Information and Inference: A Journal of the IMA (2025) **14**, iaaf022
https://doi.org/10.1093/imaiai/iaaf022

Why shallow networks struggle to approximate and learn high frequencies

SHIJUN ZHANG *
Department of Applied Mathematics, Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Kowloon, Hong Kong SAR, China
*Corresponding author. Email: shijun.zhang@polyu.edu.hk

HONGKAI ZHAO
Department of Mathematics, Duke University, 120 Science Drive, Durham, NC 27708, USA

YIMIN ZHONG
Department of Mathematics and Statistics, Auburn University, 221 Parker Hall, Auburn, AL 36849, USA

AND

HAOMIN ZHOU
School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332, USA

[Received on 24 May 2024; revised on 25 March 2025; accepted on 21 May 2025]

In this work, we present a comprehensive study combining mathematical and computational analysis to explain why a two-layer neural network struggles to handle high frequencies in both approximation and learning, especially when machine precision, numerical noise and computational cost are significant factors in practice. Specifically, we investigate the following fundamental computational issues: (1) the minimal numerical error achievable under finite precision, (2) the computational cost required to attain a given accuracy and (3) the stability of the method with respect to perturbations. The core of our analysis lies in the conditioning of the representation and its learning dynamics. Explicit answers to these questions are provided, along with supporting numerical evidence.

Keywords: shallow neural networks; low-pass filter; Gram matrix; generalized Fourier analysis; Radon transform; Rashomon set.

1. Introduction

Neural networks (NNs) are now widely used in machine learning, artificial intelligence and many other areas as a parameterized representation with certain structures for approximating an input-to-output relation, e.g. a function or map in mathematical terms. They have achieved notable successes in practice but also encountered significant challenges. More importantly, many basic and practical questions are still open. Extensive studies have been carried out to understand the properties of NNs and how they work in different perspectives, such as universal approximation property, representation capacity and optimization process (mostly based on gradient descent with different variations), often separately. For example, it has been widely known that NNs can approximate any Lipschitz function with a small error. The approximation theory has been extensively studied for various types of activation functions and

diverse structures of networks [1, 10, 13, 19, 21, 29, 35, 40, 51–56, 60, 65, 66, 69]. Recently, there are works [52, 66, 68] discussing the explicit constructions of the 'optimal' networks of multiple layers. However, one daunting issue that has not been studied systematically in the past is whether such 'optimal' approximations can be possibly attained by training the networks and more importantly, what is the approximation limit in terms of a finite machine precision, computation cost and the property of the function being approximated? Hence, an effective algorithm needs to consider all these aspects to achieve well-balanced accuracy, efficiency and stability. Due to the nonlinear nature of NN representations, this is a challenging task.

In this work, instead of a mathematical study of approximation theory, which usually does not consider the practical constraint of finite machine precision or the cost and stability of finding a good solution, we study a few basic questions from a practice point of view for both approximation and optimization for two-layer NNs. Our consideration includes both asymptotic/continuous and non-asymptotic/discrete regimes in terms of network width:

- the minimal numerical error one can achieve given a finite machine precision;
- the computation time (cost) to achieve a certain accuracy for the training process and
- stability to perturbations, e.g. noise in the data or over-fitting.

Our study shows that, in practice, a shallow NN is essentially a 'low-pass filter' due to the ill-conditioning of the representation which is explicitly characterized by the spectral decay of the Gram matrix, composed of pairwise correlation of the parameterized activation functions and asymptotic equivalence of the eigenmodes to the eigenfunctions of the Laplace operator (generalized Fourier modes) in arbitrary dimensions. More specifically, ill-conditioning of the representation means smooth modes, the number of which depends on the spectral decay rate of the Gram matrix and machine precision, can be captured and stably used for approximation. Although the universal approximation property of two-layer NNs is proved in theory, conditioning of the representation and the finite machine precision determine the achievable numerical accuracy which may be far less than the machine precision in practice, for example, when approximating functions with significant high-frequency components, such as functions with rapid changes and/or fast oscillations. Moreover, the numerical accuracy can not be further improved by increasing the amount of data or the network's width after a certain threshold since the number of eigenmodes that can be stably captured with a given machine precision does not increase. On the other hand, the low pass filter nature leads to certain stability with respect to perturbations in the high modes, e.g. noises or over-parametrization.

One of the most important features when using NNs for approximation is the capability of learning, i.e. optimizing the parameters to adapt to the underlying function manifested by data. However, the initial representation with randomized parameters can not capture those high frequency components needed to represent fine features due to the ill-conditioning and hence can not guide the optimization to achieve the adaptivity effectively. Moreover, we show that ill-conditioning of the representation causes slow learning dynamics for high frequencies, which are needed in an adaptive representation, for gradient-based optimization. Furthermore, the adaptive distribution of parameters can lead to even worse conditioning of the representation and hence even slower learning dynamics. These difficulties make approximation of high frequencies based on learning challenging if not impossible.

From a probabilistic perspective, we show that the Rashomon set, the set of parameters where accurate approximations can be achieved, for a two-layer NN has a small measure for highly oscillatory functions. The measure decreases exponentially with respect to the oscillation frequency. It implies, in practice, both low probabilities of being close to a good approximation for a random initial guess and high computational cost for finding one.

These understandings of the limit of a one-hidden layer network prompt us to study how to use multi-layer to circumvent the limit through effective smooth decomposition and composition in our future work.

1.1 *Literature review*

The approximation theory of shallow NNs has been well-known since the universal approximation theorem [2, 7, 12, 14, 31, 43]. The error bounds of approximation in L^∞ norm have been demonstrated either through explicit construction or by probabilistic proofs, see [15, 27, 31] and the references therein. However, it has been observed widely in practice that shallow NNs cannot approximate highly oscillatory functions effectively [9, 20]. Several explanations have been proposed in the past few years.

The authors of [36, 64] summarized such phenomenon as a heuristic law called *frequency principle*, that is, the training process of the NN recovers lower Fourier frequencies first. Several explanations based on the frequency principle are proposed. When the activation function σ is analytic, e.g. `Tanh` or `Sigmoid`, the authors in [64] have shown that the training of the network at the final stage can be slow due to the inability of smooth activations to pick up the high-frequency components. While for non-smooth activation functions, e.g. `ReLU`, `LReLU` and `ELU`, the authors of [36] provided an interpretation for the *frequency principle* based on the smoothing effect of the activation functions. However, the theory cannot be applied to general cases since it demands strong regularity assumptions for the activation functions and the objective function. In high dimensions, [17] and [47] claimed the bottleneck of the training of shallow networks may come from the so-called 'depth separation'of the capacity between shallow and deep NNs and explained that shallow NNs demand a width that grows exponentially in dimension to fit discontinuous functions in L^2 norm while deep NNs can fit well with a much smaller width.

Another explanation attributed the difficulty to the configuration of the training dynamics. Most of the literature focuses on two regimes: the Neural Tangent Kernel (NTK) regime and the Mean-Field regime. (1) In the NTK regime, the definition of network slightly differs from the classical one, each layer is scaled by a key factor $\frac{1}{\sqrt{n_l}}$ with n_l being the width. If the network is sufficiently wide [34, 67], the parameters of the NN almost freeze, except for the last layer. This configuration can sometimes be viewed as the final stage of training when the parameters are nearly optimal. Under such circumstances, training dynamics has been extensively investigated [16, 28, 33]. With well-distributed labeled data, the slow convergence rate relates to the fast decay rate of the eigenvalues of the neuron-tangent kernel [8, 41, 59, 61]. Studies of the eigenvalues of practical kernels have shown that the decay rates of the leading eigenvalues are closely related to the regularity near the diagonal of the kernel [5], while a fully explicit characterization of all eigenvalues is difficult. (2) In the mean-field regime, the shallow network is commonly used where an additional factor of $\frac{1}{n}$ is applied to the classical one with n being the width. Parameters can be treated as an empirical distribution function or a particle system [38, 46, 57]. As the width of the network becomes infinity, the limiting distribution obeys a gradient flow under the Wasserstein metric and the convergence is proved in [46] for C^1 activation function under the assumption that the empirical measure of particle system converges. A more general class of particle systems has been explored in [11]. However, the corresponding convergence rate is not mentioned.

In addition to the above possible explanations, initialization of parameters may also play a vital role in understanding the difficulty of training NNs to fit highly oscillatory functions. A recent work [24] considered a special setting for the two-layer `ReLU` network that the labeled points and biases are not well distributed and proved that the trained network will not converge to the desired objective function.

1.2 *Contributions*

Our main contributions are summarized below.

- We explicitly characterize the decay rate of the eigenvalues of the Gram kernel corresponding to `ReLU` activation functions (and others) in any dimensions and show that the corresponding eigenfunctions are equivalent to generalized Fourier modes. The corresponding discrete Gram matrix is also analyzed. The study implies that the approximation by a two-layer NN can only maintain a finite number of leading (smooth) modes accurately given a finite machine precision.
- We investigate the nonlinear learning dynamics based on the gradient flow for two-layer `ReLU` networks with finite width in a bounded domain. We show slow learning dynamics for high-frequency modes.
- The measure of Rashomon set, the set of parameters in parameter space that renders an approximation with a given tolerance, for two-layer NNs is characterized. The result shows that oscillatory functions are difficult to represent and learn from a probability perspective.

In this work, we mainly focus on using `ReLU` as the activation function. Our study can be extended to other activation functions as shown in the appendix. Here is the outline of this paper. First, we present a spectral analysis of the Gram matrix and least square approximation in Section 2. Then we study the training dynamics based on gradient descent in Section 3. In Section 4, the probability framework and Rashomon set are employed to show why oscillatory functions are difficult to represent and learn. Extension of the current work is briefly discussed in Section 5.

2. Gram matrix and least square approximation

We first introduce some notations and the general setup for two-layer NNs. Denote $[n] = \{1, 2, \cdots, n\}$ and $C(D)$ the continuous functions on a compact domain $D \subseteq \mathbb{R}^d$. Let $\mathcal{H}_n$ be the hypothesis space generated by shallow feed-forward networks of width n. Each $h \in \mathcal{H}_n$ has the following classical form

$$h(\boldsymbol{x}) = \sum_{i=1}^{n} a_i \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} - b_i) + c \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^d, \tag{2.1}$$

where $n \in \mathbb{N}^+$ is the width of the network, $\boldsymbol{w}_i \in \mathbb{R}^d$, $a_i, b_i, c \in \mathbb{R}$ are parameters for each $i \in [n]$. Finding the best $h \in \mathcal{H}_n$ to approximate the objective function $f(\boldsymbol{x}) \in C(D)$ is usually converted into minimizing the expected (or true) risk

$$\mathcal{L}(h,f) := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{U}(D)}\left[\ell\big(h(\boldsymbol{x}),f(\boldsymbol{x})\big)\right],$$

where $\mathcal{U}$ is some data distribution over D and $\ell(\cdot,\cdot)$ is a loss function. In practice, only finitely many samples $\big\{\big(\boldsymbol{x}_i,f(\boldsymbol{x}_i)\big)\big\}_{i=1}^{N}$ are available and the data distribution is unknown. However, one could approximate the expected risk by the empirical risk $\mathcal{L}_{\text{emp}}(h,f)$, which is given by

$$\mathcal{L}_{\text{emp}}(h,f) := \frac{1}{N}\sum_{i=1}^{N} \ell\big(h(\boldsymbol{x}_i),f(\boldsymbol{x}_i)\big).$$

In this paper, we let $\mathcal{U}$ be the uniform distribution and $\ell(y,y') = |y-y'|^2$, implying

$$\mathcal{L}(h,f) = \int_D |h(\boldsymbol{x}) - f(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x} \quad \text{and} \quad \mathcal{L}_{\text{emp}}(h,f) = \frac{1}{N}\sum_{i=1}^{N} |h(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)|^2.$$

A learning/training process is to identify $h^* \in \mathcal{H}_n$ or $\widehat{h} \in \mathcal{H}_n$ such that

$$h^* \in \arg\min_{h\in\mathcal{H}_n} \mathcal{L}(h,f) \quad \text{or} \quad \widehat{h} \in \arg\min_{h\in\mathcal{H}_n} \mathcal{L}_{\text{emp}}(h,f).$$

This study investigates the approximation capabilities of two-layer NNs within the mentioned framework. We aim to address the three basic questions outlined in the abstract that commonly arise in practical settings.

We start with a study on approximation properties of a two-layer NN as a linear representation, i.e. where the weights and biases in the hidden neurons are fixed, using least squares. In this setting, the solution can be found by solving a linear system involving the Gram matrix, the normal equation. The most fundamental question is the basis of the representation, i.e. $\{\sigma(\boldsymbol{w}_i\cdot\boldsymbol{x} - b_i), i\in[n]\}$: space the basis span and the correlation among the basis. Desirable features of a good basis for computational efficiency, accuracy and stability in practice are sparsity and well-conditioning of the Gram matrix. In other words, global interactions and strong correlations should be avoided.

We start with the most used activation function in NNs is `ReLU`: $\sigma(x) := \max(x,0)$. The general form of a shallow network in (2.1) can be simplified to

$$h(\boldsymbol{x}) = \sum_{i=1}^{n} a_i\sigma(\boldsymbol{w}_i\cdot\boldsymbol{x} - b_i) + c, \quad \boldsymbol{x}\in D\subseteq\mathbb{R}^d,\ \boldsymbol{w}_i\in\mathbb{S}^{d-1},\ a_i,b_i\in\mathbb{R}. \tag{2.2}$$

2.1 *One-dimensional case*

In one dimension, we let $D=[-1,1]$. Due to the affinity of `ReLU` and the transform $\sigma(x) = x - \sigma(-x)$, the form of a shallow network in (2.2) can be further simplified to

$$h(x) = c + vx + \sum_{i=1}^{n} a_i\sigma(x-b_i), \quad x\in D,\ c,v,a_i,b_i\in\mathbb{R}.$$

Since we only consider the approximation inside D, we may further reduce the network into $h(x) = c + \sum_{i=1}^{n} a_i\sigma(x-b_i)$ by setting $c = f(-1)$ and $b_i\in D$ as well. With fixed $b_i\in D$, `ReLU` functions $\sigma(x-b_i)$ span the same continuous piecewise linear (in each sub-intervals between b_i's) function space as the linear finite element basis (hat functions) with nodes $\{b_i\}_{i=1}^n$. Mathematically, they are the same when used in approximating a function in the domain D in the least square setting. The minimizer is the L^2 projection of $f(x)$ onto the continuous piecewise linear function space. However, the major difference in practice is the Gram matrix (mass matrix in finite element terminology or normal matrix in linear algebra terminology) of the basis, which defines the linear system one needs to solve numerically to find the best approximation. Using the finite element basis, which is local and decorrelated, the Gram matrix is sparse and well-conditioned. The condition number is proportional to the ratio between the sizes of the

maximal sub-interval and the minimal sub-interval [30]. While using `ReLU` functions, which are non-local and can be highly correlated, the Gram matrix is dense and ill-conditioned, as we will show below. As a consequence, (1) computation and memory costs involving the Gram matrix can be very expensive and (2) only those functions close to the linear space spanned by the leading eigenvectors of the Gram matrix can be approximated well. The number of the leading eigenvectors that can be used stably and accurately depends on the decay rate of the eigenvalues, machine precision and/or noise level. Now we present a spectral analysis of the Gram matrix for a set of `ReLU` functions.

Denote the Gram matrix $\boldsymbol{G} := (\boldsymbol{G}_{i,j}) \in \mathbb{R}^{n\times n}$, where

$$\boldsymbol{G}_{i,j} := \int_D \sigma(x-b_i)\sigma(x-b_j)\,\mathrm{d}x = \frac{1}{24}(2-b_i-b_j-|b_i-b_j|)^2(2-b_i-b_j+2|b_i-b_j|).$$

The correlation between two `ReLU` functions with close biases b_i, b_j is $1-O(|b_i-b_j|^2)$ and this strong correlation suggests ill-conditioning of the Gram matrix. To fully understand the spectrum property of $\boldsymbol{G}$, we define the corresponding Gram kernel function $\mathcal{G}:\mathbb{R}\times\mathbb{R}\mapsto\mathbb{R}$ as

$$\mathcal{G}(x,y) := \int_D \sigma(z-x)\sigma(z-y)\,\mathrm{d}z. \tag{2.3}$$

In particular, if we restrict $x, y \in D$,

$$\begin{aligned}\mathcal{G}(x,y) &:= \frac{1}{24}(2-x-y-|x-y|)^2(2-x-y+2|x-y|)\\ &= \frac{1}{12}|x-y|^3 + \frac{1}{12}(2-x-y)\left(2(1-x)(1-y)-(x-y)^2\right).\end{aligned} \tag{2.4}$$

First, we provide an explicit spectral characterization of the Gram kernel (2.4). Define the operator $K: L^2[-1,1]\to L^2[-1,1]$

$$Kh(x) = \int_{-1}^{1}\mathcal{G}(x,y)h(y)\,\mathrm{d}y.$$

Let μ_k, $k=1,2,\cdots$, be the eigenvalue of K in descending order and ϕ_k be the corresponding eigenfunction which satisfies

$$\int_{-1}^{1}\mathcal{G}(x,y)\phi_k(y)\,\mathrm{d}y = \mu_k\phi_k(x). \tag{2.5}$$

Taking derivatives four times on both sides, using $\frac{\partial^4}{\partial x^4}\mathcal{G}(x,y) = \delta(x-y)$ we obtain the following differential equation,

$$\phi_k^{(4)}(x) = \frac{1}{\mu_k}\phi_k(x), \tag{2.6}$$

which implies that there are constants $A_k, B_k, C_k, D_k \in \mathbb{C}$ such that

$$\phi_k(x) = A_k \cosh(w_k x) + B_k \sinh(w_k x) + C_k \cos(w_k x) + D_k \sin(w_k x)$$

for certain $w_k = \mu_k^{-\frac{1}{4}} > 0$. Furthermore, one can easily check that $\mathcal{G}(1, y) = \mathcal{G}_x(1, y) = \mathcal{G}_{xx}(-1, y) = \mathcal{G}_{xxx}(-1, y) = 0$, which implies the following boundary conditions for ϕ_k,

$$\phi_k(1) = \phi_k'(1) = \phi_k''(-1) = \phi_k'''(-1) = 0,$$

which determine A_k, B_k, C_k, D_k explicitly. We show that ϕ_k are asymptotically Fourier modes from low frequencies to high frequencies. Here we summarize the results while the detailed calculations and proofs can be found in Appendix B.

- $w_{2j+1} \in ((j+\frac{1}{4})\pi, (j+\frac{1}{2})\pi)$, $w_{2j+2} \in ((j+\frac{1}{2})\pi, (j+\frac{3}{4})\pi)$, $j \geq 0$ and $\lambda_k = w_k^{-4} \sim (\frac{k\pi}{2})^{-4}$,
- if $k = 2j+1$, $\phi_k(x) = C_k(-\frac{\cos(w_k)}{\sinh(w_k)} \sinh(w_k x) + \cos(w_k x))$, $C_k = \mathcal{O}(1)$, if $k = 2j+2$, $\phi_k(x) = D_k(-\frac{\sin(w_k)}{\cosh(w_k)} \cosh(w_k x) + \sin(w_k x))$, $D_k = \mathcal{O}(1)$
- $\{\phi_k\}_{k\geq 1}$ forms an orthonormal basis of $L^2(D)$ and

$$\|\phi_k\|_{L^\infty(D)} = \mathcal{O}(1),\ \|\phi_k'\|_{L^\infty(D)} = \mathcal{O}(k),\ \|\phi_k''\|_{L^\infty(D)} = \mathcal{O}(k^2),$$

$$\|\phi_{2j+1}(x) - \cos(w_{2j+1}x)\|_{L^2(D)} = \mathcal{O}(j^{-1/2}),\ \|\phi_{2j+2}(x) - \sin(w_{2j+2}x)\|_{L^2(D)} = \mathcal{O}(j^{-1/2}).$$

Next, we study the spectral properties of the discrete Gram matrix. Earlier work [25] studied discrete Gram matrix on the uniform grid in one dimension. We will prove results in more general settings and higher dimensions. Denote the vectors $\boldsymbol{a} := (a_i)_{i=1}^n$ and $\boldsymbol{f} := (f_i)_{i=1}^n$ that $f_i = \int_D f(x)\sigma(x - b_i)\,\mathrm{d}x$. The least-square solution $\boldsymbol{a} \in \mathbb{R}^n$, when the biases are fixed, is $\boldsymbol{a} = \boldsymbol{G}^\dagger \boldsymbol{f}$, where $\boldsymbol{G}^\dagger$ is the pseudo-inverse of $\boldsymbol{G}$. Without loss of generality, we assume that $b_i \neq b_j$ for all $i \neq j$ and the biases are sorted in ascending order, that is, $b_1 < b_2 < \cdots < b_n$. We first provide an estimate for the eigenvalue estimates of the Gram matrix $\boldsymbol{G}$. The rescaled matrix $\boldsymbol{G}_n = \frac{1}{n}\boldsymbol{G}$ is the so-called *kernel matrix* for $\mathcal{G}$, which plays an important role in kernel methods [32].

THEOREM 2.1. Suppose $\{b_i\}_{i=1}^n$ are quasi-evenly spaced on D, $b_i = -1 + \frac{2(i-1)}{n} + o\left(\frac{1}{n}\right)$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be the eigenvalues of the Gram matrix $\boldsymbol{G}$, then $|\lambda_k - \frac{n}{2}\mu_k| \leq C$ for some constant $C = \mathcal{O}(1)$, where $\mu_k = \Theta(k^{-4})$ is the kth eigenvalue of $\mathcal{G}$.

Proof. The idea of proof comes from [63]. Define the operator K^* by the kernel

$$\mathcal{G}^*(x, y) = \mathcal{G}\left(-1 + \frac{2}{n}\left\lfloor \frac{x+1}{2} \right\rfloor, -1 + \frac{2}{n}\left\lfloor \frac{y+1}{2} \right\rfloor\right)$$

where $\lfloor\cdot\rfloor$ is the floor function. We denote the eigenvalues of K^* as $\mu_1^* \geq \mu_2^* \geq \cdots$, then for the equispaced biases $b_i^* = -1 + \frac{2(i-1)}{n}$, the corresponding Gram matrix $\boldsymbol{G}^*$ has the eigenvalues exactly

$\lambda_i^* = \frac{n}{2}\mu_i^*$. Using Weyl's inequality for self-adjoint compact operators

$$|\mu_i - \mu_i^*| \leq \|K - K^*\| \leq \sqrt{\int_{-1}^{1}\int_{-1}^{1} |\mathcal{G}(x,y) - \mathcal{G}^*(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y} = \mathcal{O}(n^{-1}). \tag{2.7}$$

Now we consider perturbed $\widetilde{b}_i = b_i^* + o(\frac{1}{n})$ as mentioned in Theorem 2.1, let the corresponding Gram matrix be $\widetilde{\boldsymbol{G}}$, then by Weyl's inequality for Hermitian matrices

$$\|\lambda_i(\widetilde{\boldsymbol{G}}) - \lambda_i(\boldsymbol{G}^*)\| \leq \|\widetilde{\boldsymbol{G}} - \boldsymbol{G}^*\| \leq \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} \left|\mathcal{G}(\widetilde{b}_i, \widetilde{b}_j) - \mathcal{G}(b_i^*, b_j^*)\right|^2} = o(1).$$

Therefore $|\lambda_i(\widetilde{\boldsymbol{G}}) - \frac{n}{2}\mu_i| \leq C$ for some positive constant $C > 0$. □

THEOREM 2.2. Suppose $\{b_i\}_{i=1}^n$ are chosen as Theorem 2.1, then the condition number of the Gram matrix $\boldsymbol{G}$ satisfies

$$\kappa = \lambda_1/\lambda_n = \Omega(n^3),$$

where Ω is the big Omega notation.

Proof. The kernel $\mathcal{G}(x,y)$ permits the expansion

$$\mathcal{G}(x,y) = \sum_{k=1}^{\infty} \mu_k \phi_k(x)\phi_k(y). \tag{2.8}$$

Let $\mathcal{G}_m(x,y) := \sum_{k=1}^{m} \mu_k\phi_k(x)\phi_k(y)$ be the truncated expansion, where $m \geq 1$ is the truncation parameter. Then the Gram matrix can be decomposed into

$$\boldsymbol{G}_{i,j} = \mathcal{G}_m(b_i, b_j) + (\mathcal{G}(b_i, b_j) - \mathcal{G}_m(b_i, b_j)). \tag{2.9}$$

Define the matrix $\boldsymbol{\Phi}_m \in \mathbb{R}^{n\times m}$ with entries $(\boldsymbol{\Phi}_m)_{i,j} = \phi_j(b_i)$, $1 \leq j \leq m$, then

$$\boldsymbol{G} = \boldsymbol{\Phi}_m \boldsymbol{\Lambda}_m \boldsymbol{\Phi}_m^T + \boldsymbol{E}_m, \tag{2.10}$$

where $(\boldsymbol{\Lambda}_m)_{i,j} = \delta_{i,j}\mu_i$ and $(\boldsymbol{E}_m)_{i,j} = \sum_{k>m} \mu_k\phi_k(b_i)\phi_k(b_j)$. Let $\sigma_i(\boldsymbol{\Phi}_m\boldsymbol{\Lambda}_m\boldsymbol{\Phi}_m^T)$ denote the ith eigenvalue of $\boldsymbol{\Phi}_m\boldsymbol{\Lambda}_m\boldsymbol{\Phi}_m^T$, then by Ostrowski's theorem [26],

$$\left|\sigma_i(\boldsymbol{\Phi}_m\boldsymbol{\Lambda}_m\boldsymbol{\Phi}_m^T) - \frac{n}{2}\mu_i\right| \leq |\mu_i| \left\|\boldsymbol{\Phi}_m^T\boldsymbol{\Phi}_m - \frac{n}{2}\mathrm{Id}_m\right\|_{\mathrm{op}}, \quad 1 \leq i \leq m \leq n. \tag{2.11}$$

Therefore, using Weyl's inequality

$$\begin{aligned}\left|\lambda_i - \frac{n}{2}\mu_i\right| &\le |\lambda_i - \sigma_i(\Phi_m \Lambda_m \Phi_m^T)| + \left|\sigma_i(\Phi_m \Lambda_m \Phi_m^T) - \frac{n}{2}\mu_i\right| \\ &\le \|\boldsymbol{E}_m\|_{\mathrm{op}} + |\mu_i| \left\|\Phi_m^T \Phi_m - \frac{n}{2}\mathrm{Id}_m\right\|_{\mathrm{op}}, \qquad 1 \le i \le m \le n.\end{aligned} \tag{2.12}$$

Since the eigenfunctions ϕ_k are uniformly bounded, see Theorem B.4 in Appendix B, then

$$\|\boldsymbol{E}_m\|_{\mathrm{op}} \le Cn \sum_{k>m} \mu_k = \mathcal{O}\left(\frac{n}{m^3}\right). \tag{2.13}$$

The entries in $\Phi_m \Phi_m^T - \frac{n}{2}\mathrm{Id}_m$ can be estimated by the standard numerical quadrature analysis on abscissas $\{b_i\}_{i=1}^n$. Indeed,

$$\frac{2}{n}\sum_{i=1}^{n} \phi_j(b_i)\phi_k(b_i) - \int_{-1}^{1} \phi_j(x)\phi_k(x)\,\mathrm{d}x = \mathcal{O}\left(\frac{1}{n}\sup_{[-1,1]} (\phi_j\phi_k)'\right). \tag{2.14}$$

Hence using the estimate $\|\phi_k'\|_\infty = \mathcal{O}(k)$, $\left\|\Phi_m^T\Phi_m - \frac{n}{2}\mathrm{Id}_m\right\|_{\mathrm{op}} = \mathcal{O}(m^2)$. Combine the above estimates into (2.12) and reuse Theorem 2.1, by selecting $m = n^{1/5} i^{4/5} \ge i$,

$$\left|\lambda_i - \frac{n}{2}\mu_i\right| \le C \min\left(1, \frac{n}{m^3} + \frac{m^2}{i^4}\right) = \begin{cases} \mathcal{O}(1), & i < n^{\frac{1}{6}}, \\ \mathcal{O}(n^{2/5} i^{-12/5}), & n^{\frac{1}{6}} \le i, \le n. \end{cases} \tag{2.15}$$

It implies that $\lambda_1 = \Theta(n)$ and $\lambda_n = \mathcal{O}(n^{-2})$, which leads to the condition number estimate $\kappa = \Omega(n^3)$. □

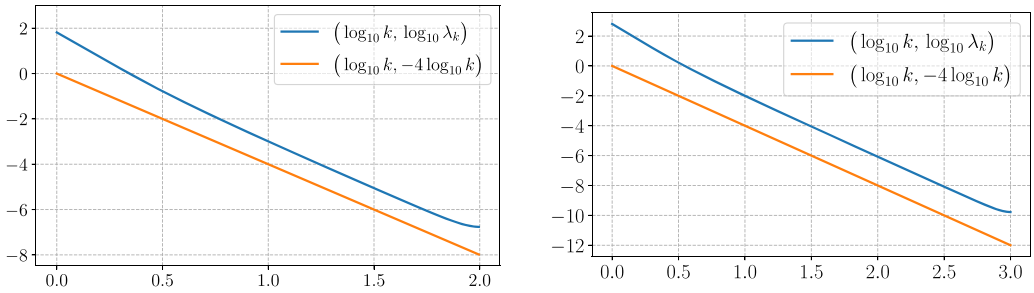
Remark 2.3. For evenly spaced biases, explicit computations for the eigenvalues in [25] show that the condition number is $\Omega(n^4)$. However, a sharp lower bound for unevenly distributed biases (grid points) is difficult. Here we use (1) Ostrowski's theorem to relate the eigenvalues between two symmetric positive definite matrices and (2) random quadrature points for integral estimation. However, we believe that for a fixed number of points in an interval, non-evenly spaced points will result in a larger condition number for the Gram matrix than equally spaced points, which seems also suggested by our numerical tests, see Figs 1 and 2 in Section 2.3.

