_j^l$ are initialized as i.i.d. normal Gaussians.

The output of each layer l at initialisation is a Gaussian process with covariance $\langle z_j^l(\mathbf{x}) z_{j'}^l(\mathbf{x}')\rangle = \delta_{jj'}\Sigma^{(l)}(\mathbf{x}, \mathbf{x}')$. By introducing the NTK's $\Theta^{(l)}(\mathbf{x}, \mathbf{x}')$ of intermediate layers, one can recursively compute [21] the both NTK and covariance

$$\begin{cases} \Sigma^{(l+1)} = \sigma_w^2 \big\langle \phi(z^l)\phi(z'^l) \big\rangle + \sigma_b^2 \\ \Theta^{(l+1)} = \Sigma^{(l+1)} + \sigma_w^2 \Theta^{(l)} \big\langle \dot\phi(z^l)\dot\phi(z'^l) \big\rangle \end{cases} \tag{11}$$

Here $z^l = z^l(\mathbf{x})$ and $z'^l = z^l(\mathbf{x}')$ are draws from the Gaussian process with covariance $\Sigma^{(l)}(\mathbf{x}, \mathbf{x}')$, and $\langle \ldots \rangle$ is the averaging. It is also convenient to parametrize the covariance at input points $\mathbf{x}, \mathbf{x}'$ in some layer l as

$$\Sigma^{(l)}(\mathbf{x}, \mathbf{x}') = \begin{pmatrix} r_l^2 & r_l r_l' \cos\varphi_l \\ r_l r_l' \cos\varphi_l & r_l'^2 \end{pmatrix} \tag{12}$$

To analyze the spectrum of the evolution operator $\widetilde{\mathcal{A}}$ given by (3) we will intensively use the explicit form of the NTK and covariance. For the ReLU activation $\phi(z) = (z)_+$ the averages in (11) can be computed analytically [11]. In the case of shallow network ($L = 2$) the result is:

$$\Sigma(\mathbf{x}, \mathbf{x}') = \tfrac{\sigma_w^2}{2\pi} r r' \big(\sin\varphi + \cos\varphi(\pi - \varphi)\big) \tag{13}$$

$$\Theta(\mathbf{x}, \mathbf{x}') = \Sigma(\mathbf{x}, \mathbf{x}') + \tfrac{\sigma_w^2}{2\pi} r r' \cos\varphi(\pi - \varphi) \tag{14}$$

Here r, r', φ are parameters from (12) with $l = 1$ (we dropped index 1). They have a clear interpretation in terms of extended input vectors $\tilde{\mathbf{x}} = (\sigma_w \mathbf{x}, \sigma_b) \in \mathbb{R}^{d+1}$. Specifically, $r = \|\tilde{\mathbf{x}}\|,\ r' = \|\tilde{\mathbf{x}'}\|$ and φ is the angle between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}'}$.

Mean Field regime. This operating regime of infinitely wide networks is naturally defined [26, 31] for shallow networks of the form $f(\mathbf{W}, \mathbf{x}) = \frac{1}{N}\sum_{i=1}^N c_i \phi(\mathbf{w}_i \cdot \mathbf{x} + b_i) = \frac{1}{N}\sum_{i=1}^N \tilde\phi(\tilde{\mathbf{w}}_i, \mathbf{x})$. Here $\tilde{\mathbf{w}}_i = (c_i, \mathbf{w}_i, b_i)$ denotes the collection of parameters, associated with a single neuron. In the infinite width limit $N \to \infty$ the evolution is described by a PDE on parameter density distribution $p(\tilde{\mathbf{w}})$:

$$\partial_t p = \nabla_{\tilde{\mathbf{w}}} \Big[p \nabla_{\tilde{\mathbf{w}}} \int K(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}')(p(\tilde{\mathbf{w}}') - p_\infty(\tilde{\mathbf{w}}')) d\tilde{\mathbf{w}}' \Big], \quad K(\tilde{\mathbf{w}}, \tilde{\mathbf{w}}') = \int \tilde\phi(\tilde{\mathbf{w}}, \mathbf{x}) \mu(\mathbf{x}) \tilde\phi(\tilde{\mathbf{w}}', \mathbf{x}) d\mathbf{x}$$

Under mild assumptions [9], solution of the MF equation converges to the global optimum $p_\infty(\tilde{\mathbf{w}}')$. At large times t the MF equation can be linearized [39] around p_∞, thus bringing the network to the lazy training regime discussed in section 3. The network NTK in this case is equal to

$$\Theta(\mathbf{x}, \mathbf{x}') = \int \nabla_{\tilde{\mathbf{w}}} \tilde\phi(\tilde{\mathbf{w}}, \mathbf{x}) p_\infty(\tilde{\mathbf{w}}) \nabla_{\tilde{\mathbf{w}}} \tilde\phi(\tilde{\mathbf{w}}, \mathbf{x}') d\tilde{\mathbf{w}}. \tag{15}$$

4.2 Singular integral operators

Consider the evolution operator $\widetilde{\mathcal{A}}$ given by Eq. (3). Under our assumptions, the kernel $\mu^{1/2}(\mathbf{x})\Theta(\mathbf{x}, \mathbf{x}')\mu^{1/2}(\mathbf{x}')$ of this operator quickly falls off at infinity and is smooth outside the diagonal $\mathbf{x} = \mathbf{x}'$, but, as we will see later, has a homogeneous singularity on this diagonal. In this setting, a general theory developed in [5] allows us to obtain the asymptotic distribution (6) of the eigenvalues with explicit constants Λ and ν.

Specifically, suppose that in a neighborhood of the diagonal the kernel Θ has a representation

$$\Theta(\mathbf{x}, \mathbf{x}') = \theta_{\mathbf{x}}(\mathbf{x} - \mathbf{x}') + \ldots, \tag{16}$$

where $\theta_{\mathbf{x}}(\cdot)$ is a (possibly $\mathbf{x}$-dependent) even ($\theta_{\mathbf{x}}(\mathbf{z}) = \theta_{\mathbf{x}}(-\mathbf{z})$) homogeneous function of degree α:

$$\theta_{\mathbf{x}}(c\mathbf{z}) = |c|^\alpha \theta_{\mathbf{x}}(\mathbf{z}), \tag{17}$$

and the dots $\ldots$ denote terms of higher smoothness.

Let N_λ denote the number of eigenvalues of $\widetilde{\mathcal{A}}$ greater than λ. Then, it is shown in [5] that for small λ, the leading term of N_λ is given by

$$N_\lambda \sim \Big(\int \gamma_{\mathbf{x}} \mu^{\frac{d}{d+\alpha}}(\mathbf{x}) d\mathbf{x}\Big) \lambda^{-\frac{d}{d+\alpha}}. \tag{18}$$

Here, $\gamma_{\mathbf{x}} = (2\pi)^{-d}|\{\mathbf{k} \in \mathbb{R}^d : \widetilde{\theta}_{\mathbf{x}}(\mathbf{k}) > 1\}|$, where $|\cdot|$ denotes the Lebesgue measure and $\widetilde{\theta}_{\mathbf{x}}$ is a suitably defined Fourier transform of the homogeneous function $\theta_{\mathbf{x}}$ (see SM, Section C).

Formula (18) can be derived as follows. Divide the domain $\mathbb{R}^d$ into multiple small subsets Ω_m, and think of $\widetilde{\mathcal{A}}$ as an operator matrix corresponding to the decomposition $L^2(\mathbb{R}^d) = \oplus_m L^2(\Omega_m)$. Using the fall off and smoothness of the kernel outside the diagonal, one can show that the leading term of N_λ is determined only by the diagonal elements of this operator matrix. Then, the leading term can be found by considering each restriction $\widetilde{\mathcal{A}}|_{L^2(\Omega_m)}$ separately and summing the respective contributions to N_λ, i.e., $N_\lambda(\widetilde{\mathcal{A}}) \sim \sum_m N_\lambda(\widetilde{\mathcal{A}}|_{L^2(\Omega_m)})$. If we decrease the size of each Ω_m by a factor M, then the number of terms in this sum increases M^d-fold, but at the same time each $N_\lambda(\widetilde{\mathcal{A}}|_{L^2(\Omega_m)})$ decreases also roughly M^d-fold, due to rescaling of eigenvalues. In the limit of infinitely small subsets Ω_m, the operator $\widetilde{\mathcal{A}}|_{L^2(\Omega_m)}$ can be approximated by the convolution with the homogeneous function $\theta_{\mathbf{x}}$. The asymptotic form of its eigenvalues can then be written in terms of the Fourier transform $\widetilde{\theta}_{\mathbf{x}}$ as given above. The power $\frac{d}{d+\alpha}$ in Eq. (18) can be deduced by observing that the volume of the $\mathbf{k}$-space corresponding to $\widetilde{\theta}_{\mathbf{x}} > \lambda$ scales as $\lambda^{-\frac{d}{d+\alpha}}$.

The formula (18) can be translated into the power law (6) by inverting the relation between λ and n (note that $N_{\lambda_n} = n$ for any n). Specifically, we find that the law (6) holds with

$$\nu = 1 + \frac{\alpha}{d}, \quad \Lambda = \Big(\int \gamma_{\mathbf{x}} \mu^{1/\nu}(\mathbf{x}) d\mathbf{x}\Big)^{\nu}. \tag{19}$$

In Section 5.2 we will show that this approach can be extended to yield the loss asymptotic (1).

5 Asymptotic analysis of wide networks

5.1 NTK operators and their singularities

In this section we demonstrate the asymptotic law (6) of the NTK spectrum as well as outputs covariance spectrum, and find constants ν, Λ . We consider a shallow ReLU network in the NTK regime and with the data distribution $\mu(\mathbf{x})$ as described in Section 3. The NTK $\Theta(\mathbf{x}, \mathbf{x}')$ and outputs covariance $\Sigma(\mathbf{x}, \mathbf{x}')$ of such network is given by (14), (13) with r, r' and φ explicitly depending on input points $\mathbf{x}, \mathbf{x}'$:

$$r(\mathbf{x}) = \sqrt{\sigma_w^2 |\mathbf{x}|^2 + \sigma_b^2}, \quad \varphi(\mathbf{x}, \mathbf{x}') = \arccos\big(\tfrac{\sigma_w^2 \mathbf{x}\cdot\mathbf{x}' + \sigma_b^2}{r(\mathbf{x}) r(\mathbf{x}')}\big). \tag{20}$$

To use the spectral theory described in section 4.2 we analyze the smoothness of the kernels Θ and Σ. Firstly, they are smooth (infinitely differentiable) functions of r, r', φ on the whole domain. Now suppose that the bias term is not absent: $\sigma_b > 0$. Then $r(\mathbf{x})$ is a smooth function for all $\mathbf{x} \in \mathbb{R}^d$. The argument of arccos in (20) is also smooth everywhere, but $\arccos(z)$ itself is smooth on $(-1, 1)$ and has divergent derivative at the end points $z = 1, -1$ corresponding to $\varphi = 0, \pi$. We see that condition $\sigma_b > 0$ implies that the case $\varphi = \pi$ is never realized, while $\varphi(\mathbf{x}, \mathbf{x}') = 0$ at all coinciding inputs. Thus we established that $\Theta(\mathbf{x}, \mathbf{x}')$ and $\Sigma(\mathbf{x}, \mathbf{x}')$ are smooth everywhere except the diagonal $\mathbf{x} = \mathbf{x}'$, where they might have a singularity.

To analyze the behavior at the diagonal, we first expand φ in Eq. (20) for small $\delta\mathbf{x} = \mathbf{x} - \mathbf{x}'$:

$$\varphi(\mathbf{x}, \mathbf{x}') = \tfrac{\sigma_w \sqrt{r^2(\mathbf{x}) - \sigma_w^2 |\mathbf{x}|^2 \cos^2\psi}}{r^2(\mathbf{x})} |\delta\mathbf{x}| + O(|\delta\mathbf{x}|^2) \tag{21}$$

Here ψ is the angle between $\delta\mathbf{x}$ and $\mathbf{x}$. If $\sigma_b > 0$, then the expression under the square root is always positive, therefore the angle $\varphi(\mathbf{x}, \mathbf{x}')$ has a homogeneous singularity of degree 1 on the diagonal. Now we expand expressions (13) and (14) for small φ and find that both NTK and covariance indeed have a singularity on the diagonal with leading singular terms.

$$\Sigma_{\text{sing}}(\mathbf{x}, \mathbf{x}') = \tfrac{\sigma_w^2}{2\pi} r^2(\mathbf{x}) \tfrac{1}{3} \varphi^3(\mathbf{x}, \mathbf{x}') \propto |\delta\mathbf{x}|^3 \tag{22}$$

$$\Theta_{\text{sing}}(\mathbf{x}, \mathbf{x}') = -\tfrac{\sigma_w^2}{2\pi} r^2(\mathbf{x}) \varphi(\mathbf{x}, \mathbf{x}') \propto |\delta\mathbf{x}| \tag{23}$$

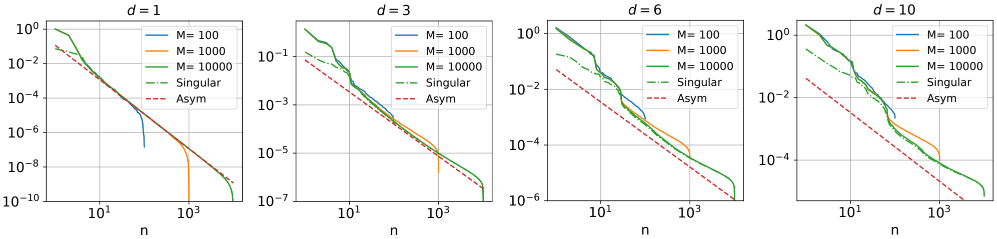


Figure 2: NTK eigenvalues λ_n for networks with different input dimension d and for different data set sizes M. *Asym* shows the theoretical power law (6) with parameters Λ, ν derived in Section 5.1; in particular, $\nu = 1 + \frac{1}{d}$. *Singular* corresponds to eigenvalues of the singular part of NTK (23). We see that for moderately big n the singular part has the same eigenvalues as the full NTK. Observe that the number n at which the spectrum converges to its asymptotic form increases with dimension d.

Now we see that both NTK Θ and covariance Σ have the form needed for application of spectral theory discussed in section 4.2. For NTK we have singularity degree $\alpha = 1$, which by Eq. (19) leads to the already announced exponent $\nu = 1 + \frac{1}{d}$ (see Figure 1). Interestingly, the singularity degree for the covariance is higher (namely, 3), resulting in a faster fall off of the corresponding eigenvalues. In the sequel (see Section 5.2), this latter degree will appear in the analysis of loss asymptotic for target functions generated by the neural network Gaussian process, and will be denoted by β.

In the case of NTK $\Theta(\mathbf{x}, \mathbf{x}')$ the singularity is essentially given by (21). The corresponding Fourier transform $\widetilde{\theta}_{\mathbf{x}}$ and the function $\gamma_{\mathbf{x}}$ can be calculated analytically (see Section C). This leads to an explicit expression for the constant Λ:

$$\Lambda = \tfrac{\sigma_w^3 \sigma_b^{\frac{1}{d}}}{(2\pi)^2} \Gamma(\tfrac{d+1}{2}) \Gamma(\tfrac{d}{2}+1)^{-(1+\frac{1}{d})} \Big\langle \mu(\mathbf{x})^{-\frac{1}{d+1}} (r(\mathbf{x}))^{\frac{d-1}{d+1}} \Big\rangle_{\mu}^{1+\frac{1}{d}}, \tag{24}$$

where $\langle u(\mathbf{x}) \rangle_\mu$ denotes the integral $\int \mu(\mathbf{x}) u(\mathbf{x}) d\mathbf{x}$. In the experiment, the value $\langle u(\mathbf{x}) \rangle_\mu$ can be computed by averaging u over a μ-distributed data set. In Figure 2 we compare theoretical and numerical NTK eigenvalue distributions for several dimensions d and data set sizes M.

5.2 The loss function

We extend now the arguments of Section 4.2 to derive the loss asymptotic (1). We consider two classes of functions g representing the initial error (4).

