UNIVERSITY OF CALIFORNIA
Los Angeles

On the Spectral Bias of Neural Networks in the Neural Tangent Kernel Regime

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Benjamin Bowman

2023

ABSTRACT OF THE DISSERTATION

On the Spectral Bias of Neural Networks in the Neural Tangent Kernel Regime

by

Benjamin Bowman
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2023
Professor Guido Francisco Montúfar Cuartas, Chair

Understanding the training dynamics of neural networks is quite difficult in general due to the highly nonlinear nature of the parameterization. A breakthrough in the theory of deep learning was the finding that in the infinite-width limit the gradient descent dynamics are characterized by a fixed kernel, coined the "Neural Tangent Kernel" (NTK). In this limiting regime the network is biased to learn the eigenvectors/eigenfunctions of the NTK at rates corresponding to their eigenvalues, a phenomenon known as "spectral bias". Considerable work has been done comparing the training dynamics of finite-width networks to the idealized infinite-width dynamics. These works typically compare the dynamics of a finite-width network to the dynamics of an infinite-width network where both networks are optimized via the empirical risk. In this work we compare a finite-width network trained on the empirical risk to an infinite-width network trained on the population risk. Consequentially, we are able to demonstrate that the finite-width network is biased towards learning the top eigenfunctions of the NTK over the entire input domain, as opposed to describing the dynamics merely on the training set. Furthermore we can demonstrate that this holds in a

regime where the network width is on the same order as the number of training samples, in contrast with prior works that require the unrealistic assumption that the network width is polynomially large in the number of samples. In a separate line of analysis, we characterize the spectrum of the NTK by expressing the NTK as a power series. We demonstrate that the NTK has a small number of large outlier eigenvalues and that the number of such eigenvalues is largely inherited from the structure of the input data. As a result we shed further insight into why the network places a preference on learning a small number of components quicker. In total, our results help classify the properties networks are biased towards in a variety of settings, which we hope will lead to more interpretable artificial intelligence in the long term.

The dissertation of Benjamin Bowman is approved.

Arash Ali Amini

Deanna M. Hunter

Stefano Soatto

Guido Francisco Montúfar Cuartas, Committee Chair

University of California, Los Angeles

2023

This work is dedicated to Wikipedia editors, contributors to online forums, contributors to open source code, and anyone who worked anonymously in the support of knowledge.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not be possible without the support of more people than I could ever hope to enumerate. I thank my high school teachers Mr. Staines and Mr. Miller for encouraging my interest in mathematics. I thank my college professor Carina Curto for encouraging me to further pursue higher mathematics and giving me the confidence to apply to graduate school, as well as helpful advice. If it were not for her encouragement I would not have had the confidence to complete this journey. I would like to thank the admin at Penn State New Kensington for going out of their way to offer me a scholarship despite the fact that I had applied after the deadline, which made the cost of attendance comparable with community college. I thank Jill Anderson for her support of my undergrad studies through the Kermit C. Anderson memorial scholarship, established by Jill in honor of her husband who died in the September 11 attacks on the World Trade Center. I thank my advisor Guido Montúfar for his guidance during my research. Guido works very hard to help his students and to promote their development as researchers. He gives his students a lot of freedom and treats them as researchers on equal footing. I thank Yonatan Dukler, Hui Jin, Pradeep Banerjee, Michael Murray, Johannes Müller for their friendship and helpful research discussions at UCLA and MPI. I thank Alessandro Achille, Stefano Soatto, Aditya Golatkar, Matthew Trager, Luca Zancato, Pramuditha Perera, and Yonatan Dukler for their collaboration and assistance during my internship at Amazon. I thank the committee members Arash A. Amini, Deanna Needell, Stefano Soatto, and Guido Montúfar for volunteering their time and guidance. I thank my family and friends for their love and support, which has provided the foundation that makes my life worthwhile.

Below I outline which chapters are based on published work, their bibliographic information, as well as the contributions of the coauthors.

- Chapter 2 is based on the following manuscript: Benjamin Bowman and Guido Montúfar. "Implicit Bias of MSE Gradient Optimization in Underparameterized Neu-

ral Networks." In *International Conference on Learning Representations*, 2022. Guido Montúfar was the principal investigator for this project. The DOI for the preprint version is `https://doi.org/10.48550/arXiv.2201.04738`

- Chapter 3 is based on the following manuscript: Benjamin Bowman and Guido Montúfar. "Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime." In *Advances in Neural Information Processing Systems*, 2022. This was joint work with Guido Montúfar who was the principal investigator. The DOI for the preprint version is `https://doi.org/10.48550/arXiv.2206.02927`

- Chapter 4 is based on the following manuscript: *Michael Murray, *Hui Jin, *Benjamin Bowman, and Guido Montúfar (*Equal Contribution). Characterizing the Spectrum of the NTK via a Power Series Expansion. In *International Conference on Learning Representations*, 2023. Michael Murray formally derived the NTK power series, contributed experiments on analyzing approximations of the NTK, and analyzed the asymptotic decay of the spectrum for nonuniform distributions. Hui Jin analyzed the asymptotic decay of the spectrum for the uniform distribution as well as providing its experimental validation. Benjamin Bowman (myself) had provided the initial idea of studying the NTK via a power series expansion and analyzed the effective rank of the NTK both theoretically and experimentally. Guido Montúfar was the principal investigator. The DOI for the preprint version is `https://doi.org/10.48550/arXiv.2211.07844`

Below I outline the funding sources that helped make my research program possible, as well as which chapters they pertain to.

- The work for Chapter 2 was completed during a visit to the Max Planck Institute for Mathematics in the Sciences in Leipzig, Germany

- The work contained in Chapters 2, 3, 4 received support from the European Research

Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 757983)

- The work in Chapter 3 received funding from UCLA FCDA
- The work in Chapter 4 was supported by NSF CAREER Grant DMS-214563

VITA

2018	Bachelors of Science, Computational Mathematics, The Pennsylvania State University
2020	Master of Arts, Applied Mathematics, UCLA
2021	Summer Researcher, Max Planck Institute for Mathematics in the Sciences, Leipzig, Saxony, Germany
2022-2023	Applied Scientist Intern, Amazon AWS AI Labs, Pasadena, CA

PUBLICATIONS

Yonatan Dukler*, Benjamin Bowman*, Alessandro Achille*, Aditya Golatkar, Ashwin Swaminathan, Stefano Soatto (*Equal Contribution). SAFE: Machine Unlearning With Shard Graphs. arXiv preprint arXiv:2304.13169, 2023. URL `https://arxiv.org/abs/2304.13169`

Yonatan Dukler, Alessandro Achille, Hao Yang, Varsha Vivek, Luca Zancato, Benjamin Bowman, Avinash Ravichandran, Charless Fowlkes, Ashwin Swaminathan, Stefano Soatto. Introspective Cross-Attention Probing for Lightweight Transfer of Pre-trained Models. arXiv preprint arXiv:2303.04105, 2023. URL `https://arxiv.org/abs/2303.04105`

Benjamin Bowman, Alessandro Achille, Luca Zancato, Matthew Trager, Pramuditha Perera, Giovanni Paolini, Stefano Soatto. À-la-carte Prompt Tuning (APT): Combining Distinct

Data Via Composable Prompting. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Bowman_A-La-Carte_Prompt_Tuning_APT_Combining_Distinct_Data_via_Composable_Prompting_CVPR_2023_paper.html

Michael Murray*, Hui Jin*, Benjamin Bowman*, and Guido Montúfar (*Equal Contribution). Characterizing the Spectrum of the NTK via a Power Series Expansion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Tvms8xrZHyR

Benjamin Bowman and Guido Montúfar. Spectral bias outside the training set for deep networks in the kernel regime. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *The 36th Conference on Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=a01PL2gb7W5

Benjamin Bowman and Guido Montúfar. Implicit bias of MSE gradient optimization in underparameterized neural networks. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=VLgmhQDVBV

OVERVIEW

In the following chapters we provide insight into the phenomenon that neural networks exhibit a preference to learn certain properties of the target function more quickly throughout training. Our analysis centers around the Neural Tangent Kernel (NTK). Specifically, we compare the optimization dynamics of finite-width networks to their corresponding infinite-width limit. We bound the difference between the trajectory of a finite-width network trained on the empirical risk to the trajectory of an infinite-width network trained on the population risk. Consequentially, we are able to demonstrate that finite-width networks exhibit a spectral bias to learn the eigenfunctions of the NTK corresponding to the large eigenvalues more quickly. In a separate line of analysis, we derive a power series expansion for the Neural Tangent Kernel to establish numerous spectral properties. Most notably, we are able to demonstrate that the NTK has a small number of outlier eigenvalues and that the structure of its spectrum is largely inherited from the input data. The fact that the NTK has a small number of outlier eigenvalues, by way of spectral bias, provides further explanation into why networks exhibit a preference to learn a small number of attributes more quickly. In total, we are able to demonstrate spectral bias holds in several distinct settings, and provide insight into the spectral bias phenomenon through various properties of the NTK spectrum. Consequentially, we make steps towards cataloging the inductive bias's of neural networks, which we hope will lead to making neural networks more intelligible in the future. Below, we provide an outline of the results contained in the following chapters.

- In Chapter 1 we introduce the Neural Tangent Kernel and describe how it arises naturally when studying the convergence of gradient descent. We introduce the spectral bias phenomenon as well as the strengths and limitations of NTK analysis.

- In Chapter 2 we introduce the "damped deviations" equation (see Lemma 2.2.3) which compares the gradient flow trajectory of a finite-width network trained on the empir-

ical loss to the corresponding infinite-width network trained on the population loss. Using this equation we can provide bounds on the trajectory whenever the network is underparameterized that are sufficient to obtain bounds on the population loss (see Theorem 2.3.5 and Corollary 2.3.6). A key element of the proof is an NTK deviation bound that holds uniformly over all inputs (see theorems 2.5.19, 2.5.25, 2.5.26), which may be of independent interest. Using a simplified version of the damped deviation equation that restricts to the training set, we can provide corresponding statements for the empirical risk that are applicable in the overparameterized regime (see Theorem 2.3.7, Theorem 2.3.8, and Corollary 2.3.9). The contents of Chapter 2 are derived from the manuscript *"Implicit bias of MSE gradient optimization in underparameterized neural networks"* which appeared in The Tenth International Conference on Learning Representations, 2022. This was joint work with Guido Montúfar.

- In Chapter 3 we expand the results in Chapter 2 to include more realistic sample complexities and deep architectures. In Corollary 3.3.7 we provide a variation of Theorem 2.3.5 without the underparameterization requirement which applies to deep networks with any combination of convolutional, residual, or fully-connected layers. Most notably, the number of samples and the width of the network can scale at the same rate to obtain vanishing bounds up to finite stopping times. Consequently, as demonstrated in Corollary 3.3.10 we can obtain bounds on the population loss in a scaling regime where the width of the network and number of samples scale at the same rate. Consequently, we are able to demonstrate that spectral bias holds for realistic scaling limits and diverse architectures. The contents of Chapter 3 correspond to the manuscript *"Spectral bias outside the training set for deep networks in the kernel regime"* which appeared in Advances in Neural Information Processing Systems, 2022. This was joint work with Guido Montúfar.

- In Chapter 4 we express the NTK as a power series to establish a number of spectral properties. Theorem 4.3.2 provides the power series coefficients for the NTK which

depend on the Hermite coefficients of the activation function as well as the depth of the network. In Theorem 4.4.1 we demonstrate that the NTK has a dominant outlier eigenvalue and that there are $O(1)$ eigenvalues on the same order of magnitude as this outlier. In Theorem 4.4.3 we demonstrate that after subtracting a rank one component, the effective rank of the NTK is upper bounded by a constant multiple of the effective rank of the input data gram. Since real world data tends to have low effective rank, the NTK also exhibits this property. In Theorem 4.4.5 we also demonstrate that this same property holds for finite-width shallow ReLU networks with high probability at initialization. In Corollary 4.4.7 and Theorem 4.4.8 we demonstrate that the asymptotic decay of the spectrum depends on the decay of the power series coefficients, with faster coefficient decay corresponding to faster eigenvalue decay. The contents of Chapter 4 are based on the manuscript *"Characterizing the Spectrum of the NTK via a Power Series Expansion"* which appeared in The Eleventh International Conference on Learning Representations, 2023. This was joint work with Michael Murray, Hui Jin, and Guido Montúfar.

CHAPTER 1

Introduction

1.1 A Brief History of the Neural Tangent Kernel

The typical operating regime in deep learning is to optimize an overparameterized network via gradient-based optimization. This has been phenomenally successful in practice but leads to a number of theoretical challenges. The first is that neural networks have a highly nonlinear parameterization which leads to optimization objectives that are nonconvex [SS89, SS91]. The nonconvexity of the optimization makes proving theoretical guarantees for gradient optimization a tall task. Furthermore overparameterized networks are able to interpolate arbitrary labels [ZBH17], and the VC-dimension of typical networks grows at least linearly with the number of parameters [BHL19, KS95]. As a consequence, classical complexity based measures from statistical learning theory such as Rademacher complexity or VC-dimension lead to vacuous generalization bounds [AB02]. Thus understanding modern deep learning will require innovations beyond the classical theories of both optimization and generalization.

The aforementioned challenges at first make the prospect of establishing a theoretical understanding of deep learning seem dismal. However, there was evidence as far back as the 1990s that overparameterized networks may be amenable to theoretical analysis. [Nea96, Wil96] demonstrated that the network outputs converge to a Gaussian process as the number of hidden units approaches infinity. This led to a line of research studying the

connection between Gaussian processes, kernel methods, neural network representations, and deep learning [CS09, DFS16, LBN18]. In a similar vein [NTS15] exhibited decreasing generalization error while increasing the network width, suggesting that overparameterized networks may have a more subtle form of capacity control.

While progress was made towards understanding neural network representations via the infinite-width limit, an understanding of the optimization dynamics was still lacking. A breakthrough emerged in 2018 when [JGH18] demonstrated that the optimization dynamics are governed via a time-dependent kernel coined the "Neural Tangent Kernel (NTK)", which in the infinite-width limit becomes constant throughout training. In this limiting setting the network parameterization becomes approximately linear [LXS19], and bounding the smallest eigenvalue of the NTK throughout training is sufficient to prove global convergence of gradient descent. In fact, almost concurrently with [JGH18] the authors in [DZP19] had used this technique to prove the first global convergence guarantee for gradient descent applied to a network trained on general data. The NTK had been studied earlier by the work [XLS17] which demonstrated that the squared loss satisfies a Polyak-Lojasiewicz (PL) inequality in any region where the smallest eigenvalue of the NTK is bounded below. This analysis ties back to a well known technique in nonconvex optimization that establishing a PL-inequality is sufficient for proving convergence of gradient descent provided that the gradient is Lipschitz [Pol63]. The innovation in [DZP19] was to prove that the gradient descent trajectory remains in a region where a PL-inequality holds, as well as an innovative technique of bounding the number of activation patterns that change for a ReLU network as a substitute for the Lipschitz property.

1.2 Global Convergence Guarantees via the NTK

1.2.1 The Neural Tangent Kernel and PL-inequalities

In this section we will briefly display how the NTK naturally emerges when studying the dynamics of gradient descent. We will focus on the regression problem. Let

$$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

denote our training data where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We will let $f(x; \theta)$ denote our neural network taking inputs $x \in \mathbb{R}^d$ with parameters $\theta \in \mathbb{R}^p$. The specific architecture will not matter for the purpose of this section. Let $\ell(z, y)$ be a loss function, e.g. $\ell(z, y) = \frac{1}{2}(z - y)^2$, and let

$$L(\theta) = \sum_{i=1}^{n} \ell(f(x_i, \theta), y_i)$$

denote our empirical risk induced by the training data $\mathcal{D}$. We note that it is not at all obvious *a priori* that gradient descent will solve

$$\min_{\theta} L(\theta),$$

because in general the loss L is nonconvex as a function of θ. Even in the case of a deep linear network, the parameterization $\theta \mapsto f(\bullet; \theta)$ is nonlinear, making this problem highly nontrivial even for the simplest networks. Furthermore for the popular ReLU activation function $\sigma(x) = \max\{0, x\}$ the gradient $\nabla_\theta L$ is non-Lipschitz, which further complicates the analysis. These difficulties together make proving convergence guarantees for neural networks highly difficult in general.