Generally speaking, if $\{b_i\}_{i=1}^n$ are distributed i.i.d with probability density function $\rho : D \mapsto \mathbb{R}$, one can reformulate the matrix-vector multiplication as a ρ weighted integral as the continuous limit. The discrete eigen system in the limit corresponds to that of the modified continuous kernel $\mathcal{G}_\rho(x,y) := \sqrt{\rho(x)}\mathcal{G}(x,y)\sqrt{\rho(y)}$. For this case, we have a similar estimate of eigenvalues if ρ is bounded from below and above by positive constants.

Lemma 2.4. Suppose $\rho(x)$ is bounded from below and above by positive constants and define

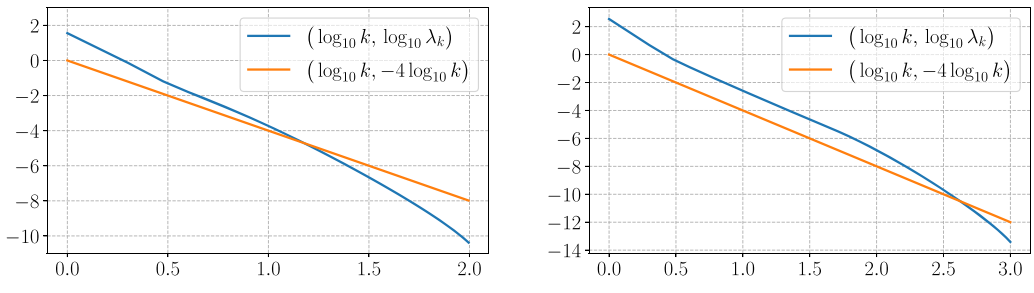
$$\mathcal{G}_\rho(x,y) := \sqrt{\rho(x)}\mathcal{G}(x,y)\sqrt{\rho(y)}.$$

Let $\widetilde{\mu}_k$ be the kth eigenvalue of $\mathcal{G}_\rho$ in descending order, then $\inf_{[-1,1]}\sqrt{\rho} \le \widetilde{\mu}_k/\mu_k \le \sup_{[-1,1]}\sqrt{\rho}$.



(a) $n = 100$. (b) $n = 1000$.

FIG. 1. Spectrum for uniform $\boldsymbol{b}$.



(a) $n = 100$. (b) $n = 1000$.

FIG. 2. Spectrum for adaptive $\boldsymbol{b}$.

Proof. By Min–Max theorem for the eigenvalues of the integral kernel $\mathcal{G}_\rho$,

$$\widetilde{\mu}_k = \max_{S_k} \min_{z \in S_k, \|z\|=1} \int_{-1}^{1} \int_{-1}^{1} \mathcal{G}_\rho(x,y) z(x) z(y) \, \mathrm{d}x \, \mathrm{d}y,$$

$$\widetilde{\mu}_k = \min_{S_{k-1}} \max_{z \in S_{k-1}^{\perp}, \|z\|=1} \int_{-1}^{1} \int_{-1}^{1} \mathcal{G}_\rho(x,y) z(x) z(y) \, \mathrm{d}x \, \mathrm{d}y,$$

where S_k is a k dimensional subspace of $L^2[-1,1]$. In the first equation, we choose the space $S_k = \mathrm{span}(\frac{\phi_1}{\sqrt{\rho}}, \cdots, \frac{\phi_k}{\sqrt{\rho}})$, where (μ_j, ϕ_j) denotes the jth eigenpair of the kernel $\mathcal{G}$. Let

$$\widehat{z} = \operatorname*{arg\,min}_{z \in S_k, \|z\|=1} \int_{-1}^{1} \int_{-1}^{1} \mathcal{G}_\rho(x,y) z(x) z(y) \, \mathrm{d}x \, \mathrm{d}y, \quad \widehat{z} = \frac{1}{\sqrt{\rho}} \sum_{j=1}^{k} c_j \phi_j, \quad \|\widehat{z}\| = 1.$$

We have

$$\widetilde{\mu}_k \geq \sum_{j=1}^{k} \mu_j c_j^2 \geq \mu_k \sum_{j=1}^{k} c_j^2 = \mu_k \|\sqrt{\rho}\widetilde{z}\| \geq \mu_k \inf \sqrt{\rho}.$$

In the second equation, we choose the space $S_{k-1} = \operatorname{span}(\sqrt{\rho}\phi_1, \cdots, \sqrt{\rho}\phi_{k-1})$ and let

$$\widetilde{z} = \underset{z \in S_{k-1}^{\perp}, \|z\|=1}{\arg\max} \int_{-1}^{1}\int_{-1}^{1} \mathcal{G}_\rho(x,y) z(x) z(y)\, \mathrm{d}x\, \mathrm{d}y, \quad \widetilde{z} = \frac{1}{\sqrt{\rho}} \sum_{j=k}^{\infty} c_j \phi_j \in S_{k-1}^{\perp}, \quad \|\widetilde{z}\| = 1,$$

then

$$\widetilde{\mu}_k \leq \sum_{j=k}^{\infty} \mu_j c_j^2 \leq \mu_k \sum_{j=k}^{\infty} c_j^2 = \mu_k \|\sqrt{\rho}\widetilde{z}\| \leq \mu_k \sup \sqrt{\rho}.$$

□

If the density function ρ is regular enough, we show a more precise characterization of the eigenvalues and that the eigenfunctions are asymptotically Fourier series. Moreover, the following probabilistic estimate for the eigenvalues of the Gram matrix can be derived.

THEOREM 2.5. Suppose $\{b_i\}_{i=1}^n$ are i.i.d with probability density function $\rho \in C^3[-1,1]$ on D such that $0 < \underline{c} \leq \rho(x) \leq \bar{c} < \infty$. Let $\widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_n \geq 0$ be the eigenvalues of the corresponding Gram matrix $\boldsymbol{G} := (\mathcal{G}(b_i, b_j))_{1 \leq i,j \leq n}$, then for sufficiently large n,

$$\left|\widetilde{\lambda}_i - \frac{n}{2}\widetilde{\mu}_i\right| = \begin{cases} \mathcal{O}\left(n^{\frac{5}{8}} i^{-3} \sqrt{\log \frac{n}{p}}\right), & i < n^{\frac{7}{8}}, \\ \mathcal{O}\left(n^{-2} \sqrt{\log \frac{n}{p}}\right), & n^{\frac{7}{8}} \leq i \leq n, \end{cases} \tag{2.16}$$

with probability $1 - p$, where $\widetilde{\mu}_i = \Theta(i^{-4})$ is the kth eigenvalue of $\mathcal{G}_\rho$.

Proof. Without loss of generality, we assume $b_1 \leq b_2 \leq \cdots \leq b_n$. Let $\phi_{\rho,k}$ be eigenfunction for the kth eigenvalue of $\mathcal{G}_\rho$. Using similar derivation as before, we obtain the differential equation

$$\frac{1}{\widetilde{\mu}_k} \sqrt{\rho(x)} \phi_{\rho,k}(x) = \frac{\mathrm{d}^4}{\mathrm{d}x^4} \left[\frac{1}{\sqrt{\rho(x)}} \phi_{\rho,k}(x)\right] \tag{2.17}$$

with boundary conditions $\phi_{\rho,k}(1) = \phi'_{\rho,k}(1) = \phi''_{\rho,k}(-1) = \phi'''_{\rho,k}(-1) = 0$. Denote $\psi_{\rho,k}(x) := \frac{1}{\sqrt{\rho(x)}} \phi_{\rho,k}(x)$, then the above equation (2.17) becomes

$$\psi_{\rho,k}^{(4)} = \frac{1}{\widetilde{\mu}_k} \rho(x) \psi_{\rho,k}. \tag{2.18}$$

Using a change of variable in the spirit of the Liouville transform,

$$t = -1 + \frac{2}{H}\int_{-1}^{x} \rho(s)^{1/4}\,\mathrm{d}s, \quad H = \int_{-1}^{1} \rho(s)^{1/4}\,\mathrm{d}s.$$

One can find that the differential equation (2.18) reduces to the form

$$\frac{\mathrm{d}^4}{\mathrm{d}t^4}\psi_{\rho,k} + p_1(t)\frac{\mathrm{d}^3}{\mathrm{d}t^3}\psi_{\rho,k} + p_2(t)\frac{\mathrm{d}^2}{\mathrm{d}t^2}\psi_{\rho,k} + p_3(t)\frac{\mathrm{d}}{\mathrm{d}t}\psi_{\rho,k} + p_4(t)\psi_{\rho,k} = \frac{H^4}{\widetilde{\mu}_k}\psi_{\rho,k}, \tag{2.19}$$

and the functions $p_k(t)$ are continuous over $[-1,1]$ since $\rho \in C^3[-1,1]$. For k sufficiently large, the asymptotic behaviours of eigenvalues $\widetilde{\mu}_k^{-1} = (\frac{k}{H}\pi)^4(1+\mathcal{O}(k^{-1}))$ can be derived based on Stone's estimate of linearly independent basis [58] and Birkhoff's method [4, 39]. Furthermore, the eigenfunction $\psi_{\rho,k}$ is asymptotically equivalent to Fourier modes in t for sufficiently large k and can be shown uniformly bounded, see detailed discussions in §4.10 of [39].

Similar to Theorem 2.2, we define the matrix $\boldsymbol{\Phi}_m \in \mathbb{R}^{n\times m}$ with entries $(\boldsymbol{\Phi}_m)_{i,j} = \phi_{\rho,j}(b_i)/\sqrt{\rho(b_i)}$, $1 \le j \le m$, then the Gram matrix with entry $\boldsymbol{G}_{i,j} = \mathcal{G}(b_i, b_j)$ equals to

$$\boldsymbol{G} = \boldsymbol{\Phi}_m \Lambda_m \boldsymbol{\Phi}_m^T + \boldsymbol{E}_m, \tag{2.20}$$

where $(\Lambda_m)_{i,j} = \delta_{i,j}\widetilde{\mu}_i$, $(\boldsymbol{E}_m)_{i,j} = \sum_{k>m}\widetilde{\mu}_k\phi_{\rho,k}(b_i)\phi_{\rho,k}(b_j)/\sqrt{\rho(b_i)\rho(b_j)}$, and the estimate (2.13) still holds. For each pair of k,j, applying the Hoeffding's inequality to (2.14), we get

$$\left|\sum_{i=1}^{n}\frac{\phi_{\rho,j}(b_i)\phi_{\rho,k}(b_i)}{\rho(b_i)} - \frac{n}{2}\int_{-1}^{1}\phi_{\rho,j}(x)\phi_{\rho,k}(x)\,\mathrm{d}x\right| = \mathcal{O}\left(\sqrt{n\log\frac{m^2}{p}}\right)$$

with probability $1-\frac{p}{m^2}$, due to the uniform boundedness of the eigenfunctions $\phi_{\rho,k}$. Then with probability at least $1-p$, we have $\|\boldsymbol{\Phi}_m^T\boldsymbol{\Phi}_m - \frac{n}{2}\mathrm{Id}_m\|_{\mathrm{op}} = \mathcal{O}\left(m\sqrt{n\log\frac{m^2}{p}}\right)$ and

$$\left|\widetilde{\lambda}_i - \frac{n}{2}\widetilde{\mu}_i\right| \le C\min_{i\le m\le n}\left(\frac{n}{m^3} + \frac{m}{i^4}\sqrt{n\log\frac{m^2}{p}}\right) = \begin{cases}\mathcal{O}\left(n^{\frac{5}{8}}i^{-3}\sqrt{\log\frac{n}{p}}\right), & i < n^{\frac{7}{8}},\\ \mathcal{O}\left(n^{-2}\sqrt{\log\frac{n}{p}}\right), & n^{\frac{7}{8}} \le i \le n,\end{cases} \tag{2.21}$$

where $m = \min(n^{\frac{1}{8}}i, n)$. □

Corollary 2.6. Under the same assumption of Corollary 2.5, with at probability $1-p$, $1 > p > ne^{-cn^{3/4}}$ for some $0 < c = \mathcal{O}(1)$, the condition number of Gram matrix $\boldsymbol{G}$ satisfies

$$\kappa = \lambda_1/\lambda_n = \Omega\left(n^3\left(\log\frac{n}{p}\right)^{-\frac{1}{2}}\right).$$

Proof. There exist positive constants C_1, C_2, C_n of $\mathcal{O}(1)$ such that

$$\lambda_1 \geq C_1 n \left(1 - C_2 n^{-3/8}\sqrt{\log \frac{n}{p}}\right), \quad \lambda_n \leq C_n n^{-2}\sqrt{\log \frac{n}{p}}.$$

Choose $c = \frac{1}{2C_2} = \mathcal{O}(1)$, then $\lambda_1 > \frac{C_1 n}{2}$. □

2.2 *Multi-dimensional case*

Now we provide the spectral analysis for `ReLU` functions in arbitrary dimensions and give a spectral estimate, although we cannot compute the eigenvalues and eigenfunctions explicitly. In this section, we consider domain $D = B_d(1)$ which is the unit ball in d-dimension. The class of NN $\mathcal{H}_n$ is

$$h(\boldsymbol{x}) = c + \sum_{i=1}^{n} a_i \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} - b_i), \quad \boldsymbol{w}_i \in \mathbb{S}^{d-1},\ b_i \in [-1, 1].$$

Denote $V = \mathbb{S}^{d-1} \times [-1, 1]$, similar to the one-dimensional setting, we consider the corresponding continuous kernel $G : L^2(V) \mapsto L^2(V)$

$$G(\boldsymbol{w}, b, \boldsymbol{w}', b') = \int_D \sigma(\boldsymbol{w} \cdot \boldsymbol{x} - b)\sigma(\boldsymbol{w}' \cdot \boldsymbol{x} - b')\,\mathrm{d}\boldsymbol{x}.$$

Let $\phi_k(\boldsymbol{w}, b)$ be an eigenfunction for eigenvalue λ_k which satisfies

$$\int_V G(\boldsymbol{w}, b, \boldsymbol{w}', b')\phi_k(\boldsymbol{w}', b')\,\mathrm{d}\boldsymbol{w}'\,\mathrm{d}b' = \lambda_k \phi_k(\boldsymbol{w}, b). \tag{2.22}$$

It is not hard to see that $\phi_k(\boldsymbol{w}, b)$ is supported on V and $\phi_k(\boldsymbol{w}, 1) = \partial_b \phi_k(\boldsymbol{w}, 1) = \partial_b^2 \phi_k(\boldsymbol{w}, 1) = 0$.

One of the useful tools to study two-layer `ReLU` networks of infinite width is the Radon transform [42, 48]. Next, we construct the theory for the Gram matrix in high dimensions using the properties of the Radon transform.

Definition 2.7. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an integrable function over all hyperplanes, the Radon transform

$$\mathcal{R}f(\boldsymbol{w}, b) = \int_{\{\boldsymbol{x} | \boldsymbol{w}\cdot\boldsymbol{x} - b = 0\}} f(\boldsymbol{x})\,\mathrm{d}H_{d-1}(\boldsymbol{x}), \quad \forall (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R},$$

where $\mathrm{d}H_{d-1}$ denotes the $(d-1)$ dimensional Lebesgue measure. The adjoint transform $\mathcal{R}^* : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}^d$ is

$$\mathcal{R}^*\Phi(\boldsymbol{x}) := \int_{\mathbb{S}^{d-1}} \Phi(\boldsymbol{w}, \boldsymbol{w} \cdot \boldsymbol{x}) d\boldsymbol{w}, \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Theorem 2.8 (Helgason [23]). The inversion formula of the Radon transform is

$$c_d f = (-\Delta)^{(d-1)/2} \mathcal{R}^* \mathcal{R} f,$$

where $c_d = (4\pi)^{(d-1)/2} \frac{\Gamma(d/2)}{\Gamma(1/2)}$.

LEMMA 2.9 (Helgason [22], Lemma 2.1). These intertwining relations hold: $\mathcal{R}\Delta = \partial_b^2\mathcal{R}$ and $\mathcal{R}^*\partial_b^2 = \Delta\mathcal{R}^*$.

Then we have the following lemma for the eigenvalues.

LEMMA 2.10. Let $u_k(\boldsymbol{x}) = \int_V \sigma(\boldsymbol{w}\cdot\boldsymbol{x} - b)\phi_k(\boldsymbol{w}, b)\,\mathrm{d}\boldsymbol{w}\,\mathrm{d}b$ and $\chi_D(\boldsymbol{x})$ be the characteristic function of D, then if d is odd,

$$c_d u_k(\boldsymbol{x}) = \lambda_k(-\Delta)^{(d+3)/2}u_k(\boldsymbol{x}), \quad \forall\,\boldsymbol{x}\in D.$$

If d is even

$$c_d(-\Delta)^{-1/2}\chi_D u_k(\boldsymbol{x}) = \lambda_k(-\Delta)^{(d+3)/2}(-\Delta)^{-1/2}\chi_D u_k(\boldsymbol{x}), \quad \forall\,\boldsymbol{x}\in D.$$

Proof. The function $u_k(\boldsymbol{x})$ satisfies

$$\Delta u_k(\boldsymbol{x}) = \int_V \delta(\boldsymbol{w}\cdot\boldsymbol{x} - b)\phi_k(\boldsymbol{w}, b)\,\mathrm{d}\boldsymbol{w}\,\mathrm{d}b = \mathcal{R}^*\phi_k(\boldsymbol{x}), \quad \forall\,\boldsymbol{x}\in D. \tag{2.23}$$

We also define

$$\widetilde{\phi}_k(\boldsymbol{w}, b) := \frac{1}{\lambda_k}\int_D u_k(\boldsymbol{x})\sigma(\boldsymbol{w}\cdot\boldsymbol{x} - b)\,\mathrm{d}\boldsymbol{x}, \quad \forall(\boldsymbol{w}, b)\in\mathbb{S}^{d-1}\times\mathbb{R}, \tag{2.24}$$

then $\phi_k(\boldsymbol{w}, b) = \widetilde{\phi}_k(\boldsymbol{w}, b)$ on V, hence $\mathcal{R}^*\phi_k = \mathcal{R}^*\widetilde{\phi}_k$ on D. Differentiate (2.24) twice in b,

$$\lambda_k\partial_b^2\widetilde{\phi}_k = \int_D u_k(\boldsymbol{x})\delta(\boldsymbol{w}\cdot\boldsymbol{x} - b)\,\mathrm{d}\boldsymbol{x} = \mathcal{R}\chi_D u_k, \quad \forall(\boldsymbol{w}, b)\in\mathbb{S}^{d-1}\times\mathbb{R}. \tag{2.25}$$

For odd d, apply $(-\Delta)^{(d-1)/2}\mathcal{R}^*$ on both sides of (2.25) and use Lemma 2.9, we obtain

$$\lambda_k(-\Delta)^{(d-1)/2}\mathcal{R}^*\partial_b^2\widetilde{\phi}_k = \lambda_k(-\Delta)^{(d-1)/2}\Delta\mathcal{R}^*\widetilde{\phi}_k = (-\Delta)^{(d-1)/2}\mathcal{R}^*(\mathcal{R}\chi_D u_k),$$

then with (2.23) and Theorem 2.8, it implies

$$c_d\chi_D u_k(\boldsymbol{x}) = \lambda_k(-\Delta)^{(d+3)/2}u_k(\boldsymbol{x}), \quad \forall\,\boldsymbol{x}\in D.$$

For even d, apply $(-\Delta)^{(d-2)/2}\mathcal{R}^*$ on both sides of (2.25) and follow the same procedure, we obtain

$$c_d\psi_k(\boldsymbol{x}) = \lambda_k(-\Delta)^{(d+3)/2}\psi_k(\boldsymbol{x}), \quad \forall\,\boldsymbol{x}\in D,$$

where $\psi_k(\boldsymbol{x}) = (-\Delta)^{-1/2}\chi_D u_k(\boldsymbol{x})$. The eigenfunctions $\phi_k(\boldsymbol{w}, b)$ can be retrieved from the relation (2.25) and boundary conditions $\phi_k(\boldsymbol{w}, 1) = \partial_b\phi_k(\boldsymbol{w}, 1) = \partial_b^2\phi_k(\boldsymbol{w}, 1) = 0$. □

The above Lemma 2.10 shows that $c_d\lambda_k^{-1}$ is the eigenvalue of $(-\Delta)^{(d+3)/2}$. From the Weyl's law for $-\Delta$, we have $\lambda_k = \Theta(k^{-(d+3)/d})$ (using Landau notation) as $k\to\infty$. Suppose the target

function $f(\boldsymbol{x})$, $\boldsymbol{x} \in D \subseteq \mathbb{R}^d$, can be represented by a superposition of ReLU functions with weight $h(\boldsymbol{w}, b)$, $(\boldsymbol{w}, b) \in V$:

$$f(\boldsymbol{x}) = \int_V \sigma(\boldsymbol{w} \cdot \boldsymbol{x} - b) h(\boldsymbol{w}, b) \, \mathrm{d}\boldsymbol{w} \, \mathrm{d}b \quad \Longrightarrow \quad \Delta f(\boldsymbol{x}) = \mathcal{R}^* h(\boldsymbol{w}, b).$$

In the one-dimensional case, the relation is further simplified to $\frac{\mathrm{d}^2}{\mathrm{d}x^2} f(x) = h(x)$. Assume $h(\boldsymbol{w}, b) = \sum_{k=1}^{\infty} \alpha_k \phi_k(\boldsymbol{w}, x)$ and use u_k defined in Lemma 2.10 and (2.23), we have

$$f(\boldsymbol{x}) = \sum_{k=1}^{\infty} \alpha_k \Delta^{-1} \mathcal{R}^* \phi_k(\boldsymbol{x}) = \sum_{k=1}^{\infty} \alpha_k u_k(\boldsymbol{x}). \tag{2.26}$$

Now we characterize the asymptotic behavior of u_k when D is a unit ball in $\mathbb{R}^d$. Due to the symmetry, u_k can be separated into $u_k(\boldsymbol{x}) = V_k(|\boldsymbol{x}|) Y_k(\widehat{\boldsymbol{x}})$, where $\widehat{\boldsymbol{x}} = \boldsymbol{x}/|\boldsymbol{x}|$ and Y_k denotes a spherical harmonics of order l on $\mathbb{S}^{d-1}$ and $V_k(r)$ satisfies the following by Lemma 2.10

$$\left(-\frac{\mathrm{d}^2}{\mathrm{d}r^2} - \frac{d-1}{r} \frac{\mathrm{d}}{\mathrm{d}r} + \frac{1}{r^2} l(l+d-2) \right)^{(d+3)/2} V_k = \frac{c_d}{\lambda_k} V_k.$$

The equation can be reduced to a set of Bessel's differential equations,

$$\left(-\frac{\mathrm{d}^2}{\mathrm{d}r^2} - \frac{d-1}{r} \frac{\mathrm{d}}{\mathrm{d}r} + \frac{1}{r^2} l(l+d-2) - \mu_{k,p}^2 \right) V_k = 0,$$

where $\mu_{k,p} = \left[\frac{c_d}{\lambda_k} \right]^{1/(d+3)} e^{2p\pi i/(d+3)}$, $p = 1, 2, \cdots, d+3$. Take a change of variable $s = \mu_{k,p} r$, the above equation becomes the standard Bessel's equation

$$\left(s^2 \frac{\mathrm{d}^2}{\mathrm{d}s^2} + (d-1) s \frac{\mathrm{d}}{\mathrm{d}s} + (s^2 - l(l+d-2)) \right) V_k = 0,$$

which implies $V_k(r) = \sum_{p=1}^{d+3} \beta_{k,p} (\mu_{k,p} r)^{-\nu} J_{l+\nu}(\mu_{k,p} r)$, $\nu = \frac{d-2}{2}$ and $\beta_j \in \mathbb{C}$ are coefficients. For sufficiently large k that $|\mu_{k,p}| \gg (l+\nu)^2$ with $\mathrm{Im}(\mu_{k,p}) \neq 0$, the asymptotical behavior is $\sqrt{\frac{2}{\pi \mu_{k,p} r}} \cos(\mu_{k,p} r - \frac{l+\nu}{2}\pi - \frac{\pi}{4})$, whose magnitude grows exponentially like $\mathcal{O}(e^{\mathrm{Im}(\mu_{k,p}) r})$ with different rates. Since the L^2 norm of $u_k(\boldsymbol{x})$ is uniformly bounded, the coefficients $\beta_{k,p} \to 0$ if $\mathrm{Im}(\mu_{k,p}) \neq 0$. That means, asymptotically, the eigenfunction u_k behaves like a usual Bessel function multiplied with spherical harmonics. For a general compact domain $D \subseteq \mathbb{R}^d$, all the arguments above are still valid except it is more difficult to explicitly write out the corresponding u_k.