Scenario 1: a discontinuous function g. We assume that g is supported on a bounded subset $\Omega \subset \mathbb{R}^d$ with a smooth boundary $\partial\Omega$ so that g has a discontinuity on this boundary but is smooth inside Ω. An obvious example of g is the indicator function of Ω. In this scenario we obtain

$$L(t) \sim \int_{\partial\Omega} |\Delta g(\mathbf{x})|^2 (\mu(\mathbf{x}) \widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{1}{d+\alpha}} dS \cdot \tfrac{1}{2\pi} \Gamma(\tfrac{1}{d+\alpha}+1) \cdot (2t)^{-\frac{1}{d+\alpha}}. \tag{25}$$

Here, $\mathbf{x}$ is the point on the surface $\partial\Omega$, $\mathbf{n}$ is the unit normal to $\partial\Omega$, and $\Delta g(\mathbf{x})$ is the size of the jump of g at $\mathbf{x}$, given by the limit of $g(\mathbf{y})$ as $\mathbf{y}$ approaches $\mathbf{x}$ from inside Ω. The eigenvalue, coefficient, and loss exponents in this scenario are, respectively,

$$\kappa = \tfrac{1}{d}, \quad \nu = 1 + \tfrac{\alpha}{d}, \quad \xi = \tfrac{\kappa}{\nu} = \tfrac{1}{d+\alpha}. \tag{26}$$

Scenario 2: g generated by a Gaussian process. Suppose that g is a realization of a Gaussian process with a covariance matrix $\Sigma(\mathbf{x}, \mathbf{x}') = \langle g(\mathbf{x}) g(\mathbf{x}') \rangle$ and that Σ has a homogeneous singularity $\zeta_{\mathbf{x}}$ of degree β at the diagonal (in the sense of Eqs. (16),(17)). In this scenario we find

$$L(t) \sim \int_{\mathbb{R}^d} \int_{|\mathbf{n}|=1} \widetilde{\zeta}_{\mathbf{x}}(\mathbf{n}) (\mu(\mathbf{x}) \widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{\beta}{d+\alpha}} d\mathbf{x} dS \cdot \tfrac{1}{2(2\pi)^d \beta} \Gamma(\tfrac{\beta}{d+\alpha}+1) \cdot (2t)^{-\frac{\beta}{d+\alpha}}. \tag{27}$$

Here, $\widetilde{\zeta}_{\mathbf{x}}$ is the Fourier transform of $\zeta_{\mathbf{x}}$ (defined similarly to the Fourier transform $\widetilde{\theta}_{\mathbf{x}}$). The eigenvalue, coefficient, and loss exponents in this scenario are, respectively,

$$\kappa = \tfrac{\beta}{d}, \quad \nu = 1 + \tfrac{\alpha}{d}, \quad \xi = \tfrac{\kappa}{\nu} = \tfrac{\beta}{d+\alpha}. \tag{28}$$

In our experiments we model GP by a large network in the NTK regime. The corresponding covariance is analyzed in Section 5.1 and has the singularity degree $\beta = 3$.

We provide the full derivations of Eqs. (25) and (27) in Section D of SM, and sketch now the main ideas. Our general strategy is to complement the localized eigenvalue analysis of Section 4.2 by the analysis of expansion coefficients. Note, however, that the simple approximation of $\widetilde{\mathcal{A}}$ by a direct sum of localized operators $\widetilde{\mathcal{A}}|_{L^2(\Omega_m)}$ (as performed in [5] and sketched in Section 4.2) is "too rough" for the study of expansion coefficients (due to a stronger effect of boundary conditions in each Ω_m). Accordingly, we replace this approximation by the short-time Fourier transform of g:

$$F(\mathbf{y},\mathbf{k}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} g(\mathbf{x})\omega(\mathbf{x}-\mathbf{y})e^{-i\mathbf{k}\cdot\mathbf{x}}d\mathbf{x},$$

where ω is a window function such that $\int \omega^2 = 1$. The coefficient $F(\mathbf{y},\mathbf{k})$ describes the component of g having the wave number $\mathbf{k}$ and localized at $\mathbf{y}$. Then at large t, using the stationary phase method,

$$g_t(\mathbf{x}) = e^{-t\widetilde{\mathcal{A}}}g(\mathbf{x}) \sim (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} F(\mathbf{y},\mathbf{k})e^{-t\mu_{\mathbf{x}}\widetilde{\theta}_{\mathbf{x}}(\mathbf{k})}\omega(\mathbf{x}-\mathbf{y})e^{i\mathbf{k}\cdot\mathbf{x}}d\mathbf{y}d\mathbf{k}.$$

The leading contribution to this integral comes from large $\mathbf{k}$. For such $\mathbf{k}$, we can write the coefficients $F(\mathbf{y},\mathbf{k})$ in an asymptotic form, primarily determined by the singularities of g. By integrating out $\mathbf{y}$ and $\mathbf{k}$, we then arrive at the desired Eqs. (25) and (27).

The above argument establishes the loss asymptotic (1) while bypassing the computation of the asymptotic (8) of the expansion coefficients aligned with the sorted "global" eigenvalues. This latter asymptotic (including the coefficient K) can be found by a similar computation, see Section D.4.

6 Extensions

The power law asymptotic of eigenvalues obtained in section 5.1 was based on the analysis of diagonal singularity of the NTK in the setting of ReLU activation, shallow depth 2, and the NTK regime. We argue now that our general approach is not restricted to this narrow setting. To show this, we separately consider several modifications of the network from Section 5.1 and an application to MNIST (see Figure 3). In this section we mostly limit ourselves to describing final results, with the derivations provided in SM, Section E.

6.1 Activations of different smoothness

Let's consider a shallow network in NTK regime with activation function $\phi_q(z) = (z)_+^q,\ q > 0$ ("a ReLU with the altered smoothness q"). Similarly to the ReLU case, one can show that NTK in current setting has a singularity on the diagonal for all values of q except half-integers $q = \frac{1}{2}, \frac{3}{2}, \ldots$. The leading singular term is

$$\Theta_q(\mathbf{x},\mathbf{x}') = \frac{\sigma_w^2}{2\pi} r^q r'^q a_q \varphi^{2q-1} \tag{29}$$

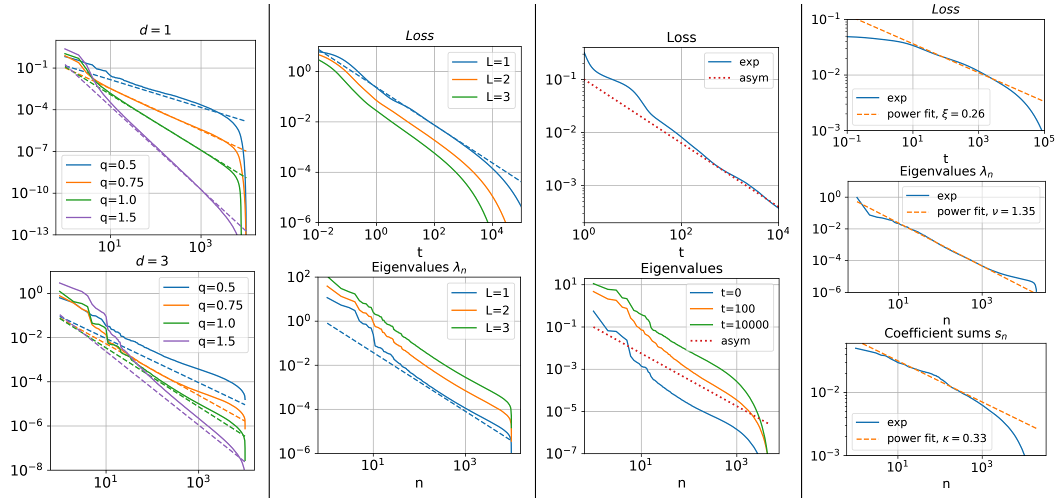
$$a_q = \frac{\Gamma^2(q)\Gamma(\frac{1}{2}-q)}{\sqrt{\pi}2^q} \tag{30}$$

Here r, r', φ are the same as in section 5.

The singularity of NTK with degree $\alpha = 2q-1$ implies the eigenvalue power law asymptotic (6) with the exponent $\nu_q = 1 + \frac{2q-1}{d}$. Thus, the singularity degree in the NTK is determined by the singularity degree of the activation function. The coefficient Λ_q can be explicitly computed as

$$\Lambda_q = \sigma_w^{\alpha+2}\sigma_b^{\frac{\alpha}{d}} q^2 (2\pi)^{d+q-2} \frac{\Gamma\big(\frac{d+\alpha}{2}\big)\Gamma^2(q)}{\big(\Gamma(\frac{d}{2}+1)\big)^{\frac{d+\alpha}{d}}} \Big\langle \mu(\mathbf{x})^{-\frac{\alpha}{d+\alpha}} r(\mathbf{x})^{\frac{2d-\alpha d-\alpha}{d+\alpha}} \Big\rangle_\mu^{\frac{d+\alpha}{d}} \tag{31}$$

In Figure 3a we compare the theoretical and numerical eigenvalue distributions for several values of d and q.



(a) Activations $(z)_+^q$.

(b) Deep networks

(c) Mean field

(d) MNIST

Figure 3: Extensions to other activations, deep networks, mean field regime, and MNIST.
(a): NTK eigenvalues for activation functions $\phi(z) = (z)_+^q$ with varying smoothness q. The theoretical distributions (dashed lines) have the exponents $\nu_q = 1 + \frac{2q-1}{d}$ (see Section 6.1).
(b): Loss evolution and NTK eigenvalue distribution for networks with varying number of hidden layers L trained on 3-dimensional data and target generated by network GP ($\beta = 3$). In agreement with theory, the numerical results (solid lines) show the scaling exponents $\xi = \frac{3}{d+1}$ and $\nu = 1 + \frac{1}{d}$ for all depths L (see Section 6.2). The theoretical predictions (dashed lines) are only shown for $L = 1$ (for other L the theoretical lines would have the same slopes but different positions determined by the coefficients C and Λ; due to computation complexity, we have found C and Λ only for $L = 1$).
(c): Loss dynamic and eigenvalue distribution at different moments of time for a network in the MF regime (Section 6.3). The network approximates GP on inputs with dimension $d = 4$. *Top:* comparison of the theoretical and experimental loss evolution. The theoretical exponent $\xi = \frac{\beta}{d+1} = \frac{3}{5}$. *Bottom:* the theoretical eigenvalue distribution and its experimental counterparts at different times. The theoretical exponent is $\nu = \frac{d+1}{d} = \frac{5}{4}$.
(d): Loss evolution, NTK eigenvalues λ_n and coefficients s_n for MNIST. All three curves are reasonably well approximated by power laws with exponents $\xi \approx 0.26$, $\nu \approx 1.35$, $\kappa \approx 0.33$ so that $\xi \approx \frac{\kappa}{\nu}$ in agreement with our theory. At the same time, our current theory does not predict specific values for the exponents ν and κ because the intrinsic dimension of the MNIST manifold is very different from the large dimension of the ambient space. See SM, Section E for a detailed discussion.

6.2 Deep networks

We consider now a network of arbitrary depth $L > 2$ in the NTK regime and with the ReLU activation function. Similarly to the shallow ReLU case, the angles φ_l (see Eq. (12)) are singular on the diagonal: $\varphi_l \propto |\mathbf{x} - \mathbf{x}'|$. Using relations (11) one can obtain recursive relations for φ_l, r_l and finally for the singular part of NTK $\Theta_{\text{sing}}^{(l)}$

$$\begin{aligned}\Theta_{\text{diag}}^{(l+1)} &= r_{l+1}^2 + \frac{\sigma_w^2}{2}\Theta_{\text{diag}}^{(l)} \\ \Theta_{\text{sing}}^{(l+1)} &= -\frac{1}{2\pi}\Theta_{\text{diag}}^{(l)}\varphi_l + \frac{\sigma_w^2}{2}\Theta_{\text{sing}}^{(l)}\end{aligned} \tag{32}$$

Here $\Theta_{\text{diag}}^{(l+1)}$ is the value of NTK on the diagonal. Since $\Theta_{\text{sing}}^{(1)} = 0$, we can see from (32) that $\Theta_{\text{sing}}^{(L)}$ is a weighted sum of $-\varphi_l$ for $l = 1, \ldots, L-1$. Thus, the singularity degree of the NTK is $\alpha = 1$ and the eigenvalue power law (6) holds with $\nu = 1 + \frac{1}{d}$. However, obtaining explicit formula for Λ is harder for this case and we leave it for future work. In Figure 3b we compare the theoretical and numerical eigenvalue distributions for NTK's of deep networks.

6.3 MF regime

In this section we consider a shallow network in the MF regime and with the ReLU activation function, with the NTK given by (15). The neural tangent kernel (14) of the shallow NTK network can be obtained from (15) with the distribution $p_\infty(c, \mathbf{w}, b)$ taken as a product of Gaussians for each variable. Therefore, neural tangent kernels in MF regime represent a broader class of kernels, containing our basic example (14). It turns out that the diagonal singularity is present for all sufficiently smooth and quickly decaying distributions p_∞ of this broader class:

$$\Theta_{\text{sing}}(\mathbf{x}, \mathbf{x}') = -\frac{\sqrt{1+|\mathbf{x}|^2}}{2} \int\limits_{\mathbf{w}\cdot\mathbf{x}=-b} \big|\mathbf{w}\cdot\delta\mathbf{x}\big| dp_{\infty,2}(\mathbf{w}, b)$$

Here $\delta\mathbf{x} = \mathbf{x} - \mathbf{x}'$ and $p_{\infty,2}$ is the second moment of the distribution p_∞ w.r.t. the variable c, i.e. $p_{\infty,2}(\mathbf{w}, b) = \int c^2 p_\infty(c, \mathbf{w}, b)dc$. Thus, the eigenvalue power law exponent is the same as in the NTK regime: $\nu = 1 + \frac{1}{d}$.

In Figure 3c we show the loss dynamic and the eigenvalue distribution at different moments of time for a network in the MF regime. Since the neural tangent kernel significantly changes during the training of the MF network, picking NTK's at different moments of training provides sufficiently distinct and general kernels of the form (15).

7 Discussion

The main results of the paper are theoretical derivations of power laws for the NTK spectrum and the loss asymptotic. Spectral properties of neural networks is an active area of current research [24, 34, 25, 28]. It has been observed (e.g., [32, 33, 14]) that, close to a global minimum, the Hessian of the loss of neural networks in practical tasks typically has eigenvalues accumulating near 0. This implies, in particular, that the Hessian is ill-conditioned, which has profound implications for network training: convergence of gradient descent in this case is much slower than for well-conditioned optimization problems. A power law spectral distribution is assumed in some recent works (e.g., for the study of generalization in [6]), but we are not aware of any previous general theoretical derivations of such a law. Our work essentially provides such a derivation (since at the global minimum the NTK spectrum coincides with the Hessian spectrum up to the eigenvalue 0). In a sense, we prove theoretically (under some reasonable assumptions) that a slow (power-law) convergence of the gradient descent is inevitable. Moreover, we precisely quantify this convergence.

Our work provides simple explicit formulas for the asymptotic eigenvalue distribution and loss evolution. This allows, in principle, to make some quantitative predictions about optimization not feasible with theoretical methods only providing upper bounds. For example, the loss exponent ξ allows to directly estimate the number of GD steps needed to decrease the loss by one order of magnitude. In principle, the full loss asymptotic (25) can be used to analyze the effect of different data distributions $\mu(x)$ on convergence speed of GD.

The key idea of our asymptotic analysis is the extraction of the diagonal singularity from the NTK kernel. This is a new and promising way to relatively easily obtain information about the model spectrum: it only requires simple Taylor expansion and, as we show, is applicable in a wide range of scenarios, extending to deep networks, the mean field regime, and different activation functions. We expect this operation to be important for future studies of network training.

Our loss exponents $\xi = \frac{1}{\alpha+d}$ and $\xi = \frac{\beta}{\alpha+d}$ tend to 0 as the dimension d increases. However, this does not contradict the practical trainability of networks for high-dimensional tasks such as image recognition, since real images occupy only tiny (i.e., effectively low-dimensional) subsets in the ambient high-dimensional space. While our current theory does not predict specific exponents in this setting, we do empirically observe power laws for eigenvalues, coefficients and the loss, and the fitted exponents satisfy our theoretical relation $\xi = \frac{\kappa}{\nu}$ (see Figure 3d). This suggests that our method can be extended to this setting, which will be the topic of a future work.

Funding transparency statement

Research was supported by Russian Science Foundation, grant 21-11-00373.

References

[1] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 74–84. PMLR, 13–18 Jul 2020.

[2] Naman Agarwal, Pranjal Awasthi, and Satyen Kale. A deep conditioning treatment of neural networks. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 249–305. PMLR, 16–19 Mar 2021.

[3] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 09–15 Jun 2019.

[4] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

[5] M Š Birman and M Z Solomjak. Asymptotic behavior of the spectrum of weakly polar integral operators. *Mathematics of the USSR-Izvestiya*, 4(5):1151–1168, oct 1970.

