Why Shallow Networks Struggle with Approximating and Learning High Frequency: A Numerical Study

Shijun Zhang SHIJUN.ZHANG@DUKE.EDU
Department of Mathematics
Duke University
Durham, NC 27708

Hongkai Zhao ZHAO@MATH.DUKE.EDU
Department of Mathematics
Duke University
Durham, NC 27708

Yimin Zhong YIMIN.ZHONG@AUBURN.EDU
Department of Mathematics and Statistics
Auburn University
Auburn, AL 36830

Haomin Zhou HMZHOU@MATH.GATECH.EDU
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

In this work, a comprehensive numerical study involving analysis and experiments shows why a two-layer neural network has difficulties handling high frequencies in approximation and learning when machine precision and computation cost are important factors in real practice. In particular, the following basic computational issues are investigated: (1) the minimal numerical error one can achieve given a finite machine precision, (2) the computation cost to achieve a given accuracy, and (3) stability with respect to perturbations. The key to the study is the conditioning of the representation and its learning dynamics. Explicit answers to the above questions with numerical verifications are presented.

Contents

1 Introduction **2**
1.1 Literature Review 4
1.2 Contributions 5

2 Gram Matrix and Least Square Approximation **6**
2.1 One-dimensional Case 7
2.2 Multi-dimensional Case 13
2.3 Numerical Experiments 15

2.3.1 Spectrum and Eigenmodes of Gram Matrix in 1D 15
2.3.2 Approximation in 1D: FEM Basis vs. ReLU Basis 16
2.3.3 Spectrum and Eigenmodes of Gram Matrix in 2D 19
2.4 Observations and Comments 20

3 Learning Dynamics **22**
3.1 Mathematical Setup of Learning Dynamics 24
3.2 Generalized Fourier Analysis of Learning Dynamics 29

4 Rashomon Set for Two-layer ReLU Neural Networks **33**

5 Further Discussions **34**

A Properties of Eigenfunctions **39**

B Generalized Fourier Modes **41**

C Improved Bounds for Learning Dynamics **43**
C.1 Part I 43
C.2 Part II 45
C.3 Part III 46
C.4 Numerical Experiments 47
C.5 Further Remarks 50
C.5.1 Initial Distribution of Biases 50
C.5.2 Activation Function 51
C.5.3 Boundedness of Weights 51

D Rashomon Set for Bounded Activation Function **51**

E Further Discussions **53**
E.1 General Case of Gram Matrix of Two-layer ReLU Networks in 1D 53
E.2 Leaky ReLU Activation Function 54
E.3 Analytic Activation Functions 56

1 Introduction

Neural networks 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. It has met with a lot of success as well as challenges in practice. More importantly, many basic and practical questions are still open. Extensive study has been carried out to understand the properties of neural networks 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 neural networks 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 (Shen et al., 2021a,b, 2022b,a; Zhang et al., 2023; Jiao

et al., 2021; Yarotsky, 2018, 2017; Shen et al., 2019, 2020; Lu et al., 2021; Zhou, 2020; Chui et al., 2018; Gribonval et al., 2022; Gühring et al., 2020; Suzuki, 2019; Nakada and Imaizumi, 2020; Chen et al., 2019; Bao et al., 2019). Recently, there are works (Shen et al., 2020; Yarotsky, 2018; Zhang, 2020) 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 neural network 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 neural networks. Our consideration includes both asymptotic and non-asymptotic regimes in term of network width:

1. the minimal numerical error one can achieve given a finite machine precision;

2. the computation time (cost) to achieve a certain accuracy for the training process;

3. stability to perturbations, e.g., noise in the data, or over-fitting.

Our study proves that, in practice, a shallow neural network is essentially a "low-pass filter" by showing an explicit spectral decay rate of the Gram matrix and asymptotic equivalence of the eigenmodes to the eigenfunctions of the Laplace operator (generalized Fourier modes) in arbitrary dimension. More specifically, the ill-conditioning of the Gram matrix means only leading (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 neural networks is proved in theory, the achievable numerical accuracy may be far less than the machine precision in practice, for example, when approximating functions with high-frequency components, such as functions with rapid changes or highly oscillatory patterns. Moreover, the numerical errors can not be further improved by increasing the number of data or the network's width after a certain stage since the number of eigenmodes that can be stably captured by a given machine precision does not increase. Approximation error for a given function is determined by the function's regularity and machine precision. On the other hand, the low pass filter nature leads to stability with respect to perturbations in the high modes, e.g., noises, or over-parametrization. These phenomena have been well observed and documented in many studies and experiments.

One of the most important features when using neural networks for approximation is the capability of learning, a.k.a. optimizing the parameters to adapt to the underlying function manifested by data. However, due to the ill-conditioning of the Gram matrix, we show that a gradient-based optimization leads to slow learning dynamics for high modes which are needed in an adaptive representation. Moreover, adaptivity may cause even worse conditioning of the resulting Gram matrix, which leads to more stiffness in the dynamics.

From a probabilistic perspective, we show that the Rashomon set, the set of parameters where accurate approximations can be achieved, for a two-layer neural network 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 higher 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 decomposition and composition in our future work.

1.1 Literature Review

The approximation theory of shallow neural networks has been well-known since the universal approximation theorem (Barron, 1993; Cybenko, 1989; Chui and Li, 1992; Park and Sandberg, 1991; Klusowski and Barron, 2018; Breiman, 1993). The error bounds of approximation in L^∞ norm have been demonstrated either through explicit construction or by probabilistic proofs, see (Domingo-Enrich and Mroueh, 2021; Klusowski and Barron, 2018; Hornik et al., 1989) and the references therein. However, it has been observed widely in practice that shallow neural networks cannot approximate highly oscillatory functions effectively (Chen et al., 2022; Hong et al., 2022; Grossmann et al., 2023). Several explanations have been proposed in the past few years.

The authors of (Luo et al., 2019; Xu et al., 2019) summarized such phenomenon as a heuristic law called *frequency principle*, that is, the training process of the neural network 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 (Xu et al., 2019) 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 (Luo et al., 2019) 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, (Eldan and Shamir, 2016) and (Safran and Shamir, 2017) claimed the bottleneck of the training of shallow networks may come from the so-called "depth separation" of the capacity between shallow and deep neural networks and explained that shallow neural networks demand a width that grows exponentially in dimension to fit discontinuous functions in L^2 norm while deep neural networks 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. (1) NTK (neural-tangent kernel) regime. In this 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 (Zhang et al., 2019; Li et al., 2020), the parameters of the neural network 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 (Du et al., 2019; Li and Liang, 2018; Jacot et al., 2018). With well-distributed labeled data, the slow convergence rate relates to the fast decay rate of the

eigenvalues of the neuron-tangent kernel (Cao et al., 2019; Su and Yang, 2019; Velikanov and Yarotsky, 2021; Nitanda and Suzuki, 2020). 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 (Birman and Solomyak, 1970), while a fully explicit characterization of all eigenvalues is difficult. (2) Mean-field regime. In this 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 (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020). 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 (Rotskoff and Vanden-Eijnden, 2018) for C^1 activation function under the assumption that the empirical measure of particle system converges. 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 neural networks to fit highly oscillatory functions. A recent work (Holzmüller and Steinwart, 2020) 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 here:

- 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 in nonuniform and nonasymptotic settings with respect to the bias distribution is also analyzed. The study implies that the approximation by a two-layer neural network 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. We show slow learning dynamics for high-frequency modes and a long training time to reach (even if possible) the optimal solution in practice.

- The measure of Rashomon set, the set of parameters in parameter space that renders an approximation with a given tolerance, for two-layer neural networks is characterized. The result shows that, due to the small correlation between an oscillatory function and the family of activation functions, 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 neural networks. 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

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

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

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

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

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

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

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

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 neural networks 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 neural network as a linear representation, i.e., where the weights and biases in the hidden neurons are fixed, using least squares. In this setting, the 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, which is formed by the inner product (or correlation) between each pair of the basis. In other words, global interactions and strong correlations should be avoided.

We start with the most used activation function in neural networks is `ReLU`: $\sigma(x) \coloneqq \max(x, 0)$. The general form of a shallow network in (1) can be simplified to

$$h(\boldsymbol{x}) = \sum_{i=1}^{n} a_i \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} - b_i) + c, \quad \boldsymbol{x} \in D \subseteq \mathbb{R}^d,\ \boldsymbol{w}_i \in \mathbb{S}^{d-1},\ a_i, b_i \in \mathbb{R}. \tag{2}$$

2.1 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) 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 $\{b_i\}_{i=1}^n$. Theoretically, 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 (Kamenski et al., 2014). 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 $\boldsymbol{G} \coloneqq (\boldsymbol{G}_{i,j}) \in \mathbb{R}^{n \times n}$, where

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

The correlation between two `ReLU` functions with close biases b_i, b_j is $1 - O(|b_i - b_j|)$ 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 $\mathcal{G} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$:

$$\mathcal{G}(x, y) \coloneqq \int_D \sigma(z - x)\sigma(z - y)dz. \tag{3}$$

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

$$\begin{aligned}\mathcal{G}(x,y) &\coloneqq \frac{1}{24}(2-x-y-|x-y|)^2(2-x-y+2|x-y|) \\ &= \frac{1}{12}|x-y|^3 + \frac{1}{12}(2-x-y)\left(2(1-x)(1-y)-(x-y)^2\right).\end{aligned} \tag{4}$$

First, we provide an explicit spectral analysis for the Gram kernel and show that its eigenvalues $\mu_k \sim k^{-4}$ and the first few leading eigenfunctions are a combination of exponential functions and Fourier modes, then followed by essentially Fourier modes, from low to high frequencies.

Lemma 1. *For the Gram kernel function* (4)*, define the operator* $K : L^2[-1,1] \to L^2[-1,1]$

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

and denote μ_j *is the* j*-th eigenvalue of* K *in descending order, then there exist positive constants* c_1 *and* c_2 *that* $c_1 j^{-4} \le \mu_j \le c_2 j^{-4}$.

Proof. Assume the eigenpairs of K are denoted by (μ_k, ψ_k), $k \ge 1$, where μ_k is sorted in descending order. For the eigenfunction ψ_k, it satisfies

$$\int_{-1}^{1} \mathcal{G}(x,y)\psi_k(y)dy = \mu_k\psi_k(x), \tag{5}$$

Taking derivatives four times on both sides, similar techniques can be found in (Bogoya et al., 2012), we obtain the following differential equation,

$$\psi^{(4)}(x) = \frac{1}{\mu_k}\psi_k(x).$$

which implies that there are constants $A_s \in \mathbb{C}$, $0 \le s \le 3$ that

$$\psi_k(x) = A_0\cosh(w_k x) + A_1\sinh(w_k x) + A_2\cos(w_k x) + A_3\sin(w_k x)$$

for certain $w_k = \mu_k^{-\frac{1}{4}} > 0$. Thus solving the eigensystem is then equivalent to solving a 4×4 matrix P that $\det(P) = 0$ for ω using (5). Observe that $\mathcal{G}(1,y) = 0$, $\mathcal{G}_x(1,y) = 0$, $\mathcal{G}_{xx}(1,y) = 1-y$ and $\mathcal{G}_{xxx}(1,y) = 1$. We find the corresponding equation as

$$\begin{pmatrix} \cosh(w_k) & \sinh(w_k) & \cos(w_k) & \sin(w_k) \\ w_k\sinh(w_k) & w_k\cosh(w_k) & -w_k\sin(w_k) & w_k\cos(w_k) \\ w_k^2\cosh(w_k) & w_k^2\sinh(w_k) & -w_k^2\cos(w_k) & -w_k^2\sin(w_k) \\ w_k^3\sinh(w_k) & w_k^3\cosh(w_k) & w_k^3\sin(w_k) & -w_k^3\cos(w_k) \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix}$$

$$= w_k^4 \begin{pmatrix} 0 \\ 0 \\ \frac{2\sinh(w_k)}{w_k}A_0 + \frac{2\sinh(w_k)-2\cosh(w_k)}{w_k^2}A_1 + \frac{2\sin(w_k)}{w_k}A_2 + \frac{2w_k\cos(w_k)-2\sin(w_k)}{w_k^2}A_3 \\ \frac{2\sinh(w_k)}{w_k}A_0 + \frac{2\sin(w_k)}{w_k}A_2 \end{pmatrix}.$$