REMARK 2.11. For the Gram matrix corresponding to activation function $\mathrm{ReLU}^m(x) = \frac{1}{m!}[\max(x, 0)]^m$ in d-dimension with $m \geq 1$, one can modify the above theorems to formally show:

- If m is odd, $(c_d \lambda_k^{-1}, \mathcal{R}^* \phi_k)$ forms an eigenpair of the operator $(-\Delta)^{\frac{d-1}{2}+(m+1)}$.

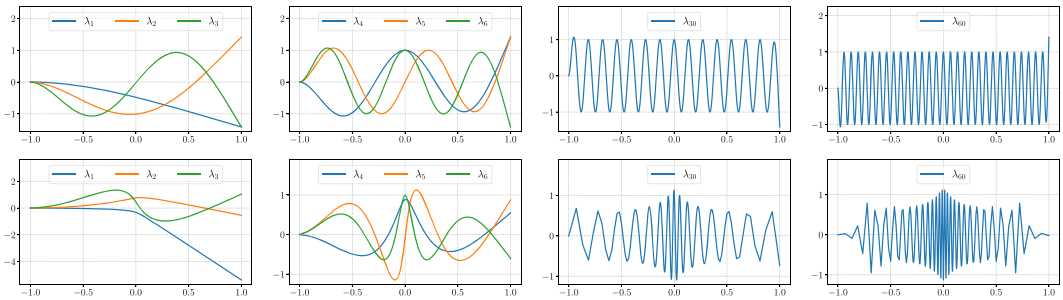


FIG. 3. Eigenmodes of λ_k for $k = \{1,2,3\},\{4,5,6\},30,60$ with $n = 1000$. The first and second rows correspond to uniform and adaptive $\boldsymbol{b}$, respectively.

- If m is even, $(c_d\lambda_k^{-1}, \mathcal{R}^*\partial_b\phi_k)$ forms an eigenpair of $(-\Delta)^{\frac{d-1}{2}+(m+1)}$.

  Using the inversion formula for the Radon transform, we have $\lambda_k = \Theta(k^{-\frac{d+2m+1}{d}})$.

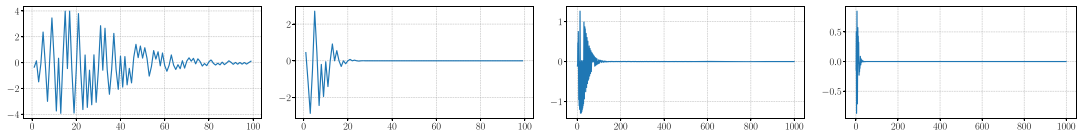
REMARK 2.12. Although the decay of eigenvalues seems slower in higher dimensions, the number of Fourier modes less than frequency ν is $\Theta(\nu^d)$. In other words, given a threshold ε for the leading singular value, no matter how wide a two-layer `ReLU` NN is, it can only resolve all Fourier modes up to frequency $\mathcal{O}(\varepsilon^{-\frac{1}{d+3}})$.

2.3 *Numerical experiments*

In this section, various numerical experiments are presented to verify our earlier analysis results: (1) the spectral property of the Gram matrix corresponding to `ReLU` functions and (2) the low-pass filter nature of two-layer networks in the least square setting.

2.3.1 *Spectrum and eigenmodes of Gram matrix in one dimension.* The first two numerical experiments show the spectrum of the Gram matrix for `ReLU` functions in the one-dimensional case. Figure 1 and the first row of Fig. 3 is the plot for eigenvalues in descending order and selected eigenmodes for uniform biases respectively, i.e. $b_j = b_1 + 2(j-1)/(n-1)$ for $j \in [n]$ with $b_1 = -1$. They agree with our analysis perfectly. Figure 2 and the second row of Fig. 3 are for non-uniform biases adaptive to the rate of change, i.e. $|f'(x)|$, of $f(x) = \arctan(25x)$ as an example. Define $F(x) = \int_{-1}^{x} |f'(t)|\,\mathrm{d}t \big/ \int_{-1}^{1} |f'(t)|\,\mathrm{d}t$, which is strictly increasing. There exists unique $b_i \in [-1,1]$ such that $F(b_i) = (i-1)/(n-1)$ for $i \in [n]$. We see that the conditioning becomes a little worse. However, the leading eigenmodes are more adaptive to the rapid change of $f(x)$ at 0. This is demonstrated further when using a least square approximation based on `ReLU` functions with uniform and adaptive biases. In Fig. 4, we show the projection of f on the leading eigenmodes corresponding to the Gram matrix. We observe that fewer leading eigenmodes are needed to represent/approximate the target function for adaptive biases compared to uniform biases.

2.3.2 *Approximation in one dimension: FEM basis vs. `ReLU` basis.* Next, we use numerical experiments to show least square approximation using FEM (finite element methods) basis vs. `ReLU` basis, two-layer NNs, in different setups when machine precision, grid resolution and the conditioning of the Gram matrix all play a role in practice. As discussed before, these two bases are equivalent mathematically.



(a) Uniform $\boldsymbol{b}$, $n = 100$. (b) Adaptive $\boldsymbol{b}$, $n = 100$. (c) Uniform $\boldsymbol{b}$, $n = 1000$. (d) Adaptive $\boldsymbol{b}$, $n = 1000$.

FIG. 4. Projection of f on the eigenmodes of the Gram matrix: coefficients vs. eigenmode index.

TABLE 1 *Error comparison for approximating* $f(x) = \arctan(25x)$ *with sufficient samples*

		float32				**float64**			
		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
		MAX	**MSE**	**MAX**	**MSE**	**MAX**	**MSE**	**MAX**	**MSE**
NN	Uniform $\boldsymbol{b}$	6.09×10^{-2}	9.58×10^{-5}	7.19×10^{-2}	1.43×10^{-4}	1.37×10^{-2}	1.70×10^{-6}	1.05×10^{-4}	1.33×10^{-10}
FEM	Uniform $\boldsymbol{b}$	1.37×10^{-2}	1.70×10^{-6}	1.05×10^{-4}	1.33×10^{-10}	1.37×10^{-2}	1.70×10^{-6}	1.05×10^{-4}	1.33×10^{-10}
NN	Adaptive $\boldsymbol{b}$	6.83×10^{-2}	7.54×10^{-5}	1.89×10^{-2}	1.06×10^{-5}	3.93×10^{-3}	1.42×10^{-6}	4.74×10^{-5}	1.17×10^{-10}
FEM	Adaptive $\boldsymbol{b}$	2.92×10^{-3}	9.95×10^{-7}	3.79×10^{-5}	1.02×10^{-10}	2.92×10^{-3}	9.95×10^{-7}	3.77×10^{-5}	1.02×10^{-10}

Numerically machine precision and grid resolution are common factors for both representations. The only difference is the Gram matrix. Table 1 summarizes the results. In particular, the final numerical mean square error (MSE) is directly related to the least square approximation. As we can see from the results, when single precision is used in the computation and $n = 100$ evenly distributed grid points (biases) are used, the grid resolution is the bottleneck for numerical accuracy near the rapid change. So the errors for FEM and NN are about the same. Numerical errors are reduced when the grid distribution is adaptive to the rapid change of the target function as described above. However, adaptive grids are much more effective for finite element basis. When a very fine grid $n = 1000$ is used, machine precision and the conditioning of the Gram matrix become the most important factors. Well-conditioned FEM (even for adaptive grid) can reach the machine precision while finite machine precision and ill-conditioned NN limit the total number of leading eigenmodes (low pass filter) and can not approximate a function with rapid change well. Moreover, increasing NN width further does not help. When double precision is used, even the NN has enough leading eigenmodes to approximate the target function well and the grid resolution becomes the limit for both FEM and NN. Hence the numerical errors for FEM and NN are similar.

Instead of using a (locally) rapid change function, we perform experiments on more and more oscillatory functions to demonstrate similar conclusions. Our target functions are $f(x) = \cos(6\pi x) - \sin(2\pi x)$ and its rescaled more oscillatory version $f(3x), f(9x)$. In these tests, instead of using the default threshold of the leading singular values of the Gram matrix based on machine precision, we introduce a manual singular value cut-off ratio[1] , η, to see the low pass filter effect for NN more clearly. The results are plotted in Fig. 5 and the approximation errors are summarized in Table 2. From both the plots and the MSE error, we can see that different cut-offs do not affect the FEM approximation since the condition number of the Gram matrix corresponding to the FEM basis on the uniform mesh is $\mathcal{O}(1)$. As a result, all eigenmodes (frequencies) that can be resolved by the grid size can be recovered accurately and stably. The slight

[1] Singular values are treated as zero if they are smaller than η times the largest singular value.

TABLE 2 *Approximation errors in Fig. 5*

	$f_1(x) = f(x)$		$f_2(x) = f(3x)$		$f_3(x) = f(9x)$	
	MAX	**MSE**	**MAX**	**MSE**	**MAX**	**MSE**
FEM (least square, $\eta = 10^{-r}$, $r = 3, 6, 9$)	4.85×10^{-5}	5.34×10^{-10}	4.38×10^{-4}	4.32×10^{-8}	3.95×10^{-3}	3.50×10^{-6}
NN (least square, $\eta = 10^{-3}$)	2.79	1.28	2.86	1.19	2.88	1.15
NN (least square, $\eta = 10^{-6}$)	2.66×10^{-1}	9.44×10^{-4}	1.74	5.39×10^{-1}	2.79	1.04
NN (least square, $\eta = 10^{-9}$)	1.52×10^{-2}	1.45×10^{-6}	6.38×10^{-2}	1.92×10^{-5}	1.13	4.65×10^{-1}

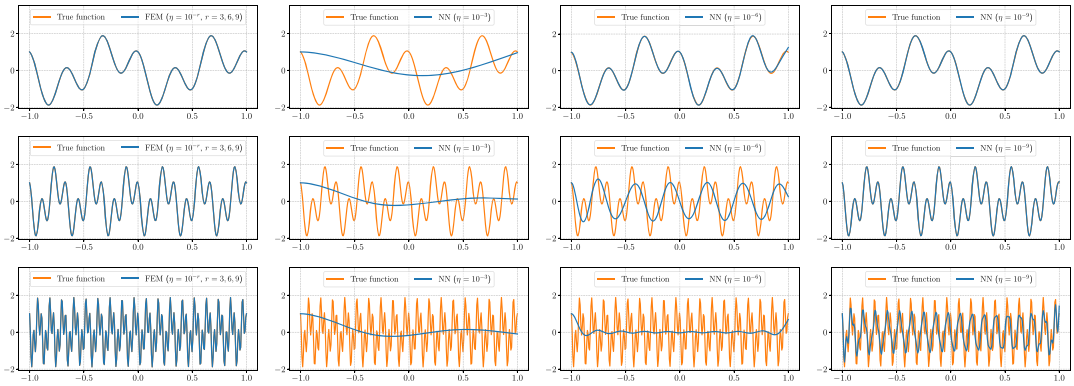


FIG. 5. Approximation comparison: FEM vs. NN with 2000 samples and $n = 2000$ equal spaced basis. The three rows correspond to f_1, f_2 and f_3, respectively. Here $f_1(x) = f(x), f_2(x) = f(3x), f_3(x) = f(9x)$, where $f(x) = \cos(6\pi x) - \sin(2\pi x)$.

decrease in accuracy as the oscillation increases is due to the fact that the least square approximation error is proportional to $h^2 \int |f''(x)|\,\mathrm{d}x$, where h is the grid size, for piecewise linear finite element basis by standard approximation theory. The dramatic effect of the cut-offs on NN is due to the fast spectral decay of the Gram matrix corresponding to the `ReLU` basis. When the cut-off ratio is $\eta = 10^{-3}$, the linear space spanned by leading eigenmodes above the threshold can not approximate even the relative smooth $f(x)$ well. When the cut-off ratio is reduced to $\eta = 10^{-9}$, there are enough leading modes above the threshold that can approximate $f(x), f(3x)$ well but not $f(9x)$.

In the following experiment, we show the low-pass filter nature of two-layer NNs vs. FEM basis with respect to noise and overfitting (or over-parametrization). The target function is $f(x) = \cos(3\pi x) - \sin(\pi x)$ with noise sampled from $\mathcal{U}(-0.5, 0.5)$ in our test. We manually select the cut-off ratio, η, for small singular values. The numerical results are shown in Fig. 6 and Table 3 for evenly distributed b_i. Since FEM has a condition number of $\mathcal{O}(1)$, all modes resolved by the grid are captured independent of the cut-off ratio η. On the other hand, NN only captures the leading eigenmodes, the number of which is determined by η. We can see that NN captures more modes as η becomes smaller (less regularized). It is also interesting to see the low pass filter effect when the Adam optimizer is used to minimize the least square, which is related to the learning dynamics analysis in Section 3. In the case of 1000 data points and 1500 degrees of freedom, Fig. 7 shows that a two-layer `ReLU` network is significantly more stable with respect to over-parametrization due to its low-pass filter nature.

2.3.3 *Spectrum and eigenmodes of Gram matrix in two dimensions.* We use a numerical example to show, Fig. 8, the spectrum of the discrete Gram matrix in two dimensions with 25600 evenly

TABLE 3 *Approximation errors in Fig. 6*

	least square				**Adam optimizer**	
	FEM ($\eta = 10^{-r}$, $r = 3, 6, 9$)	**NN** ($\eta = 10^{-3}$)	**NN** ($\eta = 10^{-6}$)	**NN** ($\eta = 10^{-9}$)	**NN (float32)**	**NN (float64)**
MAX	4.97×10^{-1}	1.81	1.15×10^{-1}	3.00×10^{-1}	1.73×10^{-1}	1.77×10^{-1}
MSE	5.90×10^{-2}	7.60×10^{-1}	2.23×10^{-3}	9.48×10^{-3}	3.68×10^{-3}	3.51×10^{-3}

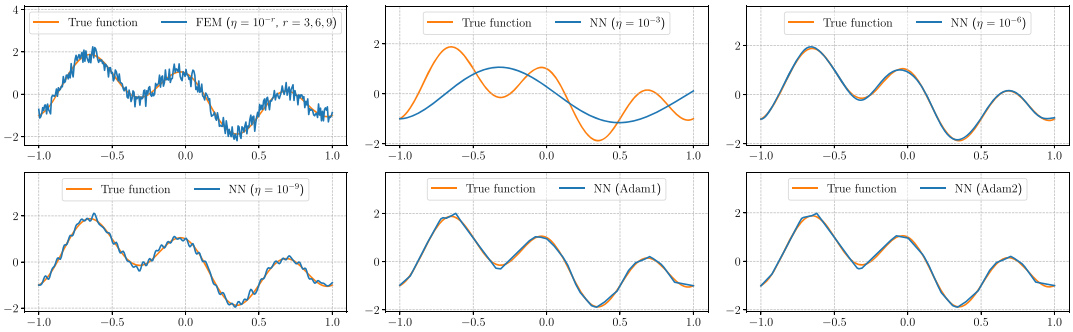


FIG. 6. Approximation stability: FEM vs. NN subject to uniform noise $\mathcal{U}(-0.5, 0.5)$ on 1000 samples and $n = 1000$ basis. The first four plots are results from the least square with different cut-off ratio η for singular values. The last two plots are the results trained by the Adam optimizer, where 'Adam1' and 'Adam2' use single and double precision, respectively.

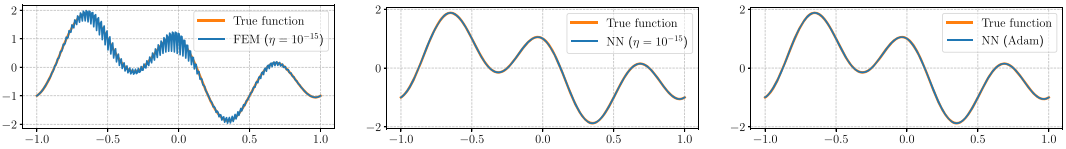


FIG. 7. Test of overfitting: FEM vs. NN with 1000 samples, $n = 1500$ basis and double precision. The first two plots are results from least square, where η is the cut-off ratio for small singular values. The last plot is the result trained by the Adam optimizer.

spaced samples for $(\boldsymbol{w}, b) \in \mathbb{S}^1 \times [-1, 1]$. The numerical experiments agree with our analysis in Section 2.2 very well. Several eigenmodes are presented in Fig. 9. In practice, those high-frequency modes whose corresponding eigenvalues are smaller than the machine precision threshold can not be captured.

2.4 *Different activation functions and scaled parameter initialization.*

Here we demonstrate the behaviours and their comparison for two-layer NNs using different activation functions, denoted generically by σ. The activation functions compared here are `ReLU`, `sine` and `tanh` and its first and second derivatives $\texttt{tanh}'$, $\texttt{tanh}''$. We limit our discussions to two-layer networks in one dimension. Figure 10(a) plots these activation functions.

A general two-layer network can be regarded as a parametrized function, denoted by $h(x; \boldsymbol{a}, \boldsymbol{w}, \boldsymbol{b})$, represented as the linear combination of a set of activation (or basis) functions parametrized by $\boldsymbol{w}$ and $\boldsymbol{b}$:

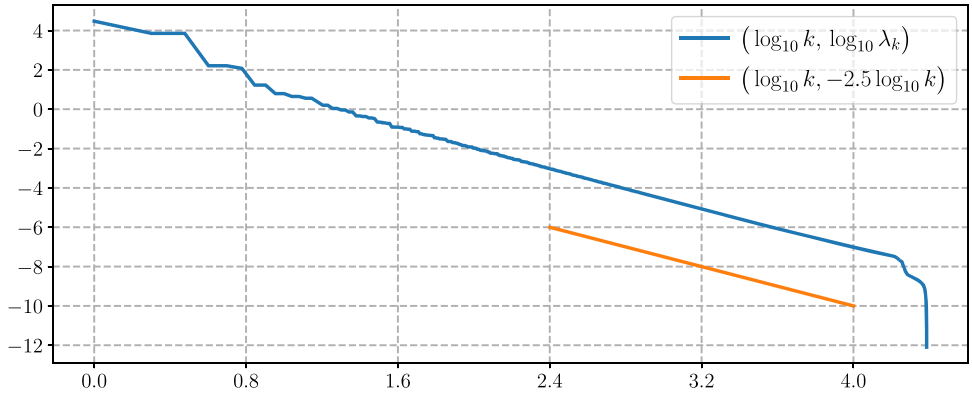
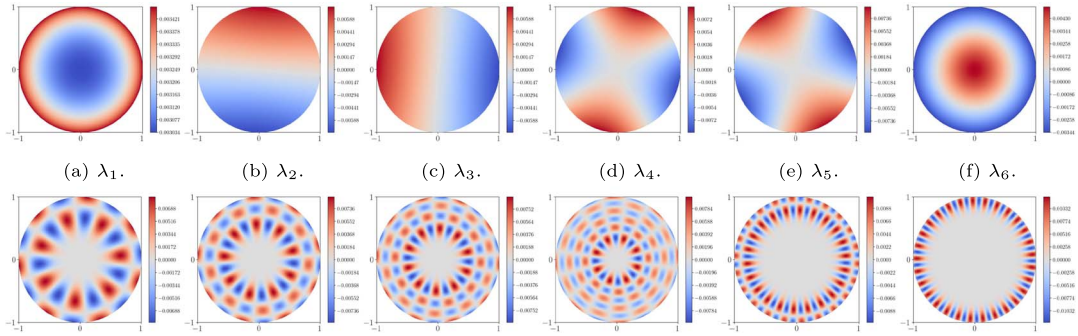


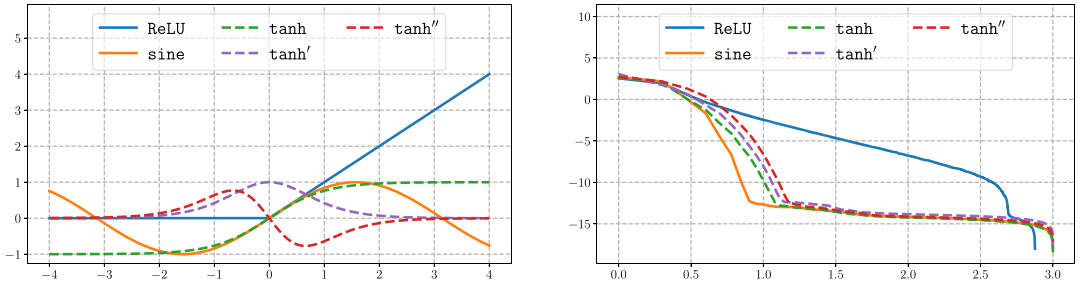
FIG. 8. Spectrum of the discrete Gram matrix in two dimensions with 25600 evenly spaced samples.



(a) λ_1. (b) λ_2. (c) λ_3. (d) λ_4. (e) λ_5. (f) λ_6.

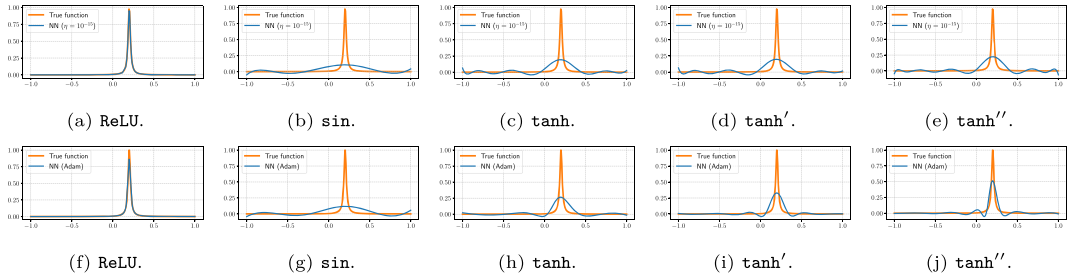
(g) λ_{50}. (h) λ_{100}. (i) λ_{150}. (j) λ_{200}. (k) λ_{250}. (l) λ_{300}.

FIG. 9. Eigenmodes of λ_k for $k \in \{1, 2, 3, 4, 5, 6, 50, 100, 150, 200, 250, 300\}$ with $n = 25600$.



(a) Activation functions. (b) Spectrum of $\boldsymbol{G}$ with $\boldsymbol{w}, \boldsymbol{b} \in \mathcal{U}(-1, 1)$.

FIG. 10. Illustrations of different activation functions and the corresponding sepctrum.



(a) ReLU. (b) sin. (c) tanh. (d) tanh′. (e) tanh′′.

(f) ReLU. (g) sin. (h) tanh. (i) tanh′. (j) tanh′′.

FIG. 11. First row: using least square approximation with $n = 512$, fixing $\boldsymbol{w}, \boldsymbol{b} \in \mathcal{U}(-1, 1)$. Second row: optimization using Adam with $n = 512$, initialization $\boldsymbol{w}, \boldsymbol{b} \in \mathcal{U}(-1, 1)$. All tests were conducted using double precision.

$$h(x; \boldsymbol{a}, \boldsymbol{w}, \boldsymbol{b}) = \sum_{i=1}^{n} a_i \sigma(w_i x + b_i). \tag{2.27}$$

As the default setting, we set $x \in [-1, 1]$ and initialize $w_i, b_i \sim \mathcal{U}(-1, 1)$ and $a_i \sim \mathcal{U}(-1/\sqrt{n}, 1/\sqrt{n})$ uniformly distributed.

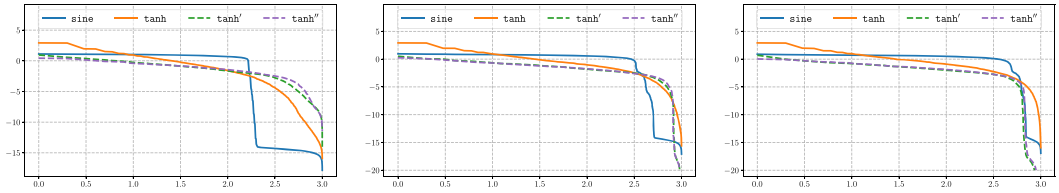
As discussed above, the conditioning of a two-layer network representation (2.27) is determined by the spectrum of the Gram Matrix $\boldsymbol{G}$, where

$$\boldsymbol{G}_{i,j} = \int_{-1}^{1} \sigma(w_i x + b_i)\sigma(w_j x + b_j)\, \mathrm{d}x, \quad w_i, b_i \sim \mathcal{U}(-1, 1).$$

The logarithmic plot of the spectra of the Gram matrices corresponding to different activation functions are shown in Fig. 10(b). Except ReLU, the Gram matrix of which has a polynomial decay as shown above, all other activation functions are analytic and hence their corresponding Gram matrices have exponential spectral decay [44, 45].

Figure 11 shows the approximation of $f(x) = \frac{1}{1+3600(x-0.2)^2}$ on $[-1, 1]$ using two-layer networks (2.27) with different activation functions. The networks have a width $n = 512$ and computations are implemented in double precision. The first row shows the least square approximation with fixed $\boldsymbol{w}, \boldsymbol{b} \in \mathcal{U}(-1, 1)$ and solving the linear system for $\boldsymbol{a}$ with Gram matrix $\boldsymbol{G}$. The second row presents the approximation results obtained by training with Adam, where the parameters (a_i, w_i, b_i) are optimized starting from uniform initialization: $\boldsymbol{a} \sim \mathcal{U}(-1/\sqrt{n}, 1/\sqrt{n})$ and $\boldsymbol{w}, \boldsymbol{b} \sim \mathcal{U}(-1, 1)$. Based on the spectral analysis, ReLU provides the best approximation due to its slowest spectral decay, or equivalently, its least bias against high frequencies among this group of activation functions. Among the remaining activation functions, tanh′′ yields better results owing to its relatively slower spectral decay.

Actually, for two-layer NNs using activation functions that are not homogeneous of degree one such as ReLU, instead of initializing w_i, b_i uniformly distributed in $(-1, 1)$, one can scale the range to be $(-s, s)$. By choosing a larger s, one can improve the representation capability of a two-layer NN (2.28). One way to see this is again through spectral analysis. The introduction of larger w_i and hence basis functions with more rapid changes (larger derivatives) leads to a slower spectral decay of the corresponding Gram matrix, $\boldsymbol{G}_{i,j} = \int_{-1}^{1} \sigma\big(w_i(x + b_i)\big)\sigma\big(w_j(x + b_j)\big)\, \mathrm{d}x$ with $\boldsymbol{w} \in \mathcal{U}(-s, s)$ and $\boldsymbol{b} \in \mathcal{U}(-1, 1)$, as shown in Fig. 12.



(a) $s = 256$. (b) $s = 512$. (c) $s = 768$.

FIG. 12. Spectrum of $\boldsymbol{G}$ with $\boldsymbol{w} \in \mathcal{U}(-s, s)$ and $\boldsymbol{b} \in \mathcal{U}(-1, 1)$, where $\boldsymbol{G}_{i,j}$ is given by $\boldsymbol{G}_{i,j} = \int_{-1}^{1} \sigma\big(w_i(x + b_i)\big)\sigma\big(w_j(x + b_j)\big)\,\mathrm{dx}$.