To make things concrete, we will for now assume $\ell(z, y) = \frac{1}{2}(z - y)^2$ is the squared loss. Furthermore we will optimize the loss via gradient flow

$$\partial_t \theta_t = -\partial_\theta L(\theta_t),$$

which is the continuous-time analog of gradient descent. Speaking loosely, one can view gradient flow as gradient descent in the limit of vanishing step sizes. A key insight of

[JGH18, DZP19] was to analyze the gradient descent dynamics in function space (i.e. the evolution of the neural network predictions) as opposed to parameter space. In this vein we will let $u_\theta, y \in \mathbb{R}^n$ be defined by

$$u_\theta = [f(x_1;\theta), f(x_2;\theta), \ldots, f(x_n;\theta)]^T,$$

$$y = [y_1, y_2, \ldots, y_n]^T.$$

u_θ denotes the neural network predictions on the training set $\mathcal{D}$ and y denotes the desired target values. To denote the predictions at time t, we will write $u_t := u_{\theta_t}$ for short. Furthermore we will let $\hat{r}_t := u_t - y$ denote the residual vector, i.e. the difference between the neural network predictions at time t and the desired labels y. Under this notation, we can write the loss at time t as

$$L(\theta(t)) = \frac{1}{2}\sum_{i=1}^{n}(f(x_i;\theta_t) - y_i)^2 = \frac{1}{2}\left\|\hat{r}_t\right\|^2.$$

We will let

$$(J_t)_{i,j} := \partial_{\theta_j} f(x_i;\theta_t)$$

be the Jacobian of u_t, i.e. $\partial_\theta u_t = J_t \in \mathbb{R}^{n\times p}$. We note that by the chain rule

$$\partial_\theta L = [\partial_\theta u_t]^T \partial_{u_t} L = J_t^T \hat{r}_t,$$

$$\partial_t \hat{r}_t = \partial_\theta u_t \cdot \partial_t \theta_t = -J_t J_t^T \hat{r}_t.$$

We define

$$H_t := J_t J_t^T.$$

The positive-semidefinite matrix H_t is called the NTK Gram matrix. It can be viewed as the Gram matrix induced by the following kernel

$$K_t(x, x') := \langle \nabla_\theta f(x;\theta_t), \nabla_\theta f(x';\theta_t)\rangle,$$

where $(H_t)_{i,j} = K_t(x_i, x_j)$. The kernel K_t is known as the time-dependent NTK. By our previous result

$$\partial_t \hat{r}_t = -J_t J_t^T \hat{r}_t = -H_t \hat{r}_t.$$

Therefore

$$\partial_t L(\theta(t)) = \partial_t \frac{1}{2} \left\| \hat{r}_t \right\|^2 = [\partial_t \hat{r}_t]^T \cdot \partial_{\hat{r}_t} \frac{1}{2} \left\| \hat{r}_t \right\|^2 = -\hat{r}_t^T H_t \hat{r}_t.$$

We now note that

$$\partial_t L(\theta(t)) = -\hat{r}_t^T H_t \hat{r}_t \leq -\lambda_{min}\left(H_t\right) \left\| \hat{r}_t \right\|^2 = -2\lambda_{min}\left(H_t\right) L(t).$$

Then by Grönwall's inequality [Gro19]

$$L(t) \leq L(0) \exp\left(-2 \int_0^t \lambda_{min}\left(H_s\right) ds \right).$$

Now assume that

$$2\lambda_{min}\left(H_t\right) \geq c > 0 \quad \forall t > 0.$$

Then we have that

$$L(t) \leq L(0) \exp(-ct). \tag{1.1}$$

Thus we have just shown that lower bounding $\lambda_{\min}\left(H_t\right)$ uniformly in time is sufficient for establishing convergence of gradient flow to a global minimum when optimizing the squared loss. The quantity c provides an estimate for the convergence rate. The bound (1.1) is analagous to linear convergence in discrete time.

Let us now consider more general loss functions $\ell(z, y)$. In general by the same calculations as before we have that

$$\partial_t L = -\left[\partial_u L\right]^T H_t \partial_u L.$$

Suppose

$$\lambda_{min}\left(H_t\right) \geq c > 0 \quad \forall t > 0.$$

Then similar to before

$$\partial_t L \leq -c \left\| \partial_u L \right\|_2^2.$$

Assuming L is bounded below it follows that

$$\liminf_{t>0} \left\| \partial_u L \right\|_2^2 = 0.$$

Suppose L is strongly convex as a function of u, i.e.

$$\langle u - u', \nabla_u L(u) - \nabla_u L(u') \rangle \geq \alpha \left\| u - u' \right\|_2^2 .$$

Then any global minimum is unique. Assume a global minimum u^* exists, then

$$\langle u_t - u^*, \nabla_u L(u_t) \rangle = \langle u_t - u^*, \nabla_u L(u_t) - \nabla_u L(u^*) \rangle \geq \alpha \left\| u_t - u^* \right\|_2^2 .$$

Thus by the Cauchy-Schwarz inequality we have

$$\left\| \nabla_u L(u_t) \right\|_2 \geq \alpha \left\| u_t - u^* \right\|_2 .$$

Thus if

$$\liminf_{t>0} \left\| \partial_u L \right\|_2^2 = 0,$$

then $\liminf_{t>0} \left\| u_t - u^* \right\|_2 = 0$. For gradient flow we have that

$$\partial_t L = - \left\| \partial_\theta L \right\|_2^2 \leq 0,$$

and thus L is nonincreasing. It follows that $\liminf_{t>0} \left\| u_t - u^* \right\|_2 = 0$ implies that

$$\lim_{t \to \infty} L(u_t) = L(u^*).$$

We have just showed that if $\lambda_{\min}(H_t) \geq c > 0$ for all $t > 0$ and $u \mapsto L(u)$ is strongly convex, then gradient flow converges to a global minimum. Another sufficient condition is that L satisfies the following PL-inequality in function space

$$\alpha |L(u) - L(u^*)|^\beta \leq \left\| \nabla_u L(u) \right\|_2 \tag{1.2}$$

for some $\alpha, \beta > 0$. Let $\sigma_{min}(J_t)$ denote the smallest singular value of J_t. Then (1.2) implies

$$\left\| \partial_\theta L \right\| = \left\| J_t^T \partial_u L \right\| \geq \sigma_{min}(J_t) \left\| \partial_u L \right\| \geq \alpha \sigma_{min}(J_t) |L(u) - L(u^*)|^\beta.$$

Thus if $\sigma_{min}(J_t) = \lambda_{\min}(H_t)^{1/2} \geq c^{1/2} > 0$ then we have a separate PL-inequality in parameter space

$$\left\| \partial_\theta L \right\|_2 \geq \alpha c^{1/2} |L(u_\theta) - L(u^*)|^\beta.$$

Since

$$\partial_t L(t) = - \left\| \partial_\theta L \right\|_2^2 ,$$

assuming L is bounded below we have that

$$\liminf_{t>0} \left\| \partial_\theta L \right\|_2^2 = 0.$$

Thus by the same reasoning as before we have that $\lim_{t\to\infty} |L(u_t) - L(u^*)| = 0$. One can reason similarly for gradient descent (as opposed to gradient flow). Specifically, if $\nabla_\theta L(\theta)$ is Lipschitz and L satisfies the PL-inequality

$$\mu(L(\theta) - L(\theta^*)) \leq \left\| \nabla_\theta L \right\|_2^2$$

where θ^* is a parameter corresponding to a global minimum, then gradient descent with constant step size converges to a global minimum [Pol63].

1.2.2 Bounding the Smallest Eigenvalue of the NTK Gram Matrix

In the previous section we demonstrated that

$$\lambda_{\min}(H_t) \geq c > 0 \quad \forall t > 0$$

is a sufficient condition for proving convergence to a global minimum. We note that proving such a bound is equivalent to bounding the smallest singular value of the network Jacobian J_t. Let us now assume again that we are dealing with the squared loss $\ell(z, y) = \frac{1}{2}(z - y)^2$. For a particular parameter θ, if we let J_θ and $\hat{r}_\theta$ denote the network Jacobian and residual respectively, then recall by the chain rule

$$\partial_\theta L = J_\theta^T \hat{r}_\theta.$$

Thus if $\sigma_{min}(J_\theta) > 0$, we have that

$$\left\| \partial_\theta L \right\| \geq \sigma_{min}(J_\theta) \left\| \hat{r}_\theta \right\| = \sigma_{min}(J_\theta) \sqrt{2L(\theta)}.$$

Consequently, wherever $\sigma_{min}(J_\theta) > 0$ we have that each critical point of the loss is a global minimum. However neural networks are known to have spurious critical points, with saddle points being particularly prevalent [DPG14, Kaw16, FA00, SGJ21]. Thus the challenge for proving convergence is to demonstrate that the gradient descent trajectory remains in a region where the smallest singular value of the Jacobian, or equivalently the smallest eigenvalue of the NTK Gram matrix, is bounded below. This was the key difficulty that was overcome in the proof in [DZP19].

It was shown in [JGH18, DZP19] that under a suitable parameterization, in the infinite-width limit the matrix H_t converges to a fixed positive-definite matrix H^∞ uniformly in time. The parameterization introduced in these works has since been called the "NTK parameterization", which we introduce below. For a fully-connected network with D hidden layers, we parameterize the network as follows. Let $\theta = vec(\{W^{(l)}, b^{(l)}\}_{l=1}^{D+1})$ where $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ and $b^{(l)} \in \mathbb{R}^{n_l}$. We can then define the network output $f(x;\theta)$ via the following relations:

$$\begin{aligned} x^{(0)} &= x \\ x^{(l)} &= \sigma\left(\frac{1}{\sqrt{n_l}} W^{(l)} x^{(l-1)} + \beta b^{(l)}\right) \quad l = 1, \dots, D \\ x^{(D+1)} &= \frac{1}{\sqrt{n_{D+1}}} W^{(D+1)} x^{(D)} + \beta b^{(D+1)} \\ f(x;\theta) &= x^{(D+1)}. \end{aligned}$$

Under this parameterization we initialize the parameters $W_{i,j}^{(l)} \sim N(0,1)$ and $b_i^{(l)} \sim N(0,1)$ independently. This is in contrast with the standard parameterization:

$$\begin{aligned} x^{(0)} &= x \\ x^{(l)} &= \sigma(W^{(l)} x^{(l-1)} + b^{(l)}) \quad l = 1, \dots, D \\ x^{(D+1)} &= W^{(D+1)} x^{(D)} + b^{(D+1)} \\ f(x;\theta) &= x^{(D+1)}, \end{aligned}$$

where the parameters are initialized $W_{i,j}^{(l)} \sim N(0, 1/n_l)$ and $b_i^{(l)} \sim N(0, \beta^2)$ independently. The two parameterizations can realize the same functions and are identical in distribution at initialization, however the gradients are different. For gradient descent, the standard parameterization and NTK parameterization are equivalent up to a parameter-dependent rescaling of the step-size [LXS19]. We also note that other parameterizations have been studied, such as the "mean-field" parameterization [SS20]. Under the NTK parameterization under fairly general assumptions

$$H_t \to H^\infty$$

in probability uniformly on $[0, T]$ where H^∞ is a fixed positive-semidefinite matrix [JGH18]. Given weak assumptions on the training data inputs $x_1, \ldots, x_n$ (e.g. no two inputs are parallel [DZP19] or they are "δ-separable" [OS20]), we have that

$$\lambda_{min}\left(H^\infty\right) > 0.$$

Consequentially, we expect that convergence of gradient flow can be guaranteed in the infinite-width limit. For the finite-width setting, the analysis is more complicated. One strategy is to bound the deviations of the NTK Gram matrix at initialization and throughout training. For example, suppose that

$$\left\|H_0 - H^\infty\right\|_{op}, \left\|H_t - H_0\right\|_{op} \leq \frac{\lambda_{min}(H^\infty)}{4}.$$

Then we have

$$\left|\lambda_{min}\left(H_t\right) - \lambda_{min}\left(H^\infty\right)\right| \leq \left\|H_t - H^\infty\right\|_{op} \leq \left\|H_0 - H^\infty\right\|_{op} + \left\|H_t - H_0\right\|_{op} \leq \frac{\lambda_{min}(H^\infty)}{2},$$

which implies that $\lambda_{min}(H_t) \geq \frac{\lambda_{min}(H^\infty)}{2}$. For simplicity, assume all layers have the same width m. At initialization it was shown in [HY20] that whenever the activation function is suitably smooth

$$\left\|H_0 - H^\infty\right\|_{op} = \tilde{\mathcal{O}}(n/\sqrt{m}) \tag{1.3}$$

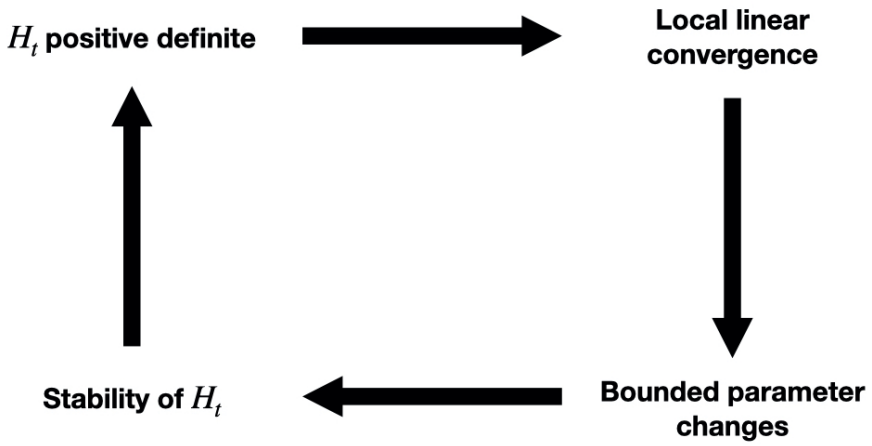

Figure 1.1: **A Seemingly Circular Argument** A sketch of the argument for proving convergence of gradient flow to a global minimum.

with high probability. Furthermore by the results in [LZB20b, LZB22] it was shown that for any $R > 0$ with high probability over the initialization that

$$\|H_t - H_0\|_{op} = \tilde{\mathcal{O}}(nR^{3D}/\sqrt{m}) \tag{1.4}$$

for any t such that $\theta_t \in B(0, R)$. Thus if we can show that θ_t remains in $B(\theta_0, R)$ for some fixed $R > 0$, then for m large enough

$$\lambda_{min}(H_t) \gtrsim \lambda_{min}(H^{\infty}).$$

Thus we need to show that there is an $R > 0$ independent of m such that $\theta_t \in B(\theta_0, R)$ for all $t > 0$.

1.2.3 Proving Global Convergence for Gradient Flow

We will sketch the following argument for convergence of gradient flow to a global minimum, which has appeared in many different variations (e.g. [DZP19, HY20, LZB22]). The proof revolves around a seemingly circular argument depicted in Figure 1.1, which can be resolved via a continuous induction argument. By (1.3) if $m \gtrsim n^2 \lambda_{\min}(H^{\infty})^{-2}$ we can assume

$$\|H_0 - H^\infty\| \leq \lambda_{\min}(H^\infty)/4.$$

Fix some value $K > 0$ and let

$$T = \sup\{t \geq 0 : \lambda_{\min}(H_t) \geq \lambda_{\min}(H^\infty)/2, \quad \|J_t\| \leq K\}.$$

We will see later that by setting K sufficiently large we can ensure that the set the supremum is taken over above is nonempty with high probability. If we can demonstrate that $T = \infty$, then we have that the smallest eigenvalue $\lambda_{\min}(H_t)$ is bounded below uniformly in time and thus we will have shown that gradient flow converges to a global minimum. Thus for the sake of contradiction assume $T < \infty$. Recall that by the results in Section 1.2.1 the bound $\lambda_{\min}(H_t) \geq \lambda_{\min}(H^\infty)/2$ implies that for $t \leq T$

$$\|\hat{r}_t\|_2^2 \leq \exp(-\lambda_{\min}(H^\infty)\, t)\, \|\hat{r}_0\|_2^2 .$$

It follows that for $t \leq T$,

$$\|\partial_t \theta_t\|_2 = \left\|J_t^T \hat{r}_t\right\|_2 \leq K \|\hat{r}_t\|_2 \leq K \exp\left(-\frac{1}{2}\lambda_{\min}(H^\infty)\, t\right) \|\hat{r}_0\|_2 .$$

Well then

$$\begin{aligned}\|\theta_T - \theta_0\|_2 \leq \int_0^T \|\partial_s \theta_s\|_2\, ds &\leq \int_0^T K \exp\left(-\frac{1}{2}\lambda_{\min}(H^\infty)\, s\right) \|\hat{r}_0\|_2\, ds \\ &\leq \frac{2K}{\lambda_{\min}(H^\infty)} \|\hat{r}_0\|_2 =: R'.\end{aligned}$$

It is not hard to show that the network outputs are bounded with high probability at initialization, thus assuming $\|y\| = O(\sqrt{n})$ we have that $\|\hat{r}_0\| = O(\sqrt{n})$. It follows then that there exists a quantity $R_{max} = O\left(\frac{K\sqrt{n}}{\lambda_{\min}(H^\infty)}\right)$ such that $R' \leq R_{max}$ with high probability. Well by Eq. (1.4) we can say with high probability for $\theta_t \in B(\theta_0, R_{max})$

$$\|H_t - H_0\|_{op} = \mathcal{O}(nR_{max}^{3L}/\sqrt{m}).$$

So if $m \gtrsim [nR_{max}^{3D}\lambda_{\min}(H^\infty)^{-1}]^2$ we can assume

$$\|H_T - H_0\|_2 \leq \lambda_{\min}(H^\infty)/8.$$

However then

$$\|H_0 - H^\infty\| \leq \lambda_{\min}(H^\infty)/4, \quad \|H_0 - H_T\| \leq \lambda_{\min}(H^\infty)/8,$$

so that

$$\|H_T - H^\infty\| \leq \frac{3}{8}\lambda_{\min}(H^\infty).$$

Well then

$$\lambda_{\min}(H_T) \geq \lambda_{\min}(H^\infty) - \|H_T - H^\infty\| \geq \frac{5}{8}\lambda_{\min}(H^\infty) > \frac{1}{2}\lambda_{\min}(H^\infty). \tag{1.5}$$

Recall the definition of T,

$$T := \sup\{t \geq 0 : \lambda_{\min}(H_t) \geq \lambda_{\min}(H^\infty)/2, \quad \|J_t\| \leq K\}.$$

By continuity and the maximality of T we must have that either $\lambda_{\min}(H_T) = \frac{1}{2}\lambda_{\min}(H^\infty)$ or $\|J_T\| = K$, however by (1.5) $\lambda_{\min}(H_T) > \frac{1}{2}\lambda_{\min}(H^\infty)$, thus it follows that $\|J_t\| = K$. However as we will see in Chapter 3 (see Lemma 3.6.12) for any $R \geq 1$ if $\sqrt{m} \geq R$ then with high probability

$$\sup_x \sup_{\theta \in \overline{B}(\theta_0, R)} \|\nabla_\theta f(x;\theta)\| = O(1).$$

Well then applying this result for $R = R_{max}$ we have that with high probability

$$\|J_t\| \leq \sqrt{n} \max_i \|\nabla_\theta f(x_i;\theta_t)\| = O(\sqrt{n})$$

for $t \leq T$. Thus by setting $K = \Theta(\sqrt{n})$ we can ensure that with high probability $\|J_t\| < K$ for all $t \leq T$, which contradicts our previous result. Thus by contradiction we conclude that $T = \infty$, and consequently we have that

$$\lambda_{\min}(H_t) \geq \frac{1}{2}\lambda_{\min}(H^\infty) \quad \forall t.$$

As we saw before this implies that

$$L(t) \leq \exp(-\lambda_{\min}(H^\infty)\, t) L(0) \quad \forall t > 0,$$

and thus we have convergence to a global minimum. The eigenvalue $\lambda_{\min}(H^{\infty})$ serves as an estimate for the convergence rate. Our requirements were that

$$m \gtrsim n^2 \lambda_{\min}(H^{\infty})^{-2},$$

and

$$m \gtrsim n^2 (R_{max})^{6D} \lambda_{\min}(H^{\infty})^{-2},$$

where

$$R_{max} = O\left(\frac{K\sqrt{n}}{\lambda_{\min}(H^{\infty})}\right) = O\left(\frac{n}{\lambda_{\min}(H^{\infty})}\right).$$

It turns out that for general inputs $x_1, \ldots, x_n$ we have that $\lambda_{\min}(H^{\infty}) = \Omega(1)$ [NMM21]. Thus we conclude that $m \gtrsim n^{O(D)}$ suffices to prove global convergence of gradient flow.