Let the determinant of the above linear equation be zero which gives the following simple relation

$$\tan(w_k)\tanh(w_k) = \pm 1.$$

Let $g(x) = \tan(x)\tanh(x)$, for $x \in (n\pi, n\pi + \frac{\pi}{2})$ or $x \in (n\pi + \frac{\pi}{2}, n\pi + \frac{3\pi}{2})$, $n \in \mathbb{Z}$, the function g is monotone and has a range of $(-\infty, \infty)$ in each interval. Therefore there is a unique $w_{2j+1} \in ((j + \frac{1}{4})\pi, (j + \frac{1}{2})\pi)$, $w_{2j+2} \in ((j + \frac{1}{2})\pi, (j + \frac{3}{4})\pi)$, $j \geq 0$, which says $\lambda_j \sim (\frac{j}{2}\pi)^{-4}$. Actually, one can show that both $c_1 = \left(\frac{2}{\pi}\right)^4 (1 + \mathcal{O}(j^{-1}))$ and $c_2 = \left(\frac{2}{\pi}\right)^4 (1 + \mathcal{O}(j^{-1}))$. Moreover, we can derive the explicit forms for the eigenfunctions:

1. If $\tan(w_k)\tanh(w_k) = 1$, $\psi_k(x) = A_2(-\frac{\cos(w_k)}{\sinh(w_k)}\sinh(w_k x) + \cos(w_k x))$,

2. If $\tan(w_k)\tanh(w_k) = -1$, $\psi_k(x) = A_3(-\frac{\sin(w_k)}{\cosh(w_k)}\cosh(w_k x) + \sin(w_k x))$.

One can observe that $A_0, A_1 = \mathcal{O}(e^{-w_k})$. A quick computation of normalization shows that A_2, A_3 are uniformly bounded, which implies that all eigenfunctions are uniformly bounded. ■

In fact, one can derive a much more accurate estimate of w_k, $k \in \mathbb{N}$.

Corollary 2. *$w_{2j+1} - (j + \frac{1}{4})\pi < e^{-2w_{2j+1}}$ and $w_{2j+2} - (j + \frac{3}{4})\pi > -e^{-2w_{2j+2}}$.*

Proof. For w_{2j+1}, use the relation $\tan(w_{2j+1}) = \coth(w_{2j+1}) = 1 + 2/(e^{2w_{2j+1}} - 1)$. We obtain that

$$w_{2j+1} - (j + \frac{1}{4})\pi < \tan(w_{2j+1} - (j + \frac{1}{4})\pi) = \frac{\tan(w_{2j+1}) - 1}{1 + \tan(w_{2j+1})} = e^{-2w_{2j+1}}.$$

The first inequality uses $x < \tan x$ for $x \in (0, \frac{\pi}{4})$. The other part of the corollary follows a similar derivation. ■

Denote the vectors $\boldsymbol{a} \coloneqq (a_i)_{i=1}^n$ and $\boldsymbol{f} \coloneqq (f_i)_{i=1}^n$ that $f_i = \int_D f(x)\sigma(x - b_i)dx$. 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 (Koltchinskii and Giné, 2000).

Theorem 3. *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 k-th eigenvalue of $\mathcal{G}$.*

Proof. The idea of proof comes from (Widom, 1958). Define the operator K^* by the kernel

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

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

$$|\mu_i - \mu_i^*| \leq \|K - K^*\| \leq \sqrt{\int_{-1}^{1}\int_{-1}^{1} |\mathcal{G}(x,y) - \mathcal{G}^*(x,y)|^2 \, dxdy} = \mathcal{O}(n^{-1}). \tag{6}$$

Now we consider perturbed $\widetilde{b}_i = b_i^* + o(\frac{1}{n})$ as mentioned in Theorem 3.1, 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 4. *Suppose $\{b_i\}_{i=1}^n$ are chosen as Theorem 3, then the condition number of the Gram matrix $\boldsymbol{G}$ satisfies*

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

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

$$\mathcal{G}(x,y) = \sum_{k=1}^{\infty} \mu_k \phi_k(x)\phi_k(y). \tag{7}$$

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

$$\boldsymbol{G}_{ij} = \mathcal{G}_m(b_i, b_j) + (\mathcal{G}(b_i, b_j) - \mathcal{G}_m(b_i, b_j)). \tag{8}$$

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

$$\boldsymbol{G} = \Phi_m \Lambda_m \Phi_m^T + \boldsymbol{E}_m, \tag{9}$$

where $(\Lambda_m)_{ij} = \delta_{ij}\mu_i$ and $(\boldsymbol{E}_m)_{ij} = \sum_{k>m} \mu_k\phi_k(b_i)\phi_k(b_j)$. Let $\sigma_i(\Phi_m\Lambda_m\Phi_m^T)$ denote the i-th eigenvalue of $\Phi_m\Lambda_m\Phi_m^T$, then by Ostrowski's theorem (Horn and Johnson, 2012),

$$\left|\sigma_i(\Phi_m\Lambda_m\Phi_m^T) - \frac{n}{2}\mu_i\right| \leq |\mu_i| \left\|\Phi_m^T\Phi_m - \frac{n}{2}\mathrm{Id}_m\right\|_{op}, \quad 1 \leq i \leq m \leq n. \tag{10}$$

Therefore, using Weyl's inequality

$$\begin{aligned}\left|\lambda_i - \frac{n}{2}\mu_i\right| &\leq |\lambda_i - \sigma_i(\Phi_m\Lambda_m\Phi_m^T)| + \left|\sigma_i(\Phi_m\Lambda_m\Phi_m^T) - \frac{n}{2}\mu_i\right| \\ &\leq \|\boldsymbol{E}_m\|_{op} + |\mu_i| \left\|\Phi_m^T\Phi_m - \frac{n}{2}\mathrm{Id}_m\right\|_{op}, \qquad 1 \leq i \leq m \leq n.\end{aligned} \tag{11}$$

Since the eigenfunctions ϕ_k are uniformly bounded, see the explicit form in Lemma 1, then

$$\|\boldsymbol{E}_m\|_{op} \leq Cn \sum_{k>m} \mu_k = \mathcal{O}\left(\frac{n}{m^3}\right). \tag{12}$$

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,

$$\frac{2}{n}\sum_{i=1}^{n}\phi_j(b_i)\phi_k(b_i) - \int_{-1}^{1}\phi_j(x)\phi_k(x)dx = \mathcal{O}\left(\frac{1}{n}\sup_{[-1,1]}(\phi_j\phi_k)'\right). \tag{13}$$

Hence using the estimate $\|\phi_k'\|_\infty = \mathcal{O}(k)$, $\left\|\Phi_m^T\Phi_m - \frac{n}{2}\mathrm{Id}_m\right\|_{op} = \mathcal{O}(m^2)$. Combine the above estimates into (11) and reuse Theorem 3, by selecting $m = n^{1/5}i^{4/5} \geq i$,

$$\left|\lambda_i - \frac{n}{2}\mu_i\right| \leq C\min\left(1, \frac{n}{m^3} + \frac{m^2}{i^4}\right) = \begin{cases}\mathcal{O}(1) & i < n^{\frac{1}{6}} \\ \mathcal{O}(n^{2/5}i^{-12/5}) & n^{\frac{1}{6}} \leq i \leq n\end{cases} \tag{14}$$

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 5. *For a given machined precision ε, when n^3 is not as large as ε^{-1}, all eigenmodes are maintained using stable least square approximation since the condition number is $\Omega(n^3)$. After that, one will only maintain the leading i mode where the ratio of the ith eigenvalue to the first eigenvalue is of order $n^{-3/5}i^{-12/5} + i^{-4} \sim \varepsilon$. When n is sufficiently large, leading $i \sim \varepsilon^{-1/4}$ modes will be maintained.*

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 $\mathcal{G}_\rho(x,y) \coloneqq \sqrt{\rho(x)}\mathcal{G}(x,y)\sqrt{\rho(y)}$. For this case, we have a similar estimate of eigenvalues if ρ is bounded from below and above by positive constants.

Lemma 6. *Suppose $\rho(x)$ is bounded from below and above by positive constants and define*

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

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

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

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

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 $\mathcal{G}$. Then let $z = \frac{1}{\sqrt{\rho}}\sum_{j=1}^k c_j\phi_j, \|z\| = 1$, we find that

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

In the second equation, we choose the space $S_{k-1} = \text{span}(\sqrt{\rho}\phi_1, \cdots, \sqrt{\rho}\phi_{k-1})$ and let $z = \frac{1}{\sqrt{\rho}} \sum_{j=k}^{\infty} c_j \phi_j \in S_{k-1}^{\perp}, \|z\| = 1$, then

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

■

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

Theorem 7. *Suppose $\{b_i\}_{i=1}^n$ are i.i.d distributed 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,*

$$\left| \widetilde{\lambda}_i - \frac{n}{2} \widetilde{\mu}_i \right| = \begin{cases} \mathcal{O}\left(n^{\frac{5}{8}} i^{-3} \sqrt{\log \frac{n}{p}} \right) & i < n^{\frac{7}{8}} \\ \mathcal{O}\left(n^{-2} \sqrt{\log \frac{n}{p}} \right) & n^{\frac{7}{8}} \leq i \leq n \end{cases} \tag{15}$$

with probability $1-p$, where $\widetilde{\mu}_i = \Theta(i^{-4})$ is the k-th 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 $\mathcal{G}_\rho$, then using the same technique as Lemma 1, we obtain the differential equation

$$\frac{1}{\widetilde{\mu}_k} \sqrt{\rho(x)} \phi_{\rho,k}(x) = \frac{d^4}{dx^4} \left[\frac{1}{\sqrt{\rho(x)}} \phi_{\rho,k}(x) \right] \tag{16}$$

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 (16) becomes

$$\psi^{(4)}_{\rho,k} = \frac{1}{\widetilde{\mu}_k} \rho(x) \psi_{\rho,k}. \tag{17}$$

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

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

One can find that the differential equation (17) reduces to the form

$$\frac{d^4}{dt^4} \psi_{\rho,k} + p_1(t) \frac{d^3}{dt^3} \psi_{\rho,k} + p_2(t) \frac{d^2}{dt^2} \psi_{\rho,k} + p_3(t) \frac{d}{dt} \psi_{\rho,k} + p_4(t) \psi_{\rho,k} = \frac{H^4}{\widetilde{\mu}_k} \psi_{\rho,k}. \tag{18}$$

and the functions $p_k(t)$ are continuous over $[-1,1]$ since $\rho \in C^3[-1,1]$. For k sufficiently large, the asymptotic behaviors 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 (Stone, 1926) and Birkhoff's method (Birkhoff,

1908; Naimark, 1967). 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 (Naimark, 1967).