Another way to look at this is that scaling $\boldsymbol{w}, \boldsymbol{b}$ by s can be regarded as scaling up the target function by s, i.e. approximating $f(x)$ on $[-1, 1]$ by a two-layer network $h(x)$

$$h(x) = \sum_{i=1}^{n} a_i \sigma\big(w_i(x + b_i)\big) \quad \text{on } [-1, 1] \text{ with } \boldsymbol{w} \in \mathcal{U}(-s, s),\ \boldsymbol{b} \in \mathcal{U}(-1, 1) \tag{2.28}$$

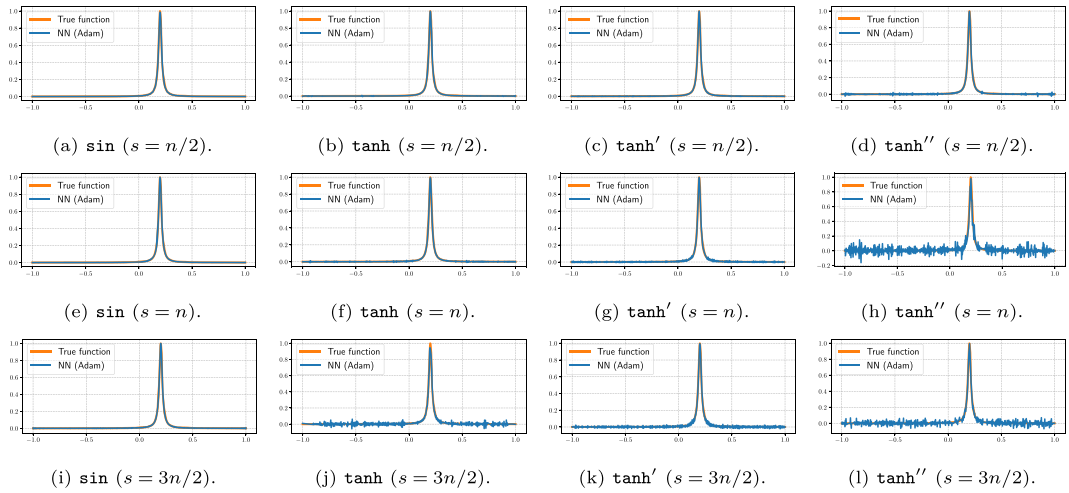
is equivalent to approximating $\widetilde{f}(x) = f(\frac{x}{s})$ on $[-s, s]$ by the two-layer network $\widetilde{h}(x) = h(\frac{x}{s})$

$$\widetilde{h}(x) = h\left(\frac{x}{s}\right) = \sum_{i=1}^{n} a_i \sigma(w_i x + b_i) \quad \text{on } [-s, s] \text{ with } \boldsymbol{w} \in \mathcal{U}(-1, 1),\ \boldsymbol{b} \in \mathcal{U}(-s, s).$$

However, in order to be able to resolve the rapid change of activation functions with derivatives proportional to s, or to have biases distributed dense enough in the interval $[-s, s]$, s can be at most proportional to the network width n. For example, if one uses `sine` as the activation function, then s can be at most $\mathcal{O}(n)$ (preferably $n/2$) so that the network can provide a set of maximally diverse random Fourier bases without missing intermediate frequencies. This is because using linear combinations of $\sin(wx + b_i), i = 1, 2$, for two different b_i can generate both $\sin(wx)$ and $\cos(wx)$. In the case of equally spaced w, using this set of parametrized activation functions is equivalent to the Fourier series with basis $\sin(mx), \cos(mx), m = -\frac{n}{2} + 1, \cdots, \frac{n}{2}$.

The experiment results shown in Fig. 13 demonstrate how scaling up $\boldsymbol{w}$ in two-layer networks as in (2.28) can enhance the representation capability and hence the approximation results. We note that different initializations may lead to different results, but the overall outcomes are largely similar. The subtle issue is how large s should be. It is interesting to observe that for activation functions `sine`, `tanh` and `tanh′`, the maximum magnitudes of their first derivatives are all bounded by 1, $s = \frac{n}{2}$ and $s = n$ work well. Once $s = \frac{3n}{2}$, the results degrade. For `tanh″`, the maximum magnitude of its first derivative is bounded by 2. It can be seen that $s = \frac{n}{2}$ works well but $s = n$ and $s = \frac{3n}{2}$ have degraded results. These tests suggest that the network size need to be able to resolve the rapid change of the activation function which is characterized by $s \cdot \sup_x |\sigma'(x)|$.

Although appropriate scaling up $\boldsymbol{w}, \boldsymbol{b}$ can improve approximation significantly over those corresponding results with normalized distribution in Fig. 11, however, our tests also suggest that the network size needs to be large enough, proportional to $s \cdot \sup_x |\sigma'(x)|$, to resolve the rapid change of the activation function. Moreover, it is difficult for a gradient descent based optimization to effectively learn adaptive $\boldsymbol{w}, \boldsymbol{b}$ to capture those fine features that are biased against in the initial representation. Hence, it implies that



(a) `sin` $(s = n/2)$. (b) `tanh` $(s = n/2)$. (c) `tanh`$'$ $(s = n/2)$. (d) `tanh`$''$ $(s = n/2)$.

(e) `sin` $(s = n)$. (f) `tanh` $(s = n)$. (g) `tanh`$'$ $(s = n)$. (h) `tanh`$''$ $(s = n)$.

(i) `sin` $(s = 3n/2)$. (j) `tanh` $(s = 3n/2)$. (k) `tanh`$'$ $(s = 3n/2)$. (l) `tanh`$''$ $(s = 3n/2)$.

FIG. 13. Illustrations of learned networks $\sum_{i=1}^{n} a_i \sigma(w_i(x + b_i))$ optimized using Adam with $n = 512$, initialization $\boldsymbol{a} \in \mathcal{U}(-1/\sqrt{n}, 1/\sqrt{n})$, $\boldsymbol{w} \in \mathcal{U}(-s, s)$ and $\boldsymbol{b} \in \mathcal{U}(-1, 1)$. All tests were conducted in double precision.

the representation capability of a two-layer network is no more than a linear representation which finds the optimal linear combination of a set of a priori given basis functions to approximate a target function, for example, the linear least square approximation. In this case, there are already well developed basis functions that provide well-conditioned representations with efficient numerical algorithms (based on solving linear systems), such as finite element bases, Fourier bases, splines, wavelets, which are usually better than using two-layer NNs when the number of basis functions is equivalent to the network width. On the other hand, the key issue is that the basis functions are given a priori, independent of the target function to be approximated. Hence, to have enough representation capability, the set of basis has to be large and diverse enough and hence suffer from the curse of dimensionality. We would like to point out that, the scaling of `ReLU` as the activation function, which is homogeneous degree of degree one, can be absorbed in $\boldsymbol{a}$. The distribution of $\boldsymbol{b}$ is equivalent to a grid for piecewise linear approximation in one dimension. Numerically, due to its ill-conditioning, the approximation is worse than using linear representation by finite element basis as discussed in earlier sections.

REMARK 2.13. For deep NNs, scaling up $\boldsymbol{w}, \boldsymbol{b}$ across all layers can cause instability for the training process.

2.5 *Observations and comments*

Based on the above analysis and numerical experiments, we give a few comments on using shallow NNs, many of which have been well observed in practice.

2.5.1 *Low-pass filter nature.* In practice (e.g. MATLAB), solving a linear system is typically approximated/regularized by its Moore-Penrose pseudo-inverse by cutting off the small eigenvalues at a threshold of $n\varepsilon\lambda_1$ to control the computer roundoff error, where n is the matrix size and ε is the machine precision or noise level in the data. For a two-layer `ReLU` network with evenly or uniformly distributed biases and width $n \geq k$, given the spectral decay of the Gram matrix, $\lambda_k = \Theta(k^{-\frac{d+3}{d}})$, only

leading modes for $k \leq m = \mathcal{O}(\varepsilon^{-\frac{d}{2d+3}})$ can be stably recovered in the least-square approximation. If the network is wide enough such that all leading modes up to m can be approximated well, the dominant numerical error is caused by the truncation of those modes higher than m. From (2.26), we see that truncation of the higher modes in the parameter space leads to a low pass filter in the approximation of the original function since u_k is equivalent to the eigenfunction of the Laplace operator asymptotically. In other words, at most all eigenmodes of the Laplace operator up to frequency $\mathcal{O}(\varepsilon^{-\frac{1}{2d+3}})$ can be captured in d dimensions no matter how wide the network is. This is because, 1) before the network width reaches a threshold $n_{\varepsilon,d} = \mathcal{O}(\varepsilon^{-\frac{d}{2d+3}})$, the network does not have a grid resolution to resolve the frequency modes of order $\mathcal{O}(\varepsilon^{-\frac{1}{2d+3}})$, 2) even as the network width passes $n_{\varepsilon,d}$, the network can only approximate leading modes k such that $\lambda_k \geq n\varepsilon\lambda_1 \Rightarrow k^{\frac{1}{d}} \leq \mathcal{O}(\varepsilon^{-\frac{1}{2d+3}})$ due to the ill-conditioning of the representation.

The machine precision for single and double precision are: $\varepsilon_1 = 2^{-23}$ and $\varepsilon_2 = 2^{-52}$ respectively. Hence, a two-layer NN can resolve about $2^{23d/(2d+3)}$ and $2^{52d/(2d+3)}$ eigenmodes respectively in d-dimensions. The number of modes in each direction that can be resolved is $2^{23/(2d+3)}$ and $2^{52/(2d+3)}$ respectively in d-dimensions, which is roughly 24 ($d = 1$), 10 ($d = 2$), 6 ($d = 3$) and 2 ($d = 10$) for single precision, and 1351 ($d = 1$), 172 ($d = 2$), 55 ($d = 3$) and 5 ($d = 10$) for double precision.

REMARK 2.14. Our analysis applies to the NTK regime where the network width goes to infinity while the biases are pretty much fixed.

2.5.2 *Approximation error.* Given a machine precision ε, the number of modes that can be captured by a two-layer `ReLU` network is at most $\mathcal{O}(\varepsilon^{-\frac{1}{2d+3}})$ (no matter how wide the network is). One can characterize the numerical error for a two-layer `ReLU` network in two regimes:

- **Small network** If the network width n is less than $\mathcal{O}(\varepsilon^{-\frac{d}{2d+3}})$, the corresponding grid resolution is less than $h = \mathcal{O}(n^{-\frac{1}{d}}) \leq \mathcal{O}(\varepsilon^{\frac{1}{2d+3}})$, which can not resolve the highest mode limited by $\varepsilon^{-\frac{1}{2d+3}}$. Hence the numerical error is dominated by the discretization error. Since the resulting approximation is continuous peicewise linear, the L^2 approximation error of a function f with bounded Sobolev norm $\|f\|_{H^2}$ is of order $h^2\|f\|_{H^2} \sim n^{-\frac{2}{d}}\|f\|_{H^2}$. In the small network regime, using a smooth activation function may be beneficial when approximating a smooth function due to the reduction of discretization error.
- **Large network** When the network is wide enough to resolve the highest mode of order $\mathcal{O}(\varepsilon^{-\frac{1}{2d+3}})$ accurately, then the numerical error is dominated by the truncation of higher modes. For a function f in Sobolev space H^p, the L^2 error due to truncation is $\mathcal{O}(\varepsilon^{\frac{p}{2d+3}}\|f\|_{H^p})$.

2.5.3 *Implications.* The ill-conditioning of two-layer NN representation and its bias against high frequencies explain why it is widely observed that shallow NNs can approximate smooth functions well, while for functions with fast transitions or rapid oscillations, one may achieve a numerical accuracy that is far from machine precision even using *wide* shallow networks for which universal approximation is proved in theory. On the other hand, the low-pass filter nature of the shallow NNs also alleviates instability with respect to noise or overfitting/over-parametrization.

The spectral decay of the Gram matrix for a set of basis depends on the smoothness of the basis. The smoother the activation function is, the faster the spectrum of the corresponding Gram matrix decays

(see Remark 2.11 and Section 2.4), and hence the fewer eigenmodes can be used for approximation in practice and the more bias against high frequencies.

However, with a fixed network width (or a fixed grid resolution), approximation order induced by the activation function also plays a role. For example, using `ReLU` results in a piecewise linear approximation, the error of which is proportional to the grid resolution squared (2nd order) if the target function is twice differentiable. If the Heaviside function (the derivative of `ReLU`) is used as the activation function, although the Gram matrix is better conditioned than using `ReLU`, the resulting piecewise constant approximation is only 1st order if the target function is differentiable, and hence may limit the numerical accuracy in practice.

In applications, if one can make the target function or map smooth under certain transformation, for instance, a linear transformation with a set of new bases, e.g. Fourier basis with high frequencies, using a NN in the transformed domain can achieve high accuracy.

2.5.4 *Gradient decent for least squares.* Before we investigate the full nonlinear learning dynamics for two-layer NNs in the next section, we demonstrate how ill-conditioning in the representation will affect the convergence of a gradient descent based optimization for a simple quadratic convex function corresponding to a linear least square approximation. Instead of finding the least square solution directly by solving the linear system (normal equation) with the Gram matrix $\boldsymbol{G}$, if one chooses to minimize the least square by using gradient descent, then the dynamics of the coefficients $\boldsymbol{a}(t) = [a_1(t), a_2(t), \cdots, a_n(t)]^T$ follow the system of ODEs

$$\frac{\mathrm{d}\boldsymbol{a}(t)}{\mathrm{d}t} = -\boldsymbol{G}\boldsymbol{a}(t) + \boldsymbol{f}, \quad \boldsymbol{G}_{i,j} = \int_D \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} - b_i)\sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} - b_j)\,\mathrm{d}\boldsymbol{x},\ \boldsymbol{f}_i = \int_D f(\boldsymbol{x})\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} - b_i)\,\mathrm{d}\boldsymbol{x}.$$

The system is stiff due to the fast spectral decay of $\boldsymbol{G}$. Let $(\lambda_k, \boldsymbol{g}_k)$ be the eigen pairs of $\boldsymbol{G}$ and define $\widehat{a}_k(t) = \boldsymbol{a}^T(t)\boldsymbol{g}_k, \widehat{f}_k = \boldsymbol{f}^T\boldsymbol{g}_k$. We have

$$\frac{\mathrm{d}\widehat{a}_k(t)}{\mathrm{d}t} = -\lambda_k\widehat{a}_k(t) + \widehat{f}_k \quad \Longrightarrow \quad \widehat{a}_k(t) = \left(\widehat{a}_k(0) - \frac{\widehat{f}_k}{\lambda_k}\right)e^{-\lambda_k t} + \frac{\widehat{f}_k}{\lambda_k}.$$

It takes at least $t > \mathcal{O}(\lambda_k^{-1})$ for the initial error in kth mode to reduce significantly. Since $\lambda_k \to 0$ as $k \to \infty$ and the corresponding eigenmode becomes more and more oscillatory, two-layer NNs bias against high frequencies in both representation and training. An appropriate stopping time can be used for the sake of computation cost or regularization for noise or machine roundoff error.

3. Learning dynamics

The key feature in machine learning using NN representation is the training process, which ideally can find the optimal parameters, i.e. an adaptive representation driven by the data. The relevant approximation theory has been studied extensively. The best-known approximation error estimates using `ReLU` as the activation function for two-layer NN [3, 42] are based on the fact that $\sigma''(t) = \delta(t)$, which implies Δf can be expressed as a standard Radon transform when viewing the parameters $\{a_i\}_{i=1}^n$ as an empirical measure of $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$. In practice, gradient-based methods (and the variants) are adopted to seek the optimal parameters.

Our previous analysis shows that a two-layer NN with fixed (or randomly sampled) biases will have a low-pass filter nature, which makes it challenging to capture high-frequency components, e.g. rapid changes or fast oscillations. Here we show that the ill-conditioning of the Gram matrix causes difficulties in the learning process as well. Intuitively, without high frequency information, there is no correct guidance in effectively and efficiently optimizing the distribution of the parameters, $(\boldsymbol{w}, b)$, from initial uniform distribution to be non-uniform adaptive to the target function. Often in practice, undesirable clustering of parameters, which reduces the effective network width, may occur during the training.

Here we study the following natural question: assuming that a gradient-based optimization can find the optimal solution, which itself is a challenging question in general, what is the training dynamics and its computation cost to attain such a solution from a random initial guess? In particular, we demonstrate that initial high-frequency component error can take a long time to correct. All these lead to numerical difficulties in achieving the optimal solution even if one assumes the learning process can find the optimal solution in theory. It implies that even if the full learning process is applied, the numerical error can still be far from the machine precision for functions with high-frequency components in practice.

Training a NN with the gradient flow can be regarded as the gradient descent method with a very small step size for finite time dynamics. We illustrate this using a one-dimensional example. Let $D = [-1, 1]$, the gradient flow of $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ follow

$$\frac{\mathrm{d}a_i}{\mathrm{d}t} = -\int_D (h(x,t) - f(x))\sigma(x - b_i)\,\mathrm{d}x \quad \text{and} \quad \frac{\mathrm{d}b_i}{\mathrm{d}t} = a_i \int_D (h(x,t) - f(x))\sigma'(x - b_i)\,\mathrm{d}x.$$

One popular way to analyze the dynamics of the neuron network is the mean-field representation in which the network is written as

$$\overline{h}(x,t) = \int_{\mathbb{R}^2} a\sigma(x - b)\mu_n(a,b,t)\,\mathrm{d}a\,\mathrm{d}b$$

with empirical measure $\mu_n(a,b,t) = \frac{1}{n}\sum_{i=1}^n \delta(a - a_i(t), b - b_i(t))$. The analysis of the limiting behavior of mean-field NNs can be found in [38, 46, 57] and the references therein. However, most of the mean-field studies assume the measure $\mu_n(\cdot,\cdot,t)$ converges as $t \to \infty$ and $n \to \infty$.

Our study works for fully discrete two-layer NNs with no convergence assumption for the measure μ_n or requiring $n \to \infty$. The main difficulty for our analysis is the possibility of biases b_i moving out of the bounded domain of interest. Define generalized Fourier modes, $\{\theta_m\}_{m\geq 1}$, which are the eigenfunctions of the Gram kernel $\mathcal{G}$ in (2.3), and $\widehat{g}$ as the generalized discrete Fourier transform of g,

$$\widehat{g}(m) = \int_D g(x)\theta_m(x)\,\mathrm{d}x.$$

In one dimension, the generalized Fourier transform is asymptotically close to the standard Fourier transform as the mode increases. The key to our study is the construction of an auxiliary function $w(x,t) \in C^2(D)$ (defined by (3.5)) that satisfies $\partial_x^2 w(x,t) = h(x,t) - f(x) = e(x,t)$, which is the approximating error at time t, with boundary conditions $w(1,t) = \partial_x w(1,t) = 0$. By studying the evolution of $\widehat{w}(m,t)$ we show at least how slow the learning dynamics can be in terms of the lower bound for time needed to reduce the initial error in mode m in half. We prove the following

statement for learning dynamics in one dimension under the following mild assumptions: 1) there exists a constant $M > 0$ that $\sup_{i=1}^{n} |a_i(t)|^2 \leq M$; 2) the initialization of biases $\{b_i(0)\}_{i=1}^{n}$ are equispaced on D.

THEOREM 3.1. If $|\widehat{w}(m,0)| \neq 0$, then it will take at least $\mathcal{O}(\frac{m^3|\widehat{w}(m,0)|}{n})$ time to reduce the generalized Fourier coefficient in half, i.e. $|\widehat{w}(m,t)| \leq \frac{1}{2}|\widehat{w}(m,0)|$.

In numerical computation, to follow the gradient flow closely, the discrete time-step is $\mathcal{O}(\frac{1}{n})$. From the relation $|\widehat{w}(m,t)| \simeq m^{-2}|\widehat{e}(m,t)|$, Theorem 3.1 says that if the initial error in mode m, $|\widehat{e}(m,0)| \neq 0$, (which means $|\widehat{w}(m,0)| > cm^{-2}$ for some $c > 0$,) it takes at least $\mathcal{O}(m)$ steps to reduce the error in mode m by half. Note that the result does not depend on the convergence of the optimization algorithm and the above estimate is a lower bound. In practice, the gradient-based training process could have an even slower learning rate. In the next theorem, the lower bound is improved in a more specific scaling regime (in terms of network width vs. frequency mode) and using better estimates (see Appendix C). It implies that it takes at least $\mathcal{O}(m^2)$ steps to reduce the error in mode m by half in numerical computation. In the special case of fixed biases (i.e. least square problem), the number of steps needed is at least $\mathcal{O}(m^4)$ if $n = \Omega(m^3)$ (see Remark C.6).

THEOREM 3.2. If the total variation of the sequence $\{a_i^2(t)\}_{i=1}^{n}$ is bounded by M'. Let $n \geq m^4$ be sufficiently large, then it will take at least $\mathcal{O}(\frac{m^4|\widehat{w}(m,0)|}{n})$ time to reduce the generalized Fourier coefficient by half, i.e. $|\widehat{w}(m,t)| \leq \frac{1}{2}|\widehat{w}(m,0)|$.

The above results show that reducing the initial error in high-frequency modes by gradient-based learning dynamics is slow. Moreover, due to the low-pass filter nature shown earlier, it is difficult to capture high-frequency modes for two-layer NNs with evenly or uniformly distributed initial biases. These two effects show why a two-layer NN struggles with high-frequency modes in approximation even if a full training process is employed.

Before we prove the above two theorems, we need to introduce the mathematical setup and prove a few lemmas and intermediate results.

3.1 *Mathematical setup of learning dynamics*

Let $D = (-1, 1)$, we consider approximating the objective function $f(x) \in C(D)$, with the shallow NN

$$h(x) = \sum_{i=1}^{n} a_i \sigma(x - b_i).$$

Generally speaking, the biases $\{b_i\}_{i=1}^{n}$ are supported on $\mathbb{R}$ during the training process. For analysis purposes, we restrict $\{b_i\}_{i=1}^{n} \subseteq \overline{D^\varepsilon}$, where $D^\varepsilon = \operatorname{supp} \chi = (-1-\varepsilon, 1)$, and modify the activation function in the two-layer NN representation as follows

$$h(x) = \sum_{i=1}^{n} a_i \psi(x, b_i), \quad \psi(x, b_i) = \chi(b_i)\sigma(x - b_i), \quad \partial_x^2 \psi(x,b) = \partial_b^2 \psi(x,b) = \delta(x-b),\ x, b \in D.$$

Here χ is an approximation to the characteristic function of D by adding a quadratic transition region:

$$\chi(x) = \begin{cases} 1, & x \in D, \\ 1 - \left(\frac{1}{\varepsilon}(x+1)\right)^2, & x \in (-1-\varepsilon, -1], \\ 0, & \text{otherwise}, \end{cases}$$

and ε is a positive parameter. The loss function is

$$\mathcal{L}(h,f) := \frac{1}{2}\int_D |h(x) - f(x)|^2 \,\mathrm{d}x.$$

Minimizing $\mathcal{L}(h,f)$ over the parameters $\{a_i\}_{i\in[n]}$ and $\{b_i\}_{i\in[n]}$ by gradient descent follows the gradient flow

$$\frac{\mathrm{d}}{\mathrm{d}t}a_i(t) = -\int_D (h(x,t) - f(x))\psi(x, b_i(t))\,\mathrm{d}x \tag{3.1}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}b_i(t) = -a_i(t)\int_D (h(x,t) - f(x))\partial_b\psi(x, b_i(t))\,\mathrm{d}x \tag{3.2}$$

with certain initial conditions $\{a_i(0)\}_{i\in[n]}$ and $\{b_i(0)\}_{i\in[n]}$. Due to the modification of the network representation, one can show that no bias can move outside $\overline{D^\varepsilon}$ at a later time.

THEOREM 3.3. If the initial biases and weights satisfy $\{b_i(0)\}_{i\in[n]} \subseteq D^\varepsilon$ and $\sup_{i\in[n]} |a_i(0)| < \infty$, we have $\{b_i(t)\}_{i\in[n]} \subseteq \overline{D^\varepsilon}$, $\forall t \geq 0$.

Proof. At any time t, using Cauchy-Schwartz inequality,

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}a_i(t)\right| &\leq \sqrt{\int_D |h(x,t) - f(x)|^2\,\mathrm{d}x}\sqrt{\int_D \psi(x, b_i(t))^2\,\mathrm{d}x} \\ &\leq \sqrt{\int_D |h(x,0) - f(x)|^2\,\mathrm{d}x}\sqrt{\int_D \psi(x, b_i(t))^2\,\mathrm{d}x} \quad < \infty.\end{aligned}$$

Therefore if initial $a_i(0)$ is finite, $a_i(t)$ is always finite, which in turn implies every derivative $\frac{\mathrm{d}}{\mathrm{d}t}b_i(t)$ is also finite. Suppose at time $t > 0$ that $b_k(t) \in \overline{D^\varepsilon}^{\complement}$, then there exists an open path that $\forall t' \in (t_0, t_1) \subseteq (0, t)$ that $\frac{\mathrm{d}}{\mathrm{d}t}b_k(t') \neq 0$ and $b_k(t') \in \overline{D^\varepsilon}^{\complement}$. However, $\partial_b\psi(x,b) = 0$ for $b \in \overline{D^\varepsilon}^{\complement}$, which is a contradiction. Therefore every $\{b_i(t)\}_{i\in[n]} \subseteq \overline{D^\varepsilon}$. □

Define the Gram kernel $\mathcal{G}(b,b')$ on $D^\varepsilon \times D^\varepsilon$ by

$$\mathcal{G}(b,b') = \int_D \psi(x,b)\psi(x,b')\,\mathrm{d}x, \tag{3.3}$$

which is a compact operator in $L^2(D^\varepsilon)$. Let $\lambda_k \geq 0,\ k = 1, 2, \cdots$, be the eigenvalues in descending order and ϕ_k be the corresponding eigenfunctions, which form an orthonormal basis in $L^2(D^\varepsilon)$. We have $\mathcal{G}(b, b') = \sum_{k\geq 1} \lambda_k \phi_k(b)\phi_k(b')$. Some properties of ϕ_k are studied in Appendix A. Expand the coefficient $a_i(t)$ along the eigenfunction ϕ_k (an orthonormal basis in $L^2(D^\varepsilon)$ in discrete version (at the nodal points $b_i(t)$), its dynamics follows

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{n} a_i(t)\phi_k(b_i(t)) &= \sum_{i=1}^{n} \phi_k(b_i(t))\frac{\mathrm{d}a_i}{\mathrm{d}t} + \sum_{i=1}^{n} a_i(t)\phi_k'(b_i(t))\frac{\mathrm{d}b_i}{\mathrm{d}t} \\
&= -\sum_{i=1}^{n} \phi_k(b_i(t)) \int_D (h(x,t) - f(x))\, \psi(x, b_i(t))\, \mathrm{d}x \\
&\quad - \sum_{i=1}^{n} \phi_k'(b_i(t))|a_i(t)|^2 \int_D (h(x,t) - f(x))\partial_b \psi(x, b_i(t))\, \mathrm{d}x \\
&= -\sum_{i=1}^{n} \phi_k(b_i(t)) \left(\sum_{j=1}^{n} \boldsymbol{G}_{i,j}(t)a_j(t) - P(b_i(t)) \right) \\
&\quad - \sum_{i=1}^{n} |a_i(t)|^2 \phi_k'(b_i(t)) \left(\sum_{j=1}^{n} K_{i,j}(t)a_j(t) - Q(b_i(t)) \right).
\end{aligned}$$

where $\boldsymbol{G} = (\boldsymbol{G}_{i,j})_{i,j\in[n]}$ and $\boldsymbol{K} = (\boldsymbol{K}_{i,j})_{i,j\in[n]}$ are

$$\boldsymbol{G}_{i,j}(t) = \mathcal{G}(b_i(t), b_j(t)), \quad \boldsymbol{K}_{i,j}(t) = \partial_b \mathcal{G}(b_i(t), b_j(t)).$$

The functions P and Q are

$$P(b) = \int_D f(x)\psi(x, b)\, \mathrm{d}x, \quad Q(b) = \int_D f(x)\partial_b \psi(x, b)\, \mathrm{d}x.$$

We can represent $P(b)$ and $Q(b)$ as sum of eigenfunctions on D^ε as well:

$$P(b) = \sum_{k\geq 1} p_k \phi_k(b), \quad Q(b) = \sum_{k\geq 1} p_k \phi_k'(b).$$

Therefore we can derive that

$$\begin{aligned}
&-\sum_{i=1}^{n} \phi_k(b_i(t)) \left(\sum_{j=1}^{n} \boldsymbol{G}_{i,j}(t)a_j(t) - P(b_i(t)) \right) \\
&= -\sum_{l=1}^{\infty}\sum_{i=1}^{n} \phi_k(b_i(t))\phi_l(b_i(t)) \left[\lambda_l \sum_{j=1}^{n} \phi_l(b_j(t))a_j(t) - p_l \right]
\end{aligned}$$

and

$$-\sum_{i=1}^{n}|a_i(t)|^2\phi_k'(b_i(t))\left(\sum_{j=1}^{n}K_{i,j}(t)a_j(t)-Q(b_i(t))\right)$$
$$=-\sum_{l=1}^{\infty}\sum_{i=1}^{n}|a_i(t)|^2\phi_k'(b_i(t))\phi_l'(b_i(t))\left[\lambda_l\sum_{j=1}^{n}\phi_l(b_j(t))a_j(t)-p_l\right].$$

Denote $\Theta_k(t)=\sum_{j=1}^{n}\phi_k(b_j(t))a_j(t)-\frac{p_k}{\lambda_k}$, we find that

$$\frac{d\Theta_k(t)}{dt}=-\sum_{l=1}^{\infty}\lambda_l\left[M_{l,k}(t)+S_{l,k}(t)\right]\Theta_l(t), \tag{3.4}$$

where the infinite matrices

$$M_{l,k}(t)=\sum_{i=1}^{n}\phi_l(b_i(t))\phi_k(b_i(t)),\quad S_{l,k}(t)=\sum_{i=1}^{n}|a_i(t)|^2\phi_k'(b_i(t))\phi_l'(b_i(t))$$

are both semi-positive.