[6] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

[7] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *arXiv preprint arXiv:2006.13198*, 2021.

[8] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1606.05340*, 2020.

[9] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[10] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

[11] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 342–350. Curran Associates, Inc., 2009.

[12] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.

[13] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020.

[14] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

[15] Eugene Golikov. Towards a general theory of infinite-width limits of neural classifiers. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3617–3626. PMLR, 13–18 Jul 2020.

[16] Eugene A. Golikov. Dynamically stable infinite-width limits of neural classifiers. *arXiv preprint arXiv:2006.06574*, 2020.

[17] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2672–2680. PMLR, 09–15 Jun 2019.

[18] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. Mean-field behaviour of neural tangent kernel for deep neural networks. *arXiv preprint arXiv:1905.13654*, 2021.

[19] Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17116–17128. Curran Associates, Inc., 2020.

[20] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[21] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[22] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[23] Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15954–15964. Curran Associates, Inc., 2020.

[24] Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

[25] Charles H Martin, Tongsu Serena Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):1–13, 2021.

[26] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[27] Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2021.

[28] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. *arXiv preprint arXiv:2102.04396*, 2021.

[29] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pages 2798–2806. PMLR, 2017.

[30] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *arXiv preprint arXiv:1606.05340*, 2016.

[31] Grant M Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

[32] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[33] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[34] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[35] Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. In H. Wallach, H. Larochelle, A. Beygelzimer, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[36] Francis Williams, Matthew Trager, Claudio Silva, Daniele Panozzo, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[37] Lechao Xiao, Jeffrey Pennington, and Samuel Schoenholz. Disentangling trainability and generalization in deep neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10462–10472. PMLR, 13–18 Jul 2020.

[38] Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. In *Eighth International Conference on Learning Representations (ICLR 2020)*, September 2019. Submitted to ICLR 2020.

[39] Dmitry Yarotsky. Collective evolution of weights in wide neural networks. *arXiv preprint arXiv:1810.03974*, 2018.

Supplementary material

A Details of experiments

In this work we have two types of experiments. Both types operate with dataset consisting of M data samples drawn from some distribution $\mu(\mathbf{x})$.

In our experiments we focus on distributions μ different from simple standard distributions such us spherical Gaussian or uniform in the cube. Highly symmetric distributions such as the spherical Gaussian make the whole problem analytically solvable (see, for example, [38]). In contrast, our theory do not rely on the symmetry, and we test it on distributions μ without any symmetry. In all our experiments, the distribution $\mu(\mathbf{x})$ is constructed as follows: we randomly choose n_g points in the cube $[-1, 1]^d$ and consider n_g symmetric Gaussian distributions with centers in these points and standard deviation σ; then $\mu(\mathbf{x})$ is defined as the average of these n_g Gaussian distributions. Although each separate Gaussian distribution is symmetric, the average of $n_g > d$ randomly located Gaussian distributions almost surely removes all symmetries w.r.t. orthogonal transformations of $\mathbb{R}^d$. The typical values used in our experiments are $n_g = 8$ and $\sigma = 0.5$.

In the first type of experiments we analytically calculate the NTK using, for example, expression (14), and then numerically diagonalize the corresponding matrix and decompose a target function over its eigenvectors. Then the linear evolution $e^{-t\widetilde{\mathcal{A}}}$ can be easily computed in obtained eigenbasis. The center and right parts of figure 1 as well as figures 2,3a,3b correspond to this type of experiments. Thus, it can be considered as an experiment with infinitely wide network, but a finite dataset. In all such experiments we take the largest M possible, which is bounded by $O(M^3)$ time cost of numerical diagonalization and $O(M^2)$ memory cost of storing the NTK. The typical value used in experiments is $M = 10000$. To calculate the constant Λ in the eigenvalue asymptotics (24) and (E.10) we draw another, rather big, set of points from the distribution $\mu(\mathbf{x})$ and use this dataset for Monte Carlo estimation of the $\langle\cdot\rangle_\mu$ averages in (24) and (E.10).

In the second type of experiments we initialize and train actual wide network (typical width $N = 3000$). The left part of figure 1 and figure 3c correspond to this type of experiments. To reach large values of time t we choose the learning rate η close to its critical value $\eta_c = \frac{2}{\lambda_0}$, above which the dynamic in the 0'th eigenspace start to diverge exponentially and the network leaves the regime of approximately constant NTK. The network in the MF regime adapts to learning rates higher than critical one at initialization, but adaptation resource is limited. Overall, experiments of this type test our theoretical predictions for roughly practical sizes of networks and datasets.

In the experiments we considered two types of target functions. The first is a draw from a Gaussian process, which we model by a very wide $N = 10^6$ shallow network with NTK parametrization. Thus, the covariance of GP is given by (13). To calculate the coefficient C in the loss asymptotic (1) we use Eq. (27) with the sphere integral taken analytically as shown in (D.18). The second type of target is an indicator of a ball of some radius r. It corresponds to two-class classification task with first class located in $|\mathbf{x}| < r$ and the second class in $|\mathbf{x}| > r$. To calculate coefficient C we sampled points uniformly on a sphere $|\mathbf{x}| = r$ and used them to calculate a Monte Carlo estimate of the integral in (25). In principle, one can choose classes with more sophisticated separation boundary, but numerical calculation of the integral in (25) will be more complicated. Note also that even if the target has a spherical symmetry, the whole problem does not, because we use a non-symmetric μ.

B Derivation of the loss asymptotic from the asymptotics of the eigenvalues λ_n and the expansion coefficients c_n

In this section we prove the loss asymptotic (9). This result is established under assumption of power law asymptotics (6), (8) for the eigenvalues λ_n and partial sums of coefficients $s_n = \sum_{k\geq n} |c_k|^2$, i.e.

$$\begin{aligned} \lambda_n &\sim \Lambda n^{-\nu}, \\ s_n &\sim K n^{-\kappa}. \end{aligned} \tag{B.1}$$

Here the asymptotic similarity sign $\sim$ denotes $a_n \sim b_n \iff a_n = b_n(1 + o(1))$.

Theorem 1. *Under the assumptions* (B.1) *on the asymptotic of eigenvalues ad coefficients, the loss* $L(t) = \frac{1}{2}\sum_n e^{-2\lambda_n t}|c_n|^2$ *has the asymptotic*

$$L(t) \sim \frac{K}{2}\Gamma\left(\frac{\kappa}{\nu}+1\right)(2\Lambda t)^{-\frac{\kappa}{\nu}} \tag{B.2}$$

Proof. The constant Λ enters the loss only in combination with t. Thus, by rescaling time and noticing that the loss is proportional to K, it is sufficient to consider the case $K = 2$ and $2\Lambda = 1$. In other words, we have to prove

$$\sum_{n=0}^{\infty} e^{-tn^{-\nu}(1+u_n)}(s_n - s_{n+1}) \sim \Gamma\left(\frac{\kappa}{\nu}+1\right)t^{-\frac{\kappa}{\nu}}, \qquad \text{with} \quad s_n = n^{-\kappa}(1+v_n) \tag{B.3}$$

Here $\lim_{n\to\infty} u_n = \lim_{n\to\infty} v_n = 0$ due to asymptotic (B.1).

The idea of the proof is that in the region of small n the sum can be neglected due to exponential factor $e^{-t\#}$, while in the region of large n the sum can be replaced by the integral, since the sum argument slowly depend on n in this region. In fact both regions greatly overlap, and we can find a common point n_t inside both regions. Such common point can be taken as $n_t = \lfloor t^\beta \rfloor$ with any β from the interval $(\frac{1}{\nu+1}, \frac{1}{\nu})$ (the reason will be seen later). Let's denote

$$u_t \equiv \sup_{n\geq n_t} |u_n|, \qquad v_t \equiv \sup_{n\geq n_t} |v_n| \tag{B.4}$$

Since $u_n, v_n \to 0$ at $n \to \infty$ and $n_t \to \infty$ at $t \to \infty$, we have $u_t, v_t \to 0$ at $t \to \infty$. The strategy of the proof is first to bound the sum for $n \leq n_t$, then calculate the sum for $n > n_t$ with $u_n = v_n = 0$, and finally add corrections from u_n and v_n.

The sum over $n \leq n_t$ is bounded as

$$\begin{aligned} \sum_{n=0}^{n_t} e^{-2t\lambda_n}(s_n - s_{n+1}) &\leq e^{-2t\lambda_{n_t}} \sum_{n=0}^{n_t}(s_n - s_{n+1}) \\ &\leq \exp\left(-tn_t^{-\nu}(1-u_t)\right) s_0 \leq \exp\left(-t^{1-\beta\nu}(1-u_t)\right) s_0 \end{aligned} \tag{B.5}$$

Since $\beta < \frac{1}{\nu}$ we have $1 - \beta\nu > 0$ and the sum goes to 0 exponentially fast as $t \to \infty$.

Now we calculate the sum over $n > n_t$ with $u_n = v_n = 0$. Due to convexity of $f(x) = x^{-\alpha}$ for all $\alpha > 0$ we have the bounds

$$\kappa(n+1)^{-\kappa-1} \leq s_n - s_{n+1} \leq \kappa n^{-\kappa-1} \tag{B.6}$$

Then we approximate the sum with the integral as

$$\begin{aligned} &\sum_{n>n_t} e^{-tn^{-\nu}}\left(n^{-\kappa} - (n+1)^{-\kappa}\right) \overset{(1)}{=} \int_{n_t}^{\infty} e^{-tx^{-\nu}}\big(1 + O(tx^{-\nu-1})\big)\kappa x^{-\kappa-1}\big(1 + O(x^{-1})\big)dx \\ &\overset{(2)}{=} \int_0^{tn_t^{-\nu}} e^{-z}\kappa\left(\frac{t}{z}\right)^{-\frac{\kappa+1}{\nu}}\frac{1}{t\nu}\left(\frac{t}{z}\right)^{1+\frac{1}{\nu}} dz\Big(1 + O(tn_t^{-\nu-1}) + O(n_t^{-1})\Big) \\ &= \frac{\kappa}{\nu}t^{-\frac{\kappa}{\nu}}\left(\int_0^{\infty} z^{\frac{\kappa}{\nu}-1}e^{-z}dz + O\Big(\exp\big(-tn_t^{-\nu}\big)\big[tn_t^{-\nu}\big]^{\frac{\kappa}{\nu}-1}\Big) + O(tn_t^{-\nu-1}) + O(n_t^{-1})\right) \\ &\overset{(3)}{=} \Gamma\Big(\frac{\kappa}{\nu}+1\Big)t^{-\frac{\kappa}{\nu}}\Big(1 + O\Big(\exp\big(-t^{1-\beta\nu}\big)\big[t^{(1-\beta\nu)}\big]^{\frac{\kappa}{\nu}-1}\Big) + O\big(t^{1-\beta(\nu+1)}\big) + O\big(t^{-\beta}\big)\Big) \end{aligned} \tag{B.7}$$

Here in (1) we used (B.6) to estimate the difference $s_n - s_{n+1}$, and then first order Taylor expansion to estimate value of integrated function at non-integer points. In (2) we made a change of variables $z = tx^{-\nu}$ and estimated "big O" terms using minimum value $x_{\text{min}} = n_t$. In (3) used the definition of Gamma function and recursive relation $z\Gamma(z) = \Gamma(z+1)$, and substituted n_t. $\beta \in (\frac{1}{\nu+1}, \frac{1}{\nu})$ implies that $1 - \beta\nu > 0$ and $1 - \beta(\nu+1) < 0$, therefore all "big O" terms go to 0 as $t \to \infty$.

The last step is to include back u_n and v_n into sum over $n > n_t$. To include u_n we use that $n^{-\nu}(1-u_t) \leq 2\lambda_n \leq n^{-\nu}(1+u_t)$. Then we make lower/upper bounds for the sum with u_n by using the result of calculation (B.7) with substitution $t \to t(1 \pm u_t)$. Finally, we bound contribution from v_n as

$$\begin{aligned}
&\sum_{n>n_t} e^{-2t\lambda_n}(v_n n^{-\kappa} - v_{n+1}(n+1)^{-\kappa}) \\
&\overset{(1)}{=} \sum_{n>n_t} \big(e^{-2t\lambda_n} - e^{-2t\lambda_{n-1}}\big) v_n n^{-\kappa} + e^{-2t\lambda_{n_t}} v_{n_t+1}(n_t+1)^{-\kappa} \\
&\overset{(2)}{\leq} v_t \sum_{n>n_t} e^{-2t\lambda_n}(n^{-\kappa} - (n+1)^{-\kappa}) + e^{-2t\lambda_{n_t}}(v_{n_t+1} - v_t)(n_t+1)^{-\kappa}
\end{aligned} \tag{B.8}$$

Here in (1) we regrouped the terms in the sum, while in (2) we used $v_n \overset{n>n_t}{\leq} v_t$ and regrouped summation terms back into original form. The second term in the last line is negative and the first term is v_t times the result of (B.7). The final expression with all contributions is

$$\begin{aligned}
&\sum_{n=0}^{\infty} e^{-2t\lambda_n}(s_n - s_{n+1}) = \Gamma\Big(\frac{\kappa}{\nu} + 1\Big) t^{-\frac{\kappa}{\nu}} \times \\
&\times \Big(1 + O\Big(\exp\big(-t^{1-\beta\nu}\big)\big[t^{(1-\beta\nu)}\big]^{\max(\frac{\kappa}{\nu}-1,0)}\Big) + O\big(t^{1-\beta(\nu+1)}\big) + O(u_t) + O(v_t)\Big)
\end{aligned} \tag{B.9}$$

Here all "big O" terms vanish in the limit $t \to \infty$ and we obtain desired answer. □

C Calculation of $\gamma_{\mathbf{x}}$

Following [5], given an even homogeneous function $\theta_{\mathbf{x}} : \mathbb{R}^d \to \mathbb{R}$ of degree α, we define its Fourier transform $\widetilde{\theta}_{\mathbf{x}}$ using the Riesz summation formula:

$$\widetilde{\theta}_{\mathbf{x}}(\mathbf{k}) = \lim_{r\to\infty} \int_{\|\mathbf{x}\|<r} (1 - \tfrac{|\mathbf{x}|^2}{r^2})^c \theta(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d\mathbf{x}. \tag{C.1}$$

For sufficiently large c the limit exists and is independent of c, and the resulting function $\widetilde{\theta}_{\mathbf{x}}$ is a homogeneous function of degree $-(d+\alpha)$.

According to formula (19) from the main text, to calculate the coefficient Λ in the eigenvalue asymptotic one needs to find the volume $\gamma_{\mathbf{x}}$ of the region defined using $\widetilde{\theta}_{\mathbf{x}}(\mathbf{k})$:

$$\gamma_{\mathbf{x}} = (2\pi)^{-d} |\{\mathbf{k} \in \mathbb{R}^d : \widetilde{\theta}_{\mathbf{x}}(\mathbf{k}) > 1\}|. \tag{C.2}$$

In this section we calculate γ_x in the case of singularities $\theta_{\mathbf{x}}(\mathbf{z}) = |\mathbf{z}|^\alpha$ and $\theta_{\mathbf{x}}(\mathbf{z}) = \varphi^\alpha(\mathbf{x}, \mathbf{x}+\mathbf{z})$ with $\varphi(\mathbf{x}, \mathbf{x}')$ defined as in the main text. The latter case is needed for obtaining coefficients Λ in Eqs. (24) and (E.10).