Similar to Theorem 4, we define the matrix $\Phi_m \in \mathbb{R}^{n\times m}$ with entries $(\Phi_m)_{ij} = \phi_{\rho,j}(b_i)/\sqrt{\rho(b_i)}$, $1 \le j \le m$, then the Gram matrix with entry $\boldsymbol{G}_{ij} = \mathcal{G}(b_i, b_j)$ equals to

$$\boldsymbol{G} = \Phi_m \Lambda_m \Phi_m^T + \boldsymbol{E}_m, \tag{19}$$

where $(\Lambda_m)_{ij} = \delta_{ij}\widetilde{\mu}_i$, $(\boldsymbol{E}_m)_{ij} = \sum_{k>m} \widetilde{\mu_k}\phi_{\rho,k}(b_i)\phi_{\rho,k}(b_j)/\sqrt{\rho(b_i)\rho(b_j)}$, and the estimate (12) still holds. Applying the Hoeffding's inequality to (13), we get

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

with probability $1 - \frac{p}{m^2}$, due to the uniform boundedness of the eigenfunctions $\phi_{\rho,k}$. Then with probability $1-p$, we have

$$\begin{aligned}\left|\widetilde{\lambda}_i - \frac{n}{2}\widetilde{\mu}_i\right| &\le C \min_{i\le m\le n}\left(\frac{n}{m^3} + \frac{m}{i^4}\sqrt{n\log\frac{m^2}{p}}\right), \qquad \text{take } m = \min(n^{\frac{1}{8}}i, n)\\ &= \begin{cases}\mathcal{O}\left(n^{\frac{5}{8}}i^{-3}\sqrt{\log\frac{n}{p}}\right) & i < n^{\frac{7}{8}}\\ \mathcal{O}\left(n^{-2}\sqrt{\log\frac{n}{p}}\right) & n^{\frac{7}{8}} \le i \le n.\end{cases}\end{aligned} \tag{20}$$

■

Corollary 8. *Under the same assumption of Corollary 7, then with probability $1-n^{-\beta}$, $1 > \beta > 0$, the condition number of Gram matrix $\boldsymbol{G}$ satisfies*

$$\kappa = \lambda_1/\lambda_n = \Omega(n^3(\log n)^{-1})$$

Proof. Choose $p = n^{-\beta}$, then with probability $1-p$, there exist constants C_1 and C_n depending on p that $\lambda_1 \ge C_1 n(1-(\log n)n^{-3/8})$ and $\lambda_n \le C_n n^{-2}\log n$. Therefore, $\kappa = \Omega\left(n^3\frac{(1-(\log n)n^{-3/8})}{\log n}\right) = \Omega(n^3(\log n)^{-1})$. ■

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 neural network $\mathcal{H}$ 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 1D setting, we consider the corresponding continuous kernel $G : L^2(V) \mapsto L^2(V)$

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

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

$$\int_V G(\boldsymbol{w}, b, \boldsymbol{w}', b')\phi_k(\boldsymbol{w}', b')d\boldsymbol{w}'db' = \lambda_k\phi_k(\boldsymbol{w}, b). \tag{21}$$

One of the useful tools to study two-layer `ReLU` networks of infinite width is the Radon transform (Ongie et al., 2019; Savarese et al., 2019). Next, we construct the theory for the Gram matrix in the high dimension using the properties of the Radon transform.

Definition 9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an integrable function over all hyperplanes, the Radon transform*

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

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

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

Theorem 10 (Helgason (Helgason et al., 2011))**.** *The inversion formula of Radon transform is*

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

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

Then we have the following lemma with Radon transform:

Lemma 11. *The eigenfunction satisfies $\lambda_k\partial_b^2\phi_k = \mathcal{R}\Delta^{-1}\mathcal{R}^*\phi_k$ in weak sense.*

Proof. Differentiate (21) with b twice and integrate against the test function

$$h \in \{\mathcal{R}g \mid \Delta^{-(d+1)/2}g \in C_0^2(D)\},$$

using $\partial_b^2\sigma(\boldsymbol{w}\cdot\boldsymbol{x} - b) = \delta(\boldsymbol{w}\cdot\boldsymbol{x} - b)$, we have

$$\int_D\int_V [\mathcal{R}^*h(\boldsymbol{x})]\,\sigma(\boldsymbol{w}'\cdot\boldsymbol{x} - b')\phi_k(\boldsymbol{w}', b')d\boldsymbol{w}'db'd\boldsymbol{x} = \lambda_k\int_V \partial_b^2\phi_k(\boldsymbol{w}, b)h(\boldsymbol{w}, b)d\boldsymbol{w}db. \tag{22}$$

Observe that the test function h satisfies

$$(-\Delta)^{-1}\mathcal{R}^*h = (-\Delta)^{-(d+1)/2}(-\Delta)^{(d-1)/2}\mathcal{R}^*h = (-\Delta)^{-(d+1)/2}g \in C_0^2(D)$$

from the inversion formula of Radon transform in Theorem 10. Using the Green's formula

$$\langle\mathcal{R}^*h(\boldsymbol{x}), \sigma(\boldsymbol{w}'\cdot\boldsymbol{x} - b')\rangle_D = \langle\Delta^{-1}\mathcal{R}^*h(\boldsymbol{x}), \Delta\sigma(\boldsymbol{w}'\cdot\boldsymbol{x} - b')\rangle_D = \mathcal{R}\Delta^{-1}\mathcal{R}^*h(\boldsymbol{w}', b'),$$

we derive that the right-hand side of (22) satisfies

$$\begin{aligned}\lambda_k\int_V \partial_b^2\phi_k(\boldsymbol{w}, b)h(\boldsymbol{w}, b)d\boldsymbol{w}db &= \int_V \mathcal{R}\Delta^{-1}\mathcal{R}^*h(\boldsymbol{w}', b')\phi_k(\boldsymbol{w}', b')d\boldsymbol{w}'db' \\ &= \int_V h(\boldsymbol{w}', b')\mathcal{R}\Delta^{-1}\mathcal{R}^*\phi_k(\boldsymbol{w}', b')d\boldsymbol{w}'db',\end{aligned}$$

where we have used the fact that $\mathcal{R}\Delta^{-1}\mathcal{R}^*$ is self-adjoint. ■

Lemma 12 (Helgason (Helgason, 2022), Lemma 2.1)**.** *These intertwining relations hold:* $\mathcal{R}\Delta = \partial_b^2\mathcal{R}$ *and* $\mathcal{R}^*\partial_b^2 = \Delta\mathcal{R}^*$.

Apply $\mathcal{R}^*$ on both sides of $\lambda_k\partial_b^2\phi_k = \mathcal{R}\Delta^{-1}\mathcal{R}^*\phi_k$, then

$$\lambda_k\mathcal{R}^*\phi_k = \Delta^{-1}\mathcal{R}^*\mathcal{R}\Delta^{-1}\mathcal{R}^*\phi_k.$$

Therefore $(\lambda_k^{-1}, \mathcal{R}^*\phi_k)$ forms a pair of eigenvalue and eigenfunction of the operator $\Delta^{-1}\mathcal{R}^*\mathcal{R}\Delta^{-1}$. Therefore we only need to study the eigenvalues of $\Delta^{-1}\mathcal{R}^*\mathcal{R}\Delta^{-1}$. In the following, we show the eigenvalues of the Gram kernel decay as $\lambda_k = \Theta(k^{-(d+3)/d})$.

Theorem 13. *There are constants* $c_1, c_2 > 0$*, depending on* D *and* d*, such that* $c_1 k^{-(d+3)/d} \leq \lambda_k \leq c_2 k^{-(d+3)/d}$*,* $(\lambda_k = \Theta(k^{-(d+3)/d})$ *using Landau notation*$)$.

Proof. Using the inversion formula in Theorem 10 and previous analysis, formally we obtain

$$\Delta^{-1}\mathcal{R}^*\mathcal{R}\Delta^{-1} = c_n(-\Delta)^{-(d+3)/2},$$

whose eigenvalues $\{\omega_k\}_{k\geq 1}$ are $\omega_k = \Theta(k^{-(d+3)/d})$ by exploiting the Weyl's law for $-\Delta$. If $(d+3)/2$ is fractional, one simply let the eigenfunction vanish outside D. ■

Remark 14. *For the Gram matrix corresponding to activation function* $\texttt{ReLU}^m(x) = \frac{1}{m!}[\max(x,0)]^m$ *in* d*-dimension with* $m \geq 1$*, one can modify the above argument to formally show:*

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

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

Remark 15. *Although the decay of eigenvalues seems slower in higher dimensions, however, 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` *neural network is, it can only resolve all Fourier modes up to frequency* $\mathcal{O}(\varepsilon^{-\frac{1}{d+3}})$*. In practice, the ill-conditioning of the Gram matrix may be less severe in higher dimensions since the dense sampling of* $\boldsymbol{w}_i$ *and* b_i *may not be affordable.*

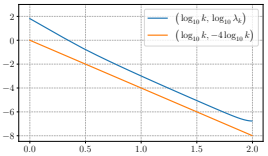
2.3 Numerical Experiments

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.

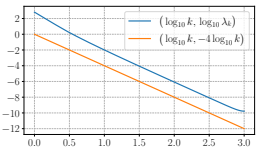
2.3.1 Spectrum and Eigenmodes of Gram Matrix in 1D

The first two numerical experiments show the spectrum of the Gram matrix for `ReLU` functions in 1D. Figure 1 and the first row of Figure 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 Figure 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 $F(x) = \int_{-1}^{x} |f'(t)|dt / \int_{-1}^{1} |f'(t)|dt$, 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 Figure 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.

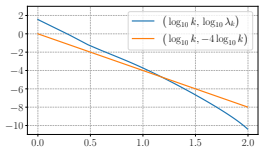


(a) $n = 100$.

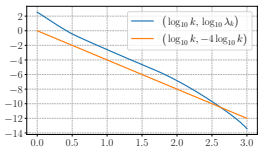


(b) $n = 1000$.

Figure 1: Spectrum for uniform $\boldsymbol{b}$.



(a) $n = 100$.



(b) $n = 1000$.

Figure 2: Spectrum for adaptive $\boldsymbol{b}$.

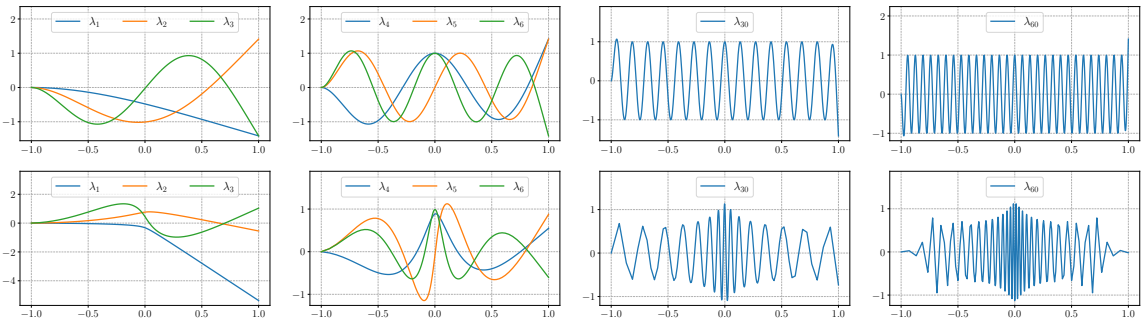
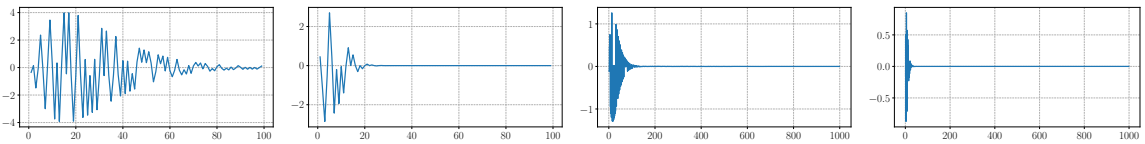


Figure 3: Eigenmodes of λ_k for $k = \{1, 2, 3\}, \{4, 5, 6\}, 30, 60$ with $n = 1000$. The first and second rows correspond to uniform and adaptive $\boldsymbol{b}$, respectively.