Consider the energy $E(t)=\frac{1}{2}\sum_{k\geq 1}\lambda_k|\Theta_k^2(t)|$, then

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t}=-\sum_{k,l=1}^{\infty}\lambda_k\lambda_l\Theta_k(t)\left[M_{l,k}(t)+S_{l,k}(t)\right]\Theta_l(t).$$

Since the matrices are both non-negative, the energy is non-increasing, thus there is a limit for $E(t)$ as $t\to\infty$. Define the auxiliary function

$$w(b,t):=\sum_{k=1}^{\infty}\lambda_k\Theta_k(t)\phi_k(b), \tag{3.5}$$

which is directly related to the error from the following relation for $b\in D$,

$$\begin{aligned}\partial_b^2 w(b,t)&=\partial_b^2\left[\sum_{k=1}^{\infty}\left(\lambda_k\sum_{j=1}^{n}\phi_k(b_j(t))a_j(t)-p_k\right)\phi_k(b)\right]\\&=\partial_b^2\left[\sum_{j=1}^{n}\mathcal{G}(b,b_j(t))a_j(t)-P(b)\right]\\&=\partial_b^2\left[\sum_{j=1}^{n}a_j(t)\int_D\psi(x,b)\psi(x,b_j(t))\,\mathrm{d}x-\int_D f(x)\psi(b)\,\mathrm{d}x\right]\\&=\sum_{j=1}^{n}a_j(t)\int_D\delta(x-b)\psi(x,b_j(t))\,\mathrm{d}x-\int_D f(x)\delta(x-b)\,\mathrm{d}x=h(b,t)-f(b).\end{aligned} \tag{3.6}$$

We first provide its regularity property which can be related to the spectral decay of the Gram matrix determined by the regularity of the activation function.

THEOREM 3.4. The auxiliary function $w \in H^1(D^\varepsilon)$ and $w \in H^2(D)$ uniformly, respectively.

Proof. From (3.6), we have $\|\partial_b^2 w\|_{L^2(D)} \leq \|h(\cdot,0) - f(\cdot)\|_{L^2(D)}$ uniformly. We also notice that each eigenfunction $\phi_k, k \in \mathbb{N}$ satisfies that $\phi_k(b,t)|_{b=1} = \partial_b \phi_k(b,t)|_{b=1} = 0$, hence $w(b,t)|_{b=1} = \partial_b w(1,t) = 0$. Using Cauchy-Schwartz inequality, we derive the following Poincaré inequalities,

$$\|w(b,t)\|^2_{L^2(D)} = \left\| \int_1^b \partial_b w(b',t)\, \mathrm{d}b' \right\|^2_{L^2(D)} \leq 2 \left\| \partial_b w(b,t) \right\|^2_{L^2(D)}$$

and

$$\left\| \partial_b w(b,t) \right\|^2_{L^2(D)} = \left\| \int_1^b \partial_b^2 w(b',t)\, \mathrm{d}b' \right\|^2_{L^2(D)} \leq 2 \left\| \partial_b^2 w(b,t) \right\|^2_{L^2(D)}.$$

Therefore $w \in H^2(D)$. Now we prove the other part of the theorem: $w \in H^1(D^\varepsilon)$ uniformly. Here we use the fact that $E(t) = \frac{1}{2} \sum_{k\in\mathbb{N}} \lambda_k |\Theta_k(t)|^2$ is uniformly bounded by $E(0)$, then

$$\|w(\cdot,t)\|^2_{L^2(D)} \leq \sum_{k\in\mathbb{N}} \lambda_k^2 |\Theta_k(t)|^2 < \lambda_1 \sum_{k\in\mathbb{N}} \lambda_k |\Theta_k(t)|^2 \leq \lambda_1 E(0).$$

Follow the same derivation of (3.6),

$$\begin{aligned}
\partial_b w(b,t) &= \partial_b \left[\sum_{k=1}^{\infty} \left(\lambda_k \sum_{j=1}^{n} \phi_k(b_j(t)) a_j(t) - p_k \right) \phi_k(b) \right] \\
&= \partial_b \left(\sum_{j=1}^{n} \mathcal{G}(b, b_j(t)) a_j(t) - P(b) \right) \\
&= \partial_b \left(\sum_{j=1}^{n} a_j(t) \int_D \psi(x,b)\psi(x,b_j(t))\, \mathrm{d}x - \int_D f(x)\psi(x,b)\, \mathrm{d}x \right) \\
&= \sum_{j=1}^{n} a_j(t) \int_D \partial_b \psi(x,b)\psi(x,b_j(t))\, \mathrm{d}x - \int_D f(x)\partial_b \psi(x,b)\, \mathrm{d}x \\
&= \int_D (h(x,t) - f(x))\, \partial_b \psi(x,b)\, \mathrm{d}x.
\end{aligned}$$

Thus using $\|h(\cdot,t) - f\|_{L^2(D)} \leq \|h(\cdot,0) - f\|_{L^2(D)}$ and $\partial_b \psi(x,b) \in C^0(D \times D^\varepsilon)$, the following inequality holds uniformly:

$$\|\partial_b w(b,t)\|^2_{L^2(D^\varepsilon)} \leq \|h(\cdot,0) - f\|^2_{L^2(D)} \int_{D\times D^\varepsilon} |\partial_b \psi(x,b)|^2\, \mathrm{d}x\, \mathrm{d}b < \infty.$$

□

Let $\{E(t_m)\}_{m\geq 1}$ be a minimizing sequence for $E(t)$, then

$$\sum_{k,l=1}^{\infty} \lambda_k \lambda_l \Theta_k(t_m)[M_{l,k}(t_m) + S_{l,k}(t_m)]\Theta_l(t_m)$$
$$= \sum_{i=1}^{n} \left(|w(b_i(t_m), t_m)|^2 + a_i^2(t_m)|\partial_b w(b_i(t_m), t_m)|^2 \right) \to 0,$$

which implies that $w(b_i(t_m), t_m) \to 0$ and $|a_i(t_m)\partial_b w(b_i(t_m), t_m)| \to 0$. Note that $H^1(D^\varepsilon)$ embeds into $L^2(D^\varepsilon)$ compactly, there exists a subsequence $\{w(\cdot, t_{m_s})\}_{s\geq 1}$ converges to $\bar{w} \in H^1(D^\varepsilon)$ strongly (which is in $C^{0,\alpha}(D^\varepsilon)$). Here are two cases:

- If the gradient dynamics $b_i(t)$ converges to b_i^* as $t \to \infty$, then we must have $\bar{w}(b_i^*) = 0$. If $\lim_{t\to\infty} a_i(t) \neq 0$ or does not exists, then $\partial_b \bar{w}(b_i^*) = 0$.
- If during the dynamics $b_i(t)$ is not converging, then there exists an open set $T \subseteq D^\varepsilon$ that $T \subseteq \limsup_{\tau\to\infty}\{b_i(t)\}_{t\geq\tau}$ or equivalently, T appears infinitely often in the dynamics. Then we must have $\bar{w}(b) = 0$, $\forall b \in T$.

Remark 3.5. In mean-field representation that $n \to \infty$ and assume the limiting measure $\int_{\mathbb{R}} \mu(da, b, t)$ has full support on D^ε, we immediately conclude that $\bar{w} \equiv 0$. Hence $\Theta_k(t) \to 0$ and $E(t) \to 0$ as $t \to \infty$. This implies the network will converge to the objective function. However, in a discrete setting, it becomes more complicated due to the existence of local minimums.

3.2 *Generalized Fourier analysis of learning dynamics*

In this section, we study the convergence of the learning dynamics in terms of frequency modes, especially the asymptotic behavior for high-frequency components. The key relation is (3.6), which implies one can study the evolution of the Fourier modes of $\partial_b^2 w(b, t)$ to understand the evolution of error $h(b, t) - f(b)$. The other key fact is $\partial_b^4 \phi(b) = \lambda_k^{-1}\phi(b)$ and the spectral decay rate $\lambda_k = \Theta(k^{-4})$. However, due to the bounded domain of interest $b \in D$ and non-periodicity at the boundary, generalized Fourier modes have to be designed for our study. We introduce an orthonormal basis $\theta_k(x) \in C(D)$ that solves the eigenvalue problem

$$\begin{aligned} \theta_k^{(4)}(x) &= p_k \theta_k(x), \\ \theta_k(1) = \theta_k'(1) &= 0, \\ \theta_k''(-1) = \theta_k'''(-1) &= 0. \end{aligned} \tag{3.7}$$

In Appendix B we provide an explicit characterization of θ_k. In a nutshell, $p_k = \Theta(k^4)$ and θ_k form a complete orthonormal basis for $L^2(D)$ and it is close to a shifted Fourier mode when k is relatively large.

Let $\widehat{w}$ denote the *generalized* discrete Fourier transform of w on D:

$$\widehat{w}(k, t) = \int_D \theta_k(b) w(b, t)\, \mathrm{d}b. \tag{3.8}$$

where θ_k is the eigenfunction defined by (3.7). We emphasize the following key relation (due to the property of `ReLU` activation function): $\partial_b^2 w(b, t) = h(b, t) - f(b)$ for $b \in D$ by (3.6). Therefore the

generalized discrete Fourier transform of $\partial_b^2 w$ will be the generalized discrete Fourier transform of the error $h(\cdot,t)-f(\cdot)$ on D. Using Lemma A.1, $\partial_b^4 w(b,t)=\sum_{k=1}^{\infty}\Theta_k(t)\phi_k(b)$ for $b\in D$.

LEMMA 3.6. The following equality holds.

$$\int_D w(b,t)\theta_k^{(4)}(b)\,\mathrm{d}b=\int_D \partial_b^2 w(b,t)\theta_k''(b)\,\mathrm{d}b.$$

Proof. Note $\theta_k''(-1)=\theta_k'''(-1)=0, w(1,t)=\partial_b w(1,t)=0$ (due to$\phi_k(1)=\phi_k'(1)=0$),

$$\begin{aligned}\int_D w(b,t)\theta_k^{(4)}(b)\,\mathrm{d}b &= w(b,t)\theta_k^{(3)}(b)\big|_{\partial D}-\int_D \partial_b w(b,t)\theta_k^{(3)}(b)\,\mathrm{d}b\\ &= -\partial_b w(b,t)\theta_k^{(2)}(b)\big|_{\partial D}+\int_D \partial_b^2 w(b,t)\theta_k''(b)\,\mathrm{d}b\\ &= \int_D \partial_b^2 w(b,t)\theta_k''(b)\,\mathrm{d}b.\end{aligned}$$

□

LEMMA 3.7. There exists a constant $c>0$ that $|\widehat{w}(k,t)|\le ck^{-2}$.

Proof. Using $\theta_k^{(4)}(b)=p_k\theta_k(b)$, we have

$$|p_k\widehat{w}(k,t)|\le\int_D|\partial_b^2 w(b,t)\theta_k''(b)|\,\mathrm{d}b\le\|h(\cdot,0)-f(\cdot)\|_{L^2(D)}\|\theta_k''\|_{L^2(D)}.$$

Since $\|\theta_k''\|_{L^2(D)}=\mathcal{O}(k^2)$ by Lemma B.4 and $p_k=\Theta(k^4)$ by Lemma B.1, there exists a constant $c>0$ that $|\widehat{w}(k,t)|\le\frac{c}{k^2}$. □

From now on, we assume that $b_i(t)$ is arranged in ascending order. Let $s=s(t)$ be the smallest index such that $\{b_j(t)\}_{j\ge s}\subseteq D$. Then use integration by parts taking into account the boundary condition of θ_k, w,

$$\begin{aligned}\int_D \theta_k(b)\partial_b^4 w(b,t)\,\mathrm{d}b &= \theta_k(b)\partial_b^3 w(b,t)\Big|_{\partial D}-\theta_k'(b)\partial_b^2 w|_{\partial D}+\int_D\theta_k''(b)\partial_b^2 w(b,t)\,\mathrm{d}b\\ &= \theta_k(b)\partial_b^3 w(b,t)\Big|_{\partial D}-\theta_k'(b)\partial_b^2 w|_{\partial D}+\int_D\theta_k^{(4)}w(b,t)\,\mathrm{d}b \qquad (3.9)\\ &= \theta_k(b)\partial_b^3 w(b,t)\Big|_{\partial D}-\theta_k'(b)\partial_b^2 w|_{\partial D}+p_k\widehat{w}(k,t).\end{aligned}$$

From the fact, $\partial_b^2 w(b,t)=h(b,t)-f(b)=\sum_{i=1}^n a_i(t)\chi(b_i(t))\sigma(b-b_i(t))-f(b)$ and $\theta_k(1)=\theta_k'(1)=0$, the two boundary terms can be explicitly written out as

$$\mathcal{H}_k(t):=\theta_k(b)\partial_b^3 w(b,t)\Big|_{\partial D}=-\theta_k(-1)\left[\sum_{i=1}^{s-1}a_i(t)\chi(b_i(t))-f'(-1)\right], \qquad (3.10)$$

$$\mathcal{J}_k(t):=-\theta_k'(b)\partial_b^2 w(b,t)\Big|_{\partial D}=-\theta_k'(-1)\left[\sum_{i=1}^{s-1}\chi(b_i(t))a_i(t)(1+b_i(t))+f(-1)\right]. \qquad (3.11)$$

Differentiating (3.9) in time and using (3.4), (3.5) and (A.1), we have

$$p_m \partial_t \widehat{w}(m,t) + \mathcal{H}'_m(t) + \mathcal{J}'_m(t) = \sum_{k=1}^{\infty} \Theta'_k(t)\widehat{\phi}_k(m)$$
$$= -n \sum_{k=1}^{\infty} \widehat{\phi}_k(m) \left(\int_{D^\varepsilon} w(b,t)\phi_k(b)\mu_0(b,t)\,\mathrm{d}b + \int_{D^\varepsilon} \partial_b w(b,t)\phi'_k(b)\mu_2(b,t)\,\mathrm{d}b \right), \tag{3.12}$$

where μ_0 and μ_2 are defined by the following positive distributions (n can be finite)

$$\begin{aligned} \mu_0(b,t) &= \frac{1}{n}\sum_{i=1}^{n} \delta(b - b_i(t)), \\ \mu_2(b,t) &= \frac{1}{n}\sum_{i=1}^{n} |a_i(t)|^2 \delta(b - b_i(t)). \end{aligned} \tag{3.13}$$

Apply the equality $\sum_{k=1}^{\infty} \phi_k(x)\phi_k(y) = \delta(x-y)$ to (3.12), we find

$$\begin{aligned} &\sum_{k=1}^{\infty} \widehat{\phi}_k(m) \int_{D^\varepsilon} w(b,t)\phi_k(b)\mu_0(b,t)\,\mathrm{d}b \\ &= \int_{D^\varepsilon}\int_{D} w(b,t) \sum_{k=1}^{\infty} \phi_k(b)\phi_k(b')\mu_0(b,t)\theta_m(b')\,\mathrm{d}b'\,\mathrm{d}b \\ &= \int_{D^\varepsilon}\int_{D} w(b,t)\delta(b-b')\mu_0(b,t)\theta_m(b')\,\mathrm{d}b'\,\mathrm{d}b \\ &= \int_{D} w(b',t)\mu_0(b',t)\theta_m(b')\,\mathrm{d}b' = \widehat{w\mu_0}(m,t). \end{aligned}$$

Similarly,

$$\begin{aligned} &\sum_{k=1}^{\infty} \widehat{\phi}_k(m) \int_{D^\varepsilon} \partial_b w(b,t)\phi'_k(b)\mu_2(b,t)\,\mathrm{d}b \\ &= \int_{D^\varepsilon}\int_{D} \partial_b w(b,t) \sum_{k=1}^{\infty} \phi'_k(b)\phi_k(b')\mu_2(b,t)\theta_m(b')\,\mathrm{d}b'\,\mathrm{d}b \\ &= \int_{D^\varepsilon}\int_{D} \partial_b w(b,t)\delta'(b-b')\mu_2(b,t)\theta_m(b')\,\mathrm{d}b'\,\mathrm{d}b \\ &= \int_{D} \theta_m(b')\delta'(b-b')\,\mathrm{d}b' \int_{D^\varepsilon} \partial_b w(b,t)\mu_2(b,t)\,\mathrm{d}b \\ &= \int_{D} \partial_b w(b',t)\mu_2(b',t)\theta'_m(b')\,\mathrm{d}b'. \end{aligned}$$

Therefore (3.12) can be further reduced to

$$\partial_t \widehat{w}(m,t) = -\frac{1}{p_m}(\mathcal{H}'_m(t) + \mathcal{J}'_m(t)) - \frac{n}{p_m}\widehat{w\mu_0}(m,t) - \frac{n}{p_m}\int_D \partial_b w(b',t)\mu_2(b',t)\theta'_m(b')\,\mathrm{d}b'. \tag{3.14}$$

Now we provide an estimate for the above equation and show a slow reduction of high-frequency modes in the initial error during the gradient flow.

THEOREM 3.8. Assume that $\sup_{1\le i\le n}|a_i(t)|^2$ is uniformly bounded by $M > 0$ and the biases $\{b_i(0)\}$ are initially equispaced on D. If $\widehat{w}(m,0) \neq 0$, then there exists a constant $\widetilde{C} > 0$ depending on the initialized loss that

$$|\widehat{w}(m,t)| > \frac{1}{2}|\widehat{w}(m,0)|, \quad 0 \le t \le \frac{p_m|\widehat{w}(m,0)|}{2n\widetilde{C}(m+1)}. \tag{3.15}$$

Especially, denote the initial error in the generalized Fourier mode θ_m by $|\widehat{w}(m,0)| > c'm^{-2}$ for certain $c' > 0$, then the error in the generalized Fourier mode θ_m takes at least $\mathcal{O}(\frac{c'm}{n})$ time to get reduced by half following gradient decent dynamics.

Proof. We estimate the contribution from each term on the right-hand side of (3.14). In particular we have $\|\theta_k\|_{L^\infty(D)} = \mathcal{O}(1)$, $\|\theta'_k\|_{L^\infty(D)} = \mathcal{O}(k)$ from Lemma B.4.

1. First, there exists a constant $C > 0$ that

$$\begin{aligned}\left|-\frac{n}{p_m}\widehat{w\mu_0}(m,t)\right| &= \left|\frac{n}{p_m}\int_D w(b',t)\mu_0(b',t)\theta_m(b')\,\mathrm{d}b'\right| \\ &\le \frac{n}{p_m}\|w(\cdot,t)\|_{C(D)}\|\theta_m\|_{C(D)} \le \frac{Cn}{p_m}.\end{aligned} \tag{3.16}$$

2. Since $w \in H^2(D)$, it can be embedded into $C^{1,\alpha}(D)$ compactly, which means $\partial_b w$ is uniformly bounded on D, hence there exists a constant C' that

$$\begin{aligned}\left|-\frac{n}{p_m}\int_D \partial_b w(b',t)\mu_2(b',t)\theta'_m(b')\,\mathrm{d}b'\right| &\le \frac{n}{p_m}\|\theta'_m\|_{C(D)}\|\partial_b w\|_{C(D)}\int_D \mu_2(b,t)\,\mathrm{d}b \\ &\le \frac{C'Mmn}{p_m}.\end{aligned}$$

3. Using the definition of $\mathcal{H}_m$ in (3.10) and $\mathcal{J}_m$ in (3.11), now we give the upper bounds of $|\mathcal{H}_m(t) - \mathcal{H}_m(0)|$ and $|\mathcal{J}_m(t) - \mathcal{J}_m(0)|$. Since initially $b_i(0) \in D$ and $\chi(b) \le 1$, we have

$$|\mathcal{H}_m(t) - \mathcal{H}_m(0)| = |\theta_m(-1)|\left|\sum_{i=1}^{s(t)-1} a_i(t)\chi(b_l(t))\right| \le C''\sqrt{M}|s(t)-1| \tag{3.17}$$

and

$$|\mathcal{J}_m(t) - \mathcal{J}_m(0)| = |\theta'_m(-1)| \left| \sum_{i=1}^{s(t)-1} a_i(t)\chi(b_i(t))(1+b_i(t)) \right| \le C'''m\sqrt{M}|s(t)-1|. \quad (3.18)$$

Now we estimate the number $s(t)$. Because the biases have a finite propagation speed

$$\left|\frac{d}{dt}b_i(t)\right| \le |a_i| \int_D |h(x,t) - f(x)||\partial_b \psi(x, b_i(t))|\,\mathrm{d}x \le K := C''''\sqrt{M}\|h(\cdot,0) - f(\cdot)\|_{L^2(D)}.$$

Therefore there are at most $\frac{1}{2}nKt$ initially evenly spaced biases moving into the transition interval. That gives $|s(t)-1| \le \frac{1}{2}nKt$.

The above estimates imply that there exists a constant $\widetilde{C} > 0$ that

$$\begin{aligned} |\widehat{w}(m,t)| - |\widehat{w}(m,0)| &\ge -\frac{n}{p_m}(C'Mm + C)t - \frac{1}{p_m}\left(C''\sqrt{M} + C'''m\sqrt{M}\right)\frac{1}{2}Knt \\ &\ge -\frac{\widetilde{C}n}{p_m}(m+1)t. \end{aligned} \quad (3.19)$$

When $\widehat{w}(m,0) \neq 0$, we solve the lower bound of the *half-reduction* time $\tau > 0$ that

$$\frac{1}{2}|\widehat{w}(m,0)| = \frac{\widetilde{C}n}{p_m}(m+1)\,\tau \quad \Longrightarrow \quad \tau = \frac{p_m|\widehat{w}(m,0)|}{2n\widetilde{C}\,(m+1)}.$$

In particular, if $|\widehat{w}(m,0)| > c'm^{-2}$ for certain $c' > 0$, the *half-reduction* time is at least $\mathcal{O}(\frac{c'm}{n})$. □

To keep a discretized gradient descent method close to the continuous gradient flow, the learning rate or step size should be small. In practice, one typically takes $\Delta t = \mathcal{O}(\frac{1}{n})$, it will take at least $\mathcal{O}(m)$ time steps to reduce the initial error $h(x,0) - f(x)$ in the generalized Fourier mode θ_m by half. It is also worth noticing that this phenomenon does not depend on the convergence of the trainable parameters.

A lower bound estimate only provides an optimistic scenario which may not be sharp. When the network width and the frequency mode satisfy different scaling laws, improved estimates can be obtained. See Appendix C for improved estimates, numerical evidence and the proof of Theorem 3.2.