Case $\theta_{\mathbf{x}}(\mathbf{z}) = |\mathbf{z}|^\alpha$. We drop index $\mathbf{x}$ since there is no dependence on it. The homogeneity of $\theta(\mathbf{x})$ with degree α implies the homogeneity of $\widetilde{\theta}(\mathbf{k})$ with degree $-\alpha - d$. This can be seen from definition (C.1) by making integration variable change $\mathbf{x} \to c\mathbf{x},\ c > 0$. Then, due to spherical symmetry of $\theta(\mathbf{z}) = |\mathbf{z}|^\alpha$, its Fourier transform has a form

$$\widetilde{\theta}(\mathbf{k}) = c_{d,\alpha} |\mathbf{k}|^{-d-\alpha} \tag{C.3}$$

To determine the coefficient $c_{d,\alpha}$ we take into account that both $\theta(\mathbf{z})$ and $\widetilde{\theta}(\mathbf{k})$ are generalized functions acting on some test functions χ. We denote the Fourier transform by F and action of generalized functions on test functions by $\langle \cdot, \cdot \rangle$ (not to be confused with averaging in Eqs. (24) and (E.10)). Then by taking a test function $\chi(\mathbf{k}) = e^{-|\mathbf{k}|^2/2}$ and its Fourier transform $\widetilde{\chi}(\mathbf{z}) = (2\pi)^{d/2} e^{-|\mathbf{z}|^2/2}$ we get

$$\begin{aligned} \langle F(\theta), \chi \rangle &= c_{d,\alpha} \int |\mathbf{k}|^{-d-\alpha} e^{-\frac{|\mathbf{k}|^2}{2}} d\mathbf{k} = c_{d,\alpha} S_{d-1} \int\limits_0^\infty k^{-1-\alpha} e^{-\frac{k^2}{2}} dk \\ &= c_{d,\alpha} S_{d-1} 2^{-\frac{2+\alpha}{2}} \int\limits_0^\infty e^{-x} x^{-\frac{2+\alpha}{2}} dx = c_{d,\alpha} S_{d-1} 2^{-\frac{2+\alpha}{2}} \Gamma\Big(-\frac{\alpha}{2}\Big) \end{aligned} \tag{C.4}$$

$$\begin{aligned} \langle \theta, F(\chi) \rangle &= (2\pi)^{\frac{d}{2}} \int |\mathbf{z}|^\alpha e^{-\frac{|\mathbf{z}|^2}{2}} d\mathbf{z} = (2\pi)^{\frac{d}{2}} S_{d-1} \int\limits_0^\infty z^{d-1+\alpha} e^{-\frac{z^2}{2}} dz \\ &= (2\pi)^{\frac{d}{2}} S_{d-1} 2^{\frac{d-2+\alpha}{2}} \int\limits_0^\infty e^{-x} x^{\frac{d-2+\alpha}{2}} dx = (2\pi)^{\frac{d}{2}} S_{d-1} 2^{\frac{d-2+\alpha}{2}} \Gamma\Big(\frac{d+\alpha}{2}\Big) \end{aligned} \tag{C.5}$$

Here in both calculations we first integrated over $d-1$-dimensional sphere with the area S_{d-1}. Then we changed variables so that the integral can be expressed in terms of Gamma function. Since $\langle F(\theta), \chi \rangle = \langle \theta, F(\chi) \rangle$ we compare expressions (C.4) and (C.5) and find

$$c_{d,\alpha} = 2^{d+\alpha} \pi^{\frac{d}{2}} \frac{\Gamma\big(\frac{d+\alpha}{2}\big)}{\Gamma\big(-\frac{\alpha}{2}\big)} \tag{C.6}$$

To find γ we notice that the volume in (C.2) is a ball with radius $c_{d,\alpha}^{1/(d+\alpha)}$. By using a volume of a unit ball in d-dimensional space $B_d = \pi^{d/2} / \Gamma\big(\frac{d}{2}+1\big)$ we get

$$\gamma = \frac{1}{\Gamma\big(\frac{d}{2}+1\big)} \left[\frac{\Gamma\big(\frac{d+\alpha}{2}\big)}{\pi^{\frac{\alpha}{2}} \Big|\Gamma\big(-\frac{\alpha}{2}\big)\Big|} \right]^{\frac{d}{d+\alpha}} \equiv \gamma_{d,\alpha} \tag{C.7}$$

Case $\theta_{\mathbf{x}}(\mathbf{z}) = \varphi^\alpha(\mathbf{x}, \mathbf{x}+\mathbf{z})$. To see that angle $\varphi(\mathbf{x}, \mathbf{x}')$ has a singularity at $\mathbf{x} = \mathbf{x}'$ we write the scalar product of $\widetilde{\mathbf{x}}, \widetilde{\mathbf{x}}'$ as

$$\cos\varphi(\mathbf{x}, \mathbf{x}')\sqrt{r(\mathbf{x})r(\mathbf{x}')} = \sigma_w^2 \mathbf{x}\cdot\mathbf{x}' + \sigma_b^2 \tag{C.8}$$

with $r^2(\mathbf{x}) = \sigma_w^2|\mathbf{x}|^2 + \sigma_b^2$ as in the paper. Expanding this expression at $\varphi = 0$ and $\delta\mathbf{x} = \mathbf{x} - \mathbf{x}' = 0$ we get

$$\varphi(\mathbf{x}, \mathbf{x}') = \sqrt{1 - \frac{\sigma_w^2|\mathbf{x}|^2}{r^2(\mathbf{x})}\cos^2\psi}\ \frac{\sigma_w|\delta\mathbf{x}|}{r(\mathbf{x})} + O\big(|\delta\mathbf{x}|^2\big), \tag{C.9}$$

which is the expression (21) from the main text. In the asymptotic analysis we need to consider only the leading singular term, therefore the homogeneous singularity has the form

$$\theta_{\mathbf{x}}(\mathbf{z}) = a(\mathbf{x})\left(\sqrt{1 - b(\mathbf{x})\frac{z_1^2}{|\mathbf{z}|^2}}\ |\mathbf{z}|\right)^\alpha \tag{C.10}$$

Here $a(\mathbf{x}) = (\sigma_w/r(\mathbf{x}))^\alpha$ and $b(\mathbf{x}) = \sigma_w^2|\mathbf{x}|^2/r^2(\mathbf{x})$ are introduced for convenience and we also omit $\mathbf{x}$ for the rest of this section. In (C.10) we oriented basis in $\mathbf{z}$ space so that the first axis is parallel to vector $\mathbf{x}$. Now we calculate Fourier transform

$$\begin{aligned}\widetilde{\theta}(\mathbf{k}) = \int \theta(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{z}}d\mathbf{z} = a\int\Big[(1-b)z_1^2 + z_2^2 + \ldots + z_d^2\Big]^{\frac{\alpha}{2}} e^{-i\mathbf{k}\cdot\mathbf{z}}d\mathbf{z}\\ \overset{(1)}{=} \frac{a}{\sqrt{1-b}}\int|\mathbf{z}'|^\alpha e^{-i\mathbf{k}'\cdot\mathbf{z}'}d\mathbf{z}' = \frac{a}{\sqrt{1-b}}c_{d,\alpha}|\mathbf{k}'|^{-d-\alpha}\end{aligned} \tag{C.11}$$

Here in (1) we changed to variables $\mathbf{z}', \mathbf{k}'$ which are the same as $\mathbf{z}, \mathbf{k}$ except the first dimension: $z_1' = z_1\sqrt{1-b}$ and $k_1' = k_1/\sqrt{1-b}$. In the original $\mathbf{k}$ space the equation $c|\mathbf{k}'|^{-d-\alpha} = 1$ defines an ellipsoid obtained from the sphere $c|\mathbf{k}|^{-d-\alpha} = 1$ by squeezing the first axis by the factor $\sqrt{1-b}$. This gives us the formula for volume γ

$$\gamma = (2\pi)^{-d}B_d\left(\frac{c_{d,\alpha}a}{\sqrt{1-b}}\right)^{\frac{d}{d+\alpha}}\sqrt{1-b} = \gamma_{d,\alpha}a^{\frac{d}{d+\alpha}}\left(\sqrt{1-b}\right)^{\frac{\alpha}{d+\alpha}} \tag{C.12}$$

Restoring $\mathbf{x}$ dependence and using $\sqrt{1-b(\mathbf{x})} = \sigma_b/r(\mathbf{x})$ we get

$$\gamma_{\mathbf{x}} = \gamma_{d,\alpha}\sigma_w^{\frac{\alpha d}{d+\alpha}}\sigma_b^{\frac{\alpha}{d+\alpha}}r(\mathbf{x})^{-\frac{\alpha d+\alpha}{d+\alpha}} \tag{C.13}$$

Case $\theta_{\mathbf{x}}(\mathbf{z}) = A(\mathbf{x})\varphi^\alpha(\mathbf{x}, \mathbf{x}+\mathbf{z})$. This case includes the singularities of shallow network NTK (23) and covariance (22), and can be easily obtained from the previous one. Since Fourier transformation (C.1) and volume calculation (C.2) are performed locally, $A(\mathbf{x})$ is effectively constant in these calculations. Thus $\gamma_{\mathbf{x}}$ from (C.13) is simply multiplied by $|A(\mathbf{x})|^{\frac{d}{d+\alpha}}$. The result is

$$\gamma_{\mathbf{x}} = |A(\mathbf{x})|^{\frac{d}{d+\alpha}}\gamma_{d,\alpha}\sigma_w^{\frac{\alpha d}{d+\alpha}}\sigma_b^{\frac{\alpha}{d+\alpha}}r(\mathbf{x})^{-\frac{\alpha d+\alpha}{d+\alpha}} \tag{C.14}$$

Equation (24) can be obtained by using $A(\mathbf{x}) = -\frac{\sigma_w^2 r(\mathbf{x})^2}{2\pi}$ from (23) and then substituting resulting $\gamma_{\mathbf{x}}$ in (19).

D Derivation of the loss asymptotic for singular evolution operators $\widetilde{\mathcal{A}}$ and specific classes of target functions g

This section expands the content of Section 5.2 of the main text. Our goal is to derive the explicit leading terms in the asymptotic of the loss $L(t) = \frac{1}{2}\|g_t\|^2$ for the evolution $g_t = e^{-t\widetilde{\mathcal{A}}}g$ when g belongs to one of the following two classes:

1. The function g is supported on a bounded subset Ω with a smooth boundary $\partial\Omega$ so that g has a discontinuity on this boundary but is smooth inside Ω. An obvious example of g is the indicator function of Ω.
2. The function g is a realization of a Gaussian process with a particular singular covariance.

Before providing these derivations, let us first recall our general setting. We consider the evolution $g_t = e^{-t\widetilde{\mathcal{A}}}g$ governed by the non-negative definite generator

$$\widetilde{\mathcal{A}}g(\mathbf{x}) = \int_{\mathbb{R}^d} \mu^{1/2}(\mathbf{x})\Theta(\mathbf{x},\mathbf{x}')\mu^{1/2}(\mathbf{x}')g(\mathbf{x}')d\mathbf{x}' \tag{D.1}$$

(see Section 3 of the main text). Here, μ is the measure corresponding to the data distribution, and Θ is the kernel associated with the neural network ansatz. This symmetric form of the evolution operator is valid in the representation in which the original functions are multiplied by $\mu^{1/2}$. Accordingly, the function g appearing in the loss formula $L(t) = \frac{1}{2}\|e^{-t\widetilde{\mathcal{A}}}g\|^2$ is given by

$$g(\mathbf{x}) = \mu^{1/2}(\mathbf{x})(\widetilde{f}(\mathbf{w}(t=0),\mathbf{x}) - f(\mathbf{x})),$$

where f is the function to be fitted by the network, and $\widetilde{f}(\mathbf{w}(t=0),\mathbf{x})$ is the initial network approximation.[1]

In scenario 1 above – a function g supported on a domain Ω and discontinuous on the boundary $\partial\Omega$ – we will show that the large-t asymptotic of the loss is given by

$$L(t) \sim \frac{1}{2\pi}\Gamma\left(\frac{1}{d+\alpha}+1\right)\int_{\partial\Omega}|\Delta g(\mathbf{x})|^2(\mu(\mathbf{x})\widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{1}{d+\alpha}}dS\cdot(2t)^{-\frac{1}{d+\alpha}}, \tag{D.2}$$

where integration is performed over the boundary, $\mathbf{x}\in\partial\Omega$ is the respective boundary point, $\mathbf{n}$ is the respective unit normal to the boundary, $\Delta g(\mathbf{x})$ is the value of discontinuity of g at $\mathbf{x}$, and $\widetilde{\theta}_{\mathbf{x}}$ is the Fourier transform of the homogeneous singularity of the kernel Θ at $\mathbf{x}=\mathbf{x}'$.

In scenario 2 – a function g generated by a Gaussian process with a homogeneous diagonal singularity of degree β – we will show that the large-t asymptotic of the loss is given by

$$L(t) \sim \frac{1}{2(2\pi)^d\beta}\Gamma\Big(\frac{\beta}{d+\alpha}+1\Big)\int_{\mathbb{R}^d}\int_{|\mathbf{n}|=1}\widetilde{\zeta}_{\mathbf{x}}(\mathbf{n})(\mu(\mathbf{x})\widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{\beta}{d+\alpha}}d\mathbf{x}dS\cdot(2t)^{-\frac{\beta}{d+\alpha}}, \tag{D.3}$$

where $\widetilde{\zeta}_{\mathbf{x}}$ is the Fourier transform of the diagonal singularity.

The general approach in obtaining these power law asymptotics is to expand g over the approximate, spatially localized eigenvectors of the evolution operator $\widetilde{\mathcal{A}}$. One can consider two slightly different versions of this approach. In one version we first find the asymptotic (8) of the cumulative distribution of the expansion coefficients, and then find the asymptotic of the loss using formula (9) of the main text. In the other version, we bypass the computation of Eq. (8), and find the asymptotic of $L(t)$ directly. (In fact, this version also uses the asymptotic relation (9) of the main text, but applies it not to the full set of eigenvalues, but rather separately to the localized eigenvector expansion at each point $\mathbf{x}$ of the domain). We find the latter approach to be somewhat more direct and efficient.

Accordingly, our derivations will be structured as follows. In Section D.1 we discuss the general ideas of localization and high-frequency asymptotics. In Section D.2 we give a direct derivation for the loss asymptotic in the case of the first (discontinuous) class g. In Section D.3 we give a direct derivation for the loss asymptotic in the case of the second (Gaussian) class g. Then, in Section D.4 we sketch the derivation of the coefficient asymptotic (8) for both classes of g.

[1] We remark in passing that we can usually control the initial approximation $\widetilde{f}$, and in some cases we can ensure that the contribution of $\mu^{1/2}\widetilde{f}$ to g is small compared to $\mu^{1/2}f$. In such cases, one can assume $g \approx -\mu^{1/2}f$.

D.1 General considerations

Recall from Section 4.2 and 5.1 of the main text that the kernel Θ has a diagonal singularity:

$$\Theta(\mathbf{x}, \mathbf{x}') = \theta_{\mathbf{x}}(\mathbf{x} - \mathbf{x}') + \dots, \tag{D.4}$$

where $\theta_{\mathbf{x}}(\cdot)$ is a (possibly $\mathbf{x}$-dependent) even ($\theta_{\mathbf{x}}(\mathbf{z}) = \theta_{\mathbf{x}}(-\mathbf{z})$) homogeneous function of degree α:

$$\theta_{\mathbf{x}}(c\mathbf{z}) = |c|^{\alpha}\theta_{\mathbf{x}}(\mathbf{z}). \tag{D.5}$$

It will be convenient to also consider the function $\psi_{\mathbf{x}}$ obtained by rescaling $\theta_{\mathbf{x}}$ by the coefficient $\mu(\mathbf{x})$:

$$\psi_{\mathbf{x}} = \mu(\mathbf{x})\theta_{\mathbf{x}}.$$

We denote by $\widetilde{\theta}_{\mathbf{x}}$ and $\widetilde{\psi}_{\mathbf{x}}$ the versions of the Fourier transforms of $\theta_{\mathbf{x}}, \psi_{\mathbf{x}}$ defined as in Eq. (C.1). Note that these functions are homogeneous with degree $-(d+\alpha)$.

To analyze the evolution $g_t = e^{-t\widetilde{\mathcal{A}}}g$, we use the short-time Fourier transform (STFT) of g:

$$F(\mathbf{y}, \mathbf{k}) = (2\pi)^{-d/2}\int_{\mathbb{R}^d} g(\mathbf{x})\omega(\mathbf{x} - \mathbf{y})e^{-i\mathbf{k}\cdot\mathbf{x}}d\mathbf{x}, \tag{D.6}$$

$$g(\mathbf{x}) = (2\pi)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} F(\mathbf{y}, \mathbf{k})\omega(\mathbf{x} - \mathbf{y})e^{i\mathbf{k}\cdot\mathbf{x}}d\mathbf{y}d\mathbf{k}. \tag{D.7}$$

Here ω is an even real smooth and compactly supported function such that $\int \omega^2 = 1$. Roughly speaking, the coefficient $F(\mathbf{y}, \mathbf{k})$ describes the component of g having wave number $\mathbf{k}$ and localized at the point $\mathbf{y}$.