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

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

2.3.2 Approximation in 1D: FEM Basis vs. `ReLU` Basis

Next, we use numerical experiments to show least square approximation using FEM basis vs. `ReLU` basis (two-layer NN) 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. 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.

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}

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, η[1], to see the low pass filter effect for NN more clearly. The results are plotted in Figure 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 decrease in accuracy as the oscillation increases is due to the fact that the least square approximation error is proportional to $h^2 \int |f''(x)|dx$, 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 $f(9x)$.

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

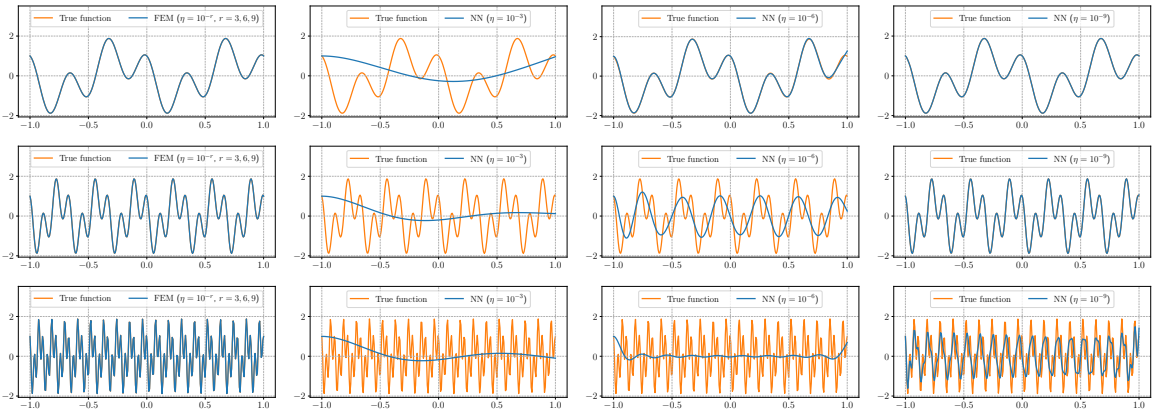


Figure 5: Approximation comparison: FEM vs. NN with 2000 samples and $n = 2000$ equal spaced basis. The three rows correspond to f_1, f_2, and f_3, respectively. Here $f_1(x) = f(x)$, $f_2(x) = f(3x)$, $f_3(x) = f(9x)$, where $f(x) = \cos(6\pi x) - \sin(2\pi x)$.

Table 2: Approximation errors in Figure 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}

In the following experiment, we show the low-pass filter nature of two-layer neural networks (NN) vs. finite element (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 Figure 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, Figure 7 shows that a two-layer `ReLU` network is significantly more stable with respect to over-parametrization due to its low-pass filter nature.

Table 3: Approximation errors in Figure 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}

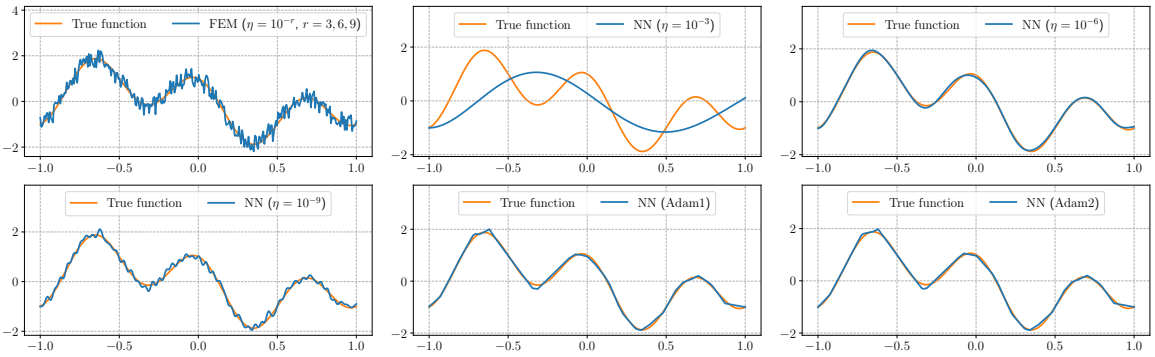


Figure 6: Approximation stability: FEM vs. NN subject to uniform noise $\mathcal{U}(-0.5, 0.5)$ on 1000 samples and $n = 1000$ basis. The first four plots are results from the least square with different cut-off ratio η for singular values. The last two plots are the results trained by the Adam optimizer,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

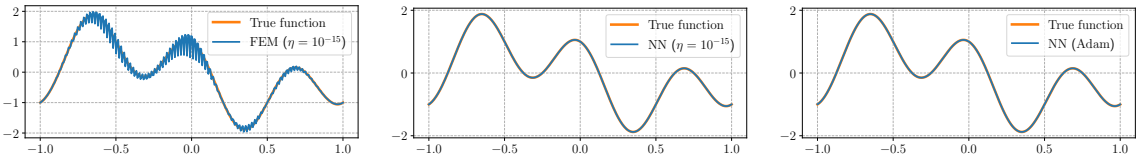


Figure 7: Test of overfitting: FEM vs. NN with 1000 samples, $n = 1500$ basis, and double precision. The first two plots are results from least square, where η is the cut-off ratio for small singular values. The last plot is the result trained by the Adam optimizer.

2.3.3 Spectrum and Eigenmodes of Gram Matrix in 2D

We use a numerical example to show, Figure 8, the spectrum of the discrete Gram matrix in 2D with 25600 evenly 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 Figure 9. In practice, those high-frequency modes whose corresponding eigenvalues are smaller than the machine precision threshold can not be captured.

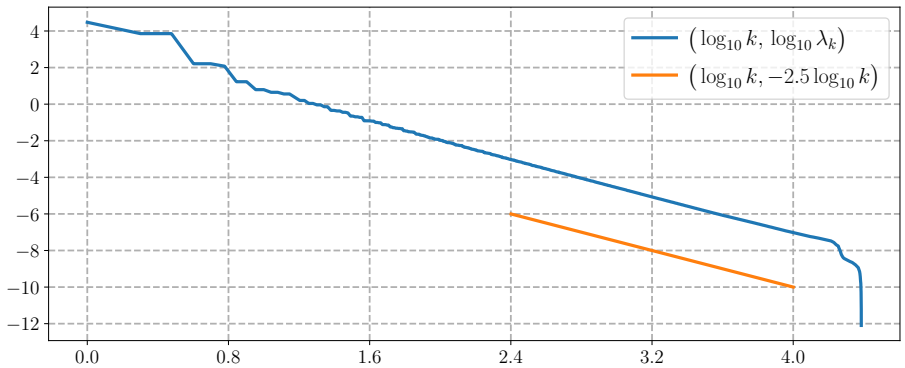
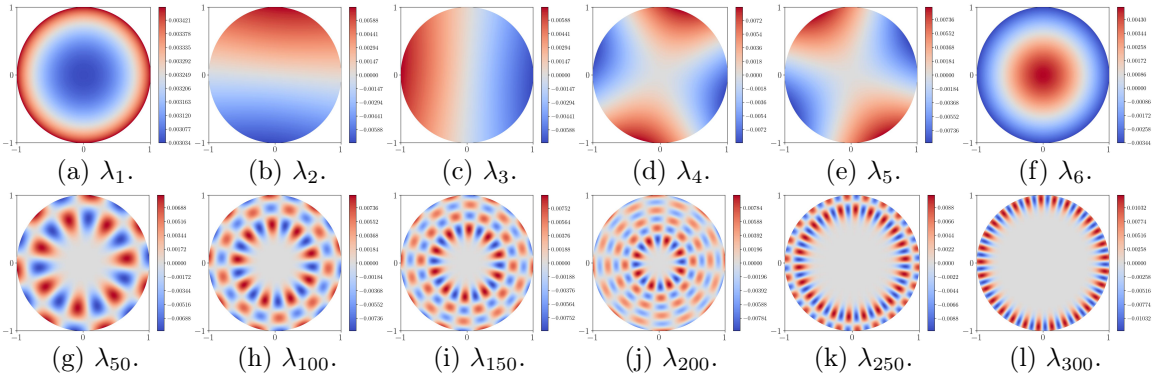


Figure 8: Spectrum of the discrete Gram matrix in 2D with 25600 evenly spaced samples.



(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}.

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

2.4 Observations and Comments

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

Low-pass filter nature

Suppose the objective function $f(\boldsymbol{x}), \boldsymbol{x} \in D \subseteq \mathbb{R}^d$, can be represented by the superposition of `ReLU` functions with weight $h(\boldsymbol{w}, b), (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R} = V$:

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

In 1D, the relation is further simplified to $\frac{d^2}{dx^2} f(x) = h(x)$. Assume $h(\boldsymbol{w}, b) = \sum_{k=1}^{\infty} \alpha_k \phi_k(\boldsymbol{w}, x)$ and hence

$$f(\boldsymbol{x}) = \sum_{k=1}^{\infty} \alpha_k \Delta^{-1} \mathcal{R}^* \phi_k. \tag{23}$$

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. 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 $\phi_k(\boldsymbol{x})$ for $k \leq m = \mathcal{O}(\varepsilon^{-\frac{d}{2d+3}})$ of $h(\boldsymbol{w}, b)$ can be stably recovered with machine precision ε in the least-square approximation. If the network is wide enough such that all leading modes up to $\phi_m(\boldsymbol{x})$ can be approximated well, the dominant numerical error is caused by the truncation of the modes higher than m. From the analysis in Section 2.2, we know that $\mathcal{R}^*\phi_k$ are eigenfunctions of the operator $\Delta^{-1}\mathcal{R}^*\mathcal{R}\Delta^{-1} = c_n(-\Delta)^{-(d+3)/2}$. In other words, (23) is the expansion of $f(\boldsymbol{x})$ in terms of the eigenfunctions of the Laplace operator. So truncation of the higher modes of $h(\boldsymbol{w}, b)$ in the parameter space leads to a low pass filter in the approximation of the original function $f(\boldsymbol{x})$. It means that 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) as the network width passes $n_{\varepsilon,d}$, the eigenvalues and eigenfunctions are very close to those of the continuous kernel and can only maintain leading modes k such that $\lambda_k \geq n\varepsilon\lambda_1 \Rightarrow k^{\frac{1}{d}} \leq \mathcal{O}(\varepsilon^{-\frac{1}{2d+3}})$.

The machine precision for single and double precision are: $\varepsilon_1 = 2^{-23}$ and $\varepsilon_2 = 2^{-52}$ respectively. Hence, a two-layer neural network can resolve about $2^{23d/(2d+3)}$ and $2^{52d/(2d+3)}$ eigenmodes respectively in d-dimensions. Since the eigenmodes are equivalent to the eigenfunctions of the Laplace operator as shown in Section 2.2, the resulting approximation behaves like a low-pass filter. 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 16. *Our analysis applies to NTK regime where the network width goes to infinity while the biases is pretty much fixed.*

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:

1. 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.

2. 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 $\mathcal{O}(\varepsilon^{\frac{p}{2d+3}}\|f\|_{H^p})$.

Implications
The low-pass filter nature and the two approximation error regimes explain why it is widely observed that shallow neural networks can approximate smooth functions well. On the other hand, for functions with sharp transitions or rapid oscillations, one may acquire a certain numerical error that is far from machine precision even using *wide* shallow networks for which universal approximation is proved in theory. The low-pass filter nature of the shallow neuron networks also implies stability with respect to noises or overfitting/over-parametrization.

From the spectral analysis for other activation functions presented in Appendix E and Remark 14, we see that the smoother the activation function is, the faster the spectrum of the corresponding Gram matrix decays, and hence the fewer eigenmodes can be used for approximation when a finite machine precision is imposed. However, with fixed network width (or a finite 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.