4. Rashomon set for two-layer `ReLU` NNs

In [49, 50], the authors claimed that the measure of the so-called Rashomon set can be used as a criterion to see if a simple model exists for the approximation problem. Let $D = B_d(1)$ be the unit ball in $\mathbb{R}^d$. The Rashomon set $\mathcal{R}_\varepsilon$ for the two-layer NN class $\mathcal{H}_n$ is defined as the following

$$\mathcal{R}_\varepsilon := \left\{h \in \mathcal{H}_n \mid \|h - f\|_{L^2(D)} \le \varepsilon\|f\|_{L^2(D)}\right\},$$

where m is the number of parameters for $\mathcal{H}_n$. If we normalize the measure for $\mathcal{H}_n$, the measure of the Rashomon set quantifies the probability that the loss is under a certain threshold for a random pick of

parameters. The main purpose of this section is to characterize the probability measure of the Rashomon set with $\mathcal{H}_n$ representing the two-layer `ReLU` NNs

$$h(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} a_j \sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} - b_j) + \boldsymbol{v}\cdot\boldsymbol{x} + c, \quad \boldsymbol{x} \in D \subseteq \mathbb{R}^d,$$

where the parameters $\{a_j\}_{j=1}^n$ are bounded by $[-A, A]$ and $\{b_i\}_{i=1}^n \subseteq [-1, 1]$. It is already known that this network can approximate a function f with an error no more than $\mathcal{O}(\sqrt{d+\log n}\, n^{-1/2-1/d})$ if the Fourier transform $\widehat{f}$ satisfies $\int_{\mathbb{R}^d} |\widehat{f}(\zeta)| \|\zeta\|_1^2 d\zeta < \infty$ [31]. The general bounds in Hölder space have been studied in [37]. Taking Laplacian on h,

$$\Delta h(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} a_j \Delta\sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} - b_j) = \frac{1}{n}\sum_{j=1}^{n} a_j \delta(\boldsymbol{w}_j \cdot \boldsymbol{x} - b_j).$$

THEOREM 4.1. Suppose $f \in C(D)$ such that there exists $g \in C_0^2(D)$ that $\Delta g = f$, then the Rashomon set $\mathcal{R}_\varepsilon \subseteq \mathcal{H}_n$ satisfies

$$\mathbb{P}(\mathcal{R}_\varepsilon) \le \exp\left(-\frac{n(1-\varepsilon)^2 \|f\|_{L^2(D)}^4}{2A^2\kappa^2}\right), \quad \kappa := \sup_{(\boldsymbol{w},b)} \int_{\{\boldsymbol{x}\in D, \boldsymbol{w}\cdot\boldsymbol{x}=b\}} g(\boldsymbol{x})\, \mathrm{d}H_{d-1}(\boldsymbol{x}).$$

Proof. Let $g \in C_0^2(D)$ that solves $\Delta g(\boldsymbol{x}) = f(\boldsymbol{x})$ with Dirichlet boundary condition.

$$\langle \Delta h(\boldsymbol{x}), g(\boldsymbol{x})\rangle = \frac{1}{n}\sum_{j=1}^{n} X_j, \quad X_j := a_j \int_{\boldsymbol{w}_j\cdot\boldsymbol{x}=b_j} g(\boldsymbol{x})\, \mathrm{d}H_{d-1}(\boldsymbol{x}).$$

The random variables X_j are i.i.d and bounded by $[-A\kappa, A\kappa]$ where κ is an absolute constant defined by

$$\kappa := \sup_{(\boldsymbol{w}_j, b_j)\in\mathbb{S}^{d-1}\times[-1,1]} \int_{\boldsymbol{w}_j\cdot\boldsymbol{x}=b_j} g(\boldsymbol{x})\, \mathrm{d}H_{d-1}(\boldsymbol{x}).$$

Then use Hoeffding's inequality,

$$\mathbb{P}\left[\frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}[X_j] \ge t\right] \le \exp\left(-\frac{nt^2}{2A^2\kappa^2}\right).$$

Especially if $\mathbb{E}(a_j) = 0$ hence $\mathbb{E}(X_j) = 0$, then

$$\mathbb{P}\left[\langle \Delta h(\boldsymbol{x}), g(\boldsymbol{x})\rangle \ge t\right] \le \exp\left(-\frac{nt^2}{2A^2\kappa^2}\right).$$

Using Green's formula,

$$\langle \Delta h(\boldsymbol{x}), g(\boldsymbol{x})\rangle - \langle h(\boldsymbol{x}), \Delta g(\boldsymbol{x})\rangle = \int_{\partial D} \left(\partial_n h(\boldsymbol{x}) f(\boldsymbol{x}) - \partial_n f(\boldsymbol{x}) h(\boldsymbol{x})\right) \mathrm{d}s = 0,$$

then using the network to approximate $\Delta g = f$ is difficult if κ is small in the sense that

$$\begin{aligned}\mathbb{P}\left[\|h(\boldsymbol{x}) - \Delta g(\boldsymbol{x})\|_{L^2(D)} \le \varepsilon \|\Delta g\|_{L^2(D)}\right] &\le \mathbb{P}\left[\langle h(\boldsymbol{x}), \Delta g(\boldsymbol{x})\rangle \ge (1-\varepsilon)\|\Delta g\|^2_{L^2(D)}\right] \\ &\le \exp\left(-\frac{n(1-\varepsilon)^2\|\Delta g\|^4_{L^2(D)}}{2A^2\kappa^2}\right).\end{aligned}$$

Here we have used the fact that $\|h(\boldsymbol{x}) - \Delta g(\boldsymbol{x})\|_{L^2(D)} \le \varepsilon\|\Delta g\|_{L^2(D)}$ implies both $\|h\|_{L^2(D)} \ge (1-\varepsilon)\|\Delta g\|_{L^2(D)}$ and

$$\langle h, \Delta g\rangle \ge \frac{1-\varepsilon^2}{2}\|\Delta g\|^2_{L^2(D)} + \frac{1}{2}\|h\|^2_{L^2(D)} \ge (1-\varepsilon)\|\Delta g\|^2_{L^2(D)}.$$

□

If f oscillates with frequency ν in every direction, then $\kappa \approx \nu^{-2}$ and the probability measure of Rashomon set scales as $\exp(-\mathcal{O}(\nu^{-4}))$ which gives another perspective why oscillatory functions are difficult to approximate by NNs in general. The proof and extension to more general activation functions are provided in Appendix D.

5. Further discussions

In this study, it is shown that the use of highly correlated activation functions in a two-layer NN makes it filter out fine features (high-frequency components) when finite machine precision is imposed, which is an implicit regularization in practice. Moreover, increasing the network width does not improve the numerical accuracy after a certain threshold is reached, although the universal approximation property is proved in theory. The smoother the activation function is, the faster the Gram matrix spectrum decays (see Remark 2.11 and Appendix E), and hence the stronger the regularization is. We plan to investigate how a multi-layer network could overcome these issues through effective decomposition and composition in our future work.

Acknowledgments

The authors gratefully acknowledge the editor and the anonymous referees for their constructive and insightful comments, which have substantially improved the quality and presentation of the paper.

Funding

S.Z. was partially supported by start-up fund P0053092 from The Hong Kong Polytechnic University. H.Z. was partially supported by NSF grants (DMS-2309551), and (DMS-2012860). Y.Z. was partially supported by NSF grant (DMS-2309530) and H.Z. was partially supported by NSF grant (DMS-2307465).

Data availability statement

No new data were generated or analyzed in support of this review.

References

1. Bao, C., Li, Q., Shen, Z., Tai, C., Lei, W. & Xiang, X. (2023) Approximation analysis of convolutional neural networks. *East Asian J. Appl. Math.*, **13**, 524–549.
2. Barron, A. R. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, **39**, 930–945.
3. Barron, A. R. & Klusowski, J. M. (2018) Approximation and estimation for high-dimensional deep learning networks *arXiv e-prints*, page arXiv:1809.03090.
4. Birkhoff, G. D. (1908) Boundary value and expansion problems of ordinary linear differential equations. *Trans. Amer. Math. Soc.*, **9**, 373–395.
5. Birman, M. S. & Solomyak, M. Z. (1970) Asymptotic behavior of the spectrum of weakly polar integral operators. *Izv. RAN. Ser. Mat.*, **34**, 1142–1158.
6. Bogoya, J. M., Böttcher, A. & Poznyak, A. (2012) Eigenvalues of hermitian toeplitz matrices with polynomially increasing entries. *J. Spectral Theory*, **2**, 267–292.
7. Breiman, L. (1993) Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. Inf. Theory*, **39**, 999–1013.
8. Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X. & Gu, Q. (2021) Towards understanding the spectral bias of deep learning *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization. 2205–2211. arXiv preprint arXiv:1912.01198.
9. Chen, M., Jiang, H., Liao, W. & Zhao, T. (2019) Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. *Advances in Neural Information Processing Systems* (Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d'., Fox, E. & Garnett, R., eds), vol. **32**. Curran Associates, Inc.
10. Chen, J., Chi, X., Yang, Z., et al. (2022) Bridging traditional and machine learning-based algorithms for solving PDEs: the random feature method. *J. Mach. Learn.*, **1**, 268–298.
11. Chizat, L. & Bach, F. (2018) On the global convergence of gradient descent for over-parameterized models using optimal transport. In: Bengio S., Wallach H., Larochelle H., Grauman K., Cesa-Bianchi N., Garnett R., (eds.), *Advances in Neural Information Processing Systems*, Curran Associates, Inc. **31**. https://proceedings.neurips.cc/paper_files/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf
12. Chui, C. K. & Li, X. (1992) Approximation by ridge functions and neural networks with one hidden layer. *J. Approx. Theory*, **70**, 131–141.
13. Chui, C. K., Lin, S.-B. & Zhou, D.-X. (2018) Construction of neural networks for realization of localized deep learning. *Front. Appl. Math. Stat.*, **4**, 14.
14. Cybenko, G. (1989) Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, **2**, 303–314.
15. Domingo-Enrich, C & Mroueh, Y. (2022) Tighter sparse approximation bounds for ReLU neural networks. *ICLR*.
16. Du, S., Lee, J., Li, H., Wang, L. & Zhai, X. (2019) Gradient descent finds global minima of deep neural networks. In: Chaudhuri, Kamalika, Salakhutdinov, Ruslan (eds.), *International Conference on Machine Learning*. PMLR, pp. 1675–1685.
17. Eldan, R. & Shamir, O. (2016) The power of depth for feedforward neural networks. In: Feldman Vitaly, Rakhlin Alexander, Shamir Ohad (eds.), *Conference on Learning Theory*, New York, USA: PMLR, pp. 907–940.
18. Frobenius, G. (1912) Über matrizen aus nicht negativen elementen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, **23**, 456–477.
19. Gribonval, R., Kutyniok, G., Nielsen, M. & Voigtlaender, F. (2022) Approximation spaces of deep neural networks. *Constr. Approx.*, **55**, 259–367.

20. Grossmann, T. G., Komorowska, U. J., Latz, J. & Schönlieb, C. -B. (2024) Can physics-informed neural networks beat the finite element method? *IMA J. Appl. Math.*, **89**, 143–174.
21. Gühring, I., Kutyniok, G. & Petersen, P. (2020) Error bounds for approximations with deep ReLU neural networks in W^{s}, p norms. *Anal. Appl.*, **18**, 803–859.
22. Helgason, S. (2011) *Integral Geometry and Radon Transforms*. New York, NY: Springer.
23. Helgason, S. (2022) *Groups and Geometric Analysis: Integral Geometry, Invariant Differential Operators, and Spherical Functions*, vol. **83**. American Mathematical Society.
24. Holzmüller, D. & Steinwart, I. (2022) Training two-layer ReLU networks with gradient descent is inconsistent. *Journal of Machine Learning Research*. **23**, 1–82.
25. Hong, Q., Tan, Q., Siegel, J. W. & Xu, J. (2022) On the activation function dependence of the spectral bias of neural networks arXiv preprint arXiv:2208.04924.
26. Horn, R. A. & Johnson, C. R. (2012) *Matrix Analysis*. Cambridge University Press.
27. Hornik, K., Stinchcombe, M. & White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.
28. Jacot, A., Gabriel, F. & Hongler, C. (2018) Neural tangent kernel: convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, **31**.
29. Jianfeng, L., Shen, Z., Yang, H. & Zhang, S. (2021) Deep network approximation for smooth functions. *SIAM J. Math. Anal.*, **53**, 5465–5506.
30. Jiao, Y., Lai, Y., Xiliang, L., Wang, F., Yang, J. Z. & Yang, Y. (2023) Deep neural networks with ReLU-sine-exponential activations break curse of dimensionality in approximation on hölder class. *SIAM J. Math. Anal.*, **55**, 3635–3649.
31. Kamenski, L., Huang, W. & Hongguo, X. (2014) Conditioning of finite element equations with arbitrary anisotropic meshes. *Math. Comp.*, **83**, 2187–2211.
32. Klusowski, J. M. & Barron, A. R. (2018) Approximation by combinations of ReLU and squared ReLU ridge functions with l^1 and l^0 controls. *IEEE Trans. Inf. Theory*, **64**, 7649–7656.
33. Koltchinskii, V., Giné, E. & Gine, E. (2000) Random matrix approximation of spectra of integral operators. *Bernoulli*, **6**, 113–167.
34. Li, Y. & Liang, Y. (2018) Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in Neural Information Processing Systems*, **31**, 6799–6810.
35. Li, Y., Ma, T. & Zhang, H. R. (2020) Learning over-parametrized two-layer neural networks beyond NTK. *Proceedings of Thirty Third Conference on Learning Theory, volume 125 of Proceedings of Machine Learning Research* (Abernethy, J. & Agarwal, S. eds). PMLR, pp. 2613–2682.
36. Lili, S. & Yang, P. (2019) On learning over-parameterized neural networks: a functional approximation perspective. *Advances in Neural Information Processing Systems*, **32**.
37. Luo, T., Ma, Z., Zhi-Qin John, X. & Zhang, Y. (2021) Theory of the frequency principle for general deep neural networks. *CSIAM Trans. Appl. Math.*, **2**, 484–507.
38. Mao, T. & Zhou, D.-X. (2023) Rates of approximation by ReLU shallow neural networks. *J. Complexity*, **79**, 101784.
39. Mei, S., Montanari, A. & Nguyen, P.-M. (2018) A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, **115**, E7665–E7671.
40. Naimark, M. A. (1967) *Linear Differential Operators*. New York: Frederick Ungar.
41. Nakada, R. & Imaizumi, M. (2020) Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, **21**, 1–38.
42. Nitanda, A. & Suzuki, T. (2021) Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *ICLR*.
43. Ongie, G., Willett, R., Soudry, D. & Srebro, N. (2020) A function space view of bounded norm infinite width ReLU nets: the multivariate case. *ICLR*.
44. Park, J. & Sandberg, I. W. (1991) Universal approximation using radial-basis-function networks. *Neural Comput.*, **3**, 246–257.
45. Reade, J. B. (1983) Eigenvalues of positive definite kernels. *SIAM J. Math. Anal.*, **14**, 152–157.

46. Reade, J. B. (1984) Eigenvalues of positive definite kernels II. *SIAM J. Math. Anal.*, **15**, 137–142.
47. Rotskoff, G. & Vanden-Eijnden, E. (2022) Trainability and accuracy of artificial neural networks: an interacting particle system approach. *Comm. Pure Appl. Math.*, **75**, 1889–1935.
48. Safran, I. & Shamir, O. (2017) Depth-width tradeoffs in approximating natural functions with neural networks. *International Conference on Machine Learning*. PMLR, pp. 2979–2987.
49. Savarese, P., Evron, I., Soudry, D. & Srebro, N. (2019) How do infinite width bounded norm networks look in function space? *Conference on Learning Theory*. PMLR, pp. 2667–2690.
50. Semenova, L., Rudin, C. & Parr, R. (2019) A study in rashomon curves and volumes: a new perspective on generalization and model simplicity in machine learning arXiv preprint arXiv:1908.01755.
51. Semenova, L., Rudin, C. & Parr, R. (2022) On the existence of simpler machine learning models. *2022 ACM Conference on Fairness, Accountability, and Transparency*. pp. 1827–1858.
52. Shen, Z., Yang, H. & Zhang, S. (2019) Nonlinear approximation via compositions. *Neural Networks*, **119**, 74–84.
53. Shen, Z., Yang, H. & Zhang, S. (2020) Deep network approximation characterized by number of neurons. *Commun. Comput. Phys.*, **28**, 1768–1811.
54. Shen, Z., Yang, H. & Zhang, S. (2021) Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Comput.*, **33**, 1005–1036.
55. Shen, Z., Yang, H. & Zhang, S. (2021) Neural network approximation: three hidden layers are enough. *Neural Networks*, **141**, 160–173.
56. Shen, Z., Yang, H. & Zhang, S. (2022) Deep network approximation: achieving arbitrary accuracy with fixed number of neurons. *J. Mach. Learn. Res.*, **23**, 1–60.
57. Shen, Z., Yang, H. & Zhang, S. (2022) Neural network architecture beyond width and depth. *Advances in Neural Information Processing Systems*, vol. **35**. (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh eds). Curran Associates, Inc., pp. 5669–5681.
58. Sirignano, J. & Spiliopoulos, K. (2020) Mean field analysis of neural networks: a central limit theorem. *Stochastic Process. Appl.*, **130**, 1820–1852.
59. Stone, M. H. (1926) A comparison of the series of Fourier and birkhoff. *Trans. Amer. Math. Soc.*, **28**, 695–761.
60. Suzuki, T. (2019) Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. *International Conference on Learning Representations*.
61. Velikanov, M. & Yarotsky, D. (2021) Explicit loss asymptotics in the gradient descent training of neural networks. *Advances in Neural Information Processing Systems*, **34**, 2570–2582.
62. Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. **47**. Cambridge University Press.
63. Widom, H. (1958) On the eigenvalues of certain hermitian operators. *Trans. Amer. Math. Soc.*, **88**, 491–522.
64. Yarotsky, D. (2017) Error bounds for approximations with deep ReLU networks. *Neural Networks*, **94**, 103–114.
65. Yarotsky, D. (2018) Optimal approximation of continuous functions by very deep ReLU networks. *Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research*. (Bubeck, S., Perchet, V. & Rigollet, P. eds), PMLR, pp. 639–649.
66. Zhang, S. (2020) *Deep Neural Network Approximation via Function Compositions* PhD Thesis. National University of Singapore.
67. Zhang, G., Martens, J. & Grosse, R. B. (2019) Fast convergence of natural gradient descent for over-parameterized neural networks. *Advances in Neural Information Processing Systems*, **32**.
68. Zhi-Qin John, X., Zhang, Y., Luo, T., Xiao, Y. & Ma, Z. (2020) Frequency principle: Fourier analysis sheds light on deep neural networks. *Commun. Comput. Phys.*, **28**, 1746–1767.
69. Zhou, D.-X. (2020) Universality of deep convolutional neural networks. *Appl. Comput. Harmon. Anal.*, **48**, 787–794.

A. Properties of eigenfunctions

We show properties of the eigenfunctions ϕ_k for the Gram kernel defined by (2.5) or equivalently by (3.3).

LEMMA A1. ϕ_k is a cubic polynomial over $(-1-\varepsilon,-1)$ and

$$\partial_b^4\phi_k(b)=\begin{cases}0, & b\in(-1-\varepsilon,-1),\\ \lambda_k^{-1}\phi_k(b), & b\in(-1,1),\end{cases} \tag{A.1}$$

where $\partial_b\phi_k$ is continuous at $b=-1$ but $\partial_b^2\phi_k$ has a jump at $b=-1$.

Proof. We use the definition of eigenfunction

$$\int_{D^\varepsilon}\mathcal{G}(b,b')\phi_k(b')\,\mathrm{d}b'=\lambda_k\phi_k(b).$$

For $b\in(-1,1)$, $\chi(b)=1$, then differentiating both sides

$$\lambda_k\partial_b^4\phi_k(b)=\partial_b^4\int_{D^\varepsilon}\int_D\sigma(x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\mathrm{d}b'=\int_{D^\varepsilon}\delta(b-b')\chi(b')\phi_k(b')\,\mathrm{d}b'=\phi_k(b).$$

For $b\in(-1-\varepsilon,-1)$, $\chi(b)$ is quadratic, $\sigma(x-b)=x-b$ for $x\in D$, we differentiate both sides

$$\lambda_k\partial_b^4\phi_k(b)=\partial_b^4\left(\chi(b)\int_{D^\varepsilon}\int_D(x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)=0.$$

Now, we compute $\partial_b\phi_k$ and $\partial_b^2\phi_k$ across $b=-1$. Note $\chi'(-1)=0$, then

$$\begin{aligned}\lim_{b\to-1^-}\partial_b\phi_k(b)&=\frac{1}{\lambda_k}\lim_{b\to-1^-}\partial_b\left(\chi(b)\int_{D^\varepsilon}\int_D(x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\&=\frac{1}{\lambda_k}\chi'(-1)\lim_{b\to-1^-}\left(\int_{D^\varepsilon}\int_D(x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\&\quad+\frac{1}{\lambda_k}\chi(-1)\lim_{b\to-1^-}\left(\int_{D^\varepsilon}\int_D(-1)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\&=\frac{1}{\lambda_k}\left(\int_{D^\varepsilon}\int_D(-1)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\end{aligned}$$

and because $\sigma'(x-b)=1$ if $b=-1$ and $x\in D$,

$$\begin{aligned}\lim_{b\to-1^+}\partial_b\phi_k(b)&=\frac{1}{\lambda_k}\lim_{b\to-1^+}\partial_b\left(\int_{D^\varepsilon}\int_D\sigma(x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\&=\frac{1}{\lambda_k}\lim_{b\to-1^+}\left(\int_{D^\varepsilon}\int_D-\sigma'(x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\&=\frac{1}{\lambda_k}\left(\int_{D^\varepsilon}\int_D(-1)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right).\end{aligned}$$

For $\partial_b^2\phi_k$ across $b=-1$, following a similar process,

$$\begin{aligned}\lim_{b\to -1^-}\partial_b^2\phi_k(b) &= \frac{1}{\lambda_k}\lim_{b\to -1^-}\partial_b^2\left(\chi(b)\int_{D^\varepsilon}\int_D (x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\ &= \frac{1}{\lambda_k}\chi''(-1)\lim_{b\to -1^-}\left(\int_{D^\varepsilon}\int_D (x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\ &= -\frac{1}{\lambda_k}\frac{2}{\varepsilon^2}\left(\int_{D^\varepsilon}\int_D (x+1)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\end{aligned}$$

and

$$\begin{aligned}\lim_{b\to -1^+}\partial_b^2\phi_k(b) &= \frac{1}{\lambda_k}\lim_{b\to -1^+}\partial_b^2\left(\int_{D^\varepsilon}\int_D \sigma(x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\ &= \frac{1}{\lambda_k}\lim_{b\to -1^+}\left(\int_{D^\varepsilon}\int_D \delta(x-b)\sigma(x-b')\chi(b')\phi_k(b')\,\mathrm{d}x\,\mathrm{d}b'\right)\\ &= \frac{1}{\lambda_k}\int_{D^\varepsilon}\sigma(-1-b')\chi(b')\phi_k(b')\,\mathrm{d}b' = \frac{1}{\lambda_k}\int_{-1-\varepsilon}^{-1}(-1-b')\chi(b')\phi_k(b')\,\mathrm{d}b'.\end{aligned}$$

As $\varepsilon\to 0$, $\lim_{b\to -1^+}\partial_b^2\phi_k(b)=O(\varepsilon)$ while $\lim_{b\to -1^-}\partial_b^2\phi_k(b)=O(\varepsilon^{-2})$. □

B. Generalized Fourier modes

In this section, we study the eigenfunctions ϕ_k for the one-dimensional continuous Gram kernel defined in (2.5), which are exactly the generalized Fourier modes θ_k defined in (3.7).

LEMMA B1. The eigenfunction θ_k satisfies

$$\theta_k(x) = A_k\cosh(c_k x)+B_k\sinh(c_k x)+C_k\cos(c_k x)+D_k\sin(c_k x),$$

where $\tanh(c_k)\tan(c_k)=\pm 1$, $p_k=c_k^4$, and $c_{2k+1}\in((k+\frac{1}{4})\pi,(k+\frac{1}{2})\pi)$, $c_{2k+2}\in((k+\frac{1}{2})\pi,(k+\frac{3}{4})\pi)$, $k\geq 0$.

Proof. The form of the eigenfunction is standard. Using the boundary conditions, $\theta_k(1)=\theta_k''(-1)=0$,

$$\begin{aligned}A_k\cosh(c_k)+B_k\sinh(c_k)+C_k\cos(c_k)+D_k\sin(c_k)=0,\\ A_k\cosh(c_k)-B_k\sinh(c_k)-C_k\cos(c_k)+D_k\sin(c_k)=0,\end{aligned}$$

which means

$$A_k\cosh(c_k)+D_k\sin(c_k)=0,\quad B_k\sinh(c_k)+C_k\cos(c_k)=0.$$

The other two boundary conditions are

$$\begin{aligned}A_k\sinh(c_k)+B_k\cosh(c_k)-C_k\sin(c_k)+D_k\cos(c_k)=0,\\ -A_k\sinh(c_k)+B_k\cosh(c_k)-C_k\sin(c_k)-D_k\cos(c_k)=0,\end{aligned}$$

which means

$$A_k\sinh(c_k)+D_k\cos(c_k)=0,\quad B_k\cosh(c_k)-C_k\sin(c_k)=0.$$

If $C_k \neq 0$, then $B_k = -C_k \frac{\cos(c_k)}{\sinh(c_k)} = C_k \frac{\sin(c_k)}{\cosh(c_k)}$, which is $\tan(c_k)\tanh(c_k) = -1$. If $D_k \neq 0$, then we can derive $\tan(c_k)\tanh(c_k) = 1$. Let $r(x) = \tan(x)\tanh(x)$, for $x \in (0, \frac{\pi}{2})$ or $x \in (n\pi + \frac{1}{2}\pi, n\pi + \frac{3\pi}{2})$, $n \in \mathbb{Z}$, the function r is monotone, therefore $c_{2k+1} \in ((k+\frac{1}{4})\pi, (k+\frac{1}{2})\pi)$, $c_{2k+2} \in ((k+\frac{1}{2})\pi, (k+\frac{3}{4})\pi)$, $k \geq 0$. □

From the above analysis, the eigenfunctions are

$$\theta_{2k+1}(x) = C_{2k+1}\left(-\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\sinh(c_{2k+1}x) + \cos(c_{2k+1}x)\right),$$
$$\theta_{2k+2}(x) = D_{2k+2}\left(-\frac{\sin(c_{2k+2})}{\cosh(c_{2k+2})}\cosh(c_{2k+2}x) + \sin(c_{2k+2}x)\right),$$

which shows $\{\theta_k\}_{k\geq 1}$ are actually the eigenfunctions of the Gram kernel $\mathcal{G}$ in (2.3).

LEMMA B2. $\{\theta_k\}_{k\geq 1}$ forms an orthonormal basis of $L^2(D)$.

Proof. Since each eigenvalue is simple from Theorem B.4 and we verify the following equality using integration by parts,

$$p_i \int_D \theta_i(x)\theta_j(x)\,\mathrm{d}x = \int_D \theta_i^{(4)}(x)\theta_j(x)\,\mathrm{d}x = \int_D \theta_i(x)\theta_j^{(4)}(x)\,\mathrm{d}x = p_j \int_D \theta_i(x)\theta_j(x)\,\mathrm{d}x.$$

Thus $\{\theta_i\}_{i\geq 1}$ forms an orthonormal basis. □

LEMMA B3. The eigenfunctions $\{\theta_k\}_{k\geq 1}$ can form a complete basis for $L^2(D)$.

Proof. Suppose $\{\theta_m\}_{m=1}^{\infty}$ is not complete, then there exists a nonzero $\gamma \in L^2(D)$ that $\widehat{\gamma}(m) = 0$ for all $m \in \mathbb{N}$, then using the fact that $\{\theta_m\}_{m\geq 1}$ are the eigenfunctions of the Gram kernel $\mathcal{G}$ in (2.3), note $\mathcal{G}(x,y) = \mathcal{G}(y,x)$, it is self-adjoint, then by Hilbert–Schmidt theorem,

$$\int_D \mathcal{G}(x,y)\gamma(y)\,\mathrm{d}y = 0, \tag{B.1}$$

and we differentiate the above equation four times on both sides, which leads to $\gamma^{(4)}(x) = 0$ which means γ should be a cubic polynomial with boundary conditions $\gamma(1) = \gamma'(1) = \gamma''(-1) = \gamma'''(-1) = 0$, which means $\gamma \equiv 0$, it is a contradiction with our assumption of $\|\gamma\|_D \neq 0$. □

Next, we show the eigenfunctions are *almost* Fourier modes.

THEOREM B4. The following statements hold

1. $(k+\frac{1}{4})\pi < c_{2k+1} < (k+\frac{1}{4})\pi + e^{-2c_{2k+1}}$ and $(k+\frac{3}{4})\pi > c_{2k+2} > (k+\frac{3}{4})\pi - e^{-2c_{2k+2}}$.
2. $A_k, B_k = \mathcal{O}(e^{-c_k})$ and $C_k, D_k = \mathcal{O}(1)$.
3. $\|\theta_k\|_{L^\infty(D)} = \mathcal{O}(1)$, $\|\theta_k'\|_{L^\infty(D)} = \mathcal{O}(k)$ and $\|\theta_k''\|_{L^\infty(D)} = \mathcal{O}(k^2)$.
4. $\|\theta_{2k+1}(x) - \cos(c_{2k+1}x)\|_{L^2(D)} = \mathcal{O}(k^{-1/2})$ and $\|\theta_{2k+2}(x) - \sin(c_{2k+2}x)\|_{L^2(D)} = \mathcal{O}(k^{-1/2})$.