1.3 Spectral Bias

In the previous section we demonstrated that bounding the smallest eigenvalue of the NTK Gram matrix is sufficient for establishing convergence, and that the bound for the eigenvalue provides an estimate for the convergence rate of gradient descent. However, in general this is a pessimistic estimate and the convergence rate along different components will vary. From the results in Section 1.2.1 that we have for the squared loss

$$\partial_t \hat{r}_t = -H_t \hat{r}_t.$$

Recall that for large width networks that $H_t \approx H^{\infty}$, and thus the gradient descent dynamics can be approximated by the evolution

$$\partial_t \hat{r}_t = -H^{\infty} \hat{r}_t,$$

which has the explicit solution

$$\exp(-H^{\infty} t)\hat{r}_0. \tag{1.6}$$

Let $u_1, \ldots, u_n$ denote the eigenvectors of H^∞ with corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then we can analyze the convergence along the direction of u_i:

$$\langle u_i, \exp(-H^\infty t)\hat{r}_0 \rangle = \exp(-\lambda_i t)\langle u_i, \hat{r}_0 \rangle.$$

We thus see that the convergence rate along the direction u_i is given by the eigenvalue λ_i, and consequently the directions corresponding to large eigenvalues will be learned much more quickly. As we will see in the later chapters (see e.g. Fig. 3.1), the NTK Gram matrix tends to have a small number of outlier eigenvalues and a long tail of small eigenvalues. In fact, in Chapter 4 we prove that there are $O(1)$ eigenvalues on the same order of magnitude as the largest eigenvalue λ_1 (see Theorem 4.4.1 and Observation 4.4.2). Consequently, there are a small number of directions that are learned much more quickly than others.

The phenomenon that eigenvectors of the NTK corresponding to large eigenvalues are learned quicker can be described as a type of "spectral bias" [CFW21]. Classically, "spectral bias" was the title given to the phenomenon that neural networks tend to learn the low Fourier frequencies quicker during training[1] [RBA19, XZX19, YAA22]. However, in special cases these two notions coincide. Specifically, if we let m denote the width of the network we can define

$$K^\infty(x, x') := \lim_{m \to \infty} \langle \nabla_\theta f(x; \theta), \nabla_\theta f(x'; \theta) \rangle$$

where the convergence is in probability over the parameter initialization [JGH18]. K^∞ is called the analytical Neural Tangent Kernel (NTK), and the matrix H^∞ introduced in Section 1.2.1 is the Gram matrix induced by this kernel and the training data, i.e.

$$H^\infty_{i,j} := K^\infty(x_i, x_j).$$

Let X denote the input domain and let ρ denote the distribution for the training data inputs, i.e. $x_i \sim \rho$. Then the kernel K^∞ induces an integral operator $T_{K^\infty} : L^2_\rho(X) \to L^2_\rho(X)$

$$T_{K^\infty} g(x) := \int_X K^\infty(x, s) g(s) d\rho(s).$$

[1] This has also been called the "Frequency Principle" [XZX19].

By Mercer's theorem [Mer09] we have the decomposition

$$K^{\infty}(x, x') = \sum_{i=1}^{\infty} \sigma_i \phi_i(x) \phi_i(x')$$

where $\{\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of $L^2_\rho(X)$ and each ϕ_i is an eigenfunction of T_{K^∞} with eigenvalue $\sigma_i \geq 0$. Whenever ρ is the uniform distribution on the sphere $X = S^{d-1}$, the eigenfunctions ϕ_i can be taken to be the spherical harmonics, which in $d = 2$ corresponds to the Fourier basis. In the work [RJK19] it was demonstrated that in the $d = 2$ case for shallow ReLU networks the large eigenvalues of T_{K^∞} correspond to the low Fourier frequencies. We note that we can consider the eigenvectors u_i of H^∞ to be empirical estimates of the eigenfunctions of T_{K^∞}. In this case "spectral bias" in the sense of learning the low Fourier frequencies faster coincides with "spectral bias" in the sense of learning the dominant eigenvectors of the NTK Gram matrix faster.

[ADH19a] had quantified the extent in which finite-width networks approximate the idealized infinite-width dynamics that are given by the evolution described in (1.6). However, this equation only describes the network on the training set $x_1, \ldots, x_n$. Let f^* be our target function so that $y_i = f^*(x_i)$. We are interested in describing the behavior of the residual $r_t(x) := f(x; \theta_t) - f^*(x)$ for an arbitrary input x. Informally speaking, in the limit of infinite data the matrix H^∞ converges to the integral operator T_{K^∞} and the empirical residual $\hat{r}_t$ converges to the full residual r_t. In this idealized setting the evolution described in (1.6) becomes

$$r_t = \exp(-T_{K^\infty} t) r_0. \tag{1.7}$$

Assuming (1.7) holds we have

$$\langle r_t, \phi_i \rangle_{L^2_\rho} = \langle \exp(-T_{K^\infty} t) r_0, \phi_i \rangle_{L^2_\rho} = \exp(-\sigma_i t) \langle r_0, \phi_i \rangle_{L^2_\rho}. \tag{1.8}$$

Thus under the evolution described in (1.7) we have that the eigenfunctions ϕ_i are learned at rates corresponding to their eigenvalues σ_i. In contrast to (1.6), (1.7) and (1.8) describe the dynamics of the residual over the entire input domain and not just the training set. Thus

in this limiting setting the network exhibits a stronger form of spectral bias that determines the behavior of the network over the entire input domain. In Chapters 2 and 3 we will quantify to what extent the finite-width network trained on finitely many samples exhibits the behavior of the idealized limit of infinite width and infinite data described in (1.7).

1.4 Limitations and Challenges of NTK Analysis

The paper that introduced the Neural Tangent Kernel [JGH18] has become one of the most highly cited works in deep learning theory, with the NTK having attracted both fanaticism and criticism [Ana21]. While the NTK has greatly enhanced the understanding of the optimization dynamics of wide networks [DZP19, DLL19, OS20, ALS19a, NM20, Ngu21, ZCZ20, ZG19, LXS19], this analysis breaks down whenever the depth of the network scales in tandem with the width [HN20], which is known to achieve better performance in practice [TL19, RKH21]. Furthermore NTK analysis is only applicable when training with small learning rates, with more moderate learning rates leading to distinct behavior [LBD20]. It is also known that in practice the NTK deviates to adapt to the target function [BGL21, ABP22, BES22], which stands in contrast to the infinite-width behavior where the NTK is constant. Establishing a theoretical framework that can handle more realistic scalings for the depth and learning rate which also makes allowances for feature learning remains an active challenge. Nevertheless, infinite-width networks achieve compelling performance and serve well as a first approximation of the average behavior of finite-width models [LSP20], suggesting that the Neural Tangent Kernel will remain a fundamental tool in deep learning theory.

CHAPTER 2

Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks

2.1 Introduction

A surprising but well established empirical fact is that neural networks optimized by gradient descent can find solutions to the empirical risk minimization (ERM) problem that generalize. This is surprising from an optimization point-of-view because the ERM problem induced by neural networks is nonconvex [SS89, SS91] and can even be NP-Complete in certain cases [BR93]. Perhaps even more surprising is that the discovered solution can generalize even when the network is able to fit arbitrary labels [ZBH17], rendering traditional complexity measures such as Rademacher complexity inadequate. How does deep learning succeed in the face of pathological behavior by the standards of classical optimization and statistical learning theory?

Towards addressing generalization, a modern line of thought that has emerged is that gradient descent performs implicit regularization, limiting the solutions one encounters in practice to a favorable subset of the model's full capacity (see e.g. [NTS15, NTS17, GWB17, WZE17]). An empirical observation is that neural networks optimized by gradient descent tend to fit the low frequencies of the target function first, and only pick up the higher frequencies later in training [RBA19, RJK19, BGG20, XZX19]. A closely related theme is gradient descent's bias towards smoothness for regression problems [WTP19, JM21]. For

classification problems, in suitable settings gradient descent provably selects max-margin solutions [SHN18, JT19]. Gradient descent is not impartial, thus understanding its bias is an important program in modern deep learning.

Generalization concerns aside, the fact that gradient descent can succeed in a nonconvex optimization landscape warrants attention on its own. A brilliant insight made by [JGH18] is that in function space the neural network follows a kernel gradient descent with respect to the "Neural Tangent Kernel" (NTK). This kernel captures how the parameterization biases the trajectory in function space, an abstraction that allows one to largely ignore parameter space and its complications. This is a profitable point-of-view, but there is a caveat. The NTK still depends on the evolution of the network parameters throughout time, and thus is in general time-dependent and complicated to analyze. However, under appropriate scaling of the parameters in the infinite-width limit it remains constant [JGH18]. Once the NTK matrix has small enough deviations to remain strictly positive definite throughout training, the optimization dynamics start to become comparable to that of a linear model [LXS19]. For wide networks (quadratic or higher polynomial dependence on the number of training data samples n and other parameters) this property holds and this has been used by a variety of works to prove global convergence guarantees for the optimization [DZP19, OS20, DLL19, ALS19a, ALS19b, ZCZ20, ZG19, SY20, DGM20][1] and to characterize the solution throughout time [ADH19a, BGG20]. The NTK has been so heavily exploited in this setting that it has become synonymous with polynomially wide networks where the NTK is strictly positive definite throughout training. This begs the question, to what extent is the NTK informative outside this regime?

While the NTK has hitherto been associated with the heavily overparameterized regime, in this Chapter we will demonstrate that refined analysis is possible in the underparameterized setting. Our theorems primarily concern a one-hidden layer network, however unlike

[1]Not all these works explicitly use that the NTK is positive definite. However, they all operate in the regime where the weights do not vary much and thus are typically associated with the NTK regime.

many NTK results appearing in the literature our network has *biases* and *both* layers are trained. In fact, the machinery we build is strong enough to extend some existing results in the overparameterized regime appearing in the literature to the case of training both layers.

2.1.1 Related Work

There has been a deluge of works on the Neural Tangent Kernel since it was introduced by [JGH18], and thus we do our best to provide a partial list. Global convergence guarantees for the optimization, and to a lesser extent generalization, for networks polynomially wide in the number of training samples n and other parameters has been addressed in several works [DZP19, OS20, DLL19, ALS19a, ALS19b, ZCZ20, ZG19, SY20, ADH19a]. To our knowledge, for the regression problem with arbitrary labels, quadratic overparameterization $m \gtrsim n^2$ is state-of-the art [OS20, SY20, NM20]. [EMW20] gave a fairly comprehensive study of optimization and generalization of shallow networks trained under the standard parameterization. Under the standard parameterization, changes in the outer layer weights are more significant, whereas under the NTK parameterization both layers have roughly equal effect. Since we study the NTK parameterization in our results, we view the analysis as complementary.

The results in this chapter are perhaps most closely connected with [ADH19a]. In Theorem 4.1 in that work they showed that for a shallow network in the polynomially overparameterized regime $m \gtrsim n^7$, the training error along eigendirections of the NTK matrix decay linearly at rates that correspond to their eigenvalues. Our main Theorem 2.3.5 can be viewed as an analogous statement for the actual risk (not the empirical risk) in the underparameterized regime: eigenfunctions of the NTK integral operator T_{K^∞} are approximately learned linearly at rates that correspond to their eigenvalues. In contrast with [ADH19a], we have that the requirements on width m and number of samples n required to learn eigenfunctions with large eigenvalues are smaller compared to those with small eigenvalues. Surprisingly the machinery we build is also strong enough to prove in our setting the direct analog of

Theorem 4.1. Note that [ADH19a] train the hidden layer of a ReLU network via gradient descent, whereas we are training both layers with biases for a network with smooth activations via gradient flow. Due to the different settings, the results are not directly comparable. This important detail notwithstanding, our overparameterization requirement ignoring logarithmic factors is smaller by a factor of $\frac{n^2}{d\delta^4}$ where n is the number of input samples, d is the input dimension, and δ is the failure probability. [BGG20] extended Theorem 4.1 in [ADH19a] to deep ReLU networks without bias where the first and last layer are fixed, with a higher overparameterization requirement than the original [ADH19a]. Since the first and last layers are fixed this cannot be specialized to get a guarantee for training both layers of a shallow network even with ReLU activations.

Although it was not our focus, the tools to prove Theorem 2.3.5 are enough to prove analogs of Theorem 4 and Corollary 2 in the work of [SY19]. Theorem 4 and Corollary 2 of [SY19] are empirical risk guarantees that show that for target functions that participate mostly in the top eigendirections of the NTK integral operator T_{K^∞}, moderate overparameterization is possible. Again in this work they train the hidden layer of a ReLU network via gradient descent, whereas we are training both layers with biases for a network with smooth activations via gradient flow. Again due to the different settings, we emphasize the results are not directly comparable. In our results the bounds and requirements are comparable to [SY19], with neither appearing better. Nevertheless we think it is important to demonstrate that these results hold for training both layers with biases, and we hope our "Damped Deviations" approach will simplify the interpretation of the aforementioned works.

Theorem 4.2 in [CFW21] provides an analogous statement to our Theorem 2.3.5 if you replace our quantities with their empirical counterparts. While our statement concerns the projections of the test residual onto the eigenfunctions of an operator associated with the Neural Tangent Kernel, their statement concerns the inner products of the empirical residual with those eigenfunctions. Their work was a crucial step towards explaining the spectral bias from gradient descent, however we view the difference between tracking the

empirical quantities versus the actual quantities to be highly nontrivial. Another difference is they consider a ReLU network whereas we consider smooth activations; also they consider gradient descent versus we consider gradient flow. Due to the different settings we would like to emphasize that the scalings of the different parameters are not directly comparable, nevertheless the networks they consider are significantly wider. They require at least $m \geq \tilde{O}(\max\{\sigma_k^{-14}, \epsilon^{-6}\})$, where σ_k is a cutoff eigenvalue and ϵ is the error tolerance. By contrast in our result, to have the projection onto the top k eigenvectors be bounded by epsilon in L2 norm requires $m = \tilde{\Omega}(\sigma_k^{-4}\epsilon^{-2})$. Another detail is their network has no bias whereas ours does.

2.1.2 Our Contributions

The key idea for our results is the concept of "Damped Deviatons", the fact that for the squared error deviations of the NTK are softened by a damping factor, with large eigendirections being damped the most. This enables the following results.

- In Theorem 2.3.5 we characterize the bias of the neural network to learn the eigenfunctions of the integral operator T_{K^∞} associated with the Neural Tangent Kernel (NTK) at rates proportional to the corresponding eigenvalues.

- In Theorem 2.3.7 we show that in the overparameterized setting the training error along different directions can be sharply characterized, showing that Theorem 4.1 in [ADH19a] holds for smooth activations when training both layers with a smaller overparameterization requirement.

- In Theorem 2.3.8 and Corollary 2.3.9 we show that moderate overparameterization is sufficient for solving the ERM problem when the target function has a compact representation in terms of eigenfunctions of T_{K^∞}. This extends the results in [SY19] to the setting of training both layers with smooth activations.

2.2 Gradient Dynamics and Damped Deviations

2.2.1 Notations

We will use $\|\bullet\|_2$ and $\langle \bullet, \bullet \rangle_2$ to denote the L^2 norm and inner product respectively (for vectors or for functions depending on context). For a symmetric matrix $A \in \mathbb{R}^{k \times k}$, $\lambda_i(A)$ denotes its ith largest eigenvalue, i.e. $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_k(A)$. For a matrix A, $\|A\|_{op} := \sup_{\|x\|_2 \leq 1} \|Ax\|_2$ is the operator norm induced by the Euclidean norm. We will let $\langle \bullet, \bullet \rangle_{\mathbb{R}^n}$ denote the standard inner product on $\mathbb{R}^n$ normalized by $\frac{1}{n}$, namely $\langle x, y \rangle_{\mathbb{R}^n} = \frac{1}{n}\langle x, y \rangle_2 = \frac{1}{n}\sum_{i=1}^n x_i y_i$. We will let $\|x\|_{\mathbb{R}^n} = \sqrt{\langle x, x \rangle_{\mathbb{R}^n}}$ be the associated norm. This normalized inner product has the convenient property that if $v \in \mathbb{R}^n$ such that $v_i = O(1)$ for each i then $\|v\|_{\mathbb{R}^n} = O(1)$, where by contrast $\|v\|_2 = O(\sqrt{n})$. This is convenient as we will often consider what happens when $n \to \infty$. $\|\bullet\|_\infty$ will denote the supremum norm with associated space L^∞. We will use the standard big O and Ω notation with $\tilde{O}$ and $\tilde{\Omega}$ hiding logarithmic terms.