One way to avoid a strong correlation among activation functions (parametrized in certain ways), one can use local activation functions, e.g., the hat function for finite element basis. The resulting Gram matrix can be well-conditioned. However, the width of the network (or degrees of freedom) has to grow exponentially with the dimension in order to cover the domain of interest. Construction of a global activation function that is almost orthogonal to each other, such as the Fourier series, poses challenges in both sampling, e.g., in the Fourier domain, and computing the transformation in high dimensions.

3 Learning Dynamics

The key feature in machine learning using neural network 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 neural network (Barron and Klusowski, 2018; Ongie et al., 2019) 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. 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?

Our previous analysis shows that a two-layer neural network with fixed (or randomly sampled) biases will have a low-pass filter nature under finite machine precision, which makes it challenging to capture high-frequency components, i.e., fine-scale features. Here we show that the ill-conditioning of the Gram matrix causes difficulties in the learning process as well. 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 complete 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 neural network with the gradient flow can be regarded as the gradient descent method with a very small step size for finite time dynamics. We use 1D as an example. Let $D = [-1, 1]$, the gradient flow of $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ follow

$$\frac{da_i}{dt} = -\int_D (h(x,t) - f(x))\sigma(x - b_i)dx \quad \text{and} \quad \frac{db_i}{dt} = a_i \int_D (h(x,t) - f(x))\sigma'(x - b_i)dx.$$

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

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

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 neural networks can be found in (Rotskoff and Vanden-Eijnden, 2018; Mei et al., 2018; Sirignano and Spiliopoulos, 2020) 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 neural networks 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 (3), and $\widehat{g}$ as the generalized discrete Fourier transform of g,

$$\widehat{g}(m) = \int_D g(x)\theta_m(x)dx.$$

In 1D, the generalized Fourier transform is asymptotically close to the standard Fourier transform as the mode becomes higher and higher. The key to our study is the construction of an auxiliary function $w(x,t) \in C^2(D)$ (defined by (28)) 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 1D 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 17. *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 17 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 36).

Theorem 18. *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 neural networks with evenly or uniformly distributed initial biases. These two effects show why a two-layer neural network 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 neural network

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

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 neural network 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

$$\mathcal{L}(h, f) \coloneqq \frac{1}{2} \int_D |h(x) - f(x)|^2 dx.$$

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

$$\frac{d}{dt} a_i(t) = -\int_D (h(x, t) - f(x)) \psi(x, b_i(t)) dx \tag{24}$$

and

$$\frac{d}{dt} b_i(t) = -a_i(t) \int_D (h(x, t) - f(x)) \partial_b \psi(x, b_i(t)) dx \tag{25}$$

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 19. *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,

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

Therefore if initial $a_i(0)$ is finite, $a_i(t)$ is always finite, which in turn implies every derivative $\frac{d}{dt}b_i(t)$ 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 $\frac{d}{dt}b_k(t') \neq 0$ 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

$$\mathcal{G}(b, b') = \int_D \psi(x, b)\psi(x, b')dx, \tag{26}$$

which is a compact operator in $L^2(D^\varepsilon)$. Let $\lambda_k \geq 0, k = 1, 2, \ldots$ 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

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

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

$$G_{ij}(t) = \mathcal{G}(b_i(t), b_j(t)), \quad K_{ij}(t) = \partial_b\mathcal{G}(b_i(t), b_j(t)).$$

The functions P and Q are

$$P(b) = \int_D f(x)\psi(x, b)dx, \quad Q(b) = \int_D f(x)\partial_b\psi(x, b)dx.$$

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

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

and

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

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

$$
\frac{d\Theta_k(t)}{dt}=-\sum_{l=1}^{\infty}\lambda_l\left[M_{lk}(t)+S_{lk}(t)\right]\Theta_l(t) \tag{27}
$$

where the infinite matrices

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

are both semi-positive.

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

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

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

$$
w(b,t)\coloneqq\sum_{k=1}^{\infty}\lambda_k\Theta_k(t)\phi_k(b), \tag{28}
$$

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

$$\begin{aligned}
\partial_b^2 w(b,t) &= \partial_b^2 \left[\sum_{k=1}^{\infty} \left(\lambda_k \sum_{j=1}^{n} \phi_k(b_j(t)) a_j(t) - p_k \right) \phi_k(b) \right] \\
&= \partial_b^2 \left[\sum_{j=1}^{n} \mathcal{G}(b, b_j(t)) a_j(t) - P(b) \right] \\
&= \partial_b^2 \left[\sum_{j=1}^{n} a_j(t) \int_D \psi(x,b)\psi(x,b_j(t))dx - \int_D f(x)\psi(b)dx \right] \\
&= \sum_{j=1}^{n} a_j(t) \int_D \delta(x-b)\psi(x,b_j(t))dx - \int_D f(x)\delta(x-b)dx = h(b,t) - f(b).
\end{aligned} \tag{29}$$

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

Proof. From (29), 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 ϕ_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,

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

and

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

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

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

Follow the same derivation of (29),

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

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:

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

■

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

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

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:

1. 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 $\partial_b\bar{w}(b_i^*) = 0$.

2. 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 21. *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 (29), 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

$$\begin{aligned} \theta_k^{(4)}(x) &= p_k \theta_k(x), \\ \theta_k(1) = \theta_k'(1) &= 0, \\ \theta_k''(-1) = \theta_k'''(-1) &= 0. \end{aligned} \tag{30}$$

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:

$$\widehat{w}(k,t) = \int_D \theta_k(b) w(b,t) db. \tag{31}$$

where θ_k is the eigenfunction defined by (30). 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 (29). 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 26, $\partial_b^4 w(b,t) = \sum_{k=1}^{\infty} \Theta_k(t)\phi_k(b)$ for $b \in D$.

Lemma 22. *The following equality holds.*

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

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

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

■

Lemma 23. *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

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

Since $\|\theta_k''\|_{L^2(D)} = \mathcal{O}(k^2)$ by Lemma 30 and $p_k = \Theta(k^4)$ by Lemma 27, 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,

$$\begin{aligned}\int_D \theta_k(b)\partial_b^4 w(b,t)db &= \theta_k(b)\partial_b^3 w(b,t)\Big|_{\partial D} - \theta_k'(b)\partial_b^2 w|_{\partial D} + \int_D \theta_k''(b)\partial_b^2 w(b,t)db \\ &= \theta_k(b)\partial_b^3 w(b,t)\Big|_{\partial D} - \theta_k'(b)\partial_b^2 w|_{\partial D} + \int_D \theta_k^{(4)} w(b,t)db \\ &= \theta_k(b)\partial_b^3 w(b,t)\Big|_{\partial D} - \theta_k'(b)\partial_b^2 w|_{\partial D} + p_k\widehat{w}(k,t).\end{aligned} \tag{32}$$

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

$$\mathcal{H}_k(t) \coloneqq \theta_k(b)\partial_b^3 w(b,t)\Big|_{\partial D} = -\theta_k(-1)\left[\sum_{i=1}^{s-1} a_i(t)\chi(b_i(t)) - f'(-1)\right]. \tag{33}$$

$$\mathcal{J}_k(t) \coloneqq -\theta_k'(b)\partial_b^2 w(b,t)\Big|_{\partial D} = -\theta_k'(-1)\left[\sum_{i=1}^{s-1} \chi(b_i(t))a_i(t)(1+b_i(t)) + f(-1)\right]. \tag{34}$$

Differentiating (32) in time and using (27), (28), and (43), we have

$$\begin{aligned}&p_m\partial_t\widehat{w}(m,t) + \mathcal{H}_m'(t) + \mathcal{J}_m'(t) = \sum_{k=1}^{\infty}\Theta_k'(t)\widehat{\phi_k}(m) \\ &= -n\sum_{k=1}^{\infty}\widehat{\phi_k}(m)\left(\int_{D^\varepsilon} w(b,t)\phi_k(b)\mu_0(b,t)db + \int_{D^\varepsilon}\partial_b w(b,t)\phi_k'(b)\mu_2(b,t)db\right),\end{aligned} \tag{35}$$

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

$$\begin{aligned}\mu_0(b,t) &= \frac{1}{n}\sum_{i=1}^n \delta(b-b_i(t)), \\ \mu_2(b,t) &= \frac{1}{n}\sum_{i=1}^n |a_i(t)|^2\delta(b-b_i(t)).\end{aligned} \tag{36}$$

Apply the equality $\sum_{k=1}^{\infty}\phi_k(x)\phi_k(y) = \delta(x-y)$ to (35), we find

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

Similarly,

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

Therefore (35) can be further reduced to

$$\partial_t \widehat{w}(m,t) = -\frac{1}{p_m}(\mathcal{H}_m'(t) + \mathcal{J}_m'(t)) - \frac{n}{p_m} \widehat{w\mu_0}(m,t) - \frac{n}{p_m} \int_D \partial_b w(b',t) \mu_2(b',t) \theta_m'(b') db'. \quad (37)$$

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 24. *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*

$$|\widehat{w}(m,t)| > \frac{1}{2}|\widehat{w}(m,0)|, \quad 0 \le t \le \frac{p_m |\widehat{w}(m,0)|}{2n\widetilde{C}(m+1)}. \quad (38)$$

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 (37). In particular we have $\|\theta_k\|_{L^\infty(D)} = \mathcal{O}(1)$, $\|\theta_k'\|_{L^\infty(D)} = \mathcal{O}(k)$ from Lemma 30.

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

$$\begin{aligned}
\left| -\frac{n}{p_m} \widehat{w\mu_0}(m,t) \right| &= \left| \frac{n}{p_m} \int_D w(b',t) \mu_0(b',t) \theta_m(b') db' \right| \\
&\le \frac{n}{p_m} \|w(\cdot,t)\|_{C(D)} \|\theta_m\|_{C(D)} \le \frac{Cn}{p_m}.
\end{aligned} \quad (39)$$

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

$$\begin{aligned}
\left| -\frac{n}{p_m} \int_D \partial_b w(b',t) \mu_2(b',t) \theta_m'(b') db' \right| &\le \frac{n}{p_m} \|\theta_m'\|_{C(D)} \|\partial_b w\|_{C(D)} \int_D \mu_2(b,t) db \\
&\le \frac{C' M m n}{p_m}.
\end{aligned}$$

3. Using the definition of $\mathcal{H}_m$ in (33) and $\mathcal{J}_m$ in (34), 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

$$|\mathcal{H}_m(t) - \mathcal{H}_m(0)| = |\theta_m(-1)| \left| \sum_{i=1}^{s(t)-1} a_i(t)\chi(b_l(t) \right| \le C''\sqrt{M}|s(t)-1| \tag{40}$$

and

$$|\mathcal{J}_m(t) - \mathcal{J}_m(0)| = |\theta_m'(-1)| \left| \sum_{i=1}^{s(t)-1} a_i(t)\chi(b_i(t)(1+b_i(t)) \right| \le C'''m\sqrt{M}|s(t)-1|. \tag{41}$$

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

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

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

$$\begin{aligned} |\widehat{w}(m,t)| - |\widehat{w}(m,0)| &\ge -\frac{n}{p_m}(C'Mm + C)t - \frac{1}{p_m}\left(C''\sqrt{M} + C'''m\sqrt{M}\right)\frac{1}{2}Knt \\ &\ge -\frac{\widetilde{C}n}{p_m}(m+1)t. \end{aligned} \tag{42}$$

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 $\mathcal{O}(\frac{c'm}{n})$. ∎

In order 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 18.