The stationary phase method (or its variant described in [5]) shows that if f is a fixed function, then the leading term in the asymptotic of the action of the operator $\widetilde{\mathcal{A}}$ on the high-frequency function $f(\mathbf{x})e^{i\mathbf{k}\cdot\mathbf{x}}$ (with $\mathbf{k} \to \infty$) can be written in terms of the Fourier transform of the diagonal singularity of the kernel:

$$\int_{\mathbf{R}^d} \mu^{1/2}(\mathbf{x})\Theta(\mathbf{x}, \mathbf{x}')\mu^{1/2}(\mathbf{x}')f(\mathbf{x}')e^{i\mathbf{k}\cdot\mathbf{x}'}d\mathbf{x}' \sim \widetilde{\psi}_{\mathbf{x}}(\mathbf{k})f(\mathbf{x})e^{i\mathbf{k}\cdot\mathbf{x}}.$$

This shows that for large $\mathbf{k}$, we can think of the functions $f(\mathbf{x})e^{i\mathbf{k}\cdot\mathbf{x}}$ as of approximate eigenvectors of the operator $\widetilde{\mathcal{A}}$ and accordingly also of the evolution $e^{-t\widetilde{\mathcal{A}}}$. Then we can write

$$\begin{aligned} g_t(\mathbf{x}) &= e^{-t\widetilde{\mathcal{A}}}g(\mathbf{x}) \\ &= (2\pi)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} F(\mathbf{y}, \mathbf{k})e^{-t\widetilde{\mathcal{A}}}[\omega(\mathbf{x} - \mathbf{y})e^{i\mathbf{k}\cdot\mathbf{x}}]d\mathbf{y}d\mathbf{k} \\ &\overset{|\mathbf{k}|\gg 1}{\sim} (2\pi)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} F(\mathbf{y}, \mathbf{k})e^{-t\widetilde{\psi}_{\mathbf{x}}(\mathbf{k})}\omega(\mathbf{x} - \mathbf{y})e^{i\mathbf{k}\cdot\mathbf{x}}d\mathbf{y}d\mathbf{k}. \end{aligned} \tag{D.8}$$

To justify the assumption $|\mathbf{k}| \gg 1$, observe that the function $e^{-t\widetilde{\psi}_{\mathbf{x}}(|\mathbf{k}|)}$ is close to 0 for $t|\mathbf{k}|^{-(d+\alpha)} \gg 1$ (i.e. for $|\mathbf{k}| \ll t^{\frac{1}{d+\alpha}}$), and is close to 1 for $t|\mathbf{k}|^{-(d+\alpha)} \ll 1$ (i.e. for $|\mathbf{k}| \gg t^{\frac{1}{d+\alpha}}$), so that at large t the integral over $\mathbf{k}$ is indeed determined by the large-$|\mathbf{k}|$ asymptotic of $F(\mathbf{y}, \mathbf{k})$.

Next, we consider separately the two classes of functions g.

D.2 Scenario 1: a discontinuous g

Suppose that g is supported and smooth on the domain Ω, and has a discontinuity Δg at the boundary $\partial\Omega$. Consider the coefficient $F(\mathbf{y}, \mathbf{k})$ defined in Eq. (D.6). If $\mathbf{y}$ is such that the support of $\omega(\cdot - \mathbf{y})$ does not intersect the boundary $\partial\Omega$, then $g(\mathbf{x}')\omega(\mathbf{x}' - \mathbf{y})$ is a smooth function of $\mathbf{x}'$, and the coefficient $F(\mathbf{y}, \mathbf{k})$ will fall off faster than any power of $|\mathbf{k}|$ as $|\mathbf{k}| \to \infty$, so the contribution of such $\mathbf{y}$ to the expansion (D.8) will be negligible at large t. Assuming that the support of ω is small, it means that only $\mathbf{y}$ belonging to a narrow neighborhood of the boundary $\partial\Omega$ will contribute to (D.8). Accordingly, the function $g_t(\mathbf{x})$ will also fall off quickly away from the boundary.

Suppose now that $\mathbf{y}$ lies near the boundary and the support of $\omega(\cdot - \mathbf{y})$ intersects $\partial\Omega$. It is convenient to consider first the one-dimensional case $d = 1$.

Case $d = 1$**.** In this case the large-k asymptotic of the coefficients $F(y, k)$ will be determined by the discontinuity of $g(x')\omega(x' - y)$ at the boundary point $x' = x_0 \in \partial\Omega$:

$$F(y, k) \sim (2\pi)^{-1/2} \frac{\Delta g(x_0)\omega(x_0 - y)}{ik} e^{-kx_0}.$$

Now we substitute this into Eq. (D.8):

$$g_t(x) = (2\pi)^{-1} \Delta g(x_0) \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{e^{-t\widetilde{\psi}_{\mathbf{x}}(k)}}{ik} \omega(x_0 - y)\omega(x - y) e^{ik(x - x_0)} dy dk,$$

where x_0 is the point of $\partial\Omega$ near which the point x is located[2]. Regarding the function $\omega(x - y)$, one observes by rescaling the variable k that at large t, only a small (size-$t^{-\frac{1}{d+\alpha}}$) neighborhood of the boundary points will contribute to the integral, so we can write $\omega(x - y) \sim \omega(x_0 - y)$ and integrate out y using the formula $\int_{\mathbb{R}} \omega^2 = 1$:

$$\begin{aligned} g_t(x) &\sim (2\pi)^{-1} \Delta g(x_0) \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{e^{-t\widetilde{\psi}_{\mathbf{x}}(k)}}{ik} \omega^2(x_0 - y) e^{ik(x - x_0)} dy dk \\ &= (2\pi)^{-1} \Delta g(x_0) \int_{\mathbb{R}} \frac{e^{-t\widetilde{\psi}_{\mathbf{x}}(k)}}{ik} e^{ik(x - x_0)} dk. \end{aligned} \tag{D.9}$$

We now want to estimate the full squared norm $\|g_t\|^2 = \int_{\mathbb{R}} |g_t(x)|^2 dx$. On the whole $\mathbb{R}$, the function g_t is given by the sum of contributions from different points $x_0 \in \partial\Omega$:

$$g_t(x) \sim \sum_{x_0 \in \partial\Omega} (2\pi)^{-1} \Delta g(x_0) \int_{\mathbb{R}} \frac{e^{-t\widetilde{\psi}_{\mathbf{x}}(k)}}{ik} e^{ik(x - x_0)} dk. \tag{D.10}$$

Observe that the functions $u_{x_0}(k) = \frac{e^{-t\widetilde{\psi}_{\mathbf{x}}(k)}}{ik} e^{-ikx_0}$ with different $x_0 \in \partial\Omega$ become orthogonal in the limit $t \to \infty$ (since the Fourier transforms of these functions are localized at size-$t^{-\frac{1}{d+\alpha}}$ neighborhoods of the respective boundary points x_0). Then, by the unitarity of Fourier transform,

$$\begin{aligned} \|g_t\|^2 &\sim \sum_{x_0 \in \partial\Omega} (2\pi)^{-1} |\Delta g(x_0)|^2 \int_{\mathbb{R}} \frac{e^{-2t\widetilde{\psi}_{\mathbf{x}}(k)}}{|k|^2} dk \\ &= \sum_{x_0 \in \partial\Omega} |\Delta g(x_0)|^2 \frac{1}{\pi} \int_0^\infty \frac{\exp(-2\widetilde{\psi}_{\mathbf{x}}(1) t k^{-(d+\alpha)})}{|k|^2} dk. \end{aligned}$$

The $t \to \infty$ asymptotic of the integral has been determined in Section B (see also formula (9) in the main text): setting $\nu = d + \alpha, \kappa = 1, \Lambda = \widetilde{\psi}_{\mathbf{x}}(1), K = 1$, we get

$$\begin{aligned} \|g_t\|^2 &\sim \sum_{x_0 \in \partial\Omega} |\Delta g(x_0)|^2 \frac{1}{\pi} K \Gamma\left(\frac{\kappa}{\nu} + 1\right) (2\Lambda t)^{-\frac{\kappa}{\nu}} \\ &= \sum_{x_0 \in \partial\Omega} |\Delta g(x_0)|^2 \frac{1}{\pi} \Gamma\Big(\frac{1}{d + \alpha} + 1\Big) (2\widetilde{\psi}_{\mathbf{x}}(1) t)^{-\frac{1}{d+\alpha}}. \end{aligned}$$

Case $d > 1$**.** Assuming that the support of ω is small enough, we can approximate the boundary $\partial\Omega$ locally, in the support of $\omega(\cdot - \mathbf{y})$, by a linear hyperplane $\{\mathbf{x} : \mathbf{n} \cdot \mathbf{x} = x_0\}$, where $\mathbf{n}$ is the inward unit normal to $\partial\Omega$, so that the function $g(\cdot)\omega(\cdot - \mathbf{y})$ can be represented as a product of the Heaviside step function in the direction $\mathbf{n}$ and the smooth function $\omega(\cdot - \mathbf{y})$:

$$g(\mathbf{x}')\omega(\mathbf{x}' - \mathbf{y}) \sim \mathbf{1}_{\mathbf{n} \cdot \mathbf{x}' \geq x_0}(\mathbf{x}')\omega(\mathbf{x}' - \mathbf{y}).$$

The coefficient $F(\mathbf{y}, \mathbf{k})$ is the Fourier transform of this function w.r.t. $\mathbf{x}'$. The Fourier transform of $\mathbf{1}_{\mathbf{n} \cdot \mathbf{x}' \geq x_0}$ is a distribution concentrated on the line $l_{\mathbf{n}} = \{\mathbf{k} : \mathbf{k} = u\mathbf{n}, u \in \mathbb{R}\}$. The Fourier transform of $\mathbf{1}_{\mathbf{n} \cdot \mathbf{x}' \geq x_0}(\mathbf{x}')\omega(\mathbf{x}' - \mathbf{y})$ w.r.t. $\mathbf{x}'$ is the convolution of this line distribution with the Fourier transform of $\omega(\mathbf{x}' - \mathbf{y})$. Then for given $\mathbf{y}$, by the smoothness of ω, the coefficients $F(\mathbf{y}, \mathbf{k})$

[2] For $d = 1$, we take Ω to be a finite union of intervals, so $\partial\Omega$ consists of finitely many points.

are concentrated in a neighborhood of the line $l_{\mathbf{n}}$ in the $\mathbf{k}$-space. Let $\mathbf{k}_{\|}$ denote the projection of vector $\mathbf{k}$ to this line. Since $\widetilde{\psi}_{\mathbf{x}}(\mathbf{k})$ is a homogeneous function, for large t we can write $\widetilde{\psi}_{\mathbf{x}}(\mathbf{k}) \approx \widetilde{\psi}_{\mathbf{x}}(\mathbf{k}_{\|})$ in the integral (D.8):

$$g_t(\mathbf{x}) \sim (2\pi)^{-d/2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} F(\mathbf{y}, \mathbf{k}) e^{-t\widetilde{\psi}_{\mathbf{x}}(\mathbf{k}_{\|})} \omega(\mathbf{x} - \mathbf{y}) e^{i\mathbf{k}\cdot\mathbf{x}} d\mathbf{y} d\mathbf{k}.$$

We can now decompose the wave number $\mathbf{k}$ and the vector $\mathbf{x}$ into components parallel and orthogonal to the normal $\mathbf{n}$:

$$\mathbf{k} = \mathbf{k}_{\|} + \mathbf{k}_{\perp} = k_{\|}\mathbf{n} + \mathbf{k}_{\perp}, \quad \mathbf{x} = \mathbf{x}_{\|} + \mathbf{x}_{\perp} = x_{\|}\mathbf{n} + \mathbf{x}_{\perp},$$

and perform integration over the component $\mathbf{k}_{\perp}$ in the above formula after substituting the expression for $F(\mathbf{y}, \mathbf{k})$:

$$\begin{aligned} g_t(\mathbf{x}) &\sim (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(\mathbf{x}') \mathbf{1}_{\mathbf{n}\cdot\mathbf{x}' \geq x_0}(\mathbf{x}') \omega(\mathbf{x}' - \mathbf{y}) e^{-t\widetilde{\psi}_{\mathbf{x}}(\mathbf{k}_{\|})} \omega(\mathbf{x} - \mathbf{y}) e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} d\mathbf{y} d\mathbf{k} d\mathbf{x}' \\ &= (2\pi)^{-d+1} \int_{\mathbb{R}^d} \int_{\mathbb{R}} \int_{\mathbb{R}} g(\widetilde{\mathbf{x}}') \mathbf{1}_{x'_{\|} \geq x_0}(x'_{\|}) \omega(\widetilde{\mathbf{x}}' - \mathbf{y}) e^{-t\widetilde{\psi}_{\mathbf{x}}(\mathbf{k}_{\|})} \omega(\mathbf{x} - \mathbf{y}) e^{ik_{\|}(x_{\|} - x'_{\|})} d\mathbf{y} dk_{\|} dx'_{\|}, \end{aligned}$$

where $\widetilde{\mathbf{x}}' = \mathbf{x}_{\perp} + x'_{\|}\mathbf{n}$. We can now proceed similarly to the previous case $d = 1$. Specifically, let $\mathbf{x}_0 = \mathbf{x}_{\perp} + x_0\mathbf{n}$ be the projection of the point $\mathbf{x}'$ to the surface $\partial\Omega$. At large t and $k_{\|}$, we can approximate $g(\widetilde{\mathbf{x}}') \approx g(\mathbf{x}_0), \widetilde{\mathbf{x}}' \approx \mathbf{x}_0, \mathbf{x} \approx \mathbf{x}_0$, and integrate out $\mathbf{y}$ using $\int \omega^2(\mathbf{x}_0 - \mathbf{y}) d\mathbf{y} = 1$:

$$\begin{aligned} g_t(\mathbf{x}) &\sim (2\pi)^{-d+1} \Delta g(\mathbf{x}_0) \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbf{1}_{x'_{\|} \geq x_0}(x'_{\|}) e^{-t\widetilde{\psi}_{\mathbf{x}}(\mathbf{k}_{\|})} e^{ik_{\|}(x_{\|} - x'_{\|})} dk_{\|} dx'_{\|} \\ &\sim (2\pi)^{-d} \Delta g(\mathbf{x}_0) \int_{\mathbb{R}} \frac{e^{-t\widetilde{\psi}_{\mathbf{x}}(\mathbf{k}_{\|})}}{ik_{\|}} e^{ik_{\|}(x_{\|} - x_0)} dk_{\|}. \end{aligned}$$

This expression is analogous to the expression (D.9), and similarly to the case $d = 1$, we can now use it to obtain the asymptotic of $\|g_t\|^2$. Recall that in the case $d = 1$, for each x near the boundary we considered its projection x_0 to the boundary, and obtained the full integral $\|g_t\|^2$ by summing the contributions from different points x_0 (see Eq. (D.10) and subsequent formulas). In the present case $d > 1$, we replace this summation by integration over the boundary $\partial\Omega$. By repeating the same steps as before, we then get

$$\begin{aligned} \|g_t\|^2 &\sim \frac{1}{\pi}\Gamma\left(\frac{1}{d+\alpha} + 1\right) \int_{\partial\Omega} |\Delta g(\mathbf{x})|^2 (2\widetilde{\psi}_{\mathbf{x}}(\mathbf{n})t)^{-\frac{1}{d+\alpha}} dS \\ &= \frac{1}{\pi}\Gamma\left(\frac{1}{d+\alpha} + 1\right) \int_{\partial\Omega} |\Delta g(\mathbf{x})|^2 (\mu(\mathbf{x})\widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{1}{d+\alpha}} dS \cdot (2t)^{-\frac{1}{d+\alpha}}, \end{aligned}$$

yielding the loss asymptotic (D.2).

D.3 Scenario 2: g generated by a Gaussian process

Suppose now that g is generated by a Gaussian process with covariance $\Sigma(\mathbf{x}, \mathbf{x}') = \langle g(\mathbf{x}) g(\mathbf{x}') \rangle$, and that Σ has a homogeneous singularity of degree β on the diagonal $\mathbf{x} = \mathbf{x}'$:

$$\Sigma(\mathbf{x}, \mathbf{x}') = \zeta_{\mathbf{x}}(\mathbf{x}' - \mathbf{x}) + \ldots,$$

where the dots denote terms of a higher smoothness, and $\zeta_{\mathbf{x}}$ is an $\mathbf{x}$-dependent even homogeneous function of degree β :

$$\zeta_{\mathbf{x}}(c\mathbf{z}) = |c|^{\beta} \zeta_{\mathbf{x}}(\mathbf{z}).$$

Similarly to the previously considered homogeneous functions, we denote by $\widetilde{\zeta}_{\mathbf{x}'}$ the Fourier transform of $\zeta_{\mathbf{x}'}$ defined using Eq. (26) of the main text.