2.2.2 Gradient Dynamics and the NTK Integral Operator

We will let $f(x; \theta)$ denote our neural network taking input $x \in \mathbb{R}^d$ and parameterized by $\theta \in \mathbb{R}^p$. The specific architecture of the network does not matter for the purposes of this section. Our training data consists of n input-label pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We focus on the setting where the labels are generated from a fixed target function f^*, i.e. $y_i = f^*(x_i)$. We will concatenate the labels into a label vector $y \in \mathbb{R}^n$, i.e. $y_i = f^*(x_i)$. We will let $\hat{r}(\theta) \in \mathbb{R}^n$ be the vector whose ith entry is equal to $f(x_i; \theta) - f^*(x_i)$. Hence $\hat{r}(\theta)$ is the residual vector that measures the difference between our neural networks predictions and the labels. We will be concerned with optimizing the squared loss

$$\Phi(\theta) = \frac{1}{2n} \|\hat{r}(\theta)\|_2^2 = \frac{1}{2} \|\hat{r}(\theta)\|_{\mathbb{R}^n}^2 .$$

Optimization will be done by gradient flow

$$\partial_t \theta_t = -\partial_\theta \Phi(\theta),$$

which is the continuous time analog of gradient descent. We will denote the residual at time t, $\hat{r}(\theta_t)$, as $\hat{r}_t$ for the sake of brevity and similarly we will let $f_t(x) = f(x; \theta_t)$. We will let $r_t(x) := f_t(x) - f^*(x)$ denote the residual off of the training set for an arbitrary input x.

We quickly recall some facts about the Neural Tangent Kernel and its connection to the gradient dynamics. For a comprehensive tutorial we suggest [JGH18]. The analytical NTK is the kernel given by

$$K^\infty(x, x') := \mathbb{E}\left[\left\langle \frac{\partial f(x;\theta)}{\partial \theta}, \frac{\partial f(x';\theta)}{\partial \theta} \right\rangle_2\right],$$

where the expectation is taken with respect to the parameter initialization for θ. We associate K^∞ with the integral operator $T_{K^\infty} : L^2_\rho(X) \to L^2_\rho(X)$ defined by

$$T_{K^\infty} f(x) := \int_X K^\infty(x, s) f(s) d\rho(s),$$

where X is our input space with probability measure ρ. Our training data $x_i \in X$ are distributed according to this measure $x_i \sim \rho$. By Mercer's theorem we can decompose

$$K^\infty(x, x') = \sum_{i=1}^{\infty} \sigma_i \phi_i(x) \phi_i(x'),$$

where $\{\phi_i\}_{i=1}^n$ is an orthonormal basis of L^2, $\{\sigma_i\}_{i=1}^\infty$ is a nonincreasing sequence of positive values, and each ϕ_i is an eigenfunction of T_{K^∞} with eigenvalue $\sigma_i > 0$. When $X = S^{d-1}$ is the unit sphere, ρ is the uniform distribution, and the weights of the network are from a rotation invariant distribution (e.g. standard Gaussian), $\{\phi_i\}_{i=1}^\infty$ are the spherical harmonics (which in $d = 2$ is the Fourier basis) due to K^∞ being rotation-invariant (see Theorem 2.2 [BZZ18]). We will let $\kappa := \max_{x \in X} K^\infty(x, x)$ which will be a relevant quantity in our later theorems. In our setting κ will always be finite as K^∞ will be continuous and X will be bounded. The training data inputs $\{x_1, \ldots, x_n\}$ induce a discretization of the integral operator T_{K^∞},

namely

$$T_n f(x) := \frac{1}{n}\sum_{i=1}^{n} K^{\infty}(x, x_i) f(x_i) = \int_X K^{\infty}(x, s) f(s) d\rho_n(s),$$

where $\rho_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ is the empirical measure. We recall the definition of the time-dependent NTK[2],

$$K_t(x, x') := \left\langle \frac{\partial f(x;\theta_t)}{\partial \theta}, \frac{\partial f(x';\theta_t)}{\partial \theta} \right\rangle_2.$$

We can look at the version of T_n corresponding to K_t, namely

$$T_n^t f(x) := \frac{1}{n}\sum_{i=1}^{n} K_t(x, x_i) f(x_i) = \int_X K_t(x, s) f(s) d\rho_n(s).$$

We recall that the residual $r_t(x) := f(x;\theta) - f^*(x)$ follows the update rule

$$\partial_t r_t(x) = -\frac{1}{n}\sum_{i=1}^{n} K_t(x, x_i) r_t(x_i) = -T_n^t r_t.$$

We will let $(H_t)_{i,j} := K_t(x_i, x_j)$ and $H^{\infty}_{i,j} := K^{\infty}(x_i, x_j)$ denote the Gram matrices induced by these kernels and we will let $G_t := \frac{1}{n}H_t$ and $G^{\infty} := \frac{1}{n}H^{\infty}$ be their normalized versions[3]. Throughout we will let $u_1, \ldots, u_n$ denote the eigenvectors of G^{∞} with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. The $u_1, \ldots, u_n$ are chosen to be orthonormal with respect to the inner product $\langle \bullet, \bullet \rangle_{\mathbb{R}^n}$. When restricted to the training set we have the update rule

$$\partial_t \hat{r}_t = -\frac{1}{n} H_t \hat{r}_t = -G_t \hat{r}_t.$$

2.2.3 Damped Deviations

The concept of damped deviations comes from the very simple lemma that follows (the proof is provided in Section 2.5.4). The lemma compares the dynamics of the residual $\hat{r}(t)$ on the training set to the dynamics of an arbitrary kernel regression $\exp(-Gt)\hat{r}(0)$:

---

[2]The NTK is an overloaded term. To help ameliorate the confusion, we will use NTK to describe K^{∞} and NTK (italic font) to describe the time-dependent version K_t.

[3]G_t and G^{∞} are the natural matrices to work with when working with the mean squared error as opposed to the unnormalized squared error. Also G^{∞}'s spectra concentrates around the spectrum of the associated integral operator $T_{K^{\infty}}$ and is thus a more convenient choice in our setting.

Lemma 2.2.1. *Let $G \in \mathbb{R}^{n\times n}$ be an arbitrary positive semidefinite matrix and let G_s be the time dependent NTK matrix at time s. Then*

$$\hat{r}_t = \exp(-Gt)\hat{r}_0 + \int_0^t \exp(-G(t-s))(G - G_s)\hat{r}_s ds.$$

Let's specialize the lemma to the case where $G = G^\infty$. In this case the first term is $\exp(-G^\infty t)\hat{r}_0$, which is exactly the dynamics of the residual in the exact NTK regime when $G_t = G^\infty$ for all t. The second term is a correction term that weights the NTK deviations $(G^\infty - G_s)$ by the damping factor $\exp(-G^\infty(t-s))$. We see that damping is largest along the large eigendirections of G^∞. The equation becomes most interpretable when projected along a specific eigenvector. Fix an eigenvector u_i of G^∞ corresponding to eigenvalue λ_i. Then the equation along this component becomes

$$\langle \hat{r}_t, u_i\rangle_{\mathbb{R}^n} = \exp(-\lambda_i t)\langle \hat{r}_0, u_i\rangle_{\mathbb{R}^n} + \int_0^t \langle \exp(-\lambda_i(t-s))(G^\infty - G_s)\hat{r}_s, u_i\rangle_{\mathbb{R}^n} ds.$$

The first term above converges to zero at rate λ_i. The second term is a correction term that weights the deviatiations of the NTK matrix G_s from G^∞ by the damping factor $\exp(-\lambda_i(t-s))$. The second term can be upper bounded by

$$\begin{aligned}\left|\int_0^t \langle \exp(-\lambda_i(t-s))(G^\infty - G_s)\hat{r}_s, u_i\rangle_{\mathbb{R}^n} ds\right| &\leq \int_0^t \exp(-\lambda_i(t-s))\left\|G^\infty - G_s\right\|_{op}\left\|\hat{r}_s\right\|_{\mathbb{R}^n} ds \\ &\leq \frac{[1-\exp(-\lambda_i t)]}{\lambda_i}\sup_{s\in[0,t]}\left\|G^\infty - G_s\right\|_{op}\left\|\hat{r}_0\right\|_{\mathbb{R}^n},\end{aligned}$$

where we have used the property $\|\hat{r}_s\|_{\mathbb{R}^n} \leq \|\hat{r}_0\|_{\mathbb{R}^n}$ from gradient flow. When $f^* = O(1)$ we have that $\|\hat{r}_0\|_{\mathbb{R}^n} = O(1)$, thus whenever $\|G^\infty - G_s\|_{op}$ is small relative to λ_i this term is negligible. It has been identified that the NTK matrices tend to have a small number of outlier large eigenvalues and exhibit a low rank structure [OFL19, ADH19a]. In light of this, the dependence of the above bound on the magnitude of λ_i is particularly interesting. We reach following important conclusion.

Observation 2.2.2. *The dynamics in function space will be similar to the NTK regime dynamics along eigendirections whose eigenvalues are large relative to the deviations of the time-dependent NTK matrix from the analytical NTK matrix.*

The equation in Lemma 2.2.1 concerns the residual restricted to the training set, but we will be interested in the residual for arbitrary inputs. Recall that $r_t(x) = f(x;\theta_t) - f^*(x)$ denotes the residual at time t for an arbitrary input. Then more generally we have the following damped deviations lemma for the whole residual (proved in Section 2.5.3.3).

Lemma 2.2.3. *Let $K(x,x')$ be an arbitrary continuous, symmetric, positive-definite kernel. Let $[T_K h](\bullet) = \int_X K(\bullet,s)h(s)d\rho(s)$ be the integral operator associated with K and let $[T_n^s h](\bullet) = \frac{1}{n}\sum_{i=1}^n K_s(\bullet,x_i)h(x_i)$ denote the operator associated with the time-dependent NTK K_s. Then*

$$r_t = \exp(-T_K t)r_0 + \int_0^t \exp(-T_K(t-s))(T_K - T_n^s)r_s ds,$$

where the equality is in the L^2 sense.

For our main results we will specialize the above lemma to the case where $K = K^\infty$. However there are other natural kernels to compare against, say K_0 or the kernel corresponding to some subset of parameters. We will elaborate further on this point after we introduce the main theorem. When specializing Lemma 2.2.3 to the case $K = K^\infty$, we have that T_{K^∞} and T_n^s are the operator analogs of G^∞ and G_s respectively. From this statement the same concepts holds as before, the dynamics of r_t will be similar to that of $\exp(-T_{K^\infty}t)r_0$ along eigendirections whose eigenvalues are large relative to the deviations $(T_{K^\infty} - T_n^s)$. In the underparameterized regime we can bound the second term and make it negligible (Theorem 2.3.5) and thus demonstrate that the eigenfunctions ϕ_i of T_{K^∞} with eigenvalues σ_i will be learned at rate σ_i. When the input data are distributed uniformly on the sphere S^{d-1} and the network weights are from a rotation-invariant distribution, the eigenfunctions of T_{K^∞} are the spherical harmonics (which is the Fourier basis when $d = 2$). In this case the network is biased towards learning the spherical harmonics that correspond to large eigenvalues of T_{K^∞}. It is in this vein that we will demonstrate a spectral bias.

2.3 Main Results

Our theorems will concern the shallow neural network

$$f(x;\theta) = \frac{1}{\sqrt{m}}\sum_{\ell=1}^{m} a_\ell \sigma(\langle w_\ell, x\rangle_2 + b_\ell) + b_0 = \frac{1}{\sqrt{m}} a^T \sigma(Wx+b) + b_0,$$

where $W \in \mathbb{R}^{m\times d}$, $a, b \in \mathbb{R}^m$ and $b_0 \in \mathbb{R}$ and $w_\ell = W_{\ell,:}$ denotes the ℓth row of W and $\sigma : \mathbb{R} \to \mathbb{R}$ is applied entry-wise. $\theta = (a^T, vec(W)^T, b^T, b_0)^T \in \mathbb{R}^p$ where $p = md + 2m + 1$ is the total number of parameters. Here we are utilizing the NTK parameterization [JGH18]. For a thorough analysis using the standard parameterization we suggest [EMW20]. We will consider two parameter initialization schemes. The first initializes $W_{i,j}(0) \sim \mathcal{W}$, $b_\ell(0) \sim \mathcal{B}$, $a_\ell(0) \sim \mathcal{A}$, $b_0 \sim \mathcal{B}'$ i.i.d., where $\mathcal{W}, \mathcal{B}, \mathcal{A}, \mathcal{B}'$ represent zero-mean unit variance subgaussian distributions. In the second initialization scheme we initialize the parameters according to the first scheme and then perform the following swaps $W(0) \to \begin{bmatrix} W(0) \\ W(0) \end{bmatrix}$, $b(0) \to \begin{bmatrix} b(0) \\ b(0) \end{bmatrix}$, $a(0) \to \begin{bmatrix} a(0) \\ -a(0) \end{bmatrix}$, $b_0 \to 0$ and replace the $\frac{1}{\sqrt{m}}$ factor in the parameterization with $\frac{1}{\sqrt{2m}}$. This is called the "doubling trick" [COB19, ZXL20] and ensures that the network is identically zero $f(x;\theta_0) \equiv 0$ at initialization. We will explicitly state where we use the second scheme and otherwise will be using the first scheme.

The following assumptions will persist throughout the rest of the paper:

Assumption 2.3.1. *σ is a C^2 function satisfying $\|\sigma'\|_\infty, \|\sigma''\|_\infty < \infty$.*

Assumption 2.3.2. *The inputs satisfy $\|x\|_2 \le M$.*

The following assumptions will be used in most, but not all theorems. We will explicitly state when they apply.

Assumption 2.3.3. *The input domain X is compact with strictly positive Borel measure ρ.*

Assumption 2.3.4. *T_{K^∞} is strictly positive, i.e., $\langle f, T_{K^\infty} f\rangle_2 > 0$ for $f \neq 0$.*

Most activation functions other than ReLU satisfy Assumption 2.3.1, such as Softplus $\sigma(x) = \ln(1 + e^x)$, Sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$, and Tanh $\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Assumption 2.3.2 is a mild assumption which is satisfied for instance for RGB images and has been commonly used [DZP19, DLL19, OS20]. Assumption 2.3.3 is so that Mercer's decomposition holds, which is often assumed implicitly. Assumption 2.3.4 is again a mild assumption that is satisfied for a broad family of parameter initializations (e.g. Gaussian) anytime σ is not a polynomial function, as we will show in Section 2.5.7. Assumption 2.3.4 is not strictly necessary but it simplifies the presentation by ensuring T_{K^∞} has no zero eigenvalues.

We will track most constants that depend on parameters of our theorems such as M, the activation function σ, and the target function f^*. However, constants appearing in concentration inequalities such as Hoeffding's or Bernstein's inequality or constants arising from $\delta/2$ or $\delta/3$ arguments will not be tracked. We will reserve $c, C > 0$ for untracked constants whose precise meaning can vary from statement to statement. In the proofs in Section 3.6 it will be explicit which constants are involved.

2.3.1 Underparameterized Regime

Our main result compares the dynamics of the residual $r_t(x) = f(x; \theta_t) - f^*(x)$ to that of $\exp(-T_{K^\infty} t) r_0$ in the underparameterized setting. Note that $\langle \exp(-T_{K^\infty} t) r_0, \phi_i \rangle_2 = \exp(-\sigma_i t) \langle r_0, \phi_i \rangle_2$, thus $\exp(-T_{K^\infty} t) r_0$ learns the eigenfunctions ϕ_i of T_{K^∞} at rate σ_i. Therefore $\exp(-T_{K^\infty} t) r_0$ exhibits a bias to learn the eigenfunctions of T_{K^∞} corresponding to large eigenvalues more quickly. To our knowledge no one has been able to rigorously relate the dynamics in function space of the residual r_t to $\exp(-T_{K^\infty} t) r_0$, although that seems to be what is suggested by [RJK19, BGG20]. The existing works we are aware of [ADH19a, BGG20, CFW21] characterize the bias of the empirical residual primarily in the heavily overparameterized regime ([CFW21] stands out as requiring wide but not necessarily overparameterized networks). By contrast, we characterize the bias of the whole residual in the underparameterized regime.