4 Rashomon Set for Two-layer `ReLU` Neural Networks

In (Semenova et al., 2019, 2022), 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 $\mathcal{R}_\varepsilon \subseteq \mathbb{R}^m$ for the two-layer neural network class $\mathcal{H}_n$ is defined as the following

$$\mathcal{R}_\varepsilon \coloneqq \{h \in \mathcal{H}_n \mid \|h - f\|_{L^2(D)} \leq \varepsilon \|f\|_{L^2(D)}\},$$

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` neural networks

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

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 an admissible function with an error no more than $\mathcal{O}(\sqrt{d + \log n}\; n^{-1/2-1/d})$ (Barron and Klusowski, 2018). 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 25. *Suppose $f \in C(D)$ such that there exists $g \in C_0^2(D)$ that $\Delta g = f$, then the Rashomon set $\mathcal{R}_\varepsilon \subseteq \mathcal{H}$ satisfies*

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

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

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

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

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

Then use Hoeffding's inequality,

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

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

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

Using Green's formula,

$$\langle \Delta h(\boldsymbol{x}), g(\boldsymbol{x})\rangle - \langle h(\boldsymbol{x}), \Delta g(\boldsymbol{x})\rangle = \int_{\partial D} (\partial_n h(\boldsymbol{x}) f(\boldsymbol{x}) - \partial_n f(\boldsymbol{x}) h(\boldsymbol{x}))\, ds = 0,$$

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 neural networks 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 neural network 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 14 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.

Acknowledgment

H. Zhao was partially supported by NSF grants DMS-2309551, and DMS-2012860. Y. Zhong was partially supported by NSF grant DMS-2309530 and H. Zhou was partially supported by NSF grant DMS-2307465.

References

Chenglong Bao, Qianxiao Li, Zuowei Shen, Cheng Tai, Lei Wu, and Xueshuang Xiang. Approximation analysis of convolutional neural networks. *Semantic Scholar e-Preprint*, art. Corpus ID: 204762668, 2019.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993. ISSN 0018-9448. DOI: 10.1109/18.256500.

Andrew R. Barron and Jason M. Klusowski. Approximation and estimation for high-dimensional deep learning networks. *arXiv e-prints*, art. arXiv:1809.03090, September 2018.

George D Birkhoff. Boundary value and expansion problems of ordinary linear differential equations. *Transactions of the American Mathematical Society*, 9(4):373–395, 1908.

Mikhail Shlemovich Birman and Mikhail Zakharovich Solomyak. Asymptotic behavior of the spectrum of weakly polar integral operators. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 34(5):1142–1158, 1970.

Johan M Bogoya, Albrecht Böttcher, and Alexander Poznyak. Eigenvalues of hermitian toeplitz matrices with polynomially increasing entries. *Journal of Spectral Theory*, 2(3): 267–292, 2012.

Leo Breiman. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory*, 39(3):999–1013, 1993.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

Jingrun Chen, Xurong Chi, Zhouwang Yang, et al. Bridging traditional and machine learning-based algorithms for solving pdes: The random feature method. *arXiv preprint arXiv:2207.13380*, 2022.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: `https://proceedings.neurips.cc/paper/2019/file/fd95ec8df5dbeea25aa8e6c808bad583-Paper.pdf`.

Charles K Chui and Xin Li. Approximation by ridge functions and neural networks with one hidden layer. *Journal of Approximation Theory*, 70(2):131–141, 1992.

Charles K. Chui, Shao-Bo Lin, and Ding-Xuan Zhou. Construction of neural networks for realization of localized deep learning. *Frontiers in Applied Mathematics and Statistics*, 4: 14, 2018. ISSN 2297-4687. DOI: 10.3389/fams.2018.00014.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989. DOI: 10.1007/BF02551274.

Carles Domingo-Enrich and Youssef Mroueh. Tighter sparse approximation bounds for relu neural networks. *arXiv preprint arXiv:2110.03673*, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

Georg Frobenius. Über matrizen aus nicht negativen elementen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, 23:456–477, 1912.

Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. *Constructive Approximation*, 55:259–367, 2022.

Tamara G Grossmann, Urszula Julia Komorowska, Jonas Latz, and Carola-Bibiane Schönlieb. Can physics-informed neural networks beat the finite element method? *arXiv preprint arXiv:2302.04107*, 2023.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(05):803–859, 2020. DOI: 10.1142/S0219530519410021.

Sigurdur Helgason. *Groups and geometric analysis: integral geometry, invariant differential operators, and spherical functions*, volume 83. American Mathematical Society, 2022.

Sigurdur Helgason et al. *Integral geometry and Radon transforms*. Springer, 2011.

David Holzmüller and Ingo Steinwart. Training two-layer relu networks with gradient descent is inconsistent. *arXiv preprint arXiv:2002.04861*, 2020.

Qingguo Hong, Qinyang Tan, Jonathan W Siegel, and Jinchao Xu. On the activation function dependence of the spectral bias of neural networks. *arXiv preprint arXiv:2208.04924*, 2022.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Yuling Jiao, Yanming Lai, Xiliang Lu, Fengru Wang, Jerry Zhijian Yang, and Yuanyuan Yang. Deep neural networks with ReLU-Sine-Exponential activations break curse of dimensionality on hölder class. *arXiv e-prints*, art. arXiv:2103.00542, February 2021.

Lennard Kamenski, Weizhang Huang, and Hongguo Xu. Conditioning of finite element equations with arbitrary anisotropic meshes. *Mathematics of Computation*, 83(289): 2187–2211, 2014.

Jason M Klusowski and Andrew R Barron. Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000. ISSN 13507265. URL: `http://www.jstor.org/stable/3318636`.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in neural information processing systems*, 31, 2018.

Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer relu neural networks beyond ntk. *arXiv preprint arXiv:2007.04596*, 2020.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021. DOI: 10.1137/20M134695X.

Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the frequency principle for general deep neural networks. *arXiv preprint arXiv:1906.09235*, 2019.

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

M.A Naimark. *Linear Differential Operators*. Frederick Ungar, New York, 1967.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21 (174):1–38, 2020. URL: `http://jmlr.org/papers/v21/20-002.html`.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv preprint arXiv:2006.12297*, 2020.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

Jooyoung Park and Irwin W Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991.

JB Reade. Eigenvalues of positive definite kernels. *SIAM Journal on Mathematical Analysis*, 14(1):152–157, 1983.

JB Reade. Eigenvalues of positive definite kernels ii. *SIAM journal on mathematical analysis*, 15(1):137–142, 1984.

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

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International conference on machine learning*, pages 2979–2987. PMLR, 2017.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019. ISSN 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2019.07.011.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020. ISSN 1991-7120. DOI: 10.4208/cicp.OA-2020-0149.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*, 33(4): 1005–1036, 03 2021a. ISSN 0899-7667. DOI: 10.1162/neco_a_01364.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 141:160–173, 2021b. ISSN 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2021.04.011.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *Journal of Machine Learning Research*, 23(276):1–60, 2022a. URL: `http://jmlr.org/papers/v23/21-1404.html`.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Neural network architecture beyond width and depth. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5669–5681. Curran Associates, Inc., 2022b. URL: `https://proceedings.neurips.cc/paper_files/paper/2022/file/257be12f31dfa7cc158dda99822c6fd1-Paper-Conference.pdf`.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Marshall H Stone. A comparison of the series of fourier and birkhoff. *Transactions of the American Mathematical Society*, 28(4):695–761, 1926.

Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. *Advances in Neural Information Processing Systems*, 32, 2019.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=H1ebTsActm.

Maksim Velikanov and Dmitry Yarotsky. Explicit loss asymptotics in the gradient descent training of neural networks. *Advances in Neural Information Processing Systems*, 34: 2570–2582, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Harold Widom. On the eigenvalues of certain hermitian operators. *Transactions of the American Mathematical Society*, 88(2):491–522, 1958.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. ISSN 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2017.07.002.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649. PMLR, 06–09 Jul 2018. URL: http://proceedings.mlr.press/v75/yarotsky18a.html.

Guodong Zhang, James Martens, and Roger B Grosse. Fast convergence of natural gradient descent for over-parameterized neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Shijun Zhang. Deep neural network approximation via function compositions. *PhD Thesis, National University of Singapore*, 2020. URL: https://scholarbank.nus.edu.sg/handle/10635/186064.

Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. On enhancing expressive power via compositions of single fixed-size relu network. *arXiv e-prints*, art. arXiv:2301.12353, January 2023. DOI: 10.48550/arXiv.2301.12353.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020. ISSN 1063-5203. DOI: https://doi.org/10.1016/j.acha.2019.06.004.

Appendix A. Properties of Eigenfunctions

We show properties of the eigenfunctions ϕ_k for the Gram kernel defined by (26).

Lemma 26. *ϕ_k is a cubic polynomial over $(-1-\varepsilon,-1)$ and*

$$\partial_b^4\phi_k(b)=\begin{cases}0, & b\in(-1-\varepsilon,-1)\\ \lambda_k^{-1}\phi_k(b), & b\in(-1,1)\end{cases} \tag{43}$$

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

Appendix B. Generalized Fourier Modes

In this section, we study the generalized Fourier modes θ_k defined in (30)

Lemma 27. *The eigenfunction θ_k satisfies*

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

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 (3).

Lemma 28. *$\{\theta_k\}_{k\geq 1}$ forms an orthonormal basis of $L^2(D)$.*

Proof. Since each eigenvalue is simple from Corollary 2 and we verify the following equality using integration by parts,

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

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

Lemma 29. *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 (3), note $\mathcal{G}(x,y) = \mathcal{G}(y,x)$, it is self-adjoint, then by Hilbert-Schmidt theorem,

$$\int_D \mathcal{G}(x,y)\gamma(y)dy = 0 \tag{44}$$

and we differentiate the above equation 4 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 with a shifted frequency.

Theorem 30. *The following statements hold*

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

Proof. We only prove for θ_{2k+1}, the proof is similar for θ_{2k+2}. Using the symmetry of the domain D,

$$\begin{aligned}
&\int_D \left(-\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\sinh(c_{2k+1}x)+\cos(c_{2k+1}x)\right)^2 dx \\
&= \int_D \left(\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\sinh(c_{2k+1}x)\right)^2 dx + \int_D \cos^2(c_{2k+1}x)dx \\
&\quad - 2\int_D \frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\sinh(c_{2k+1}x)\cos(c_{2k+1}x)dx \\
&= \left(\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\right)^2\left[\frac{\sinh(2c_{2k+1})}{2c_{2k+1}}-1\right]+\left[\frac{\sin(2c_{2k+1})}{2c_{2k+1}}+1\right] \\
&= \frac{\cos^2(c_{2k+1})\coth(c_{2k+1})}{c_{2k+1}}+\frac{\sin(2c_{2k+1})}{2c_{2k+1}}+1-\left(\frac{\cos(c_{2k+1})}{\sinh(c_{2k+1})}\right)^2 \\
&= 1+\mathcal{O}(k^{-1}).
\end{aligned} \tag{45}$$

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 (45) that

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

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

Appendix C. Improved Bounds for Learning Dynamics

C.1 Part I

In the first step, we provide an improved estimate for

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

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,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \partial_b w(b_i(t),t)|a_i(t)|^2\theta'_m(b_i(t)) - \int_D \partial_b w(b,t)\mathcal{A}(b,t)\theta'_m(b)db \\
&= -\int_D \mathtt{disc}(b,t)\partial_b\left[\partial_b w(b,t)\mathcal{A}(b,t)\theta'_m(b)\right]db,
\end{aligned}$$

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 31. *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 $\mathcal{A}$:

$$V(\mathcal{A}) = a_1^2(t) + \sum_{i=1}^{n-1}|a_i^2(t)-a_{i+1}^2(t)| + a_n^2(t). \tag{46}$$

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 32. *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*

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

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

$$\begin{aligned}
&\left|\int_D \partial_b w(b,t)\mathcal{A}(b,t)\theta_m'(b)db\right| \\
&\le \left|\int_D \theta_m(b)\partial_b\left(\partial_b w(b,t)\mathcal{A}(b,t)\right)db\right| + \left|\theta_m(b)\partial_b w(b,t)\mathcal{A}(b,t)\Big|_{\partial D}\right| \\
&\le L\left[\int_D |\partial_b^2 w(b,t)\mathcal{A}(b,t)|db + \int_D |\partial_b w(b,t)||\partial_b\mathcal{A}(b,t)|db + M\|\partial_b w(\cdot,t)\|_{C(D)}\right] \\
&\le L\left[\|\partial_b^2 w(\cdot,t)\|_{L^2(D)}\|\mathcal{A}(\cdot,t)\|_{L^2(D)} + \|\partial_b w(\cdot,t)\|_{C(D)}V(\mathcal{A}) + M\|\partial_b w(\cdot,t)\|_{C(D)}\right] \\
&\le C'\|h(\cdot,t)-f(\cdot)\|_{L^2(D)}(M+M').
\end{aligned} \tag{47}$$

Therefore we have a new estimate:

$$\begin{aligned}
\left|\int_D \partial_b w(b,t)\mu_2(b,t)\theta_m'(b)db\right| &\le C'\|h(\cdot,t)-f(\cdot)\|_{L^2(D)}(M+M') \\
&\quad + \left|\int_D \texttt{disc}(b,t)\partial_b\left[\partial_b w(b,t)\mathcal{A}(b,t)\theta_m'(b)\right]db\right|.
\end{aligned}$$

Notice that this bound will be better than the constant bound in Theorem 24 if the second term on the right-hand side is relatively small. Next, we characterize the discrepancy function $\texttt{disc}(b,t)$.