To analyze the asymptotic of $\|g_t\|^2$, we use again the representation (D.8) in which we substitute the expansion for $F(\mathbf{y}, \mathbf{k})$:

$$g_t(\mathbf{x}) \sim (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(\mathbf{x}') \omega(\mathbf{x}' - \mathbf{y}) e^{-t\widetilde{\psi}_{\mathbf{x}}(\mathbf{k})} \omega(\mathbf{x} - \mathbf{y}) e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} d\mathbf{y} d\mathbf{k} d\mathbf{x}'. \qquad \text{(D.11)}$$

Using as before the argument with rescaling, we see that the leading contribution to this integral comes at large $\mathbf{k}$ and small $\mathbf{x}-\mathbf{x}'$. In particular, we can write $\omega(\mathbf{x}'-\mathbf{y}) \approx \omega(\mathbf{x}-\mathbf{y})$ and integrate $\mathbf{y}$ out:

$$g_t(\mathbf{x}) \sim (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(\mathbf{x}') e^{-t\widetilde{\psi}_{\mathbf{x}'}(\mathbf{k})} e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} d\mathbf{k} d\mathbf{x}'. \tag{D.12}$$

Also, for further convenience, we have replaced $\mathbf{x}$ by $\mathbf{x}'$ in $\widetilde{\psi}_{\mathbf{x}}(\mathbf{k})$.

We now approximate $\|g_t\|^2$ by its expectation over the target functions g generated by the Gaussian process[3]:

$$\begin{aligned}
\|g_t\|^2 &\approx \langle \|g_t\|^2 \rangle && \text{(D.13)}\\
&= \int_{\mathbb{R}^d} \langle g_t^2(\mathbf{x}) \rangle d\mathbf{x} && \text{(D.14)}\\
&\sim (2\pi)^{-2d} \int_{\mathbb{R}^{5d}} \langle g(\mathbf{x}') g(\widetilde{\mathbf{x}}') \rangle e^{-t\widetilde{\psi}_{\mathbf{x}'}(\mathbf{k})} e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')} e^{-t\widetilde{\psi}_{\widetilde{\mathbf{x}}'}(\widetilde{\mathbf{k}})} e^{-i\widetilde{\mathbf{k}}\cdot(\mathbf{x}-\widetilde{\mathbf{x}}')} d\mathbf{k} d\mathbf{x}' d\widetilde{\mathbf{k}} d\widetilde{\mathbf{x}}' d\mathbf{x}. && \text{(D.15)}
\end{aligned}$$

We integrate out $\mathbf{x}$ using the identity $\int_{\mathbb{R}^d} e^{i(\mathbf{k}-\widetilde{\mathbf{k}})\cdot\mathbf{x}} d\mathbf{x} = (2\pi)^d \delta(\mathbf{k}-\widetilde{\mathbf{k}})$:

$$\|g_t\|^2 \sim (2\pi)^{-d} \int_{\mathbb{R}^{3d}} \Sigma(\mathbf{x}', \widetilde{\mathbf{x}}') e^{-t\widetilde{\psi}_{\mathbf{x}'}(\mathbf{k}) - t\widetilde{\psi}_{\widetilde{\mathbf{x}}'}(\mathbf{k})} e^{i\mathbf{k}\cdot(\widetilde{\mathbf{x}}'-\mathbf{x}')} d\mathbf{k} d\mathbf{x}' d\widetilde{\mathbf{x}}'. \tag{D.16}$$

We isolate now the singularity and apply the stationary phase method, obtaining the high-frequency approximation

$$\int_{\mathbb{R}^d} \Sigma(\mathbf{x}', \widetilde{\mathbf{x}}') e^{-t\widetilde{\psi}_{\mathbf{x}'}(\mathbf{k}) - t\widetilde{\psi}_{\widetilde{\mathbf{x}}'}(\mathbf{k})} e^{i\mathbf{k}\cdot(\widetilde{\mathbf{x}}'-\mathbf{x}')} d\widetilde{\mathbf{x}}' \sim \widetilde{\zeta}_{\mathbf{x}'}(\mathbf{k}) e^{-2t\widetilde{\psi}_{\mathbf{x}'}(\mathbf{k})}, \quad |\mathbf{k}| \gg 1.$$

This leads to

$$\|g_t\|^2 \sim (2\pi)^{-d} \int_{\mathbb{R}^{2d}} \widetilde{\zeta}_{\mathbf{x}'}(\mathbf{k}) e^{-2t\widetilde{\psi}_{\mathbf{x}'}(\mathbf{k})} d\mathbf{k} d\mathbf{x}'. \tag{D.17}$$

To analyze the asymptotic of this integral at large t, we represent $\mathbf{k}$ as $|\mathbf{k}|\mathbf{n}$, where $\mathbf{n}$ is a unit vector. Then, using the large-$\mathbf{k}$ asymptotics $\widetilde{\zeta}_{\mathbf{x}'}(\mathbf{k}) = \widetilde{\zeta}_{\mathbf{x}'}(\mathbf{n})|\mathbf{k}|^{-(d+\beta)}$ and $\widetilde{\psi}_{\mathbf{x}'}(\mathbf{k}) = \widetilde{\psi}_{\mathbf{x}'}(\mathbf{n})|\mathbf{k}|^{-(d+\alpha)}$,

$$\begin{aligned}
\|g_t\|^2 &\sim (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{|\mathbf{n}|=1} \int_0^\infty \widetilde{\zeta}_{\mathbf{x}'}(\mathbf{n}) r^{-(d+\beta)} e^{-2t\widetilde{\psi}_{\mathbf{x}'}(\mathbf{n}) r^{-(d+\alpha)}} r^{d-1} d\mathbf{x}' dS dr\\
&= (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{|\mathbf{n}|=1} \widetilde{\zeta}_{\mathbf{x}'}(\mathbf{n}) \int_0^\infty r^{-(1+\beta)} e^{-2t\widetilde{\psi}_{\mathbf{x}'}(\mathbf{n}) r^{-(d+\alpha)}} d\mathbf{x}' dS dr\\
&\sim \frac{1}{(2\pi)^d \beta} \Gamma\Big(\frac{\beta}{d+\alpha}+1\Big) \int_{\mathbb{R}^d} \int_{|\mathbf{n}|=1} \widetilde{\zeta}_{\mathbf{x}'}(\mathbf{n}) \widetilde{\psi}_{\mathbf{x}'}^{-\frac{\beta}{d+\alpha}}(\mathbf{n}) d\mathbf{x}' dS \cdot (2t)^{-\frac{\beta}{d+\alpha}}\\
&= \frac{1}{(2\pi)^d \beta} \Gamma\Big(\frac{\beta}{d+\alpha}+1\Big) \int_{\mathbb{R}^d} \int_{|\mathbf{n}|=1} \widetilde{\zeta}_{\mathbf{x}}(\mathbf{n}) (\mu(\mathbf{x}) \widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{\beta}{d+\alpha}} d\mathbf{x} dS \cdot (2t)^{-\frac{\beta}{d+\alpha}},
\end{aligned}$$

which yields the loss asymptotic (D.3).

In the case when both GP and operator $\mathcal{A}$ originate from the same shallow ReLU network, Fourier transforms of diagonal singularities have similar angular dependence and integration over sphere can be performed analytically. We know from section C that if the kernel singularities are based on the angle $\varphi(\mathbf{x}, \mathbf{x}')$ between the input points $\mathbf{x}, \mathbf{x}'$, then $\widetilde{\zeta}_{\mathbf{x}}(\mathbf{z}) \propto \varphi^\beta(\mathbf{x}, \mathbf{x}+\mathbf{z})$ and $\widetilde{\theta}_{\mathbf{x}}(\mathbf{z}) \propto \varphi^\alpha(\mathbf{x}, \mathbf{x}+\mathbf{z})$. Then the respective Fourier transforms admit the forms $\widetilde{\zeta}(\mathbf{n}) = P(\mathbf{x})|\mathbf{n}'|^{-d-\beta}$ and

[3] In general (if the Gaussian process does not have a small correlation length), a sampled value of $\|g_t\|^2$ need not be close to the expectation $\langle \|g_t\|^2 \rangle$. However, one can show using the Wick-Isserlis formula that the variance $\langle (\|g_t\|^2 - \langle \|g_t\|^2 \rangle)^2 \rangle$ scales with t as $t^{-(d+2\beta)/(d+\alpha)}$, i.e. becomes asymptotically negligible compared to $\langle \|g_t\|^2 \rangle^2$, which scales as $t^{-2\beta/(d+\alpha)}$. We plan to return to this point in a subsequent publication.

$\widetilde{\theta}(\mathbf{n}) = Q(\mathbf{x})|\mathbf{n}'|^{-d-\alpha}$, with $\mathbf{n}'$ the same as $\mathbf{n}$ except the first dimension: $n_1' = \frac{r(\mathbf{x})}{\sigma_b} n_1$. We write the sphere integral as

$$\begin{aligned}
\int_{|\mathbf{n}|=1} dS |\mathbf{n}'|^{-d-\beta} \left(|\mathbf{n}'|^{-d-\alpha}\right)^{-\frac{\beta}{d+\alpha}} &= \int_{|\mathbf{n}|=1} dS |\mathbf{n}'|^{-d} \\
&= \int_{|\mathbf{n}|=1} dS \left(\frac{r^2}{\sigma_b^2} n_1^2 + (n_2^2 + \ldots + n_d^2)\right)^{-\frac{d}{2}} \\
&\overset{(1)}{=} \int_0^\pi d\rho (\sin\rho)^{d-2} \int_{|\widetilde{\mathbf{n}}|=1} d\widetilde{S} \left(\frac{r^2}{\sigma_b^2} \cos^2(\rho) + \sin^2(\rho)\right)^{-\frac{d}{2}} \\
&= S_{d-2} \int_0^\pi d\rho (\sin\rho)^{d-2} \left(\frac{r^2}{\sigma_b^2} \cos^2(\rho) + \sin^2(\rho)\right)^{-\frac{d}{2}} \\
&= S_{d-2} \int_{-\infty}^{+\infty} d(\cot\rho) \left(\frac{r^2}{\sigma_b^2} \cot^2(\rho) + 1\right)^{-\frac{d}{2}} \\
&= \frac{\sigma_b}{r} S_{d-2} \int_{-\infty}^{+\infty} dz (z^2+1)^{-\frac{d}{2}} \\
&\overset{(2)}{=} \frac{\sigma_b}{r} S_{d-1} \qquad \text{(D.18)}
\end{aligned}$$

Here in (1) we split integration over sphere $|\mathbf{n}| = 1$ over the first axis and remaining $d-2$ dimensional sphere $|\widetilde{\mathbf{n}}| = 1$: $n_1 = \cos\rho$ and $(n_2, \ldots, n_d) = \widetilde{\mathbf{n}} \sin\rho$. Finally in (2) the value of the integral over z equals S_{d-1}/S_{d-2}, which can be inferred from the spherically symmetric case $r/\sigma_b = 1$.

D.4 The coefficient distributions

The derivations given above bypass the explicit computation of the cumulative distribution function s_n for the coefficients c_n of the expansion of g w.r.t. the eigenbasis of the operator $\widetilde{\mathcal{A}}$ (see Eqs. (7),(8)). These can be derived (at least heuristically) using essentially the same approach based on localized approximate eigendecomposition, but this time accompanied by the count of the total contribution of the coefficients corresponding to the given eigenvalue threshold from all the points of the domain.

It is convenient to introduce the partial sum $Q(\lambda)$ of the coefficients $|c_n|^2$ defined as in (7) but expressed in terms of the eigenvalue threshold λ:

$$Q(\lambda) = \sum_{n:\lambda_n<\lambda} |c_n|^2. \qquad \text{(D.19)}$$

The large-n asymptotic of s_n corresponds to the small-λ asymptotic of $Q(\lambda)$.

For Scenario 1 (discontinuous g), the resulting expression is

$$Q(\lambda) \sim \frac{1}{\pi} \int_{\partial\Omega} |\Delta g(\mathbf{x})|^2 (\mu(\mathbf{x})\widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{1}{d+\alpha}} dS \cdot \lambda^{\frac{1}{d+\alpha}}. \qquad \text{(D.20)}$$

For Scenario 2 (Gaussian g), the resulting expression is

$$Q(\lambda) \sim \frac{1}{(2\pi)^d \beta} \int_{\mathbb{R}^d} \int_{|\mathbf{n}|=1} \widetilde{\zeta}_{\mathbf{x}}(\mathbf{n}) (\mu(\mathbf{x})\widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{\beta}{d+\alpha}} d\mathbf{x} dS \cdot \lambda^{\frac{\beta}{d+\alpha}}. \qquad \text{(D.21)}$$

Since we have already found the loss asymptotics for both scenarios (Eqs. (D.2),(D.3)), we can establish the above expressions by showing that in either case $Q(\lambda) \sim a\lambda^b$ with some specific exponent b; the coefficient a can then be deduced from the respective loss coefficient.

To find the exponent b, we consider again the STFT representation (D.6),(D.7). Suppose that the function ω lives on a small scale $\frac{1}{M}$:

$$\omega(\mathbf{x}) = M^{d/2} \omega_0(M\mathbf{x}),$$

and suppose that the domain is accordingly decomposed into $\propto M^d$ independent "$\mathbf{y}$–cells". We can think of the respective STFT coefficients $F(\mathbf{y}, \mathbf{k})$ as representing the actual coefficients in the eigenvector expansion. Consider now separately the two scenarios.

Scenario 1: a discontinuous g. The coefficients $F(\mathbf{y},\mathbf{k})$ are negligible for cells not intersecting the boundary $\partial\Omega$. Suppose that the cell intersects $\partial\Omega$. Then the coefficients $F(\mathbf{y},\mathbf{k})$ in this cell vanish outside the line $\mathbf{k}=u\mathbf{n}$ in the $\mathbf{k}$-space, where $\mathbf{n}$ is the unit normal to $\partial\Omega$. For $\mathbf{k}=u\mathbf{n}$, we have

$$|F(\mathbf{y},\mathbf{k})| \propto M^{d/2}M^{1-d}\tfrac{|\Delta g(\mathbf{y})|}{|\mathbf{k}|} = M^{1-d/2}\tfrac{|\Delta g(\mathbf{y})|}{|\mathbf{k}|}.$$

We assume now that for suitable discrete wave numbers $\mathbf{k}$ the coefficients $F(\mathbf{y},\mathbf{k})$ are associated to respective approximate eigenvectors of $\widetilde{\mathcal{A}}$ in the $\mathbf{y}$-cell, and estimate the respective contribution of the coefficients to the sum $S(\lambda)$. The discreteness results from the finite size of the support of ω: the density of the eigenvalues scales with M as M^{-d}. The respective discrete constants u for the relation $\mathbf{k}=u\mathbf{n}$ scale as integer multiples of M, i.e. $u_l\sim lM$. Accordingly, the contribution of the coefficients $F(\mathbf{y},\mathbf{k})$ in the $\mathbf{y}$-cell to $Q(\lambda)$ can be estimated as

$$\begin{aligned}\sum_{\mathbf{k}:\mathbf{k}=u_l\mathbf{n},\lambda_{\mathbf{k}}<\lambda}|F(\mathbf{y},\mathbf{k})|^2 &\sim M^{-1}\int_{k_0}^{\infty}M^{2-d}\tfrac{|\Delta g|^2}{k^2}dk\\ &\sim M^{1-d}\tfrac{|\Delta g|^2}{k_0},\end{aligned}\tag{D.22}$$

where the wave number $k_0=|u\mathbf{n}|$ corresponds to the eigenvalue λ:

$$\widetilde{\theta}_{\mathbf{x}}(u\mathbf{n})=\lambda.$$

We can find k_0 using the homogeneity of $\widetilde{\theta}_{\mathbf{x}}$: since $\widetilde{\theta}_{\mathbf{x}}(u\mathbf{n})=|u|^{-d-\alpha}\widetilde{\theta}_{\mathbf{x}}(\mathbf{n})$, we have

$$k_0=|u|=(\tfrac{\widetilde{\theta}_{\mathbf{x}}(\mathbf{n})}{\lambda})^{1/(d+\alpha)}.$$

Substituting this into Eq. (D.22) and taking into account that there are $\propto M^{d-1}$ cells intersecting $\partial\Omega$, we find Eq. (D.20) up to a coefficient.