Theorem 2.3.5. *Assume that Assumptions 2.3.3 and 2.3.4 hold. Let P_k be the orthogonal projection in L^2 onto span$\{\phi_1, \ldots, \phi_k\}$ and let $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$. If we are doing the doubling trick set $S' = 0$ and otherwise set $S' = O\left(\sqrt{\tilde{O}(d) + \log(c/\delta)}\right)$, $S = \|f^*\|_\infty + S'$. Also let $T > 0$. Assume $m \geq D^2 \|y\|_{\mathbb{R}^n}^2 T^2$, and*

$$m \geq O(\log(c/\delta) + \tilde{O}(d)) \max\left\{T^2, 1\right\}.$$

Then with probability at least $1 - \delta$ we have that for all $t \leq T$ and $k \in \mathbb{N}$

$$\|P_k(r_t - \exp(-T_{K^\infty}t)r_0)\|_2 \leq \frac{1 - \exp(-\sigma_k t)}{\sigma_k}\tilde{O}\left(S\left[1 + tS\right]\frac{\sqrt{d}}{\sqrt{m}} + S(1+T)\frac{\sqrt{p}}{\sqrt{n}}\right),$$

and

$$\|r_t - \exp(-T_{K^\infty}t)r_0\|_2 \leq t\tilde{O}\left(S\left[1 + tS\right]\frac{\sqrt{d}}{\sqrt{m}} + S(1+T)\frac{\sqrt{p}}{\sqrt{n}}\right).$$

Theorem 2.3.5 will be proved in Section 2.5.3. The proof uses the uniform deviation bounds for the NTK to bound $T_n - T_n^s$ and tools from empirical process theory to show convergence of T_n to T_{K^∞} uniformly over a class of functions corresponding to networks with bounded parameter norms.

To interpret the results, we observe that to track the dynamics for eigenfunctions corresponding to eigenvalue σ_k and above, the expression under the $\tilde{O}$ needs to be small relative to $\frac{1}{\sigma_k}$. Thus the bias towards learning the eigenfunctions corresponding to large eigenvalues appears more pronounced. When $t = \log(\|r_0\|_2/\epsilon)/\sigma_k$, we have that $\|P_k \exp(-T_{K^\infty}t)r_0\|_2 \leq \epsilon$. Thus by applying this stopping time we get that to learn the eigenfunctions corresponding to eigenvalue σ_k and above up to ϵ accuracy we need $\frac{t^2}{\sqrt{m}} \lesssim \epsilon$ and $\frac{t^2\sqrt{p}}{\sqrt{n}} \lesssim \epsilon$ which translates to $m \gtrsim \sigma_k^{-4}\epsilon^{-2}$ and $n \gtrsim p\sigma_k^{-4}\epsilon^{-2}$. In typical NTK works the width m needs to be polynomially large relative to the number of samples n, where by contrast here the width depends on the inverse of the eigenvalues for the relevant components of the target function. From an approximation point-of-view this makes sense; the more complicated the target function the more expressive the model must be. We believe future works can adopt more precise requirements on the width m that do not require growth relative to the number of samples

n. To further illustrate the scaling of the parameters required by Theorem 2.3.5, we can apply Theorem 2.3.5 for an appropriate stopping time to get a bound on the test error.

Corollary 2.3.6. *Assume Assumptions 2.3.3 and 2.3.4 hold. Suppose that $f^* = O(1)$ and assume we are performing the doubling trick where $f_0 \equiv 0$ so that $r_0 = -f^*$. Let $k \in \mathbb{N}$ and let P_k be the orthogonal projection onto span$\{\phi_1, \ldots, \phi_k\}$. Set $t = \frac{\log(\sqrt{2}\|P_k f^*\|_2/\epsilon^{1/2})}{\sigma_k}$ Then we have that $m = \tilde{\Omega}(\frac{d}{\epsilon\sigma_k^4})$ and $n = \tilde{\Omega}\left(\frac{p}{\sigma_k^4\epsilon}\right)$ suffices to ensure with probability at least $1 - \delta$*

$$\frac{1}{2}\|r_t\|_2^2 \leq 2\epsilon + 2\|(I - P_k)f^*\|_2^2.$$

If one specialized to the case where f^* is a finite sum of eigenfunctions of T_{K^∞} (when the data is uniformly distributed on the sphere S^{d-1} and the network weights are from a rotation invariant distribution this corresponds to a finite sum of spherical harmonics, which in $d = 2$ is equivalently a bandlimited function) one can choose k such that $\|(I - P_k)f^*\|_2^2 = 0$. It is interesting to note that in this special case gradient flow with early stopping achieves essentially the same rates with respect to m and n (up to constants and logarithms) as the estimated network in the classical approximation theory paper by [Bar94]. It is also interesting to note that the approximation results by [Bar94] depend on the decay in frequency domain of the target function f^* via their constant C_{f^*}, and similarly for us the constant $1/\sigma_k^4$ grows with the bandwidth of the target function in the case of uniform distribution on the sphere S^1 which we mentioned parenthetically above.

While in Theorem 2.3.5 we compared the dynamics of r_t against that of $\exp(-T_{K^\infty}t)r_0$, the damped deviations equation given by Lemma 2.2.3 enables you to compare against $\exp(-T_K t)r_0$ for an arbitrary kernel K. There are other natural choices for K besides $K = K^\infty$, the most obvious being $K = K_0$. In Section 2.5.3.8 we prove a version of Theorem 2.3.5 where $K = K_0$ and θ_0 is an arbitrary deterministic parameter initialization. This could be interesting in scenarios where the parameters are initialized from a pretrained network or one has a priori knowledge that informs the selection of θ_0. One could let K be the kernel corresponding to some subset of parameters, such as the random feature kernel [RR08a]

corresponding to the outer layer. This would compare the dynamics of training all layers to that of training a subset of the parameters. If one wanted to account for adaptations of the kernel K_t one could try to set $K = K_{t_0}$ for some $t_0 > 0$. However since θ_{t_0} depends on the training data it is not obvious how one could produce a bound for $T_n^s - K_{t_0}$. Nevertheless we leave the suggestion open as a possibility for future work.

2.3.2 Overparameterized Regime

Once one has deviation bounds for the NTK so that the quantity $\|G^\infty - G_s\|_{op}$ is controlled, the damped deviations equation (Lemma 2.2.1) allows one to control the dynamics of the empirical risk. In this section we will demonstrate three such results that follow from this approach. The following is our analog of Theorem 4.1 from [ADH19a] in our setting, proved in Section 2.5.5. The result demonstrates that when the network is heavily overparameterized, the dynamics of the residual $\hat{r}_t$ follow the NTK regime dynamics $\exp(-G^\infty t)\hat{r}_0$.

Theorem 2.3.7. *Assume $m = \tilde{\Omega}(dn^5\epsilon^{-2}\lambda_n(H^\infty)^{-4})$ and $m \geq O(\log(c/\delta) + \tilde{O}(d))$ and $f^* = O(1)$. Assume we are performing the doubling trick so that $\hat{r}_0 = -y$. Let $v_1, \ldots, v_n$ denote the eigenvectors of G^∞ normalized to have unit L2 norm $\|v_i\|_2 = 1$. Then with probability at least $1 - \delta$*

$$\hat{r}_t = \exp(-G^\infty t)(-y) + \delta(t),$$

where $\sup_{t\geq 0} \|\delta(t)\|_2 \leq \epsilon$. In particular

$$\|\hat{r}_t\|_2 = \sqrt{\sum_{i=1}^{n} \exp(-2\lambda_i t)|\langle y, v_i\rangle_2|^2} \pm \epsilon.$$

In the work of [ADH19a] the requirement is $m = \Omega(\frac{n^7}{\lambda_n(H^\infty)^4\kappa^2\delta^4\epsilon^2})$ and $\kappa = O(\frac{\epsilon\delta}{\sqrt{n}})$ where $w_\ell \sim N(0, \kappa^2 I)$ (not to be confused with our definition of $\kappa := \max_x K^\infty(x,x)$). By contrast our weights have unit variance, which for Gaussian initialization corresponds to $w_\ell \sim N(0, I)$. They require κ to be small to ensure the neural network is small in magnitude at initialization. To achieve the same effect we can perform antisymmetric initializeation to ensure

the network is equivalently 0 at initializeation. Our overparameterization requirement ignoring logarithmic factors is smaller by a factor of $\frac{n^2}{d\delta^4}$. Again due to the different settings we do not claim superiority over this work.

The following is our analog of Theorem 4 in [SY19] proved in Section 2.5.6. This shows that when the target function has a compact representation in terms of eigenfunctions of T_{K^∞}, a more moderate overparameterization is sufficient to approximately solve the ERM problem.

Theorem 2.3.8. *Assume Assumptions 2.3.3 and 2.3.4 hold. Furthermore assume* $m = \tilde{\Omega}\left(\epsilon^{-2} dT^2 \left\|f^*\right\|_\infty^2 (1 + T\left\|f^*\right\|_\infty)^2\right)$ *where* $T > 0$ *is a time parameter and* $m \geq O(\log(c/\delta) + \tilde{O}(d))$ *and* $n \geq \frac{128\kappa^2 \log(2/\delta)}{(\sigma_k - \sigma_{k+1})^2}$. *Also assume* $f^* \in L^\infty(X) \subset L^2(X)$ *and let* $P^{T_{K^\infty}}$ *be the orthogonal projection onto the eigenspaces of* T_{K^∞} *corresponding to the eigenvalue* $\alpha \in \sigma(T_{K^\infty})$ *and higher. Assume that* $\left\|(I - P^{T_{K^\infty}})f^*\right\|_\infty \leq \epsilon'$ *for some* $\epsilon' \geq 0$. *Pick* k *so that* $\sigma_k = \alpha$ *and* $\sigma_{k+1} < \alpha$, *i.e.* k *is the index of the last repeated eigenvalue corresponding to* α *in the ordered sequence* $\{\sigma_i\}_i$. *Also assume we are performing the doubling trick so that* $\hat{r}(0) = -y$. *Then we have with probability at least* $1 - 3\delta$ *over the sampling of* $x_1, \ldots, x_n$ *and* θ_0 *that for* $t \leq T$

$$\left\|\hat{r}_t\right\|_{\mathbb{R}^n} \leq \exp(-\lambda_k t)\left\|y\right\|_{\mathbb{R}^n} + \frac{4\kappa\left\|f^*\right\|_2\sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}} + 2\epsilon' + \epsilon.$$

[SY19] have $\left\|f^*\right\|_2 \leq \left\|f^*\right\|_\infty \leq 1$, $\kappa \leq \frac{1}{2}$ and they treat d as a constant. Taking these into account we do not see the overparameterization requirements or bounds of either work being superior to the other. From Theorem 2.3.8, setting $\epsilon = \frac{4\kappa\|f^*\|_2\sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}}$ and $\epsilon' = 0$ we immediately get the analog of Corollary 2 in the work of [SY19]. This explains how in the special case that the target function is a finite sum of eigenfunctions of T_{K^∞}, the width m and the number of samples n can grow at the same rate, up to logarithms, and still solve the ERM problem. This is an ERM guarantee for $m = \tilde{\Omega}(n)$ and thus attains moderate overparameterization.

Corollary 2.3.9. *Assume Assumptions 2.3.3 and 2.3.4 hold. Furhtermore assume* $m = \tilde{\Omega}\left(\frac{n(\sigma_k - \sigma_{k+1})^2 d\|f^*\|_\infty^2 (1 + \lambda_k^{-1}\|f^*\|_\infty)^2}{\kappa^2\|f^*\|_2^2\lambda_k^2}\right)$ $m \geq O(\log(c/\delta) + \tilde{O}(d))$ $n \geq \frac{128\kappa^2\log(2/\delta)}{(\sigma_k - \sigma_{k+1})^2}$. *Let* f^*, $P^{T_{K^\infty}}$,

and k be the same as in the hypothesis of Theorem 2.3.8. Furthermore assume that

$$\left\|(I - P^{T_{K^\infty}})f^*\right\|_\infty = 0.$$

Also assume we are performing the doubling trick so that $\hat{r}(0) = -y$. Set

$$T = \log(\sqrt{n}\,\|\hat{r}(0)\|_{\mathbb{R}^n})/\lambda_k.$$

Then we have with probability at least $1 - 3\delta$ over the sampling of $x_1, \ldots, x_n$ and θ_0 that for $t \leq T$

$$\|\hat{r}_t\|_{\mathbb{R}^n} \leq \exp(-\lambda_k t)\,\|y\|_{\mathbb{R}^n} + \frac{8\kappa\,\|f^*\|_2\,\sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}}.$$

Note [SY19] are training only the hidden layer of a ReLU network by gradient descent, by contrast we are training both layers with biases of a network with smooth activations by gradient flow. For Corollary 2 in [SY19] they have the overparameterization requirement $m \gtrsim n \log n \left(\frac{1}{\lambda_k^4} + \frac{\log^4 n \log^2(1/\delta)}{(\lambda_k - \lambda_{k+1})^2 n^2 \lambda_k^4}\right)$. Thus both bounds scale like $\frac{n}{\lambda_k^4}$. Our bound has the extra factor $(\sigma_k - \sigma_{k+1})^2$ in front which could make it appear smaller at first glance but their Theorem 4 is strong enough to include this factor in the corollary they just chose not to. Thus we view both overparameterization requirements as comparable with neither superior to the other.

2.4 Conclusion and Future Directions

The damped deviations equation allows one to compare the dynamics when optimizing the squared error to that of an arbitrary kernel regression. We showed how this simple equation can be used to track the dynamics of the test residual in the underparameterized regime and extend existing results in the overparameterized setting. In the underparameterized setting the neural network learns eigenfunctions of the integral operator T_{K^∞} determined by the Neural Tangent Kernel at rates corresponding to their eigenvalues. In the overparameterized setting the damped deviations equation combined with NTK deviation bounds allows one

to track the dynamics of the empirical risk. In this fashion we extended existing work to the setting of a network with smooth activations where all parameters are trained as in practice. We hope damped deviations offers a simple interpretation of the MSE dynamics and encourages others to compare against other kernels in future work.

2.5 Appendix

2.5.1 Additional Notations

We let $[k] := \{1, 2, 3, \ldots, k\}$. For a set A we let $|A|$ denote its cardinality. $\|\bullet\|_F$ denotes the Frobenius norm for matrices, and for two matrices $A, B \in \mathbb{R}^{n\times m}$ we will let $\langle A, B\rangle = Tr(A^T B) = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}B_{i,j}$ denote the Frobenius or entry-wise inner product. We will let $B_R := \{x : \|x\|_2 \leq R\}$ to be the Euclidean ball of radius $R > 0$.

2.5.2 NTK Deviation and Parameter Norm Bounds

Let $\Gamma > 1$. At the end of this section we will prove a high probability bound of the form

$$\sup_{(x,x')\in B_M\times B_M} |K_t(x, x') - K^{\infty}(x, x')| = \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 \left\|\hat{r}(0)\right\|_{\mathbb{R}^n}\right]\right).$$

Ideally we would like to use the results in [HY20] where they prove for a deep feedforward network without biases:

$$\sup_{1\leq i,j\leq n} |K_t(x_i, x_j) - K^{\infty}(x_i, x_j)| = \tilde{O}\left(\frac{t^2}{m} + \frac{1}{\sqrt{m}}\right).$$

However there are three problems that prevent this. The first is that the have a constant under the $\tilde{O}$ above that depends on the training data. Specifically their Assumption 2.2 requires that the smallest singular value of the data matrix $[x_{\alpha_1}, \ldots, x_{\alpha_r}]$ is greater than $c_r > 0$ where $1 \leq \alpha_1, \ldots, \alpha_r \leq n$ are arbitrary distinct indices. As you send the number of samples to infinity you will have $c_r \to 0$, thus it is not clear how the bound will scale in the large sample regime. The second is that their bound only holds on the training data, whereas

we need a bound that is uniform over all inputs. The final one is their network does not have biases. In the following section we will overcome these issues. The main difference between our argument and theirs is how we prove convergence at initialization. In their argument for convergence at initialization they make repeated use of a Gaussian conditioning lemma as they pass through the layers, and this relies on their Assumption 2.2. By contrast we will use Lipschitzness of the NTK and convergence over an ϵ net to prove convergence at initialization. As we see it, our deviation bounds for the time derivative $\partial_t K_t$ are proved in a very similar fashion and the rest of the argument is very much inspired by their approach.

At the time of submission of this manuscript, we were made aware of the work by [LZB20b] that provides an alternative uniform NTK deviation bound by providing a uniform bound on the operator norm of the Hessian. Their work is very nice, and it opens the door to extending the results of this paper to the other architectures they consider. Nevertheless, we proceed with our original analysis below.

This section is conceptually simple but technical. We will take care to outline the high level structure of each section to prevent the technicalities from obfuscating the overall simplicity of the approach. Our argument runs through the following steps:

- Control parameter norms throughout training.
- Bound the Lipschitz constant of the NTK with respect to spatial inputs.
- Use concentration of subexponential random variables (Bernstein's inequality) to show that $|K^\infty(z') - K_0(z')| = \tilde{O}(1/\sqrt{m})$ (roughly) for all z' in an ϵ net of the spatial domain. Combine with the Lipschitz property of the NTK to show convergence over *all* inputs, namely $\sup_{z \in B_M} |K^\infty(z) - K_0(z)| = \tilde{O}(1/\sqrt{m})$ (roughly).
- Produce the bound $\sup_{z \in B_M \times B_M} |\partial_t K_t(z)| = \tilde{O}(1/\sqrt{m})$ (roughly).
- Conclude that $\sup_{z \in B_M \times B_M} |K_t(z) - K_0(z)| = \tilde{O}(t/\sqrt{m})$ (roughly).