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

Because

$$\left|\frac{d}{dt}b_i(t)\right| \le |a_i(t)|\int_D |h(x,t)-f(x)||\partial_b\psi(x,b_i(t))|dx \le K \coloneqq C'''\sqrt{M}\|h(x,0)-f(x)\|_{L^2(D)}. \tag{48}$$

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

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

■

Remark 33. *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 24. In fact, if $V(\mathcal{A})$ becomes $\mathcal{O}(m)$, we may also have to return to the Theorem 24.*

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 34. *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 $|\frac{d}{dt}b_i(t)| \le K \coloneqq C'''\sqrt{M}\|h(\cdot,0) - f(\cdot)\|_{L^2(D)}$ (see (48)). 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

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

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 35. *Suppose* $\sup_{1\le i\le n}|a_i(t)|^2$ *is uniformly bounded by* M*, the total variation of* $\mathcal{A}$ (46) *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 32, there exists a constant $C'' > 0$ that

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

We only consider the case that $n \ge m$, otherwise we return to Theorem 24. Recall that in Theorem 24 that $|\widehat{w\mu_0}(m,t)|$ is uniformly bounded (see (39)) and $|\mathcal{H}_m(t) - \mathcal{H}_m(0)| \le \frac{1}{2}|\theta_m(-1)|\sqrt{M}Knt$ (see (40) and (41)). Combine the estimate in Theorem 34 and follow the same process as (42), we can find a constant $C''' > 0$ that

$$|\widehat{w}(m,t)| - |\widehat{w}(m,0)| \le -C'''\frac{n}{p_m}\left(\left(1 + \frac{m^2}{n}\right)t + (m+m^2)t^2\right). \tag{49}$$

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

$$\frac{1}{2}|\widehat{w}(m,0)| = \frac{n}{p_m}C'''\left(\left(1 + \frac{m^2}{n}\right)\tau + (m+m^2)\tau^2\right). \tag{50}$$

The solution satisfies

$$\tau \ge \frac{p_m}{2C'''n}\frac{|\widehat{w}(m,0)|}{\sqrt{(1+m^2/n)^2 + 2(m+m^2)p_m|\widehat{w}(m,0)|/(C'''n)}}. \tag{51}$$

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 24. ■

Remark 36. *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* (37) *without the boundary terms and the last term involving θ'_m. Since $w(b,t)$ is H^2, one gets $\widehat{\mu_0 w}(m,t) \leq C(\frac{1}{m^2} + m\,disc(\{b_i(0)\}_{i=1}^n))$, for $n \succ m^3$ and equispaced $\{b_i\}_{i=1}^n$ that $disc(\{b_i(0)\}_{i=1}^n) = \mathcal{O}(\frac{1}{n})$, it takes $\mathcal{O}(m^4)$ steps to reduce the initial error in mode m by half. Thus it takes at least $\mathcal{O}(m^3)$ 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 16 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. $a_i(0) = \frac{1}{2}\cos(i)$.

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

$$E_k(t) = \left|\int_D (h_k(x,t) - f(x))f(x)dx\right|.$$

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 4 and 5. 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 10, 11, 12 for initialization (A) and Fig 13, 14, 15 for initialization (B). One possible explanation is the overestimate using the total variation in (47) which might be unnecessary. As we mentioned in Remark 36, we record the training dynamics with *fixed* biases, initialization (A), and choose $n = k^4$, see Fig 16. The result matches the argument in Remark 36. 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 16.

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

Table 4: 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 5: 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	$\mathcal{O}(n)$	$\mathcal{O}(k^{1.59})$	$\mathcal{O}(k^{1.75})$	$\mathcal{O}(k^{1.90})$

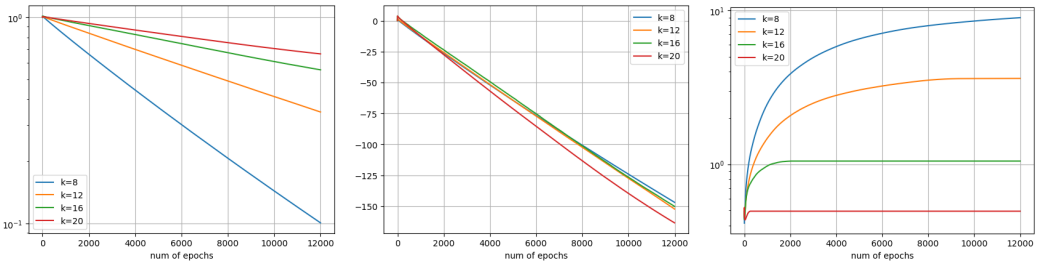


Figure 10: Experiment for initialization (A) and $\beta = 4$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.87} \ln(E_k(t))$. Right: the total variations of $\mathcal{A}$.

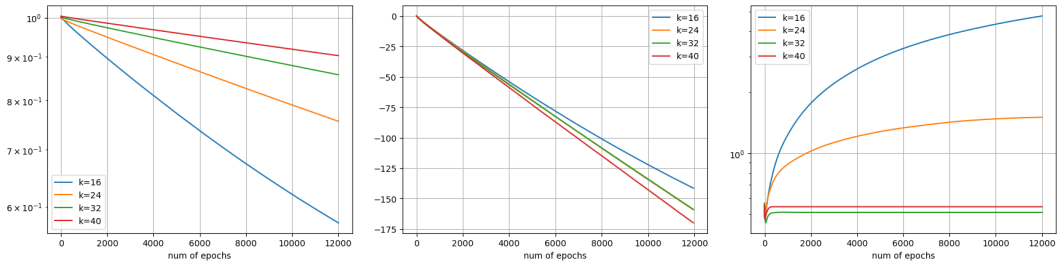


Figure 11: Experiment for initialization (A) and $\beta = 3$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.82} \ln(E_k(t))$. Right: the total variations of $\mathcal{A}$.

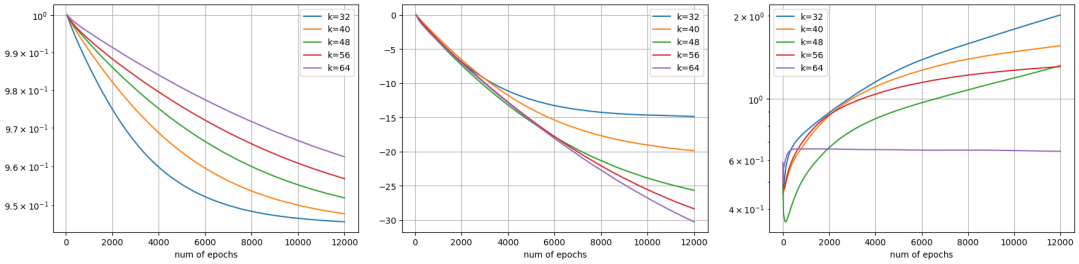


Figure 12: Experiment for initialization (A) and $\beta = 2$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.61} \ln(E_k(t))$ (fitting first 2000 epochs only). Right: the total variations of $\mathcal{A}$.

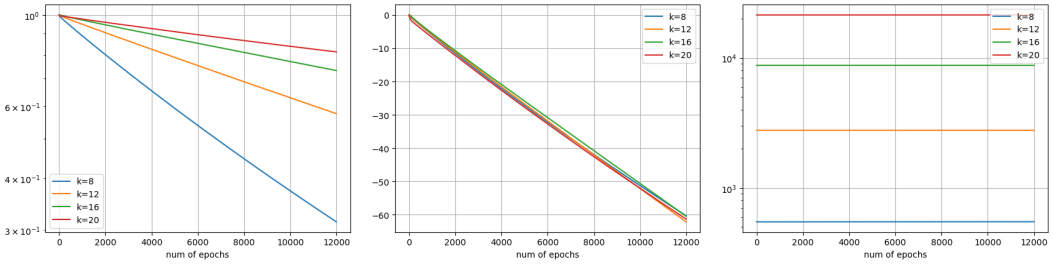


Figure 13: Experiment for initialization (B) and $\beta = 4$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.90} \ln(E_k(t))$ (fitting first 2000 epochs only). Right: the total variations of $\mathcal{A}$.

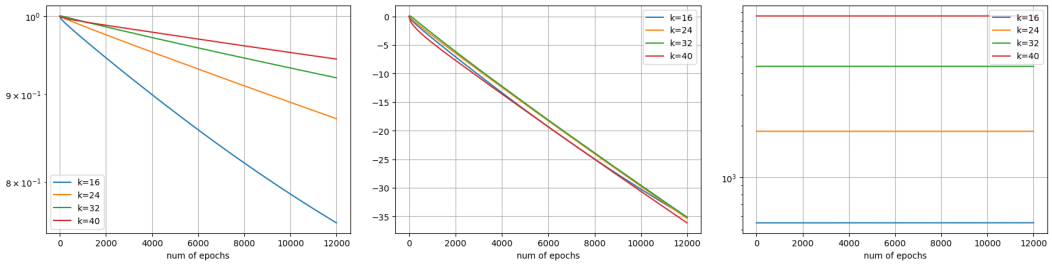


Figure 14: Experiment for initialization (B) and $\beta = 3$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.75} \ln(E_k(t))$. Right: the total variations of $\mathcal{A}$.

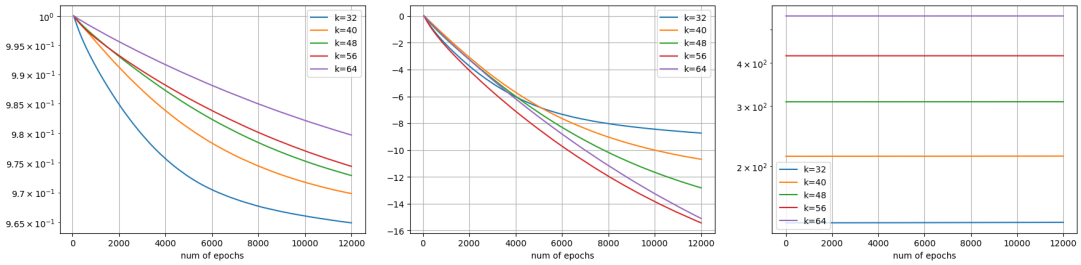


Figure 15: Experiment for initialization (B) and $\beta = 2$. Left: the graphs of $E_k(t)$. Middle: the graphs of $k^{1.59} \ln(E_k(t))$ (fitting first 2000 epochs only). Right: the total variations of $\mathcal{A}$.