Proof. For c_{2k+1}, from the relation $\tan(c_{2k+1}) = \coth(c_{2k+1}) = 1 + 2/(e^{2c_{2k+1}} - 1)$, we obtain that

$$c_{2k+1} - \left(k + \frac{1}{4}\right)\pi < \tan\left(c_{2k+1} - \left(k + \frac{1}{4}\right)\pi\right) = \frac{\tan(c_{2k+1}) - 1}{1 + \tan(c_{2k+1})} = e^{-2c_{2k+1}}.$$

The first inequality uses $x < \tan x$ for $x \in (0, \frac{\pi}{4})$. The result for c_{2k+2} follows a similar derivation.

We only prove for θ_{2k+1}, the proof is similar for θ_{2k+2}.

$$\begin{aligned}
&\int_D \left(-\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\sinh(c_{2k+1}x) + \cos(c_{2k+1}x)\right)^2 \mathrm{d}x \\
&= \int_D \left(\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\sinh(c_{2k+1}x)\right)^2 \mathrm{d}x + \int_D \cos^2(c_{2k+1}x)\,\mathrm{d}x \\
&\quad - 2\int_D \frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\sinh(c_{2k+1}x)\cos(c_{2k+1}x)\,\mathrm{d}x \qquad \text{(B.2)} \\
&= \left(\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\right)^2 \left[\frac{\sinh(2c_{2k+1})}{2c_{2k+1}} - 1\right] + \left[\frac{\sin(2c_{2k+1})}{2c_{2k+1}} + 1\right] \\
&= \frac{\cos^2(c_{2k+1})\coth(c_{2k+1})}{c_{2k+1}} + \frac{\sin(2c_{2k+1})}{2c_{2k+1}} + 1 - \left(\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\right)^2 = 1 + \mathcal{O}(k^{-1}).
\end{aligned}$$

Therefore $C_{2k+1} = 1 + \mathcal{O}(k^{-1})$ and $|B_{2k+1}| \le \sinh(c_{2k+1})^{-1}|C_{2k+1}| = \mathcal{O}(e^{-c_{2k+1}})$. The norm $\|\theta_{2k+1}\|_{L^\infty(D)} \le 2C_{2k+1} = \mathcal{O}(1)$ and

$$\theta'_{2k+1}(x) = C_{2k+1}c_{2k+1}\left(-\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\cosh(c_{2k+1}x) - \sin(c_{2k+1}x)\right),$$

which means $\|\theta'_{2k+1}\|_{L^\infty(D)} = \mathcal{O}(k)$ and similarly, $\|\theta''_{2k+1}\|_{L^\infty(D)} = \mathcal{O}(k^2)$. We also can see from (B.2) that

$$\int_D |\theta_{2k+1}(x) - C_{2k+1}\cos(c_{2k+1}x)|^2\,\mathrm{d}x = \mathcal{O}(k^{-1}),$$

which means $\theta_k \sim C_{2k+1}\cos(c_{2k+1}x) + \mathcal{O}(k^{-1/2}) = \cos(c_{2k+1}x) + \mathcal{O}(k^{-1/2})$. □

C. Improved bounds for learning dynamics

C.1 *Part I*

In the first step, we provide an improved estimate for

$$\int_D \partial_b w(b', t)\mu_2(b', t)\theta'_m(b')\,\mathrm{d}b' = \frac{1}{n}\sum_{i=1}^n \partial_b w(b_i(t), t)|a_i(t)|^2\theta'_m(b_i(t)).$$

Here we slightly abuse the notation and treat $\partial_b w$ as zero outside D. Define the piecewise linear continuous function $\mathcal{A}(\cdot, t) \in C(D)$ that $\mathcal{A}(b_i(t), t) = |a_i(t)|^2$, then we have the following equation

using integration by parts,

$$\frac{1}{n}\sum_{i=1}^{n}\partial_b w(b_i(t),t)|a_i(t)|^2\theta_m'(b_i(t)) - \int_D \partial_b w(b,t)\mathcal{A}(b,t)\theta_m'(b)\,\mathrm{d}b$$
$$= -\int_D \texttt{disc}(b,t)\partial_b\left[\partial_b w(b,t)\mathcal{A}(b,t)\theta_m'(b)\right]\mathrm{d}b,$$

where `disc` is the discrepancy function

$$\texttt{disc}(b,t) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[-1,b)}(b_i(t)) - b$$

and $\mathbb{1}_S$ denotes the characteristic function on the set S.

LEMMA C1. Let $\mathcal{V}(t)$ be the total variation of $\partial_b w(b,t)\mathcal{A}(b,t)\theta_m'(b)$, then there exists an absolute constant $C > 0$ that

$$\mathcal{V}(t) \le CV(\mathcal{A})m^2\|h(\cdot,t) - f(\cdot)\|_{L^2(D)},$$

where $V(\mathcal{A})$ is the total variation of $\mathcal{A}$:

$$V(\mathcal{A}) = a_1^2(t) + \sum_{i=1}^{n-1}|a_i^2(t) - a_{i+1}^2(t)| + a_n^2(t). \tag{C.1}$$

Proof. The result comes from the fact $\partial_b^2 w(b,t) = h(b,t) - f(b)$, $|\theta_m''(b)| = O(m^2)$, and bounded variation functions form a Banach algebra. □

LEMMA C2. Suppose $\sup_{1\le i\le n}|a_i(t)|^2$ is uniformly bounded by M, the total variation of $\mathcal{A}$ is bounded by M', and the initial biases are equispaced distributed, then there exists a constant $C > 0$ such that

$$\left|\int_D \partial_b w(b',t)\mu_2(b',t)\theta_m'(b')\,\mathrm{d}b'\right| \le C\|h(\cdot,0) - f(\cdot)\|_{L^2(D)}\left(1 + \left(\frac{2}{n} + Kt\right)m^2\right).$$

The constant $K = \sqrt{2M}\|h(x,0) - f(x)\|_{L^2(D)}$.

Proof. Let $L = \|\theta_m\|_{C(D)} = \mathcal{O}(1)$, we apply integration by parts and there exists a constant $C' > 0$ such that

$$\begin{aligned}
&\left|\int_D \partial_b w(b,t)\mathcal{A}(b,t)\theta_m'(b)\,\mathrm{d}b\right| \\
&\le \left|\int_D \theta_m(b)\partial_b\left(\partial_b w(b,t)\mathcal{A}(b,t)\right)\,\mathrm{d}b\right| + \left|\theta_m(b)\partial_b w(b,t)\mathcal{A}(b,t)\Big|_{\partial D}\right| \\
&\le L\left[\int_D |\partial_b^2 w(b,t)\mathcal{A}(b,t)|\,\mathrm{d}b + \int_D |\partial_b w(b,t)||\partial_b\mathcal{A}(b,t)|\,\mathrm{d}b + M\|\partial_b w(\cdot,t)\|_{C(D)}\right] \\
&\le L\left[\|\partial_b^2 w(\cdot,t)\|_{L^2(D)}\|\mathcal{A}(\cdot,t)\|_{L^2(D)} + \|\partial_b w(\cdot,t)\|_{C(D)}V(\mathcal{A}) + M\|\partial_b w(\cdot,t)\|_{C(D)}\right] \\
&\le C'\|h(\cdot,t) - f(\cdot)\|_{L^2(D)}(M + M').
\end{aligned} \tag{C.2}$$

Therefore we have a new estimate:

$$\left|\int_D \partial_b w(b,t)\mu_2(b,t)\theta_m'(b)\,\mathrm{d}b\right| \le C'\|h(\cdot,t)-f(\cdot)\|_{L^2(D)}(M+M') + \left|\int_D \mathtt{disc}(b,t)\partial_b\left[\partial_b w(b,t)\mathcal{A}(b,t)\theta_m'(b)\right]\mathrm{d}b\right|.$$

Notice that this bound will be better than the constant bound in Theorem 3.8 if the second term on the right-hand side is relatively small. Next, we characterize the discrepancy function $\mathtt{disc}(b,t)$.

$$\begin{aligned}|\mathtt{disc}(b,t)-\mathtt{disc}(b,0)| &= \left|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[-1,b)}(b_i(t))-b-\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[-1,b)}(b_i(0))-b\right)\right| \\ &= \frac{1}{n}\left|\sum_{i=1}^n\left(\mathbb{1}_{[-1,b)}(b_i(t))-\mathbb{1}_{[-1,b)}(b_i(0))\right)\right|.\end{aligned}$$

Because

$$\begin{aligned}\left|\frac{d}{dt}b_i(t)\right| &\le |a_i(t)|\int_D |h(x,t)-f(x)||\partial_b\psi(x,b_i(t))|\,\mathrm{d}x \\ &\le K := C'''\sqrt{M}\|h(x,0)-f(x)\|_{L^2(D)},\end{aligned} \tag{C.3}$$

and within time t, the maximum distance of propagation is Kt for each bias. Therefore,

$$\begin{aligned}\frac{1}{n}\left|\sum_{i=1}^n\left(\mathbb{1}_{[-1,b)}(b_i(t))-\mathbb{1}_{[-1,b)}(b_i(0))\right)\right| &\le \frac{1}{n}\left|\sum_{i=1}^n\left(\mathbb{1}_{[-1+Kt,b-Kt)}(b_i(0))-\mathbb{1}_{[-1,b)}(b_i(0))\right)\right| \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[-1,-1+Kt)\cup[b-Kt,b)}(b_i(0)) \le Kt+\frac{1}{n}.\end{aligned}$$

Therefore, using $|\mathtt{disc}(b,0)| \le \frac{1}{n}$ for equispaced biases,

$$\begin{aligned}\left|\int_D \partial_b w(b,t)\mu_2(b,t)\theta_m'(b)\,\mathrm{d}b\right| &\le C'\|h(x,t)-f(x)\|_{L^2(D)}(M+M')+\left(\frac{2}{n}+Kt\right)\mathcal{V}(t) \\ &\le \|h(x,0)-f\|_{L^2(D)}(M+M')\left(C'+\left(\frac{2}{n}+Kt\right)C''m^2\right).\end{aligned}$$

□

REMARK C3. If the bound of $V(\mathcal{A})$ can be as large as $\mathcal{O}(nM)$ at the worst case, then we may return to the Theorem 3.8. In fact, if $V(\mathcal{A})$ becomes $\mathcal{O}(m)$, we may also have to return to the Theorem 3.8.

C.2 *Part II*

In this part, we try to optimize the bound for $\mathcal{J}_m(t)-\mathcal{J}_m(0)$:

$$\mathcal{J}_m(t)-\mathcal{J}_m(0)=\theta_m'(-1)h(-1,t)=\theta_m'(-1)\sum_{i=1}^{s-1}a_i(t)\chi(b_i(t))(-1-b_i(t)).$$

The main result is the following estimate.

THEOREM C4. Assume that $\sup_{1\le i=1\le n}|a_i(t)|^2$ is uniformly bounded by $M>0$ and the biases $\{b_i(0)\}_{i=1}^n$ are initially equispaced on D, then there exists a constant $C>0$ that

$$|\mathcal{J}_m(t)-\mathcal{J}_m(0)|\le Cm\sqrt{M}K^2nt^2.$$

Proof. The biases are propagating with finite speed that $|\frac{d}{dt}b_i(t)|\le K:=C'''\sqrt{M}\|h(\cdot,0)-f(\cdot)\|_{L^2(D)}$ (see (C.3)). If the initial biases are *equispaced* distributed on D, then $s(t)-1\le\frac{Knt}{2}$. For each $1\le i\le s(t)-1$, the bias $b_i(t)$ satisfies

$$|b_i(t)-b_i(0)|\le Kt,\quad \text{and}\quad |b_i(0)-(-1)|\le Kt.$$

Therefore

$$\begin{aligned}\sum_{i=1}^{s(t)-1}\chi(b_i(t))|-1-b_i(t)|&\le\sum_{i=1}^{s(t)-1}|-1-(b_i(0)-Kt)|\\&\le\sum_{i=1}^{s(t)-1}|-1-b_i(0)|+Kt(s(t)-1)\\&\le K^2nt^2.\end{aligned}$$

Therefore $|\mathcal{J}_m-\mathcal{J}_m(0)|\le|\theta_m'(-1)|\sqrt{M}K^2nt^2\le Cm\sqrt{M}K^2nt^2$. □

C.3 *Part III*

Now we can prove the following theorem with the improved bounds. In the following, we assume $n\ge m$. In other words, it only makes sense to study the learning dynamics for those frequency modes which can be resolved by the grid resolution corresponding to the network width.

THEOREM C5. Suppose $\sup_{1\le i\le n}|a_i(t)|^2$ is uniformly bounded by M, the total variation of $\mathcal{A}$ (C.1) is bounded by M', and the initial biases are equally spaced. Let $n\ge m^4$ be sufficiently large, then it takes at least $\mathcal{O}(\frac{m^4|\widehat{w}(m,0)|}{n})$ to reduce the initial error in generalized Fourier mode $|\widehat{w}(m,t)|\le\frac{1}{2}|\widehat{w}(m,0)|$. Especially when $|\widehat{w}(m,0)|>c'm^{-2}$, the half-reduction time is at least $\mathcal{O}(\frac{m^2c'}{n})$.

Proof. Using Lemma C2, there exists a constant $C''>0$ that

$$\left|\int_D\partial_b w(b',t)\mu_2(b',t)\theta_m'(b')\,\mathrm{d}b'\right|\le C''\left(1+\left(\frac{1}{n}+t\right)m^2\right).$$

We only consider the case that $n\ge m$, otherwise we return to Theorem 38. Recall that in Theorem 3.8 that $|\widehat{w\mu_0}(m,t)|$ is uniformly bounded (see (3.16)) and $|\mathcal{H}_m(t)-\mathcal{H}_m(0)|\le\frac{1}{2}|\theta_m(-1)|\sqrt{M}Knt$ (see (3.17) and (3.18)). Combine the estimate in Theorem C4 and follow the same process as (3.19), we can find a constant $C'''>0$ that

$$|\widehat{w}(m,t)|-|\widehat{w}(m,0)|\le-C'''\frac{n}{p_m}\left(\left(1+\frac{m^2}{n}\right)t+(m+m^2)t^2\right).\tag{C.4}$$

We solve an upper bound for the half-reduction time τ from the quadratic equation

$$\frac{1}{2}|\widehat{w}(m,0)| = \frac{n}{p_m}C'''\left(\left(1+\frac{m^2}{n}\right)\tau + (m+m^2)\tau^2\right). \tag{C.5}$$

The solution satisfies

$$\tau \geq \frac{p_m}{2C'''n}\frac{|\widehat{w}(m,0)|}{\sqrt{(1+m^2/n)^2 + 2(m+m^2)p_m|\widehat{w}(m,0)|/(C'''n)}}. \tag{C.6}$$

Let's use the notations $A \prec B$ and $A \succ B$ to represent the relation $A = \mathcal{O}(B)$ and $B = \mathcal{O}(A)$, respective. We find the following regimes:

1. $n \succ m^2$ and $|\widehat{w}(m,0)| \succ \frac{n}{m^6}$, then $\tau = \mathcal{O}\left(m\sqrt{\frac{\widehat{w}(m,0)}{n}}\right)$.
2. $n \succ m^2$ and $|\widehat{w}(m,0)| \prec \frac{n}{m^6}$, then $\tau = \mathcal{O}\left(\frac{m^4}{n}|\widehat{w}(m,0)|\right)$.
3. $m \prec n \prec m^2$ and $|\widehat{w}(m,0)| \succ \frac{1}{nm^2}$, then $\tau = \mathcal{O}\left(m\sqrt{\frac{\widehat{w}(m,0)}{n}}\right)$.
4. $m \prec n \prec m^2$ and $|\widehat{w}(m,0)| \prec \frac{1}{nm^2}$, then $\tau = \mathcal{O}(m^2|\widehat{w}(m,0)|)$.

In the second case, when $n \succ m^4$ is sufficiently large and $|\widehat{w}(m,0)| > c'm^{-2}$, the half-reduction time becomes $\mathcal{O}(\frac{c'm^2}{n})$, which is a better bound than the one in Theorem 3.8. □

REMARK C6. If the biases b_i are equally spaced and fixed, the learning dynamics become the gradient flow for the least square problem. Using a standard Fourier basis one gets a simpler version of (3.14) without the boundary terms and the last term involving θ'_m. Since $w(b,t)$ is H^2, one gets $\widehat{\mu_0 w}(m,t) \leq C(\frac{1}{m^2} + m\text{disc}(\{b_i(0)\}_{i=1}^n))$, for $n \succ m^3$ and equispaced $\{b_i\}_{i=1}^n$ that $\text{disc}(\{b_i(0)\}_{i=1}^n) = \mathcal{O}(\frac{1}{n})$, it takes $\mathcal{O}(m^4)$ steps to reduce the initial error in mode m by half. Hence the full learning dynamics, i.e. involving the bias, while requiring more computation cost in each step, may speed up the convergence. See Fig. C7 for an example.

C.4 *Numerical experiments*

In the following, we perform numerical experiments to demonstrate the scaling laws with a different total variation of $|a_i(t)|^2$. The objective function is

$$f(x) = \sin(k\pi x)$$

with k chosen from selected high frequencies. We set the number of neurons $n = k^\beta$, $\beta \in \{2,3,4\}$ for a selected frequency k. The learning rate is selected as n^{-1}. This set up is regime 1 above, where $\widehat{w}(k,0) = k^{-2}$, and hence the number of iterations should be of order $\mathcal{O}(n\tau) = \mathcal{O}(\sqrt{n})$.

The initialization of biases $\{b_i(0)\}_{i=1}^n$ are equispaced and ordered ascendingly. The weights $\{a_i(0)\}_{i=1}^n$ are initialized in the following ways.

(A) $a_i(0) = \frac{1}{2}(-1)^i$.

(B) $a_i(0) = \frac{1}{2}\cos(i)$.

TABLE C1 *Theoretical lower bounds*

	$TV(\mathcal{A})$	$\beta = 2$	$\beta = 3$	$\beta = 4$
Init A	$\mathcal{O}(1)$	$\mathcal{O}(k)$	$\mathcal{O}(k^{1.5})$	$\mathcal{O}(k^2)$
Init B	$\mathcal{O}(n)$	$\mathcal{O}(k)$	$\mathcal{O}(k)$	$\mathcal{O}(k)$

TABLE C2 *Experimental fitted results*

	$TV(\mathcal{A})$	$\beta = 2$	$\beta = 3$	$\beta = 4$
Init A	$\mathcal{O}(1)$	$\mathcal{O}(k^{1.61})$	$\mathcal{O}(k^{1.82})$	$\mathcal{O}(k^{1.87})$
Init B	$\mathcal{O}(n)$	$\mathcal{O}(k^{1.59})$	$\mathcal{O}(k^{1.75})$	$\mathcal{O}(k^{1.90})$

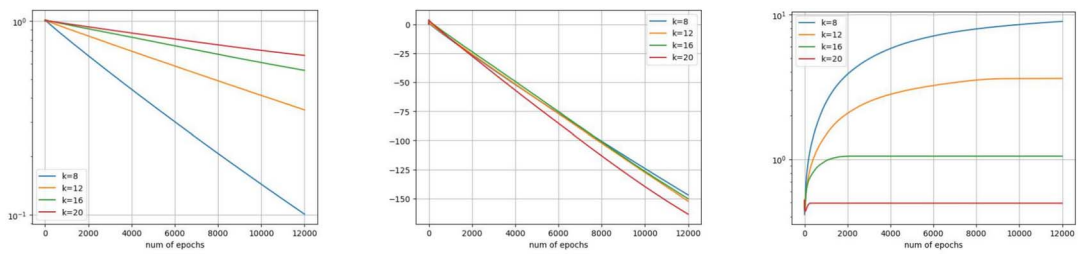


FIG. C1. Experiment for initialization (A) and $\beta = 4$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.87}\ln(E_k(t))$. Right: the total variations of $\mathcal{A}$.

We record the dynamics of the network (denoted by h_k) at exactly frequency k through the projection

$$E_k(t) = \left| \int_D (h_k(x,t) - f(x)) f(x)\, \mathrm{d}x \right|.$$

As we will see in the experiments, the total variation in $\mathcal{A}$ is slowly varying in time, so we can summarize the *a priori* theoretical lower bounds and the experimental results of the number of epochs in the following tables C1 and C2. The initialization (A) has a relatively small total variation, and the experiments agree with the theoretical bounds quite well. However, similar results are observed for initialization (B), which has a relatively large total variation during the training, see Fig. C1, C2, C3 for initialization (A) and Fig. C4, C5, C6 for initialization (B). One possible explanation is the overestimate using the total variation in (C2) which might be unnecessary. As we mentioned in Remark C6, we record the training dynamics with *fixed* biases, initialization (A), and choose $n = k^4$, see Fig. C7. The result matches the argument in Remark C6. For comparison purposes, we additionally demonstrate an example with `Adam` optimizer, which shows a similar scaling relation as the `GD` optimizer at the initial training stage, see Fig. C7.

All these tests show a consistent phenomenon, the higher the frequency, the slower the learning dynamics.

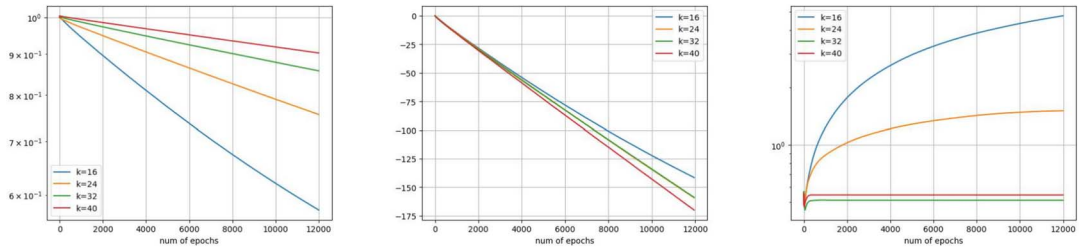


FIG. C2. Experiment for initialization (A) and $\beta = 3$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.82} \ln(E_k(t))$. Right: the total variations of $\mathcal{A}$.

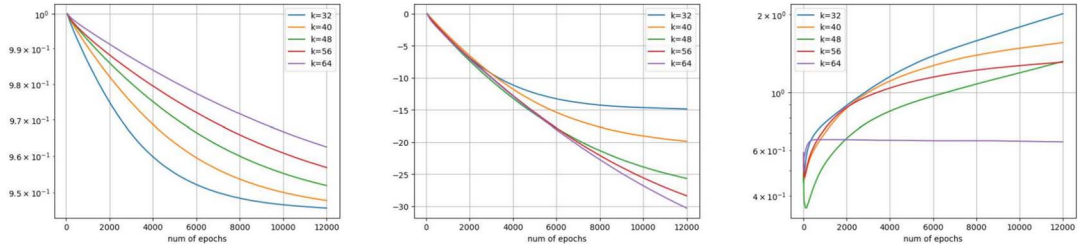


FIG. C3. Experiment for initialization (A) and $\beta = 2$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.61} \ln(E_k(t))$ (fitting first 2000 epochs only). Right: the total variations of $\mathcal{A}$.

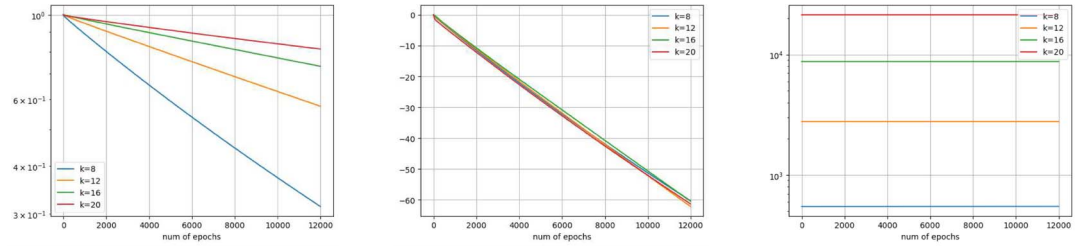


FIG. C4. Experiment for initialization (B) and $\beta = 4$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.90} \ln(E_k(t))$ (fitting first 2000 epochs only). Right: the total variations of $\mathcal{A}$.

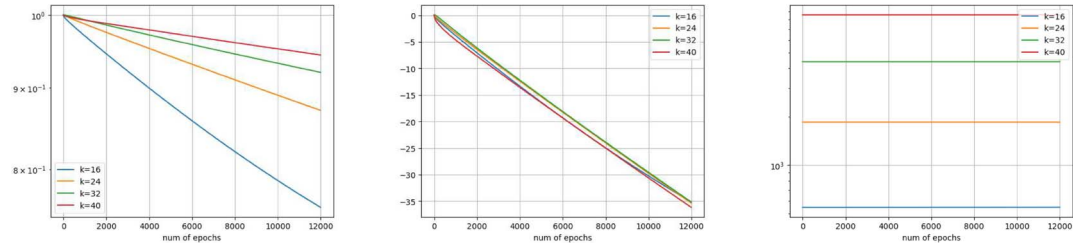


FIG. C5. Experiment for initialization (B) and $\beta = 3$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.75} \ln(E_k(t))$. Right: the total variations of $\mathcal{A}$.

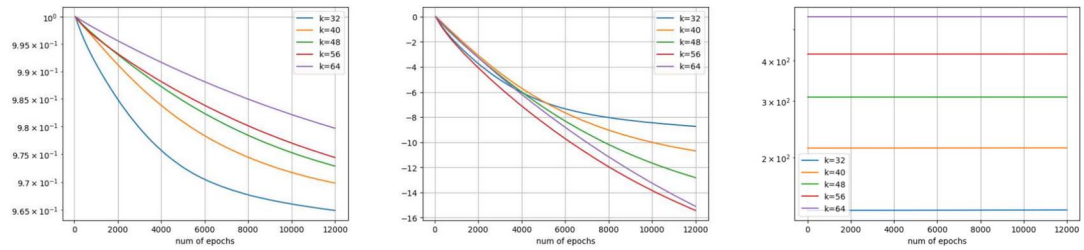


FIG. C6. Experiment for initialization (B) and $\beta = 2$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.59}\ln(E_k(t))$ (fitting first 2000 epochs only). Right: the total variations of $\mathcal{A}$.

C.5 *Further remarks*

C.5.1 *Initial distribution of biases.* If the initial biases are uniformly distributed instead of equispaced, the previous estimates need to be modified. In particular, the upper bound of $s(t) - 1$ will become $\frac{Knt}{2} + \mathcal{O}_p(\sqrt{Knt})$ using the Chebyshev inequality. The discrepancy of $\{b_i(0)\}_{i=1}^n$ will be also updated to $\mathcal{O}_p(n^{-1/2})$. Then we have the following modified estimates:

$$|\mathcal{H}_m(t) - \mathcal{H}_m(0)| \le \frac{1}{2}|\theta_m(-1)|\sqrt{M}(nKt + \mathcal{O}_p(\sqrt{nKt})),$$

$$|\mathcal{J}_m(t) - \mathcal{J}_m(0)| \le |\theta'_m(-1)|\sqrt{M}Kt(Knt + \mathcal{O}_p(\sqrt{Knt})),$$

$$\left|\int_D \partial_b w(b',t)\mu_2(b',t)\theta'_m(b')\,\mathrm{d}b'\right| \le C''\left(1 + \left(t + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right)\right)m^2\right),$$

and (C.4) becomes

$$|\widehat{w}(m,t)| - |\widehat{w}(m,0)| \underset{p}{\le} -C'''\frac{n}{p_m}\left(\left(1 + \frac{m^2}{\sqrt{n}}\right)t + (m+m^2)t^2 + (1+mt)\sqrt{\frac{t}{n}}\right).$$

In particular, when $n \succ m^4$ and $|\widehat{w}(m,0)| > c'm^{-2}$, the half-reduction time is still the same $\mathcal{O}_p(m^2c'/n)$ as Theorem C5. A similar probabilistic estimate can be derived for initial biases sampled from a continuous probability density function which is bounded from below and above by positive constants.