2.5.2.1 Important Equations

The following list contains the equations that are relevant for this section. We found it easier to read the following proofs by keeping these equations on a separate piece of paper or in a separate tab. We write $a \otimes x = ax^T$. Also throughout this section the training data will be considered fixed and thus the randomness of the inputs is not relevant to this section. The randomness will come entirely from the parameter initialization θ_0.

$$f(x;\theta) = \frac{1}{\sqrt{m}} a^T \sigma(Wx + b) + b_0$$

$$x^{(1)} := \frac{1}{\sqrt{m}} \sigma(Wx + b)$$

$$\sigma_1'(x) := diag(\sigma'(Wx + b))$$

$$\sigma_1''(x) := diag(\sigma''(Wx + b))$$

$$\partial_a f(x;\theta) = \frac{1}{\sqrt{m}} \sigma(Wx + b) = x^{(1)}$$

$$\partial_W f(x;\theta) = \frac{1}{\sqrt{m}} \sigma_1'(x) a \otimes x$$

$$\partial_b f(x;\theta) = \frac{1}{\sqrt{m}} \sigma_1'(x) a$$

$$\partial_{b_0} f(x;\theta) = 1$$

$$\partial_t a = -\frac{1}{n} \sum_{i=1}^{n} \hat{r}_i x_i^{(1)}$$

$$\partial_t W = -\frac{1}{n} \sum_{i=1}^{n} \hat{r}_i \frac{1}{\sqrt{m}} \sigma_1'(x_i) a \otimes x_i$$

$$\partial_t b = -\frac{1}{n} \sum_{i=1}^{n} \hat{r}_i \frac{1}{\sqrt{m}} \sigma_1'(x_i) a$$

$$\partial_t b_0 = -\frac{1}{n} \sum_{i=1}^{n} \hat{r}_i$$

$$\partial_t x^{(1)} = \partial_t \frac{1}{\sqrt{m}}\sigma(Wx+b) = \frac{1}{\sqrt{m}}\sigma_1'(x)[\partial_t W x + \partial_t b]$$
$$= -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \left[\frac{1}{\sqrt{m}}\sigma_1'(x)\sigma_1'(x_i)\frac{a}{\sqrt{m}}\right][\langle x, x_i\rangle_2 + 1]$$

$$\partial_t \sigma_1'(x) = \partial_t \sigma'(Wx+b) = \sigma_1''(x) diag(\partial_t W x + \partial_t b)$$
$$= -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \frac{1}{\sqrt{m}}\sigma_1''(x)\sigma_1'(x_i) diag(a)[\langle x, x_i\rangle_2 + 1]$$

2.5.2.2 A Priori Parameter Norm Bounds

In this section we will provide bounds for the following quantities:

$$\xi(t) = \max\{\frac{1}{\sqrt{m}}\left\|W(t)\right\|_{op}, \frac{1}{\sqrt{m}}\left\|b(t)\right\|_2, \frac{1}{\sqrt{m}}\left\|a(t)\right\|_2, 1\},$$
$$\tilde{\xi}(t) = \max\{\max_{\ell\in[m]}\left\|w_\ell(t)\right\|_2, \left\|a(t)\right\|_\infty, \left\|b(t)\right\|_\infty, 1\}.$$

Here $w_\ell = W_{\ell,:} \in \mathbb{R}^d$ is the vector of input weights to the ℓth unit. These quantities appear repeatedly throughout the rest of the proofs of this section and thus need to be controlled. The parameter norm bounds will also be useful for the purpose of the covering number argument in Section 2.5.3.6. This section is broken down as follows:

- Prove Lemma 2.5.1
- Bound $\xi(t)$
- Bound $\tilde{\xi}(t)$

The time derivatives throughout will repeatedly be of the form $\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i(t)v_i$. Lemma 2.5.1 provides a simple bound that we will use over and over again.

Lemma 2.5.1. *Let $\|\bullet\|$ be any norm over a vector space V. Then for any $v_1, \ldots, v_n \in V$ we have*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i(t)v_i\right\| \leq \max_{i\in[n]}\|v_i\|\left\|\hat{r}(t)\right\|_{\mathbb{R}^n} \leq \max_{i\in[n]}\|v_i\|\left\|\hat{r}(0)\right\|_{\mathbb{R}^n}.$$

Proof. Note that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i(t)v_i\right\| \le \frac{1}{n}\sum_{i=1}^{n}|\hat{r}_i(t)|\,\|v_i\| \le \max_{i\in[n]}\|v_i\|\,\frac{1}{n}\sum_{i=1}^{n}|\hat{r}_i(t)|$$
$$\le \max_{i\in[n]}\|v_i\|\,\frac{1}{\sqrt{n}}\,\|\hat{r}(t)\|_2 = \max_{i\in[n]}\|v_i\|\,\|\hat{r}(t)\|_{\mathbb{R}^n} \le \max_{i\in[n]}\|v_i\|\,\|\hat{r}(0)\|_{\mathbb{R}^n},$$

where the last inequality follows from $\|\hat{r}(t)\|_{\mathbb{R}^n} \le \|\hat{r}(0)\|_{\mathbb{R}^n}$ from gradient flow. □

We now proceed to bound $\xi(t)$.

Lemma 2.5.2. *Let* $\xi(t) = \max\{\frac{1}{\sqrt{m}}\|W(t)\|_{op}, \frac{1}{\sqrt{m}}\|b(t)\|_2, \frac{1}{\sqrt{m}}\|a(t)\|_2, 1\}$ *and*

$$D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}.$$

Then for any initial conditions $W(0)$, $b(0)$, $a(0)$ *we have for all* t

$$\xi(t) \le \exp\left(\frac{D}{\sqrt{m}}\int_0^t \|\hat{r}(s)\|_{\mathbb{R}^n}\,ds\right)\xi(0) \le \exp\left(\frac{D}{\sqrt{m}}\|\hat{r}(0)\|_{\mathbb{R}^n}\,t\right)\xi(0).$$

Proof. Recall that

$$\partial_t a = -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i x_i^{(1)}$$
$$\partial_t W = -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\frac{1}{\sqrt{m}}\sigma_1'(x_i)a\otimes x_i$$
$$\partial_t b = -\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\frac{1}{\sqrt{m}}\sigma_1'(x_i)a.$$

We will show that each of the above derivatives is $\lesssim \|\hat{r}(t)\|_{\mathbb{R}^n}\,\xi(t)$ then apply Grönwall's inequality. By Lemma 2.5.1 it suffices to show that the terms multiplied by $\hat{r}_i$ in the above sums are $\lesssim \xi(t)$. First we note that

$$\left\|x_i^{(1)}\right\|_2 = \left\|\frac{1}{\sqrt{m}}\sigma(Wx_i+b)\right\|_2 \le |\sigma(0)| + \frac{1}{\sqrt{m}}\|\sigma'\|_\infty\|Wx_i+b\|_2$$
$$\le |\sigma(0)| + \frac{1}{\sqrt{m}}\|\sigma'\|_\infty\left[\|W\|_{op}\|x_i\|_2 + \|b\|_2\right] \le |\sigma(0)| + \frac{1}{\sqrt{m}}\|\sigma'\|_\infty\left[\|W\|_{op}M + \|b\|_2\right]$$
$$\le D\xi.$$

Second we have that

$$\left\| \frac{1}{\sqrt{m}} \sigma_1'(x_i) a \otimes x_i \right\|_{op} = \left\| \frac{1}{\sqrt{m}} \sigma_1'(x_i) a \right\|_2 \|x_i\|_2 \leq M \|\sigma'\|_\infty \frac{1}{\sqrt{m}} \|a\|_2 \leq D\xi.$$

Finally we have that

$$\left\| \frac{1}{\sqrt{m}} \sigma_1'(x_i) a \right\| \leq \|\sigma'\|_\infty \frac{1}{\sqrt{m}} \|a\|_2 \leq D\xi.$$

Thus by Lemma 2.5.1 and the above bounds we have

$$\|\partial_t W(t)\|_{op}, \|\partial_t a(t)\|_2, \|\partial_t b(t)\|_2 \leq D \|\hat{r}(t)\|_{\mathbb{R}^n} \xi(t).$$

Let $v(t)$ be a placeholder for one of the functions $\frac{1}{\sqrt{m}} a(t), \frac{1}{\sqrt{m}} W(t), \frac{1}{\sqrt{m}} b(t)$ with corresponding norm $\|\bullet\|$. Then we have that

$$\begin{aligned} \|v(t)\| &\leq \|v(0)\| + \|v(t) - v(0)\| = \|v(0)\| + \left\| \int_0^t \partial_s v(s) ds \right\| \\ &\leq \|v(0)\| + \int_0^t \|\partial_s v(s)\| \, ds \leq \xi(0) + \int_0^t \frac{\|\hat{r}(s)\|_{\mathbb{R}^n}}{\sqrt{m}} D\xi(s) ds. \end{aligned}$$

This inequality holds for any of the three choices of v thus we get that

$$\xi(t) \leq \xi(0) + \int_0^t \frac{\|\hat{r}(s)\|_{\mathbb{R}^n}}{\sqrt{m}} D\xi(s) ds.$$

Therefore by Grönwall's inequality we get that

$$\xi(t) \leq \exp\left(\frac{D}{\sqrt{m}} \int_0^t \|\hat{r}(s)\|_{\mathbb{R}^n} \, ds \right) \xi(0) \leq \exp\left(\frac{D}{\sqrt{m}} \|\hat{r}(0)\|_{\mathbb{R}^n} \, t \right) \xi(0).$$

□

We will now bound $\tilde{\xi(t)}$ using essentially the same argument as in the previous lemma.

Lemma 2.5.3. *Let* $\tilde{\xi}(t) = \max\{\max_{\ell \in [m]} \|w_\ell(t)\|_2, \|a(t)\|_\infty, \|b(t)\|_\infty, 1\}$ *and*

$$D = 3 \max\{|\sigma(0)|, M \|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}.$$

Then for any initial conditions $W(0)$, $b(0)$, $a(0)$ *we have for all* t

$$\tilde{\xi}(t) \leq \exp\left(\frac{D}{\sqrt{m}} \int_0^t \|\hat{r}(s)\|_{\mathbb{R}^n} \, ds \right) \tilde{\xi}(0) \leq \exp\left(\frac{D}{\sqrt{m}} \|\hat{r}(0)\|_{\mathbb{R}^n} \, t \right) \tilde{\xi}(0).$$

Proof. The proof is basically the same as Lemma 2.5.2. We have that

$$\partial_t w_\ell = -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \frac{a_\ell}{\sqrt{m}} \sigma'(\langle w_\ell, x_i\rangle_2 + b_\ell) x_i.$$

Now note

$$\left\| \frac{a_\ell}{\sqrt{m}} \sigma'(\langle w_\ell, x_i\rangle_2 + b_\ell) x_i \right\|_2 \leq \frac{1}{\sqrt{m}} \|a\|_\infty \|\sigma'\|_\infty M \leq \frac{D}{\sqrt{m}} \tilde{\xi}.$$

Thus by Lemma 2.5.1 we have that

$$\|\partial_t w_\ell(t)\|_2 \leq \frac{D}{\sqrt{m}} \|\hat{r}(t)\|_{\mathbb{R}^n} \tilde{\xi}(t).$$

On the other hand

$$\partial_t a = -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i x_i^{(1)},$$

with

$$\begin{aligned} \left\| x_i^{(1)} \right\|_\infty = \left\| \frac{1}{\sqrt{m}} \sigma(W x_i + b) \right\|_\infty &\leq \frac{1}{\sqrt{m}} \left[|\sigma(0)| + \|\sigma'\|_\infty \|W x_i + b\|_\infty \right] \\ &\leq \frac{1}{\sqrt{m}} \left[|\sigma(0)| + \|\sigma'\|_\infty \left(M \max_\ell \|w_\ell\|_2 + \|b\|_\infty \right) \right] \leq \frac{D}{\sqrt{m}} \tilde{\xi}. \end{aligned}$$

Thus again by Lemma 2.5.1 we have

$$\|\partial_t a(t)\|_\infty \leq \frac{D}{\sqrt{m}} \|\hat{r}(t)\|_{\mathbb{R}^n} \tilde{\xi}(t).$$

Finally we have

$$\partial_t b = -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \frac{1}{\sqrt{m}} \sigma_1'(x_i) a,$$

with

$$\left\| \frac{1}{\sqrt{m}} \sigma_1'(x_i) a \right\|_\infty \leq \frac{1}{\sqrt{m}} \|\sigma'\|_\infty \|a\|_\infty \leq \frac{D}{\sqrt{m}} \tilde{\xi}.$$

Again applying Lemma 2.5.1 one last time we get

$$\|\partial_t b(t)\|_\infty \leq \frac{D}{\sqrt{m}} \|\hat{r}(t)\|_{\mathbb{R}^n} \tilde{\xi}(t).$$

Therefore by the same argument as in Lemma 2.5.2 using Grönwall's inequality we get that

$$\tilde{\xi}(t) \leq \exp\left(\frac{D}{\sqrt{m}} \int_0^t \|\hat{r}(s)\|_{\mathbb{R}^n} \, ds \right) \tilde{\xi}(0) \leq \exp\left(\frac{D}{\sqrt{m}} \|\hat{r}(0)\|_{\mathbb{R}^n} \, t \right) \tilde{\xi}(0).$$

□

2.5.2.3 *NTK* is Lipschitz with Respect to Spatial Inputs

The NTK being Lipschitz with respect to spatial inputs is essential to our proof. The Lipschitz property means that to show convergence uniformly for *all* inputs it suffices to show convergence on an ϵ net of the spatial domain. Since the parameters are changing throughout time, the Lipschitz constant of the NTK will change throughout time. We will see that the Lipschitz constant depends on the quantities $\xi(t)$ and $\tilde{\xi}(t)$ from the previous Section 2.5.2.2.

The NTK $K_t(x, x')$ is a sum of terms of the form $g(x)^T g(x')$ where g is one of the derivatives $\partial_a f(x;\theta_t), \partial_b f(x;\theta_t), \partial_W f(x;\theta_t), \partial_{b_0} f(x;\theta_t)$. Since $\partial_{b_0} f(x;\theta_t) \equiv 1$ this term can be ignored for the rest of the section. The upcomming Lemma 2.5.4 shows that if g is Lipschitz and bounded then $(x, x') \mapsto g(x)^T g(x')$ is Lipschitz. This lemma guides the structure of this section:

- Prove Lemma 2.5.4
- Show that $\partial_a f(x;\theta), \partial_b f(x;\theta), \partial_W f(x;\theta)$ are bounded and Lipschitz
- Conclude the NTK is Lipschitz

Lemma 2.5.4. *Let $g : \mathbb{R}^k \to \mathbb{R}^l$ be L-Lipschitz with respect to the 2-norm, i.e.*

$$\|g(x) - g(z)\|_2 \leq L \|x - z\|_2$$

and satisfy $\|g(x)\|_2 \leq M$ for all x in some set $\mathcal{X}$. Then $K_g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$K_g(x, x') := g(x)^T g(x')$$

is ML-Lipschitz with respect to the norm

$$\|(x, x')\| := \|x\|_2 + \|x'\|_2 .$$

Proof. We have

$$\begin{aligned}
|K_g(x,x') - K_g(z,z')| &= |g(x)^T g(x') - g(z)^T g(z')| \\
&= |g(x)^T (g(x') - g(z'))| + |(g(x) - g(z))^T g(z')| \\
&\leq \|g(x)\|_2 \|g(x') - g(z')\|_2 + \|g(x) - g(z)\|_2 \|g(z')\|_2 \\
&\leq ML \|x' - z'\|_2 + ML \|x - z\|_2 \leq ML \|(x,x') - (z,z')\|.
\end{aligned}$$

□

By the previous Lemma 2.5.4, to show that the NTK is Lipschitz it suffices to show that $\partial_a f(x;\theta), \partial_b f(x;\theta), \partial_W f(x;\theta), \partial_{b_0} f(x;\theta)$ are bounded and Lipschitz. The following lemma bounds the norms of the derivatives $\partial_a f(x;\theta), \partial_W f(x;\theta), \partial_b f(x;\theta)$.

Lemma 2.5.5. *Let* $D = 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$ *and*

$$\xi = \max\{\frac{1}{\sqrt{m}}\|W\|_{op}, \frac{1}{\sqrt{m}}\|b\|_2, \frac{1}{\sqrt{m}}\|a\|_2, 1\}.$$

Then

$$\|\partial_a f(x;\theta)\|_2, \|\partial_W f(x;\theta)\|_F, \|\partial_b f(x;\theta)\|_2 \leq D\xi.$$

Proof. We have

$$\begin{aligned}
\|\partial_a f(x;\theta)\|_2 = \left\|\frac{1}{\sqrt{m}}\sigma(Wx+b)\right\|_2 &\leq |\sigma(0)| + \frac{1}{\sqrt{m}}\|\sigma'\|_\infty \|Wx+b\|_2 \\
&\leq |\sigma(0)| + \frac{1}{\sqrt{m}}\|\sigma'\|_\infty \left[\|W\|_{op}\|x\|_2 + \|b\|_2\right] \\
&\leq |\sigma(0)| + \frac{1}{\sqrt{m}}\|\sigma'\|_\infty \left[\|W\|_{op} M + \|b\|_2\right] \leq D\xi,
\end{aligned}$$

$$\|\partial_W f(x;\theta)\|_F = \left\|\frac{1}{\sqrt{m}}\sigma_1'(x)a \otimes x\right\|_F = \left\|\frac{1}{\sqrt{m}}\sigma_1'(x)a\right\|_2 \|x\|_2 \leq \frac{M}{\sqrt{m}}\|\sigma'\|_\infty \|a\|_2 \leq D\xi,$$

$$\|\partial_b f(x;\theta)\|_2 = \left\|\frac{1}{\sqrt{m}}\sigma_1'(x)a\right\|_2 \leq \frac{1}{\sqrt{m}}\|\sigma'\|_\infty \|a\|_2 \leq D\xi.$$

□

The following lemma demonstrates that $\partial_a f(x;\theta)$, $\partial_W f(x;\theta)$, and $\partial_b f(x;\theta)$ are Lipschitz as functions of the input x.