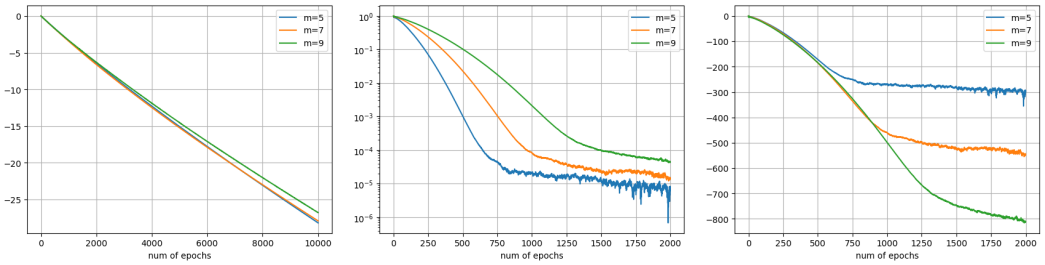


Figure 16: Additional experiment for initialization (A) and $\beta = 4$. Left: Graphs of $k^{3.5} \ln(E_k(t))$ for $k = 5, 7, 9$ with *fixed* equispaced biases, trained by `GD`. Middle: Graphs of $E_k(t)$ for $k = 5, 7, 9$, trained by `Adam`. Right: Graphs of $k^2 \ln E_k(t)$ for $k = 5, 7, 9$, trained by `Adam`.

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:

$$\begin{aligned} |\mathcal{H}_m(t) - \mathcal{H}_m(0)| &\le \frac{1}{2}|\theta_m(-1)|\sqrt{M}(nKt + \mathcal{O}_p(\sqrt{nKt})), \\ |\mathcal{J}_m(t) - \mathcal{J}_m(0)| &\le |\theta_m'(-1)|\sqrt{M}Kt(Knt + \mathcal{O}_p(\sqrt{Knt})), \\ \left| \int_D \partial_b w(b', t)\mu_2(b', t)\theta_m'(b')db' \right| &\le C'' \left(1 + \left(t + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right)\right) m^2\right), \end{aligned}$$

and (49) 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 35. 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 $\frac{1}{p!}\sigma^p(x)$, $p \geq 1$, 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 35, the *half-reduction* time of "frequency" m is at least $\mathcal{O}(\frac{m^{2p+2}}{n}|\widehat{w}(m,0)|)$ in the Theorem 35. This also implies that using a shallow neuron network with smooth activation functions will be even worse to learn and approximate 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 $\texttt{ReLU}^p(x) \coloneqq \frac{1}{p!}\sigma^p(x)$, $p = 1, 2$ to train the shallow neuron network $h_{m,p}(x,t)$ to approximate $f(x)$, respectively. The number of neurons $n = 10^4$. We record the the *error* of the Fourier mode

$$E_{m,p}(t) \coloneqq \left| \int_D (h(x,t) - f(x)) f(x) dx \right|$$

at each iteration. The errors are shown in Figure 17 for $m = 5, 7, 9$. We can observe that the decay rates of `ReLU` and $\texttt{ReLU}^2$ are $\mathcal{O}(m^{-2})$ and $\mathcal{O}(m^{-3})$ roughly.

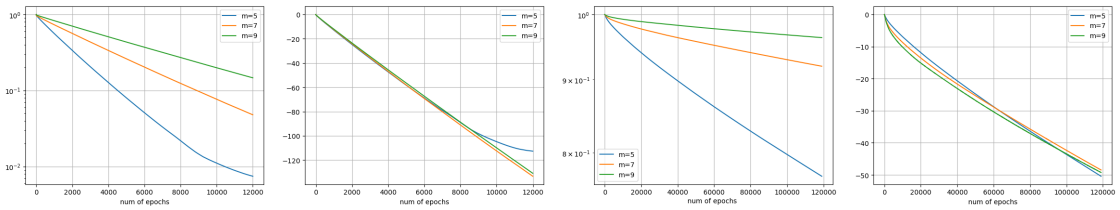


Figure 17: The comparison of `ReLU`, $\texttt{ReLU}^2$ for the approximation to $f(x) = \sin(m\pi x)$. From left to right, the figures are $E_{m,1}(t)$, $m^{1.95}\ln(E_{m,1}(t))$, $E_{m,2}(t)$, $m^{3.25}\ln(E_{m,2}(t))$.

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 24 or Theorem 35. When $n \succ m$ in Theorem 24 or $n \succ m^2$ in Theorem 35, such requirement can be relaxed to only the initial condition $\sup_{i\geq 1}|a_i(0)|^2 \leq M$.

Appendix 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 25.

Theorem 37. *Assuming the same network structure as Theorem 25 and σ as a bounded activation function that*

$$-1 \leq \sigma(t) \leq 1, \quad \sigma'(t) > 0, \quad \forall\, t \in \mathbb{R}, \tag{52}$$

then the Rashomon set's measure

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

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

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

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

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

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

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

Since the random variable a_i is sub-Gaussian, then X_i is also sub-Gaussian by $|X_i| \leq 2a_i\ell\kappa$. One can apply the Hoeffding's inequality (see Theorem 2.6.2 (Vershynin, 2018)) that

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

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 (52) can not feel oscillations in f, i.e., $\langle f, \sigma\rangle$ is small due to cancellation. According to the heuristic argument in (Semenova et al., 2022), it also implies that it is relatively difficult to find a shallow neural network with activation function (52) that can approximate highly oscillatory functions well. Or in other words, the optimal set of parameters only occupies an extremely small measure of the parameter space for highly oscillatory functions.

Appendix E. Further Discussions

E.1 General Case of Gram Matrix of Two-layer `ReLU` Networks in 1D

When the two-layer `ReLU` network in 1D is

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

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

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

where the sub-kernels are

$$\begin{aligned}
\mathcal{G}^{++}(x,y) &= \frac{1}{24}(2 - x - y - |x-y|)^2(2 - x - y + 2|x-y|) \\
\mathcal{G}^{--}(x,y) &= \frac{1}{24}(2 + x + y - |x-y|)^2(2 + x + y + 2|x-y|) \\
\mathcal{G}^{+-}(x,y) &= \frac{1}{48}\left[|x-y| - (x-y)\right]^3, \\
\mathcal{G}^{-+}(x,y) &= \frac{1}{48}\left[|x-y| + (x-y)\right]^3.
\end{aligned}$$

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:

1. 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.
2. 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 38. *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 39. *The kernel $K_\alpha = |x-y|^\alpha$ has been studied in (Bogoya et al., 2012) 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*

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

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*

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

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 differentiation technique in Lemma 1, we arrive at the same form:

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

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 Lemma 1 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:

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

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

Theorem 40. *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* k*-th eigenvalue of* $\mathcal{G}_\alpha$*.*

Proof. Let the kernel $\mathcal{G}_\alpha(x,y) \coloneqq \mathcal{G}(x,y) + \alpha^2\mathcal{G}(-x,-y) - \frac{\alpha}{6}|x-y|^3$, then following the same idea as Lemma 1, 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 $w_{\alpha,k}$:

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

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

$$
\begin{aligned}
P_{0,\alpha}(x) &= \frac{3}{2} + 6\alpha^2 - 6\alpha^3 + \frac{3}{2}\alpha^4 + A_\alpha(x) + B_\alpha(x) \\
P_{1,\alpha}(x) &= P_{0,\alpha}(x) - 3 \\
P_{2,\alpha}(x) &= -3\alpha^2(\alpha^2 - 3\alpha + 3)x + 2\alpha^2(\alpha - 2)(2 - 9\alpha^2 + 6\alpha^3)x^3 \\
Q_{0,\alpha}(x) &= 2\alpha^2 x(-18 + 15\alpha - 42\alpha^2 - 57\alpha^3 + 18\alpha^4 + (12\alpha^2 - 14\alpha^3 + 4\alpha^4)x^2) \\
Q_{1,\alpha}(x) &= 2\alpha^2 x(9 - 6\alpha + 45\alpha^2 - 18\alpha^4 + (8 + 4\alpha + 24\alpha^2 - 28\alpha^3 + 8\alpha^4)x^2) \\
Q_{2,\alpha}(x) &= -12\alpha^2 x^2(-4 + 3\alpha - 10\alpha^2 + 13\alpha^2 + 4\alpha^4) \\
R_{0,\alpha}(x) &= \frac{1}{2}\left[3 - 24\alpha^2 + 12\alpha^3 + 111\alpha^4 - 144\alpha^5 + 48\alpha^6\right] - A_\alpha(x) + B_\alpha(x) \\
R_{1,\alpha}(x) &= \frac{1}{2}\left[-3 + 48\alpha^2 - 36\alpha^2 - 105\alpha^4 + 144\alpha^5 - 48\alpha^6\right] - A_\alpha(x) + B_\alpha(x) \\
R_{2,\alpha}(x) &= \alpha^2(27 - 21\alpha - 87\alpha^2 + 114\alpha^3 - 36\alpha^4)x + \alpha^2(8 - 4\alpha - 12\alpha^2 + 14\alpha^3 - 4\alpha^4)x^2.
\end{aligned}
$$

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 7. It shows that the kth eigenvalue grows as $\mathcal{O}((\alpha - 1)^2 k^{-4})$ when k is sufficiently large.

Corollary 41. *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 k-th 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 (Frobenius, 1912) (or Krein-Rutman theorem for positive compact operators), one can estimate both $\mu_{\alpha,1}$ from below.

Corollary 42. *For any $\alpha \in (0, 1)$, the leading eigenvalue $\mu_{\alpha,1} \in [0.941, 2.754]$.*

Proof. First, we compute that

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

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

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

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

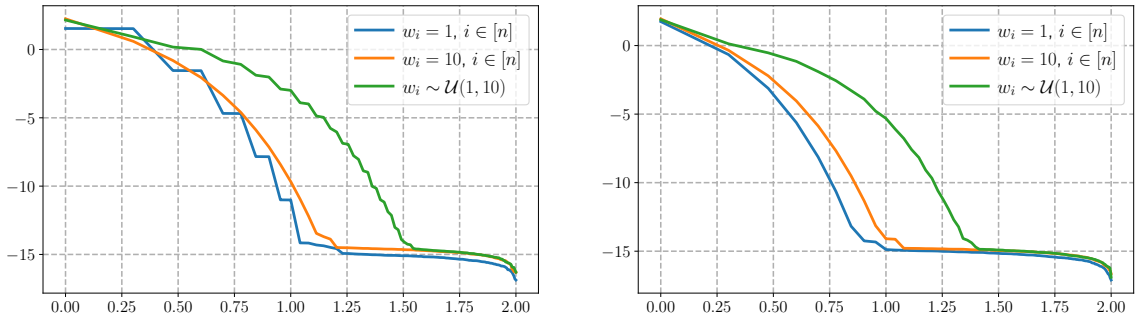
For analytic activation functions such as `Tanh` or `Sigmoid`, the Gram matrix is formed by

$$G_{ij} = \int_D \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} - b_i)\sigma(\boldsymbol{w}_j \cdot \boldsymbol{x} - b_j)d\boldsymbol{x}$$

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 $\boldsymbol{z} \coloneqq (\boldsymbol{w}, -1) \in \mathbb{R}^{d+1}$. Since the kernel is analytic in both $\boldsymbol{x}$ and $\boldsymbol{y}$, the eigenvalues of the kernel are decaying faster than any polynomial rate (Reade, 1983, 1984). Two examples in 1D are provided in Figure 18.



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

Figure 18: 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 $\mathcal{U}(1, 10)$.