C.5.2 *Activation function.* The regularity of the activation function plays a crucial role in the analysis. In general, using a smoother activation function, which leads to a faster spectrum decay of the corresponding Gram matrix, will take an even longer time to eliminate higher frequencies. For instance, if the activation function is chosen as $\frac{1}{p!}\sigma^p(x)$, $p \ge 1$, where σ is the `ReLU` activation function, then a similar analysis will show that if $n \succ m^{p+3}$, then under the assumption of Theorem C5, the *half-reduction* time of 'frequency' m is at least $\mathcal{O}(\frac{m^{2p+2}}{n}|\widehat{w}(m,0)|)$ in the Theorem C5. This also implies that using a shallow neuron network with smooth activation functions will be even worse for learning and approximating high-frequency information by minimizing the L^2 error or mean-squared error (`MSE`).

In the following, we perform a simple numerical experiment to validate our conclusion. We set the objective function as the Fourier mode $f(x) = \sin(m\pi x)$ on D, $m \in \mathbb{N}$ and use the activation functions $\texttt{ReLU}^p(x) := \frac{1}{p!}\sigma^p(x)$, $p = 1, 2$ to train the shallow neuron network $h_{m,p}(x,t)$ to approximate $f(x)$,

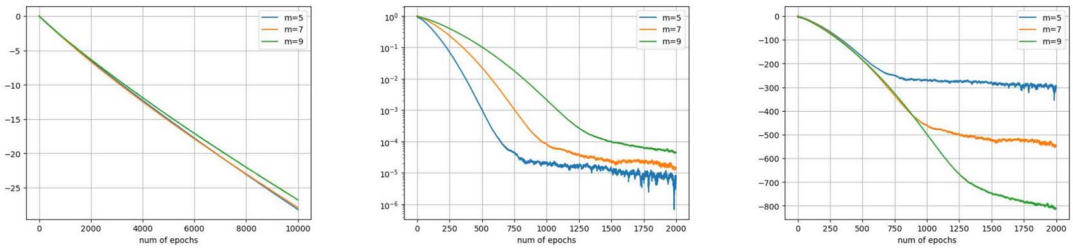


FIG. C7. Additional experiment for initialization (A) and $\beta = 4$. Left: Graphs of $k^{3.5}\ln(E_k(t))$ for $k = 5, 7, 9$ with *fixed* equispaced biases, trained by GD. Middle: Graphs of $E_k(t)$ for $k = 5, 7, 9$, trained by Adam. Right: Graphs of $k^2 \ln E_k(t)$ for $k = 5, 7, 9$, trained by Adam.

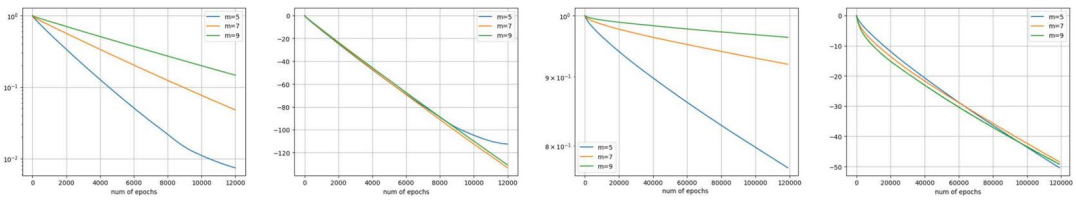


FIG. C8. The comparison of ReLU, ReLU^2 for the approximation to $f(x) = \sin(m\pi x)$. From left to right, the figures are $E_{m,1}(t)$, $m^{1.95}\ln(E_{m,1}(t))$, $E_{m,2}(t)$, $m^{3.25}\ln(E_{m,2}(t))$.

respectively. The number of neurons $n = 10^4$. We record the *error* of the Fourier mode

$$E_{m,p}(t) := \left| \int_D (h(x,t) - f(x)) f(x)\, \mathrm{d}x \right|$$

at each iteration. The errors are shown in Fig. C8 for $m = 5, 7, 9$. We can observe that the decay rates of ReLU and ReLU^2 are $\mathcal{O}(m^{-2})$ and $\mathcal{O}(m^{-3})$ roughly.

C.5.3 *Boundedness of weights.* One may notice that the requirement of weights $\sup_{i\geq 1} |a_i(t)|^2 \leq M$ for all $t > 0$ can be relaxed to $0 \leq t \leq \tau$, where τ denotes the lower bound of half-reduction time in Theorem 3.8 or Theorem C5. When $n \succ m$ in Theorem 38 or $n \succ m^2$ in Theorem C5, such requirement can be relaxed to only the initial condition $\sup_{i\geq 1} |a_i(0)|^2 \leq M$.

D. Rashomon set for bounded activation function

In this section, we characterize the Rashomon set for a general bounded activation function instead of ReLU. We consider a more general setting of the parameter space: a_i are mean-zero i.i.d sub-Gaussian random variables that

$$\mathbb{P}[|a_i| > t] < 2e^{-mt^2}, \quad i \in [n]$$

for some $m > 0$. Then we have the following estimate for the Rashomon set by following a similar idea for the proof of Theorem 41.

THEOREM D1. Assuming the same network structure as Theorem 41 and σ as a bounded activation function that

$$-1 \le \sigma(t) \le 1, \quad \sigma'(t) > 0, \quad \forall t \in \mathbb{R}, \tag{D.1}$$

then the Rashomon set's measure

$$\mathbb{P}[\|h(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L^2(D)} \le \varepsilon \|f\|_{L^2(D)}] \le 2\exp\left(-\frac{Cn(1-\varepsilon)^2\|f\|_{L^2(D)}^4}{4\kappa^2\ell^2\theta^2}\right),$$

where $\theta = \|a_i\|_{\psi_2}$ is the *Orlicz* norm, $\ell = \mathrm{diam}(D)$, and κ denotes

$$\kappa := \sup_{(\boldsymbol{w},t)\in\mathbb{S}^{d-1}\times\mathbb{R}} \left| \int_{\{\boldsymbol{x}\cdot\boldsymbol{w}=t, \boldsymbol{x}\in D\}} f(\boldsymbol{x}) dH_{d-1}\boldsymbol{x} \right|.$$

Proof. We denote $r_i := |\boldsymbol{w}_i|$ and $\ell = \mathrm{diam}(D)$, then let

$$X_i := a_i \int_D f(\boldsymbol{x})\sigma(\boldsymbol{w}_i\cdot\boldsymbol{x} - b_i)d\boldsymbol{x} = \frac{a_i}{r_i}\int_{-\ell r_i}^{\ell r_i} \sigma(s-b_i)\int_{\{\boldsymbol{x}\cdot\boldsymbol{w}_i=s\}} f(\boldsymbol{x})dH_{d-1}(\boldsymbol{x})ds.$$

Then $\langle h, f\rangle = \frac{1}{n}\sum_{i=1}^n X_i$ and

$$\mathbb{P}\left[\|h-f\|_{L^2(D)}^2 \le \varepsilon^2\|f\|_{L^2(D)}^2\right] \le \mathbb{P}\left[(1-\varepsilon)\|f\|_{L^2(D)}^2 \le \frac{1}{n}\sum_{i=1}^n X_i\right].$$

Since the random variable a_i is sub-Gaussian, then X_i is also sub-Gaussian by $|X_i| \le 2a_i\ell\kappa$. One can apply the Hoeffding's inequality (see Theorem 2.6.2 [62]) that

$$\mathbb{P}\left[(1-\varepsilon)\|f\|_{L^2(D)}^2 \le \frac{1}{n}\sum_{i=1}^n X_i\right] \le 2\exp\left(\frac{-Cn(1-\varepsilon)^2\|f\|_{L^2(D)}^4}{4\kappa^2\ell^2\theta^2}\right).$$

for certain absolute constant C. □

The constant κ stands for the largest possible average of f on every hyperplane $\{\boldsymbol{x}\cdot\boldsymbol{w} = t\}$, $t \in \mathbb{R}$. When $f(\boldsymbol{x})$ is oscillatory in all directions, the constant κ becomes small. More intuitively speaking, an activation function of the form (D.1) can not feel oscillations in f, i.e. $\langle f, \sigma\rangle$ is small due to cancellation. According to the heuristic argument in [50], it also implies that it is relatively difficult to find a shallow neural network with activation function (D.1) that can approximate highly oscillatory functions well. In other words, the optimal set of parameters only occupies an extremely small measure of the parameter space for highly oscillatory functions.

E. Further discussions

E.1 *General case of Gram matrix of two-layer* ReLU *networks in one dimension*

When the two-layer ReLU network in one dimension is

$$f(x) = c + \sum_{i=1}^{n} a_i \sigma(w_i x - b_i), \quad x \in D := [-1, 1],$$

where $w_i \in \{+1, -1\}$ obeys the Bernoulli distribution with $p = \frac{1}{2}$. Then the corresponding Gram matrix has the following block structure of continuous kernels

$$\mathcal{G} := \begin{pmatrix} \mathcal{G}^{++}(x,y) & \mathcal{G}^{+-}(x,y) \\ \mathcal{G}^{+-}(x,y) & \mathcal{G}^{--}(x,y) \end{pmatrix}$$

where the sub-kernels are

$$\mathcal{G}^{++}(x,y) = \frac{1}{24}(2 - x - y - |x-y|)^2(2 - x - y + 2|x-y|),$$

$$\mathcal{G}^{--}(x,y) = \frac{1}{24}(2 + x + y - |x-y|)^2(2 + x + y + 2|x-y|),$$

$$\mathcal{G}^{+-}(x,y) = \frac{1}{48}\left[|x-y| - (x-y)\right]^3,$$

$$\mathcal{G}^{-+}(x,y) = \frac{1}{48}\left[|x-y| + (x-y)\right]^3.$$

Note the kernels $\mathcal{G}^{+-}(x,y) = \mathcal{G}^{-+}(-x,-y)$ and $\mathcal{G}^{++}(x,y) = \mathcal{G}^{--}(-x,-y)$, suppose (g_k^+, g_k^-) is an eigenfunction of $\mathcal{G}$ for eigenvalue λ_k, we can obtain:

- If $\phi_k(x) = g_k^+(x) + g_k^-(-x) \not\equiv 0$, then it is an eigenfunction of $\mathcal{G}_\phi = \mathcal{G}^{++}(x,y) + \mathcal{G}^{+-}(x,-y)$ for eigenvalue λ_k.
- If $\psi_k(x) = g_k^+(x) - g_k^-(-x) \not\equiv 0$, then it is an eigenfunction of $\mathcal{G}_\psi = \mathcal{G}^{++}(x,y) - \mathcal{G}^{+-}(x,-y)$ for eigenvalue λ_k.

The kernel $\mathcal{G}_\phi \in C^2(D \times D)$ is

$$\mathcal{G}_\phi(x,y) = \frac{1}{12}(|x-y|^3 + |x+y|^3 + 4 + 12xy - 6(x+y) - 6xy(x+y)).$$

Based on the above observation, it is straightforward to derive the following theorem.

THEOREM E1. If the two kernels $\mathcal{G}_\phi$ and $\mathcal{G}_\psi$ do not allow common eigenvalues, then (g_k^+, g_k^-) is an eigenfuntion of $\mathcal{G}$, then they satisfy either $g_k^+(x) = g_k^-(x)$ or $g_k^+(x) = g_k^-(-x)$.

REMARK E2. The kernel $K_\alpha = |x-y|^\alpha$ has been studied in [6] for $\alpha = 1$, where the eigenvalues have a leading positive term and all of the rest eigenvalues are negative and decay as $\frac{c}{(2k+1)^2}$. Consider the eigenvalue for $|x-y|^3$, we need to find

$$\lambda h(x) = \int_{-1}^{x} (x-y)^3 h(y)\,\mathrm{d}y + \int_{x}^{1} (y-x)^3 h(y)\,\mathrm{d}y$$

by differentiating the above equation 4 times, we get $f^{(4)} = \frac{12}{\lambda}h(x)$, let $\omega^4 = \frac{12}{\lambda}$, then the solution consists of the basis

$$\sum_{k=0}^{3} A_k \exp(\omega e^{\frac{2\pi ik}{4}} x).$$

Thus solving the eigensystem is equivalent to solving a 4×4 matrix $\det(M) = 0$ for ω. Similar arguments hold for the Hankel kernel $|x+y|^3$, they share the same basis. The exact values are quite expensive to compute.

Now we apply the same idea for $\mathcal{G}_\phi$, by the same differentiation technique used in deriving (2.6), we arrive at the same form:

$$\phi_k(x) = \sum_{l=0}^{3} c_l \exp(\omega e^{\frac{2\pi il}{4}} x),$$

where $\omega^4 = \frac{2}{\lambda}$, here $\lambda > 0$, thus we choose $\omega \in \mathbb{R}^+$ and the basis are more explicit:

$$\phi_k(x) = c_0 \cosh(\omega x) + c_1 \sinh(\omega x) + c_2 \cos(\omega x) + c_3 \sin(\omega x).$$

The eigenvalues λ_k can be computed in a similar way as in Theorem B4 and there are constants $c_1, c_2 > 0$ that $c_1 k^{-4} \le \lambda_k \le c_2 k^{-4}$.

E.2 *Leaky* `ReLU` *activation function*

For the leaky `ReLU` activation function with parameter $\alpha \in (0,1)$, $\sigma_\alpha(x) = \sigma(x) - \alpha\sigma(-x)$, we can derive the Gram matrix G_α:

$$\begin{aligned} G_{\alpha,ij} &= \int_{-1}^{1} \sigma_\alpha(x-b_i)\sigma_\alpha(x-b_i)\,\mathrm{d}x \\ &= \int_{-1}^{1} (\sigma(x-b_i) - \alpha\sigma(-x+b_i))(\sigma(x-b_j) - \alpha\sigma(-x+b_j))\,\mathrm{d}x \\ &= \mathcal{G}(b_i,b_j) + \alpha^2\mathcal{G}(-b_i,-b_j) - \frac{\alpha}{6}|b_i-b_j|^3. \end{aligned}$$

Then we can derive the following estimate for the eigenvalue for G_α. Let the kernel $\mathcal{G}_\alpha(x,y) := \mathcal{G}(x,y) + \alpha^2\mathcal{G}(-x,-y) - \frac{\alpha}{6}|x-y|^3$.

THEOREM E3. Suppose b_i are quasi-evenly spaced on $[-1,1]$, $b_i = -1 + \frac{2(i-1)}{n} + o\left(\frac{1}{n}\right)$. Let $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0$ be the eigenvalues of the Gram matrix G_α then $|\lambda_k - \frac{n}{2}\mu_{\alpha,k}| \le C$ for some constant $C = \mathcal{O}(1)$, where $\mu_{\alpha,k} = \mathcal{O}(\frac{(\alpha-1)^2}{k^4})$ is the kth eigenvalue of $\mathcal{G}_\alpha$.

Proof. Let the kernel $\mathcal{G}_\alpha(x,y) := \mathcal{G}(x,y) + \alpha^2\mathcal{G}(-x,-y) - \frac{\alpha}{6}|x-y|^3$, then using the same differentiation technique in deriving (2.6), if $\psi_{\alpha,k}$ is an eigenfunction of $\mathcal{G}_\alpha$ for the eigenvalue $\mu_{k,\alpha}$, we have

$$\psi^{(4)}_{\alpha,k} = \frac{(\alpha-1)^2}{\mu_k}\psi_{\alpha,k}.$$

Let $w_{\alpha,k} = \sqrt{|1-\alpha|}\mu_k^{-\frac{1}{4}}$, then equivalently we obtain the following equation for $w_{\alpha,k}$:

$$\begin{aligned}&\left(P_{0,\alpha}(w_{\alpha,k}) + P_{1,\alpha}(w_{\alpha,k})\cos(2w_{\alpha,k}) + P_{2,\alpha}(w_{\alpha,k})\sin(2w_{\alpha,k})\right)\\ &+ \tanh(w_{\alpha,k})\left(Q_{0,\alpha}(w_{\alpha,k}) + Q_{1,\alpha}(w_{\alpha,k})\cos(2w_{\alpha,k}) + Q_{2,\alpha}(w_{\alpha,k})\sin(2w_{\alpha,k})\right)\\ &+ \tanh^2(w_{\alpha,k})\left(R_{0,\alpha}(w_{\alpha,k}) + R_{1,\alpha}(w_{\alpha,k})\cos(2w_{\alpha,k}) + R_{2,\alpha}(w_{\alpha,k})\sin(2w_{\alpha,k})\right) = 0,\end{aligned}$$

where $P_{i,\alpha}, Q_{i,\alpha}, R_{i,\alpha}$, $i = 0, 1, 2$ are polynomials of $w_{\alpha,k}$ of degree ≤ 4. Set $A_\alpha(x) = (-36\alpha^4 + 42\alpha^5 - 12\alpha^6)x^2$ and $B_\alpha(x) = (8\alpha^4 - 8\alpha^5 + 2\alpha^6)x^4$, then

$$\begin{aligned}P_{0,\alpha}(x) &= \frac{3}{2} + 6\alpha^2 - 6\alpha^3 + \frac{3}{2}\alpha^4 + A_\alpha(x) + B_\alpha(x),\\ P_{1,\alpha}(x) &= P_{0,\alpha}(x) - 3,\\ P_{2,\alpha}(x) &= -3\alpha^2(\alpha^2 - 3\alpha + 3)x + 2\alpha^2(\alpha-2)(2 - 9\alpha^2 + 6\alpha^3)x^3,\end{aligned}$$

$$\begin{aligned}Q_{0,\alpha}(x) &= 2\alpha^2 x(-18 + 15\alpha - 42\alpha^2 - 57\alpha^3 + 18\alpha^4 + (12\alpha^2 - 14\alpha^3 + 4\alpha^4)x^2),\\ Q_{1,\alpha}(x) &= 2\alpha^2 x(9 - 6\alpha + 45\alpha^2 - 18\alpha^4 + (8 + 4\alpha + 24\alpha^2 - 28\alpha^3 + 8\alpha^4)x^2),\\ Q_{2,\alpha}(x) &= -12\alpha^2 x^2(-4 + 3\alpha - 10\alpha^2 + 13\alpha^2 + 4\alpha^4),\end{aligned}$$

$$\begin{aligned}R_{0,\alpha}(x) &= \frac{1}{2}\left[3 - 24\alpha^2 + 12\alpha^3 + 111\alpha^4 - 144\alpha^5 + 48\alpha^6\right] - A_\alpha(x) + B_\alpha(x),\\ R_{1,\alpha}(x) &= \frac{1}{2}\left[-3 + 48\alpha^2 - 36\alpha^2 - 105\alpha^4 + 144\alpha^5 - 48\alpha^6\right] - A_\alpha(x) + B_\alpha(x),\\ R_{2,\alpha}(x) &= \alpha^2(27 - 21\alpha - 87\alpha^2 + 114\alpha^3 - 36\alpha^4)x + \alpha^2(8 - 4\alpha - 12\alpha^2 + 14\alpha^3 - 4\alpha^4)x^2.\end{aligned}$$

We can rewrite the equation as

$$Z_{0,\alpha}(w_{\alpha,k}) + Z_{1,\alpha}(w_{\alpha,k})\cos(2w_{\alpha,k}) + Z_{2,\alpha}(w_{\alpha,k})\sin(2w_{\alpha,k}) = 0,$$

where $Z_{0,\alpha} = P_{0,\alpha} + \tanh(w_{\alpha,k})Q_{0,\alpha} + \tanh^2(w_{\alpha,k})R_{0,\alpha}$, $Z_{1,\alpha} = P_{1,\alpha} + \tanh(w_{\alpha,k})Q_{1,\alpha} + \tanh^2(w_{\alpha,k})R_{1,\alpha}$ and $Z_{2,\alpha} = P_{2,\alpha} + \tanh(w_{\alpha,k})Q_{2,\alpha} + \tanh^2(w_{\alpha,k})R_{2,\alpha}$. It is not hard to show that as $w_{\alpha,k} > 0$, we have $0 < 1 - \tanh(w_{\alpha,k}) \leq 2e^{-2w_{\alpha,k}}$, and there exists a constant $c > 0$ that $\forall x > c$,

$$Z_{0,\alpha}(x) + Z_{1,\alpha}(x) > 0, \quad \text{and} \quad Z_{0,\alpha}(x) - Z_{1,\alpha}(x) < 0$$

by computing the sign of leading power in x, which implies that there exist roots on the intervals $[n\pi, (n+\frac{1}{2})\pi]$ and $[(n+\frac{1}{2})\pi, (n+1)\pi]$, respectively for sufficiently large n. □

The following corollary can be derived by using the Corollary 2.5. It shows that the kth eigenvalue grows as $\mathcal{O}((\alpha-1)^2 k^{-4})$ when k is sufficiently large.

COROLLARY E4. When $\{b_i\}_{i=1}^n$ are i.i.d uniformly distributed on $[-1, 1]$, then with probability $1-p$ that

$$|\lambda_k - \frac{n}{2}\mu_{\alpha,k}| = \begin{cases}\mathcal{O}\left(n^{\frac{5}{8}} i^{-3}\sqrt{\log\frac{n}{p}}\right), & i < n^{\frac{7}{8}},\\ \mathcal{O}\left(n^{-2}\sqrt{\log\frac{n}{p}}\right), & n^{\frac{7}{8}} \leq i \leq n,\end{cases}$$

for certain constant $C > 0$, where $\mu_{\alpha,k} = \mathcal{O}(\frac{(\alpha-1)^2}{k^4})$ is the kth eigenvalue of $\mathcal{G}_\alpha$.

The leading eigenvalue $\mu_{\alpha,1}$ can be estimated from above using the Hilbert–Schmidt norm of $\mathcal{G}_\alpha$ and also using the Perron-Frobenious theorem [18] (or Krein-Rutman theorem for positive compact operators), one can estimate both $\mu_{\alpha,1}$ from below.

COROLLARY E5. For any $\alpha \in (0, 1)$, the leading eigenvalue $\mu_{\alpha,1} \in [0.941, 2.754]$.

Proof. First, we compute that

$$\begin{aligned} v(x) &:= \int_{-1}^{1} \mathcal{G}_\alpha(x, y)\, \mathrm{d}y \\ &= \frac{(x-1)^2(x^2+6x+17) - 2\alpha(1+6x^2+x^4) + \alpha^2(x+1)^2(x^2-6x+17)}{24}, \end{aligned}$$

then the leading eigenvalue satisfies

$$\mu_{\alpha,1} \geq \frac{\|v\|_{L^2[-1,1]}}{\sqrt{2}} \geq 2\sqrt{\frac{(728 - 323\alpha + 450\alpha^2 - 323\alpha^3 + 728\alpha^4)}{2835}}.$$

Minimize the right-hand side, we find that $\mu_{\alpha,1} \geq 0.941$, $\forall \alpha \in [0, 1]$, the minimum is achieved around $\alpha = 0.351$. For the upper bound, we compute the Hilbert–Schmidt norm

$$\sqrt{\int_{-1}^{1}\int_{-1}^{1} |\mathcal{G}_\alpha(x, y)|^2\, \mathrm{d}x\, \mathrm{d}y} = \sqrt{\beta(\Upsilon \cdot (1, \alpha, \alpha^2, \cdots, \alpha^8))} \leq \frac{32}{3\sqrt{15}} \approx 2.754, \quad \alpha \in [0, 1],$$

where $\beta = \frac{512}{212837625}$ and

$$\Upsilon = (1370738, -172283, 394834, -98757, 164086, -98757, 394834, -172283, 1370738) \in \mathbb{R}^9,$$

and the maximum is achieved at $\alpha = 1$. □

We should observe that the leading eigenvalue $\mu_{\alpha,1}$ actually is uniformly bounded from below if $\alpha \in (0, 1)$. Therefore the decay of eigenvalues is even worse for α close to 1. This somewhat is straightforward since $\alpha \sim 1$ means the loss of nonlinearity and the eigenvalues collapse to zeros except the leading one.

E.3 *Analytic activation functions*

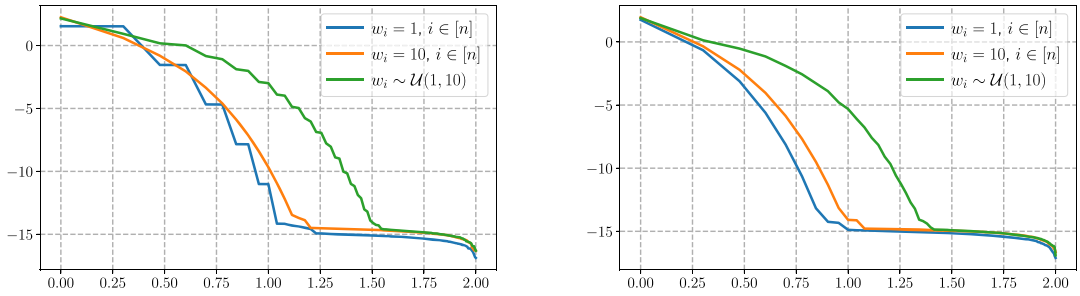
For analytic activation functions such as `Tanh` or `Sigmoid`, the Gram matrix is formed by

$$\boldsymbol{G}_{i,j} = \int_D \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} - b_i)\sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} - b_j) d\boldsymbol{x}.$$

Particularly, if the weights $|\boldsymbol{w}_i| \leq A < \infty$, then the kernel function can be viewed as

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{y}) = \int_D \sigma(\boldsymbol{x} \cdot \boldsymbol{z})\sigma(\boldsymbol{y} \cdot \boldsymbol{z}) d\boldsymbol{w}, \quad \boldsymbol{x}, \boldsymbol{y} \in [-A, A] \times [-1, 1],$$

where $\boldsymbol{z} := (\boldsymbol{w}, -1) \in \mathbb{R}^{d+1}$. Since the kernel is analytic in both $\boldsymbol{x}$ and $\boldsymbol{y}$, the eigenvalues of the kernel are decaying faster than any polynomial rate [44, 45]. Two examples in one dimension are provided in Fig. E1.



(a) `Tanh`: $\sigma(x) = (e^x - e^{-x})/(e^x + e^{-x})$. (b) `Sigmoid`: $\sigma(x) = 1/(1 + e^{-x})$.

FIG. E1. Illustrations of the spectrum of Gram matrices in the one-dimensional case with $n = 100$ for `Tanh` and `Sigmoid` activation functions. The x-axis and y-axis correspond to $\log_{10} k$ and $\log_{10} \lambda_k$, respectively, for $k \in [n]$. Here, $(b_i)_{i=1}^n$ is evenly spaced in the interval $[-1, 1]$ and $(w_i)_{i=1}^n$ is chosen from one of three cases: $w_i = 1$ for all i, $w_i = 10$ for all i, or w_i randomly sampled from a uniform distribution $\mathcal{U}(1, 10)$.