Lemma 2.5.6. *Let*

$$\xi = \max\{\frac{1}{\sqrt{m}}\left\|W\right\|_{op}, \frac{1}{\sqrt{m}}\left\|b\right\|_2, \frac{1}{\sqrt{m}}\left\|a\right\|_2, 1\},$$

$$\tilde{\xi} = \max\{\max_{\ell\in[m]}\left\|w_\ell\right\|_2, \left\|a\right\|_\infty, \left\|b\right\|_\infty, 1\},$$

$$D' = \max\{\left\|\sigma'\right\|_\infty, M\left\|\sigma''\right\|_\infty, \left\|\sigma''\right\|_\infty\},$$

$$L = 2\xi\tilde{\xi}D'.$$

Then $\partial_a f(x;\theta)$, $\partial_b f(x;\theta)$, $\partial_W f(x;\theta)$ are all L-Lipschitz with respect to the Euclidean norm $\left\|\bullet\right\|_2$. In symbols:

$$\left\|\partial_a f(x;\theta) - \partial_a f(y;\theta)\right\|_2 \le L\left\|x-y\right\|_2,$$

$$\left\|\partial_W f(x;\theta) - \partial_W f(y;\theta)\right\|_F \le L\left\|x-y\right\|_2,$$

$$\left\|\partial_b f(x;\theta) - \partial_b f(y;\theta)\right\|_2 \le L\left\|x-y\right\|_2.$$

Proof. We have

$$\left\|\partial_a f(x;\theta) - \partial_a f(y;\theta)\right\|_2 = \left\|\frac{1}{\sqrt{m}}(\sigma(Wx+b) - \sigma(Wy+b))\right\|_2$$
$$\le \frac{1}{\sqrt{m}}\left\|\sigma'\right\|_\infty\left\|W(x-y)\right\|_2 \le \frac{1}{\sqrt{m}}\left\|\sigma'\right\|_\infty\left\|W\right\|_{op}\left\|x-y\right\|_2 \le L\left\|x-y\right\|_2,$$

$$
\begin{aligned}
\|\partial_W f(x;\theta) - \partial_W f(y;\theta)\|_F &= \left\| \frac{1}{\sqrt{m}}\sigma_1'(x)a \otimes x - \frac{1}{\sqrt{m}}\sigma_1'(y)a \otimes y \right\|_F \\
&\leq \left\| \frac{1}{\sqrt{m}}\sigma_1'(x)a \otimes [x-y] \right\|_F + \left\| \frac{1}{\sqrt{m}}[\sigma_1'(x)a - \sigma_1'(y)a] \otimes y \right\|_F \\
&\leq \frac{1}{\sqrt{m}} \|\sigma_1'(x)a\|_2 \|x-y\|_2 + \frac{1}{\sqrt{m}} \|[\sigma_1'(x) - \sigma_1'(y)]a\|_2 \|y\|_2 \\
\leq \frac{1}{\sqrt{m}} \|\sigma'\|_\infty \|a\|_2 \|x-y\|_2 + \frac{1}{\sqrt{m}} \|\sigma'(Wx+b) - \sigma'(Wy+b)\|_\infty \|a\|_2 M \\
&\leq \frac{1}{\sqrt{m}} \|\sigma'\|_\infty \|a\|_2 \|x-y\|_2 + \frac{1}{\sqrt{m}} \|\sigma''\|_\infty \|W(x-y)\|_\infty \|a\|_2 M \\
\leq \frac{1}{\sqrt{m}} \|\sigma'\|_\infty \|a\|_2 \|x-y\|_2 + \frac{1}{\sqrt{m}} \|\sigma''\|_\infty \max_{\ell\in[m]} \|w_\ell\|_2 \|x-y\|_2 \|a\|_2 M \\
&\leq L \|x-y\|_2 ,
\end{aligned}
$$

$$
\begin{aligned}
\|\partial_b f(x;\theta) - \partial_b f(y;\theta)\|_2 &= \left\| \frac{1}{\sqrt{m}}\sigma_1'(x)a - \frac{1}{\sqrt{m}}\sigma_1'(y)a \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \|\sigma'(Wx+b) - \sigma'(Wy+b)\|_\infty \|a\|_2 \\
&\leq \frac{1}{\sqrt{m}} \|\sigma''\|_\infty \|W(x-y)\|_\infty \|a\|_2 \\
\leq \frac{1}{\sqrt{m}} \|\sigma''\|_\infty \max_{\ell\in[m]} \|w_\ell\|_2 \|x-y\|_2 \|a\|_2 &\leq L \|x-y\|_2 .
\end{aligned}
$$

□

Finally we can prove that the Neural Tangent Kernel is Lipschitz.

Theorem 2.5.7. *Consider the Neural Tangent Kernel*

$$K(x,y) = \langle \partial_a f(x;\theta), \partial_a f(y;\theta) \rangle_2 + \langle \partial_b f(x;\theta), \partial_b f(y;\theta) \rangle_2 + \langle \partial_W f(x;\theta), \partial_W f(y;\theta) \rangle + 1$$

and let

$$\xi = \max\{ \frac{1}{\sqrt{m}} \|W\|_{op}, \frac{1}{\sqrt{m}} \|b\|_2, \frac{1}{\sqrt{m}} \|a\|_2, 1 \},$$

$$\tilde{\xi} = \max\{\max_{\ell \in [m]} \|w_\ell\|_2, \|a\|_\infty, \|b\|_\infty, 1\},$$

$$D = 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\},$$

$$D' = \max\{\|\sigma'\|_\infty, M\|\sigma''\|_\infty, \|\sigma''\|_\infty\}.$$

Then the Neural Tangent Kernel is Lipschitz with respect to the norm

$$\|(x,y)\| := \|x\|_2 + \|y\|_2$$

with Lipschitz constant $L := 6DD'\xi^2\tilde{\xi}$. *In symbols:*

$$|K(x,y) - K(x',y')| \leq L\|(x,y) - (x',y')\|.$$

Proof. By Lemma 2.5.5, we have that the gradients are bounded

$$\|\partial_a f(x;\theta)\|_2, \|\partial_W f(x;\theta)\|_F, \|\partial_b f(x;\theta)\|_2 \leq D\xi.$$

Also by Lemma 2.5.6 the gradients are Lipschitz with Lipschitz constant $2\xi\tilde{\xi}D'$. Thus these two facts combined with Lemma 2.5.4 tell us that each of the three terms $\langle \partial_a f(x;\theta), \partial_a f(y;\theta)\rangle$, $\langle \partial_b f(x;\theta), \partial_b f(y;\theta)\rangle$, and $\langle \partial_W f(x;\theta), \partial_W f(y;\theta)\rangle$ are individually Lipschitz with constant $(D\xi)\cdot(2\xi\tilde{\xi}D')$. Thus the Lipschitz constant of the NTK itself is bounded by the sum of the 3 Lipschitz constants, for a total of $6DD'\xi^2\tilde{\xi}$. □

Using that the NTK at time zero $K_0(x,y)$ is Lipschitz we can prove that the analytical NTK $K^\infty = \mathbb{E}[K_0(x,y)]$ is Lipschitz. We will use this primarily as a qualitative statement, meaning that the estimate that we derive for the Lipschitz constant will not be used as it is not very explicit. Rather, in theorems where we use the fact that K^∞ is Lipschitz we will simply take the Lipschitz constant of K^∞ as an external parameter.

Theorem 2.5.8. *Assume that* $W_{i,j}(0) \sim \mathcal{W}$, $b_\ell(0) \sim \mathcal{B}$, $a_\ell(0) \sim \mathcal{A}$ *are all i.i.d. zero-mean, unit variance subgaussian random variables. Let*

$$\xi(0) = \max\{\frac{1}{\sqrt{m}}\|W(0)\|_{op}, \frac{1}{\sqrt{m}}\|b(0)\|_2, \frac{1}{\sqrt{m}}\|a(0)\|_2, 1\},$$

$$\tilde{\xi}(0) = \max\{\max_{\ell\in[m]} \|w_\ell(0)\|_2, \|a(0)\|_\infty, \|b(0)\|_\infty, 1\},$$
$$D = 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\},$$
$$D' = \max\{\|\sigma'\|_\infty, M\|\sigma''\|_\infty, \|\sigma''\|_\infty\}.$$

Then the analytical Neural Tangent Kernel $K^\infty(x,y) = \mathbb{E}[K_0(x,y)]$ *is Lipschitz with respect to the norm*

$$\|(x,y)\| := \|x\|_2 + \|y\|_2$$

with Lipschitz constant $\leq 6DD'\mathbb{E}[\xi^2\tilde{\xi}] < \infty$*. If one instead does the doubling trick then the same conclusion holds.*

Proof. First assume we are not doing the doubling trick. We note that

$$|K^\infty(x,y) - K^\infty(x',y')| = |\mathbb{E}[K_0(x,y)] - \mathbb{E}[K_0(x',y')]|$$
$$\leq \mathbb{E}\,|K_0(x,y) - K_0(x',y')| \leq 6DD'\mathbb{E}[\xi^2\tilde{\xi}]\,\|(x,y) - (x',y')\|$$

where the last line follows from the Lipschitzness of K_0 provided by Theorem 2.5.7. Using that $\|W(0)\|_{op} \leq \|W(0)\|_F$ and the fact that the Euclidean norm of a vector with i.i.d. subgaussian entries is subgaussian (Theorem 3.1.1 [Ver18]), we have that $\xi(0)$ and $\tilde{\xi}(0)$ are maximums of subgaussian random variables. Since a maximum of subgaussian random variables is subgaussian, we have that $\xi(0)$ and $\tilde{\xi}(0)$ are subgaussian. From the inequality $ab \leq \frac{1}{2}(a^2+b^2)$ we get $\mathbb{E}[\xi^2\tilde{\xi}] \leq \frac{1}{2}\mathbb{E}[\xi^4] + \frac{1}{2}\mathbb{E}[\tilde{\xi}^2] < \infty$ since moments of subgaussian random variables are all finite. Since the doubling trick does not change the distribution of K_0, the same conclusion holds under that initialization scheme. □

2.5.2.4 *NTK* Convergence at Initialization

In this section we prove that $\sup_{z\in B_M\times B_M} |K_0(z) - K^\infty(z)| = \tilde{O}(1/\sqrt{m})$. Our argument traces the following steps:

- Show that K_0 is sum of averages of m independent subexponential random variables

- Use subexponential concentration to show that $\sup_{z' \in \Delta} |K_0(z') - K^\infty(z')| = \tilde{O}(1/\sqrt{m})$ for all z' in an ϵ net Δ of $B_M \times B_M$

- Use that K_0 is Lipschitz and convergence over the epsilon net Δ to show that

$$\sup_{z \in B_M \times B_M} |K_0(z) - K^\infty(z)| = \tilde{O}(1/\sqrt{m}) \text{ (roughly)}$$

We recall the following definitions 2.5.6 and 2.7.5 from [Ver18].

Definition 2.5.9. *([Ver18]) Let Y be a random variable. Then we define the subgaussian norm of Y to be*

$$\|Y\|_{\psi_2} = \inf\{t > 0 : \mathbb{E} \exp(Y^2/t^2) \leq 2\}.$$

If $\|Y\|_{\psi_2} < \infty$, then we say Y is subgaussian.

Definition 2.5.10. *([Ver18]) Let Y be a random variable. Then we define the subexponential norm of Y to be*

$$\|Y\|_{\psi_1} = \inf\{t > 0 : \mathbb{E} \exp(|Y|/t) \leq 2\}.$$

If $\|Y\|_{\psi_1} < \infty$, then we say Y is subexponential.

We also recall the following useful lemma (Lemma 2.7.7 [Ver18]).

Lemma 2.5.11. *([Ver18]) Let X and Y be subgaussian random variables. Then XY is subexponential. Moreover*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

We recall one last definition (Definition 3.4.1 [Ver18])

Definition 2.5.12. *([Ver18]) A random vector $Y \in \mathbb{R}^k$ is called subgaussian if the one dimensional marginals $\langle Y, x \rangle$ are subgaussian random variables for all $x \in \mathbb{R}^k$. The subgaussian norm of Y is defined as*

$$\|Y\|_{\psi_2} = \sup_{x \in S^{k-1}} \|\langle Y, x \rangle\|_{\psi_2}.$$

The typical example of a subgaussian random vector is a random vector with independent subgaussian coordinates. The following lemma demonstrates that the NTK at initializeation is a sum of terms that are averages of independent subexponential random variables, which will enable us to use concentration arguments later.

Theorem 2.5.13. *Let w_ℓ, b_ℓ, a_ℓ all be independent subgaussian random variables with subgaussian norms satisfying $\|\bullet\|_{\psi_2} \leq K$. Furthermore assume $\|1\|_{\psi_2} \leq K$. Also let*

$$D = 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}.$$

Then for fixed x, y, each of the following

$$\langle \partial_a f(x;\theta), \partial_a f(y;\theta)\rangle, \langle \partial_b f(x;\theta), \partial_b f(y;\theta)\rangle, \langle \partial_W f(x;\theta), \partial_W f(y;\theta)\rangle$$

is an average of m independent subexponential random variables with subexponential norms bounded by D^2K^2.

Proof. We first observe that

$$\begin{aligned}\langle \partial_a f(x;\theta), \partial_a f(y;\theta)\rangle_2 &= \frac{1}{m}\langle \sigma(Wx+b), \sigma(Wy+b)\rangle_2 \\ &= \frac{1}{m}\sum_{\ell=1}^m \sigma(\langle w_\ell, x\rangle_2 + b_\ell)\sigma(\langle w_\ell, y\rangle_2 + b_\ell),\end{aligned}$$

$$\begin{aligned}\langle \partial_b f(x;\theta), \partial_b f(y;\theta)\rangle_2 &= \langle \frac{1}{\sqrt{m}}\sigma_1'(x)a, \frac{1}{\sqrt{m}}\sigma_1'(y)a\rangle_2 \\ &= \frac{1}{m}\sum_{\ell=1}^m a_\ell^2 \sigma'(\langle w_\ell, x\rangle_2 + b_\ell)\sigma'(\langle w_\ell, y\rangle_2 + b_\ell),\end{aligned}$$

$$\begin{aligned}\langle \partial_W f(x;\theta), \partial_W f(y;\theta)\rangle &= \langle \frac{1}{\sqrt{m}}\sigma_1'(x)a\otimes x, \frac{1}{\sqrt{m}}\sigma_1'(y)a\otimes y\rangle_2 \\ = \frac{1}{m}\langle \sigma_1'(x)a, \sigma_1'(y)a\rangle_2\langle x, y\rangle_2 &= \frac{\langle x, y\rangle_2}{m}\sum_{\ell=1}^m a_\ell^2\sigma'(\langle w_\ell, x\rangle_2 + b_\ell)\sigma'(\langle w_\ell, y\rangle_2 + b_\ell).\end{aligned}$$

Note that

$$|\sigma(\langle w_\ell, x\rangle_2 + b_\ell)| \leq |\sigma(0)| + \|\sigma'\|_\infty [|\langle w_\ell, x\rangle| + |b_\ell|].$$

Thus

$$\begin{aligned}\|\sigma(\langle w_\ell, x\rangle_2 + b_\ell)\|_{\psi_2} &\leq |\sigma(0)| \, \|1\|_{\psi_2} + \|\sigma'\|_\infty [\||\langle w_\ell, x\rangle|\|_{\psi_2} + \||b_\ell|\|_{\psi_2}] \\ &\leq |\sigma(0)| \, \|1\|_{\psi_2} + \|\sigma'\|_\infty [M \, \|w_\ell\|_{\psi_2} + \||b_\ell|\|_{\psi_2}] \\ &\leq 3 \max\{|\sigma(0)|, M \, \|\sigma'\|_\infty, \|\sigma'\|_\infty\} K \leq DK.\end{aligned}$$

Also

$$|a_\ell \sigma'(\langle w_\ell, x\rangle_2 + b_\ell)| \leq |a_\ell| \, \|\sigma'\|_\infty \leq D|a_\ell|,$$

therefore

$$\|a_\ell \sigma'(\langle w_\ell, x\rangle_2 + b_\ell)\|_{\psi_2} \leq D \, \|a_\ell\|_{\psi_2} \leq DK.$$

Finally

$$\big\||\langle x, y\rangle_2|^{1/2} a_\ell \sigma'(\langle w_\ell, x\rangle_2 + b_\ell)\big\|_{\psi_2} \leq M \, \|\sigma'\|_\infty \, \|a_\ell\|_{\psi_2} \leq DK.$$

It follows by Lemma 2.5.11 that each of $\langle \partial_a f(x;\theta), \partial_a f(y;\theta)\rangle$, $\langle \partial_W f(x;\theta), \partial_W f(y;\theta)\rangle$, and $\langle \partial_b f(x;\theta), \partial_b f(y;\theta)\rangle$ is an average of m independent subexponential random variables with subexponential norm $\|\bullet\|_{\psi_1} \leq D^2K^2$. □

We now recall the following Theorem from (Theorem 5.39 [Ver12]) which will be useful.

Lemma 2.5.14 ([Ver12])**.** *Let A be an $N \times n$ matrix whose rows A_i are independent subgaussian isotropic random vectors in $\mathbb{R}^n$. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-ct^2)$ one has the following bounds on the singular values*

$$\sqrt{N} - C\sqrt{n} - t \leq s_{min}(A) \leq s_{max}(A) \leq \sqrt{N} + C\sqrt{n} + t.$$

Here $C = C_K > 0$ depends only on the subgaussian norms $K = \max_i \|A_i\|_{\psi_2}$ of the rows.