Scenario 2: g generated by a Gaussian process. As before, finding the exponent b can be reduced to estimating the asymptotic of $|F(\mathbf{y},\mathbf{k})|^2$ at a fixed $\mathbf{y}$ and large $\mathbf{k}$. Computing the expectation, we get

$$\begin{aligned}\langle|F(\mathbf{y},\mathbf{k})|^2\rangle = (2\pi)^{-d}&\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\omega(\mathbf{x}-\mathbf{y})\omega(\mathbf{x}'-\mathbf{y})\\ &\times\Sigma(\mathbf{x},\mathbf{x}')e^{-i\mathbf{k}\cdot(\mathbf{x}-\mathbf{x}')}d\mathbf{x}d\mathbf{x}'\\ \propto{}& |\mathbf{k}|^{-(d+\beta)}.\end{aligned}$$

Since the wave vector $\mathbf{k}$ corresponds to an eigenvalue $\lambda\propto|\mathbf{k}|^{-(d+\alpha)}$ and since the density of the wave numbers $\mathbf{k}\in\mathbb{R}^d$ associated with localized eigenvectors of $\widetilde{\mathcal{A}}$ scales as M^{-d}, we can write

$$\begin{aligned}\sum_{\mathbf{k}:\lambda_{\mathbf{k}}<\lambda}|F(\mathbf{y},\mathbf{k})|^2 &\propto M^{-d}\int_{|\mathbf{k}|<\lambda^{-1/(d+\alpha)}}|\mathbf{k}|^{-(d+\beta)}d\mathbf{k}\\ &\propto M^{-d}\lambda^{\beta/(d+\alpha)}.\end{aligned}$$

Collecting the contributions to $Q(\lambda)$ from all $\sim M^d$ $\mathbf{y}$-cells, we thus get

$$Q(\lambda)\sim a\lambda^{\beta/(d+\alpha)},$$

as claimed.

By expressing λ through n with the help of the eigenvalue asymptotic (6) and Eq. (19), we can also cast the obtained formulas for $Q(\lambda)$ in the form $s_n\sim Kn^{-\kappa}$ as in Eq. (8).

For Scenario 1 (discontinuous g), the resulting coefficient and exponent are

$$\begin{aligned}\kappa &= \tfrac{1}{d},\\ K &= \Big(\tfrac{1}{\pi}\int_{\partial\Omega}|\Delta g(\mathbf{x})|^2(\mu(\mathbf{x})\widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{1}{d+\alpha}}dS\Big)\Big(\int\gamma_{\mathbf{x}}\mu^{\frac{d}{d+\alpha}}(\mathbf{x})d\mathbf{x}\Big)^{1/d}.\end{aligned}$$

For Scenario 2 (Gaussian g), the resulting coefficient and exponent are

$$\begin{aligned}\kappa &= \tfrac{1}{d},\\ K &= \Big(\tfrac{1}{(2\pi)^d\beta}\int_{\mathbb{R}^d}\int_{|\mathbf{n}|=1}\widetilde{\zeta}_{\mathbf{x}}(\mathbf{n})(\mu(\mathbf{x})\widetilde{\theta}_{\mathbf{x}}(\mathbf{n}))^{-\frac{\beta}{d+\alpha}}d\mathbf{x}dS\Big)\Big(\int\gamma_{\mathbf{x}}\mu^{\frac{d}{d+\alpha}}(\mathbf{x})d\mathbf{x}\Big)^{1/d}.\end{aligned}$$

E Extensions

In this section we derive results of section 6 in the paper.

E.1 Activations of different smoothness

We consider a shallow network in NTK regime with activation function $\phi_q(z) = (z)_+^q, \; q > 0$. Output covariance Σ_q and NTK Θ_q for such network can be written as

$$\Sigma_q(\mathbf{x}, \mathbf{x}') = \sigma_w^2 \Big\langle \big(z(\mathbf{x})\big)_+^q \big(z(\mathbf{x}')\big)_+^q \Big\rangle \tag{E.1}$$

$$\Theta_q(\mathbf{x}, \mathbf{x}') = \Sigma_q(\mathbf{x}, \mathbf{x}') + \sigma_w^2(\sigma_w^2 \mathbf{x}\cdot\mathbf{x}' + \sigma_b^2) q^2 \Big\langle \big(z(\mathbf{x})\big)_+^{q-1} \big(z(\mathbf{x}')\big)_+^{q-1} \Big\rangle \tag{E.2}$$

Here the average is taken w.r.t. pair of Gaussian random variables $z(\mathbf{x}), z(\mathbf{x}')$ with zero mean and covariance

$$\big\langle (z(\mathbf{x}), z(\mathbf{x}')^T (z(\mathbf{x}), z(\mathbf{x}')\big\rangle = \begin{pmatrix} r^2(\mathbf{x}) & r(\mathbf{x})r(\mathbf{x}')\varphi(\mathbf{x},\mathbf{x}') \\ r(\mathbf{x})r(\mathbf{x}')\varphi(\mathbf{x},\mathbf{x}') & r^2(\mathbf{x}') \end{pmatrix} \tag{E.3}$$

Such averages were calculated in [11] for integer q, but we take intermediate integral representation (eqs. (3), (16)) from this paper, which we will analyze for general q. As usual, we omit explicit $\mathbf{x}, \mathbf{x}'$ dependence for brevity.

$$\big\langle (z)_+^q (z')_+^q \big\rangle = \frac{1}{2\pi} r^q r'^q \Gamma(q+1) \big(\sin\varphi\big)^{2q+1} \int\limits_0^{\frac{\pi}{2}} \frac{\big(\cos\psi\big)^q}{\big(1 - \cos\varphi\cos\psi\big)^{q+1}} d\psi \tag{E.4}$$

Let's denote the integral in (E.4) by $I_q(\varphi)$. We will transform it so it has the form of integral representation of the hypergeometric function ${}_2F_1$

$$\begin{aligned} I_q(\varphi) &= \int\limits_0^{\frac{\pi}{2}} \frac{\big(\cos\psi\big)^q}{\big(1-\cos\varphi\cos\psi\big)^{q+1}} d\psi \\ &= \int\limits_0^1 \frac{y^q}{\sqrt{1-y^2}(1-\cos\varphi y)^{q+1}} dy, \quad y = \cos\psi \\ &= \frac{1}{(1-\cos\varphi)^{q+1}} \int\limits_0^1 t^q (1-t)^{-\frac{1}{2}} \left(1 + \frac{1+\cos\varphi}{1-\cos\varphi} t\right)^{-\frac{1}{2}} dt, \quad t = \frac{y(1-\cos\varphi)}{1-\cos\varphi y} \end{aligned} \tag{E.5}$$

The hypergeometric function ${}_2F_1(a, b; c; z)$ has the following integral representation and asymptotic expansion at $z = -\infty$:

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int\limits_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a} dt \tag{E.6}$$

$${}_2F_1(a,b;c;-z) = z^{-b} \frac{\Gamma(a-b)\Gamma(c)}{\Gamma(a)\Gamma(-b+c)} \left(1 + \sum_{n\geq 1} g_n z^{-n}\right) + z^{-a} \sum_{n\geq 0} f_n z^{-n} \tag{E.7}$$

Here g_n are f_n are coefficients in the asymptotic expansion. Comparing our integral I_q with integral representation of ${}_2F_1$ we see that it is indeed hypergeometric function with parameters $a = \frac{1}{2}, b = q+1, c = q + \frac{3}{2}$ and argument $z = -\frac{1+\cos\varphi}{1-\cos\varphi}$. Singularity at $\varphi = 0$ is located at hypergeometric function argument $z = -\infty$, therefore we need exactly asymptotic (E.7) to analyze singularity. Substituting our values of ${}_2F_1$ parameters we obtain the following asymptotic expansion at $\varphi = 0$

$$(\sin\varphi)^{2q+1} I_q(\varphi) = \frac{\Gamma(q+1)\Gamma(-\frac{1}{2}-q)}{\sqrt{\pi}} \frac{(\sin\varphi)^{2q+1}}{(1+\cos\varphi)^{q+1}} \left[1 + \sum_{n\geq 1} g_n \Big(\frac{1-\cos\varphi}{1+\cos\varphi}\Big)^n\right] + \frac{(\sin\varphi)^{2q+1}}{(1-\cos\varphi)^{q+\frac{1}{2}}(1+\cos\varphi)^{\frac{1}{2}}} \sum_{n\geq 0} f'_n \Big(\frac{1-\cos\varphi}{1+\cos\varphi}\Big)^n \tag{E.8}$$

As it is written now, the asymptotic expansion above is not an expansion in powers φ, but it can be turned into one by replacing functions of φ with their Taylor expansions. In particular, $\sin\varphi = \varphi + O(\varphi^3)$, $1-\cos\varphi = \frac{1}{2}\varphi^2 + O(\varphi^4)$ and $1+\cos\varphi = 2 + O(\varphi^2)$. In the asymptotic expansion (E.8) the second term starting from φ^0 is the leading one. However, it contains only even powers φ^{2n}, which are all regular. On the contrary, the first term starts with φ^{2q+1} and it is singular for all q except half-integers. Taking the leading singular term from (E.8) we obtain the leading singular term of NTK (E.2)

$$\Theta_{q,\text{sing}} = \frac{\sigma_w^2}{2\pi} r^{2q} q^2 \frac{\Gamma^2(q)\Gamma(\frac{1}{2}-q)}{\sqrt{\pi}2^q} \varphi^{2q-1} \tag{E.9}$$

Combining this with the γ coefficient from (C.14) with $\alpha = 2q-1$ and $A(\mathbf{x})$ deduced from (E.9) we get eigenvalue asymptotic coefficient Λ_q

$$\Lambda_q = \sigma_w^{\alpha+2}\sigma_b^{\frac{\alpha}{d}} q^2 (2\pi)^{d+q-2} \frac{\Gamma\big(\frac{d+\alpha}{2}\big)\Gamma^2\big(\frac{\alpha+1}{2}\big)}{\Big(\Gamma(\frac{d}{2}+1)\Big)^{\frac{d+\alpha}{d}}} \Big\langle \mu(\mathbf{x})^{-\frac{\alpha}{d+\alpha}} r(\mathbf{x})^{\frac{2d-\alpha d-\alpha}{d+\alpha}} \Big\rangle_\mu^{\frac{d+\alpha}{d}} \tag{E.10}$$

In the case of half-integer q the coefficient in (E.9) diverges due to gamma function $\Gamma(\frac{1}{2}-q)$ having simple poles at positive half integer q. Quite interestingly, the same delta function is found in $\gamma_{d,\alpha}^{(d+\alpha)/d}$ and they cancel. Therefore, the final constant Λ_q formally has a meaningful limit at half integer q. However, existence of a limit does not prove that at half-integer q eigenvalues have an asymptotic with constant (E.10). The half integer case should be studied separately and we leave it for the future work.

E.2 Deep networks

We consider deep network $L > 2$ in the NTK regime and with ReLU activation function. Covariances and NTK's of intermediate layers are calculated as

$$\begin{cases} \Sigma^{(l+1)}(\mathbf{x},\mathbf{x}') = \sigma_w^2 \big\langle \phi(z^l(\mathbf{x}))\phi(z^l(\mathbf{x}')) \big\rangle + \sigma_b^2 \\ \Theta^{(l+1)}(\mathbf{x},\mathbf{x}') = \Sigma^{(l+1)}(\mathbf{x},\mathbf{x}') + \sigma_w^2 \Theta^{(l)}(\mathbf{x},\mathbf{x}') \big\langle \dot\phi(z^l(\mathbf{x}))\dot\phi(z^l(\mathbf{x}')) \big\rangle \end{cases} \tag{E.11}$$

Here, as in the paper, $z^l(\mathbf{x})$ is a GP with covariance $\Sigma_l(\mathbf{x},\mathbf{x}')$. Parametrizing covariance as

$$\Sigma^{(l)}(\mathbf{x},\mathbf{x}') = \begin{pmatrix} r_l(\mathbf{x})^2 & r_l(\mathbf{x}) r_l(\mathbf{x}') \cos\varphi_l(\mathbf{x},\mathbf{x}') \\ r_l(\mathbf{x}) r_l(\mathbf{x})' \cos\varphi_l & r_l(\mathbf{x}')^2 \end{pmatrix} \tag{E.12}$$

From this point we again drop $\mathbf{x},\mathbf{x}'$ dependence. Using parametrization (E.12) we rewrite recursive relations (E.11)

$$\begin{cases} \Sigma^{(l+1)} = \frac{\sigma_w^2}{2\pi} r_l r'_l \big(\sin\varphi_l + \cos\varphi_l(\pi-\varphi_l)\big) + \sigma_b^2 \\ \Theta^{(l+1)} = \Sigma^{(l+1)} + \Theta^{(l)} \frac{\sigma_w^2}{2\pi}(\pi-\varphi_l) \end{cases} \tag{E.13}$$

We see that NTK's $\Theta^{(l)}$ can be fully expressed through φ_l and r_l. From (E.13) the recursive relations for φ_l, r_l are

$$\begin{cases} r_{l+1}^2 = \frac{\sigma_w^2}{2} r_l^2 + \sigma_b^2 \\ \cos\varphi_{l+1} = \frac{1}{r_{l+1} r'_{l+1}} \Big[\frac{\sigma_w^2}{2\pi} r_l r'_l \big(\sin\varphi_l + \cos\varphi_l(\pi-\varphi_l)\big) + \sigma_b^2\Big] \end{cases} \tag{E.14}$$

From these equations we see that $\varphi_{l+1} = 0$ only when $\varphi_l = 0$ and $r_l = r'_l$. This, in turn, happens only when $\mathbf{x} = \mathbf{x}'$. Using starting values $r_1^2(\mathbf{x}) = \sigma_w^2|\mathbf{x}|^2 + \sigma_b^2$ and $\varphi(\mathbf{x}, \mathbf{x}')$ defined in (C.8), and the fact that $\arccos(z)$ is smooth everywhere except $z = -1, 1$, one can see that $r_l(\mathbf{x})$ is smooth everywhere and $\varphi(\mathbf{x}, \mathbf{x}')$ is smooth everywhere except the diagonal $\mathbf{x} = \mathbf{x}'$. Therefore, NTKs $\Theta^{(L)}$ are smooth away from diagonal and might have a singularity on it. From (E.13) and smoothness of $r_l(\mathbf{x})$ we see that the only source of singularity are φ_l.

To find singular expression for φ_l let's assume that $\varphi_l = O(|\mathbf{x} - \mathbf{x}'|)$ and carefully expand second equation in (E.14) up to second order in $|\mathbf{x} - \mathbf{x}'|$. To do this we note that $|r_l - r'_l| = O(|\mathbf{x} - \mathbf{x}'|)$ and $\sin\varphi_l + \cos\varphi_l(\pi - \varphi_l) = \pi(1 - \varphi_l^2/2) + O(\varphi_l^3)$

$$\varphi_{l+1}^2 = \frac{\sigma_w^2 r_l^2}{2r_{l+1}^2}\Big(\varphi_l^2 + (r_l - r'_l)^2\frac{\sigma_b^2}{r_{l+1}^2}\Big) + O(|\mathbf{x} - \mathbf{x}'|^3) \tag{E.15}$$

Thus we confirmed our assumption $\varphi_l = O(|\mathbf{x} - \mathbf{x}'|)$. In the leading order both $(r_l - r'_l)^2$ and φ_1^2 are homogeneous functions of degree 2. Combining this with (E.15) we see that all φ_l are homogeneous functions of degree 1 in the leading order of $\mathbf{x} - \mathbf{x}'$.

Now we find the leading singular part of $\Theta^{(l+1)}$. We will see that the leading singular part has degree 1, therefore we assume the following NTK expansion

$$\Theta^{(l)} = \Theta_{\text{diag}}^{(l)} - \sum_{m=1}^{l-1} a_m^{(l)}\varphi_m + O(|\mathbf{x} - \mathbf{x}|^2) \tag{E.16}$$

Here $\Theta_{\text{diag}}^{(l)}$ is the value of NTK at the diagonal, $a_m^{(l)}$ are constants and the sum represents leading singular part of the NTK $\Theta_{\text{sing}}^{(l)}$. The recursion relation (E.13) can be now written as

$$\Theta_{\text{diag}}^{(l+1)} + \Theta_{\text{sing}}^{(l+1)} + O(|\mathbf{x} - \mathbf{x}|^2) = r_{l+1}^2 + \big(\Theta_{\text{diag}}^{(l)} + \Theta_{\text{sing}}^{(l)}\big)\frac{\sigma_w^2}{2\pi}(\pi - \varphi_l) + O(|\mathbf{x} - \mathbf{x}|^2) \tag{E.17}$$

From this we extract recursive relations for diagonal and singular parts of NTK

$$\begin{aligned}\Theta_{\text{diag}}^{(l+1)} &= r_{l+1}^2 + \frac{\sigma_w^2}{2}\Theta_{\text{diag}}^{(l)}\\ \Theta_{\text{sing}}^{(l+1)} &= -\frac{1}{2\pi}\Theta_{\text{diag}}^{(l)}\varphi_l + \frac{\sigma_w^2}{2}\Theta_{\text{sing}}^{(l)}\end{aligned} \tag{E.18}$$

Constants $a_m^{(l)}$ can be explicitly extracted from this relations. Since $\Theta_{\text{diag}}^{(l)} > 0$ we can see that all $a_m^{(l)} > 0$. It means that the leading singular terms of order $O(|\mathbf{x} - \mathbf{x}|)$ will not cancel each over, thus confirming that the leading singularity in NTK has homogeneity degree 1.

E.3 MF regime

The NTK of network in MF regime is given by

$$\Theta(\mathbf{x}, \mathbf{x}') = \int \nabla_{\tilde{\mathbf{w}}}\tilde{\phi}(\tilde{\mathbf{w}}, \mathbf{x})p(\tilde{\mathbf{w}})\nabla_{\tilde{\mathbf{w}}}\tilde{\phi}(\tilde{\mathbf{w}}, \mathbf{x}')d\tilde{\mathbf{w}} \tag{E.19}$$

Here $\tilde{\phi}(\tilde{\mathbf{w}}, \mathbf{x}) = c\phi(\mathbf{w}\cdot\mathbf{x} + b)$ is a computation of a single neuron in shallow network and $p(\tilde{\mathbf{w}}) = p(c, \mathbf{w}, b)$ is a distribution of a neuron parameters. In the case of ReLU activation $\phi(z) = (z)_+$ we rewrite it as

$$\begin{aligned}\Theta(\mathbf{x}, \mathbf{x}') = &\int (\mathbf{w}\cdot\mathbf{x} + b)_+(\mathbf{w}\cdot\mathbf{x}' + b)_+ dp_0(\mathbf{w}, b)\\ &+ (1 + \mathbf{x}\cdot\mathbf{x}')\int H(\mathbf{w}\cdot\mathbf{x} + b)H(\mathbf{w}\cdot\mathbf{x}' + b)dp_2(\mathbf{w}, b)\\ &= I_1(\mathbf{x}, \mathbf{x}') + (1 + \mathbf{x}\cdot\mathbf{x}')I_2(\mathbf{x}, \mathbf{x}')\end{aligned} \tag{E.20}$$

Here the first integral I_1 corresponds to taking the gradient w.r.t. c in (E.19) and the second integral I_2 corresponds to taking the gradients w.r.t. $\mathbf{w}$ and b; $p_0(\mathbf{w}, b)$ and $p_2(\mathbf{w}, b)$ are the 0'th and 2'nd moment of the distribution $p(c, \mathbf{w}, b)$ w.r.t. the variable c; $H(z)$ is the Heaviside step function.

Now suppose that p_0 and p_2 are sufficiently smooth and fall off quickly at infinity. To analyze the smoothness of the NTK (E.20) let's differentiate it using that $\frac{d}{dz}(z)_+ = H(z)$ and $\frac{d}{dz}H(z) = \delta(z)$. We start with the second integral I_2. The first derivative produces one delta function, and, together with the left Heaviside function, they are located on hyperplanes $\mathbf{w}\cdot\mathbf{x}+b=0$ and $\mathbf{w}\cdot\mathbf{x}'+b=0$ in the $(\mathbf{w},b)$ space. First we consider a neighborhood of points $\mathbf{x}\neq\mathbf{x}'$ - the corresponding integral continuously depend on $\mathbf{x},\mathbf{x}'$. If we differentiate the second time we will have two delta function, which restrict the integral to $d-1$ dimensional subspace of $(\mathbf{w},b)$ space, which is also continuously depend on $\mathbf{x},\mathbf{x}'$. Further derivatives can be translated to differentiating $p_2(\mathbf{w},b)$, with the result being continuous as long as p_2 is sufficiently smooth. Thus we established that the second integral is smooth at $\mathbf{x}\neq\mathbf{x}'$. Now let's turn to the diagonal $\mathbf{x}=\mathbf{x}'$, where two hyperplanes coincide, and consider the first derivative. Without loss of generality assume that the derivative is over $\mathbf{x}'$, then it is discontinuous, because infinitely small change of $\mathbf{x}$ will change which half of hyperplane $\mathbf{w}\cdot\mathbf{x}'+b=0$ (corresponding to delta function) is located in the halfspace "allowed" by the Heaviside function: $\mathbf{w}\cdot\mathbf{x}+b>0$. To summarize, $I_2(\mathbf{x},\mathbf{x}')$ is smooth outside of the diagonal $\mathbf{x}=\mathbf{x}'$ and has a first order singularity on it. The first integral I_1 is treated similarly, except that one has to differentiate 3 times instead of 1.

Since the order of singularity is higher in I_1, the leading singular term comes from I_2. We focus now on deriving its behavior near the diagonal. Consider $\mathbf{x}'=\mathbf{x}+a\hat{\mathbf{n}}$ with small $a>0$ and unit vector $\hat{\mathbf{n}}$. We write I_2 as

$$I_2(\mathbf{x},\mathbf{x}') = \int\limits_{\mathbf{w}\cdot\mathbf{x}+b>0} p_2(\mathbf{w},b)d\mathbf{w}db \quad - \int\limits_{0<\mathbf{w}\cdot\mathbf{x}+b<-a\mathbf{w}\cdot\hat{\mathbf{n}}} p_2(\mathbf{w},b)d\mathbf{w}db \tag{E.21}$$

The first integral here is the values of I_2 on the diagonal, and the second integral is singular, because, e.g., it doesn't change sing with $\hat{\mathbf{n}}\to-\hat{\mathbf{n}}$ due to being non-negative. We calculate the second integral in (E.21) up to the first order in $|\mathbf{x}-\mathbf{x}'|=a$. The integration is taken in the thin region adjacent the the half of the hyperplane $\mathbf{w}\cdot\mathbf{x}+b=0$ specified by $-\mathbf{w}\cdot\hat{\mathbf{n}}>0$. The thickness of this region at point $\mathbf{w}$ can be calculated using geometric reasoning. The answer is $a(-\mathbf{w}\cdot\hat{\mathbf{n}})_+/\sqrt{|\mathbf{x}|^2+1}$. This gives us

$$I_2(\mathbf{x},\mathbf{x}') - I_2(\mathbf{x},\mathbf{x}) = -\frac{a}{\sqrt{|\mathbf{x}|^2+1}} \int\limits_{\mathbf{w}\cdot\mathbf{x}+b=0} (-\mathbf{w}\cdot\hat{\mathbf{n}})_+ p_2(\mathbf{w},b)d\mathbf{w}db \; + O(|\mathbf{x}-\mathbf{x}'|^2) \tag{E.22}$$

However, the expression (E.22) in principle contains both regular and singular parts. We can extract the singular part using the fact that it doesn't change sign under $\hat{\mathbf{n}}\to-\hat{\mathbf{n}}$, while the regular part does. Since the leading singular part of NTK comes from I_2, we can write it as

$$\Theta_{\text{sing}}(\mathbf{x},\mathbf{x}') = \frac{\sqrt{|\mathbf{x}|^2+1}}{2} \int\limits_{\mathbf{w}\cdot\mathbf{x}+b=0} \big|\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}')\big| p_2(\mathbf{w},b)d\mathbf{w}db. \tag{E.23}$$

We see that the singularity is of homogeneous type with degree 1, as for the network in NTK regime.

E.4 MNIST and other high dimensional data

Almost all experimental tests presented in the main text involve relatively low-dimensional approximation problems (with d ranging from 1 to 4). A natural question is whether the developed theory is applicable to high dimensional problems, such as image recognition and the likes, where the ambient dimension can be 10^3 or higher. We should, however, point out a crucial difference between these problems and the setting of the present paper. Specifically, whereas in our setting we assume that the approximated function f is defined on $\mathbb{R}^d$ or a d-dimensional subset thereof, in image recognition problems (and many other solvable high-dimensional problems) the function f is only defined on a low-dimensional data manifold occupying a tiny portion in the ambient space. Accordingly, there are two completely different dimensions: the dimension of the ambient space (e.g., 28 × 28 for MNIST) and the much lower (approximate) intrinsic dimension of the data manifold. This data sparsity is quite a general property of solvable high-dimensional problems: without it, because of the curse of dimensionality, predictive models would need to be extremely complex. In contrast to the ambient dimension, the intrinsic dimension can only be roughly estimated by the metric properties of data manifold. In any case, we expect this dimension to be primarily associated with some natural deformations of the considered objects (e.g., shifts, rotations, and other transformations of MNIST digits) and therefore to be not too large.

Since the data manifold can have a nontrivial curvature and other peculiarities not covered by the theoretical setting of the present paper, our theoretical predictions have only a limited applicability to this more general setting. A complete respective theory is beyond the scope of the present paper and will be a subject of a future study.

Nevertheless, in this paper we present experimental results on the eigenvalue, coefficient sum and loss asymptotics for MNIST (see Figure 3d). In this experiment, we used a shallow network in the NTK regime, with MSE loss and one-hot encoded classes as targets. The dataset size was 20000 points. We make several nontrivial observations about the experimental curves, suggesting that our method can indeed be generalized to a high-dimensional setting.

1. The eigenvalue and coefficient distributions λ_n and s_n, as well as the loss function $L(t)$ are well approximated by the power laws (6),(8),(1) with exponents $\nu = 1.35, \kappa = 0.33, \xi = 0.26$.
2. The theoretically expected relation $\xi = \kappa/\nu$ is approximately satisfied for experimentally found values $\nu = 1.35, \kappa = 0.33, \xi = 0.26$. This is not very surprising since the derivation of the formula (9) is based on the assumptions (6) and (8), which are verified experimentally.
3. The third, and probably most interesting observation, concerns the effective dimension d_{eff} obtained by comparing the experimental values of the three exponents with the theoretical expressions (26), i.e. $\nu = 1 + \frac{1}{d_{\text{eff}}}, \kappa = \frac{1}{d_{\text{eff}}}, \xi = \frac{1}{d_{\text{eff}}+1}$ (assuming they remain valid). A priori, the effective dimensions inferred from κ and ν do not have to coincide because κ and ν are not directly related. However, in our MNIST experiment, all three formulas give approximately the same value $d_{\text{eff}} \approx 3$.

We leave a detailed study of these effects and their full theoretical explanation for a future work.

F Additional experiments

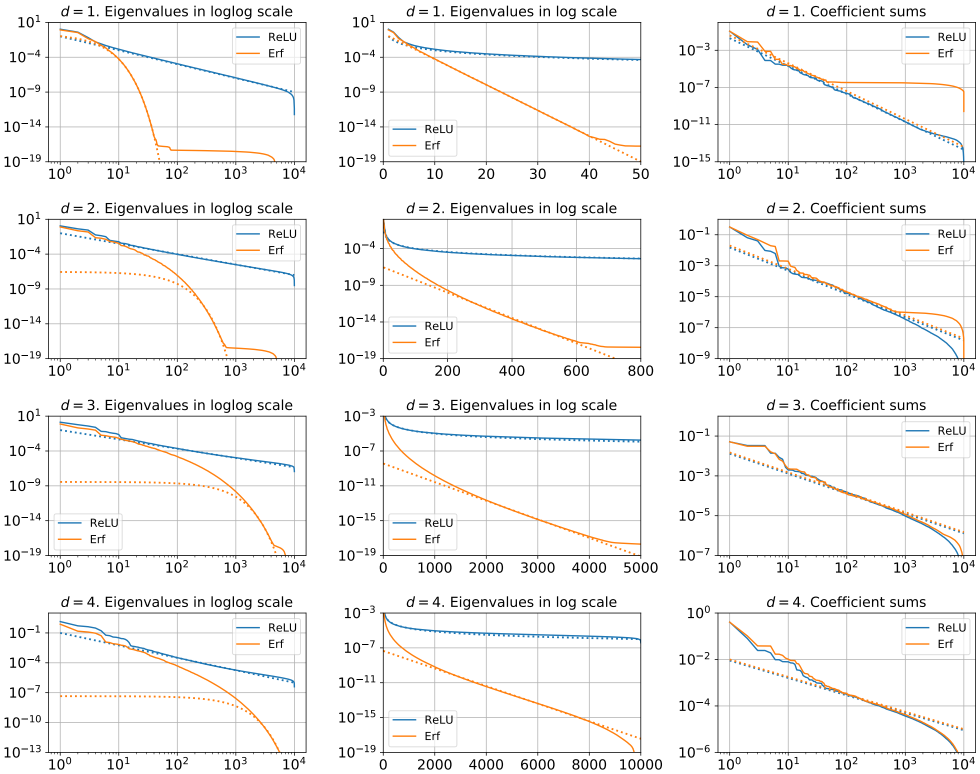


Figure 4: Distribution of eigenvalues λ_n and coefficient partial sums s_n for activation functions $\phi(z) = (z)_+$ and $\phi(z) = \mathrm{Erf}(z)$. Target functions in both cases are draws from Gaussian Process modeled by shallow network with NTK parametrization. Dotted lines show analytic expressions fitting the experimental distributions. Eigenvalues for the ReLU activation are fitted with the power law $\lambda_n = \Lambda n^{-1-\frac{1}{d}}$, eigenvalues for the Erf activation are fitted with the exponential law $\lambda_n \sim \Lambda e^{-an}$, and coefficients are fitted with the power law $s_n = K n^{-\frac{3}{d}}$.

In the paper we considered only discontinuous activation functions $\phi(z) = (z)_+^q$. The exponent in power law asymptotic in this case depends on activation smoothness q as $\nu = 1 + \frac{2q-1}{d}$, which means that smoother activations produce NTKs with more quickly decreasing eigenvalues. The natural question is what would be the asymptotic of eigenvalues for smooth activation, although our theory does not apply to such activations. As an example of smooth activation we consider the error function $\phi(z) = \mathrm{Erf}(z)$. The NTK of shallow network with the error function activation is calculated in [21], and it is smooth everywhere.

In Figure 4 we can see eigenvalues λ_n and coefficient partial sums s_n for NTKs based on the ReLU and Erf activations. We see that the eigenvalues in the Erf case decrease much faster and can be approximately fitted by an exponential law $\lambda_n \sim \Lambda e^{-an}$. Quite interestingly, for both activations the coefficient distributions s_n behave almost identically, which suggests that the eigenvectors are asymptotically represented by highly oscillating functions regardless of NTK type.

The second experiment is about data distribution. As we mentioned in Section A, the use of symmetric distributions $\mu(\mathbf{x})$ makes the evolution operator $\widetilde{\mathcal{A}}$ also symmetric. In Figure 5 we illustrate this point by plotting eigenvalue distribution for normal Gaussian distribution, uniform distribution on $[-1, 1]^d$ and average of randomly chosen Gaussians as described in Section A. We see that in the case of symmetric data distribution, the eigenvalue distribution has a staircase-like shape, especially for higher dimensions. This is explained by the high degeneracy of the eigenvalues of symmetric operators. However, for distribution $\mu(\mathbf{x})$ made of randomly chosen Gaussians the staircase-like

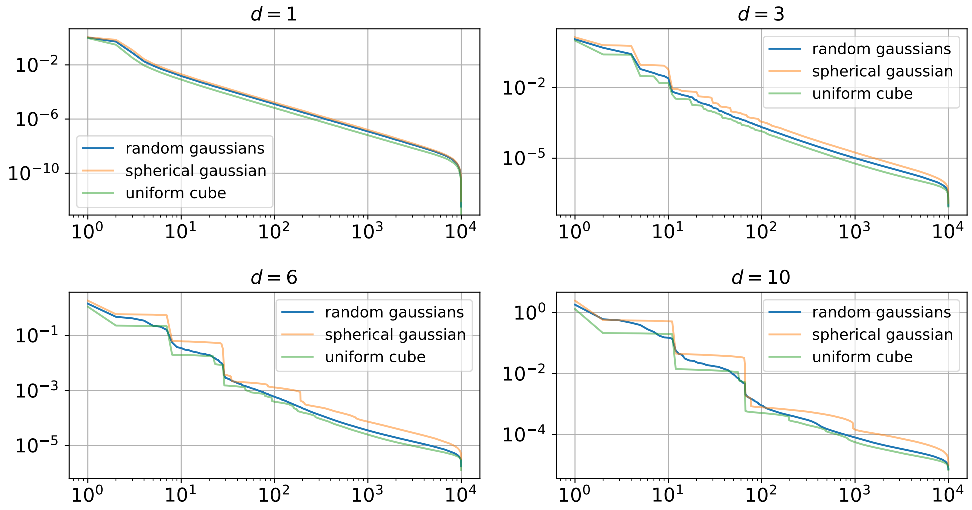


Figure 5: Eigenvalue distribution for different data distribution $\mu(\mathbf{x})$

shape is significantly smoothed, which indicates that such data distribution sufficiently eliminates all the symmetry-based features of corresponding linear operator.