Also the following special case of (Lemma 5.5 [Ver12]) will be useful for us.

Lemma 2.5.15 ([Ver12])**.** *Let Y be subgaussian. Then*

$$\mathbb{P}\left(|Y| > t\right) \leq C\exp(-ct^2/\left\|Y\right\|_{\psi_2}^2).$$

It will be useful to remind the reader that $C, c > 0$ denote absolute constants whose meaning will vary from statement-to-statement, as this abuse of notation becomes especially prevalent during the concentration of measure arguments of the rest of the section. The following lemma provides a concentration inequality for the maximum of subgaussian random variables which will be useful for bounding ξ and $\tilde{\xi}$ later which is necessary for bounding the Lipschitz constant of K_0.

Lemma 2.5.16. *Let $Y_1, \ldots, Y_n$ be subgaussian random variables with $\left\|Y_i\right\|_{\psi_2} \leq K$ for $i \in [n]$. Then there exists absolute constants $c, c', C > 0$ such that*

$$\mathbb{P}\left(\max_{i\in[n]} |Y_i| > t + K\sqrt{c' \log n}\right) \leq C\exp(-ct^2/K^2).$$

Proof. Since each Y_i is subgaussian we have for any $t \geq 0$ (Lemma 2.5.15)

$$\mathbb{P}\left(|Y_i| > t\right) \leq C\exp\left(-ct^2/\left\|Y_i\right\|_{\psi_2}^2\right).$$

By the union bound,

$$\begin{aligned}\mathbb{P}\left(\max_{i\in[n]} |Y_i| > t + K\sqrt{c^{-1}\log n}\right) &\leq \sum_{i=1}^{n} \mathbb{P}\left(|Y_i| > t + K\sqrt{c^{-1}\log n}\right) \\ &\leq nC\exp\left(-c\left[t + K\sqrt{c^{-1}\log n}\right]^2/K^2\right) = C\exp(-ct^2/K^2).\end{aligned}$$

Thus by setting $c' := c^{-1}$ we get the desired result. □

We now introduce a high probability bound for ξ.

Lemma 2.5.17. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\left\|w_\ell\right\|_{\psi_2}, \left\|a_\ell\right\|_{\psi_2}, \left\|b_\ell\right\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Let*

$$\xi = \max\{\frac{1}{\sqrt{m}}\left\|W\right\|_{op}, \frac{1}{\sqrt{m}}\left\|b\right\|_2, \frac{1}{\sqrt{m}}\left\|a\right\|_2\}.$$

Then with probability at least $1-\delta$

$$\xi \leq 1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}.$$

Proof. Note that by setting $t = \sqrt{c^{-1}\log(2/\delta)}$ in Lemma 2.5.14 we have that with probability at least $1-\delta$

$$\frac{1}{\sqrt{m}}\|W\|_{op} \leq 1 + \frac{C\sqrt{d}}{\sqrt{m}} + \frac{\sqrt{c^{-1}\log(2/\delta)}}{\sqrt{m}}.$$

Also by Theorem 3.1.1 in [Ver18]

$$\left\|\|a\|_2 - \sqrt{m}\right\|_{\psi_2} \leq CK^2,$$

$$\left\|\|b\|_2 - \sqrt{m}\right\|_{\psi_2} \leq CK^2.$$

Thus by Lemma 2.5.15 and a union bound we have with probability at least $1-2\delta$,

$$\frac{1}{\sqrt{m}}\|a\|_2, \frac{1}{\sqrt{m}}\|b\|_2 \leq 1 + \frac{C}{\sqrt{m}}K^2\sqrt{\log(c/\delta)}.$$

Thus by replacing every δ in the above arguments with $\delta/3$ and using the union bound we have with probability at least $1-\delta$

$$\xi \leq 1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}.$$

□

Similarly we now introduce a high probability bound for $\tilde{\xi}$.

Lemma 2.5.18. *Assume that* $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ *are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume* $\|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ *for each* $\ell \in [m]$ *where* $K \geq 1$. *Let*

$$\tilde{\xi} = \max\{\max_{\ell\in[m]}\|w_\ell\|_2, \|a\|_\infty, \|b\|_\infty\}.$$

Then with probability at least $1-\delta$ *we have*

$$\tilde{\xi} \leq \sqrt{d} + CK^2\left[\sqrt{\log(c/\delta)} + \sqrt{\log m}\right].$$

Proof. By Theorem 3.1.1 in [Ver18] we have

$$\left\| \|w_\ell\|_2 - \sqrt{d} \right\|_{\psi_2} \leq CK^2.$$

Well then by Lemma 2.5.16 there is a constant $c' > 0$ so that

$$\mathbb{P}\left(\max_{\ell \in [m]} \|w_\ell\|_2 - \sqrt{d} > t + CK^2\sqrt{c' \log m} \right)$$
$$\leq \mathbb{P}\left(\max_{\ell \in [m]} \left| \|w_\ell\|_2 - \sqrt{d} \right| > t + CK^2\sqrt{c' \log m} \right) \leq C\exp(-ct^2/K^4).$$

Thus by setting $t = CK^2\sqrt{\log(c/\delta)}$ we have with probability at least $1 - \delta$

$$\max_{\ell \in [m]} \|w_\ell\|_2 \leq \sqrt{d} + CK^2\sqrt{\log(c/\delta)} + CK^2\sqrt{\log m},$$

where we have absorbed the constant $\sqrt{c'}$ into C. Similarly by Lemma 2.5.16 and a union bound we get with probability at least $1 - 2\delta$ that

$$\|a\|_\infty, \|b\|_\infty \leq CK\sqrt{\log(c/\delta)} + CK\sqrt{\log m}.$$

Thus by replacing each δ with $\delta/3$ in the above arguments and using the union bound we get with probability at least $1 - \delta$

$$\tilde{\xi} \leq \sqrt{d} + CK^2\left[\sqrt{\log(c/\delta)} + \sqrt{\log m} \right].$$

□

We are now finally ready to prove the main theorem of this section.

Theorem 2.5.19. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Let*

$$D = 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\},$$

$$D' = \max\{\|\sigma'\|_\infty, M\|\sigma''\|_\infty, \|\sigma''\|_\infty\}.$$

Define

$$\rho(M,\sigma,d,K,\delta,m) :=$$
$$CDD'\left\{1+C\frac{\sqrt{d}+K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2\left\{\sqrt{d}+CK^2\left[\sqrt{\log(c/\delta)}+\sqrt{\log m}\right]\right\}.$$

Let $L(K^\infty)$ denote the Lipschitz constant of K^∞. If

$$m \geq C[\log(c/\delta)+2d\log(CM\max\{\rho,L(K^\infty)\}\sqrt{m})],$$

then with probability at least $1-\delta$

$$\sup_{z\in B_M\times B_M}|K_0(z)-K^\infty(z)| \leq$$
$$\frac{1}{\sqrt{m}}\left[1+CD^2K^2\sqrt{\log(c/\delta)+2d\log(CM\max\{\rho,L(K^\infty)\}\sqrt{m})}\right].$$

If one instead does the doubling trick then the same conclusion holds.

Proof. First assume we are not doing the doubling trick. Recall that by Theorem 2.5.7 that K_0 is Lipschitz with constant at most

$$CDD'\xi(0)^2\tilde{\xi(}0),$$

where ξ and $\tilde{\xi}$ are defined as in the theorem. Well then by Lemmas 2.5.17, 2.5.18 and a union bound we have with probability at least $1-2\delta$

$$\xi(0)^2\tilde{\xi}(0) \leq \left\{1+C\frac{\sqrt{d}+K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2\left\{\sqrt{d}+CK^2\left[\sqrt{\log(c/\delta)}+\sqrt{\log m}\right]\right\}.$$

Let $L(K_0), L(K^\infty)$ denote the Lipschitz constant of K_0 and K^∞ respectively. Then assuming the above inequality holds we have that

$$L(K_0) \leq \rho(M,\sigma,d,K,\delta,m). \tag{2.1}$$

For conciseness from now on we will suppress the arguments of ρ. Now set

$$\gamma := \frac{1}{2\max\{\rho,L(K^\infty)\}\sqrt{m}}.$$

Let $\mathcal{N}_\gamma(B_M)$ be the cardinality of a maximal γ-net of the ball $B_M = \{x : \|x\|_2 \leq M\}$ with respect to the L2 norm $\|\bullet\|_2$. By a standard volume argument we have that

$$\mathcal{N}_\gamma(B_M) \leq \left(\frac{CM}{\gamma}\right)^d.$$

By taking the product of two $\gamma/2$ nets of B_M it follows that we can choose a γ net of $B_M \times B_M$, say Δ, with respect to the norm

$$\|(x,y)\| = \|x\|_2 + \|y\|_2$$

such that

$$|\Delta| \leq \left|\mathcal{N}_{\gamma/2}(B_M)\right|^2 \leq \left(\frac{CM}{\gamma}\right)^{2d} =: \mathcal{M}_\gamma.$$

By Theorem 2.5.13 for $(x,y) \in B_M \times B_M$ fixed each of the following

$$\langle \partial_a f(x;\theta), \partial_a f(y;\theta)\rangle_2, \langle \partial_b f(x;\theta), \partial_b f(y;\theta)\rangle_2, \langle \partial_W f(x;\theta), \partial_W f(y;\theta)\rangle$$

is an average of m subexponential random variables with subexponential norm at most D^2K^2. Therefore separately from the randomness discussed before by Bernstein's inequality (Theorem 2.8.1 [Ver18]) and a union bound we have

$$\mathbb{P}(|K_0(x,y) - K^\infty(x,y)| > t) \leq 3 \times 2\exp\left(-c\min\left\{\frac{mt^2}{D^4K^4}, \frac{mt}{D^2K^2}\right\}\right).$$

Thus for $t \leq D^2K^2$ we have

$$\mathbb{P}(|K_0(x,y) - K^\infty(x,y)| > t) \leq 6\exp\left(-c\frac{mt^2}{D^4K^4}\right).$$

Then by a union bound and the previous inequality we have that for $t \leq D^2K^2$

$$\mathbb{P}\left(\max_{z'\in\Delta}|K_0(z') - K^\infty(z')| > t\right) \leq 6\mathcal{M}_\gamma \exp\left(-c\frac{mt^2}{D^4K^4}\right).$$

Thus by setting $t = CD^2K^2\frac{\sqrt{\log(c/\delta)+\log\mathcal{M}_\gamma}}{\sqrt{m}}$ (note that the condition on m in the hypothesis ensures that $t \leq D^2K^2$) we get that with probability $1-\delta$

$$\max_{z'\in\Delta}|K_0(z') - K^\infty(z')| \leq t.$$

Now fix $z \in B_M \times B_M$ and choose $z' \in \Delta$ such that $\|z - z'\| \leq \gamma$. Then

$$\begin{aligned} |K_0(z) - K^\infty(z)| &\leq |K_0(z) - K_0(z')| \\ &+ |K_0(z') - K^\infty(z')| + |K^\infty(z') - K^\infty(z)| \\ &\leq 2\max\{L(K_0), L(K^\infty)\}\gamma + t. \end{aligned} \tag{2.2}$$

Note that this argument runs through for any $z \in B_M \times B_M$ therefore

$$\sup_{z \in B_M \times B_M} |K_0(z) - K^\infty(z)| \leq 2\max\{L(K_0), L(K^\infty)\}\gamma + t.$$

Well by replacing δ with $\delta/3$ in the previous arguments by taking a union bound we can assume that equations (2.1) and (2.2) hold simultaneously. In which case

$$\begin{aligned} &\sup_{z \in B_M \times B_M} |K_0(z) - K^\infty(z)| \leq 2\max\{L(K_0), L(K^\infty)\}\gamma + t \leq 2\max\{\rho, L(K^\infty)\}\gamma + t \\ &\leq \frac{1}{\sqrt{m}} + CD^2K^2\frac{\sqrt{\log(c/\delta) + \log \mathcal{M}_\gamma}}{\sqrt{m}} = \frac{1}{\sqrt{m}} + CD^2K^2\frac{\sqrt{\log(c/\delta) + 2d\log(CM/\gamma)}}{\sqrt{m}} \\ &= \frac{1}{\sqrt{m}}\left[1 + CD^2K^2\sqrt{\log(c/\delta) + 2d\log(CM\max\{\rho, L(K^\infty)\}\sqrt{m})}\right], \end{aligned}$$

where we have used the definition of $\mathcal{M}_\gamma$ in the second-to-last equality and the definition of γ in the last equality. Since the doubling trick does not change the distribution of K_0, the same conclusion holds under that initialization scheme. □

We immediately get the following corollary.

Corollary 2.5.20. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Then*

$$m \geq C[\log(c/\delta) + \tilde{O}(d)]$$

suffices to ensure that with probability at least $1 - \delta$

$$\sup_{z \in B_M \times B_M} |K_0(z) - K^\infty(z)| = \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\right).$$

If one instead does the doubling trick then the same conclusion holds.

2.5.2.5 Control of Network at Initialization

Many of our previous results depend on the quantity $\|\hat{r}(0)\|_{\mathbb{R}^n}$ which depends on the network at initialization. Before we proceed we must control the infinity norm of the network at initialization and work out a few consequences of this. The following lemma controls $\|f(\bullet;\theta_0)\|_\infty$.

Lemma 2.5.21. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$, $b_0 \sim \mathcal{B}'$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\|1\|_{\psi_2}, \|w_\ell\|_{\psi_2}$, $\|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Let*

$$D = 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\},$$

$$L(m,\sigma,d,K,\delta) := \sqrt{m}\|\sigma'\|_\infty \left\{1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2.$$

Assume that

$$m \geq C[\log(c/\delta) + d\log(CML)].$$

Then with probability at least $1-\delta$

$$\sup_{x\in B_M} |f(x;\theta_0)| \leq CDK^2\sqrt{d\log(CML) + \log(c/\delta)} = \tilde{O}(\sqrt{d}).$$

Proof. First we note that

$$\left|\frac{a^T}{\sqrt{m}}\sigma(Wx+b) - \frac{a^T}{\sqrt{m}}\sigma(Wy+b)\right| \leq \frac{\|a\|_2}{\sqrt{m}}\|\sigma(Wx+b) - \sigma(Wy+b)\|_2$$
$$\leq \frac{\|a\|_2}{\sqrt{m}}\|\sigma'\|_\infty \|W(x-y)\|_2 \leq \frac{\|a\|_2}{\sqrt{m}}\|\sigma'\|_\infty \|W\|_{op}\|x-y\|_2 \leq \sqrt{m}\|\sigma'\|_\infty \xi(0)^2\|x-y\|_2,$$

where $\xi(0)$ is defined as in Lemma 2.5.17. Thus $f(\bullet;\theta_0)$ is Lipschitz with constant $L = \sqrt{m}\|\sigma'\|_\infty \xi(0)^2$. Well then by Lemma 2.5.17 we have with probability at least $1-\delta$

$$\xi(0)^2 \leq \left\{1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2. \tag{2.3}$$

When the above holds we have that $f(\bullet;\theta_0)$ is Lipschitz with constant

$$L := \sqrt{m}\,\|\sigma'\|_\infty \left\{1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2.$$

On the other hand note that

$$|\sigma(\langle w_\ell, x\rangle_2 + b_\ell)| \le |\sigma(0)| + \|\sigma'\|_\infty [|\langle w_\ell, x\rangle_2| + |b_\ell|].$$

Thus

$$\begin{aligned}\|\sigma(\langle w_\ell, x\rangle_2 + b_\ell)\|_{\psi_2} &\le |\sigma(0)|\,\|1\|_{\psi_2} + \|\sigma'\|_\infty [\||\langle w_\ell, x\rangle_2|\|_{\psi_2} + \||b_\ell|\|_{\psi_2}] \\ &\le |\sigma(0)|\,\|1\|_{\psi_2} + \|\sigma'\|_\infty [M\,\|w_\ell\|_{\psi_2} + \||b_\ell|\|_{\psi_2}] \\ &\le 3\max\{|\sigma(0)|, M\,\|\sigma'\|_\infty, \|\sigma'\|_\infty\} K \le DK.\end{aligned}$$

Therefore by Lemma 2.5.11 we have

$$\|a_\ell\sigma(\langle w_\ell, x\rangle_2 + b_\ell)\|_{\psi_1} \le DK^2.$$

Thus for each x fixed we have by Bernstein's inequality (Theorem 2.8.1 [Ver18]).

$$\mathbb{P}\left(\left|\sum_{\ell=1}^m a_\ell\sigma(\langle w_\ell, x\rangle_2 + b_\ell)\right| > t\sqrt{m}\right) \le 2\exp\left(-c\min\left[\frac{t^2}{[DK^2]^2}, \frac{t\sqrt{m}}{DK^2}\right]\right).$$

Thus for $t \le \sqrt{m}DK^2$ this simplifies to

$$\mathbb{P}\left(\left|\sum_{\ell=1}^m a_\ell\sigma(\langle w_\ell, x\rangle_2 + b_\ell)\right| > t\sqrt{m}\right) \le 2\exp\left(-c\frac{t^2}{D^2K^4}\right).$$

Let Δ be a γ net of the ball $B_M = \{x : \|x\|_2 \le M\}$ with respect to the Euclidean $\|\bullet\|_2$ norm. Then by a standard volume argument we have that

$$|\Delta| \le \left(\frac{CM}{\gamma}\right)^d =: \mathcal{M}_\gamma.$$

Thus by a union bound we have for $t \le \sqrt{m}DK^2$

$$\mathbb{P}\left(\max_{x\in\Delta}\left|\sum_{\ell=1}^m a_\ell\sigma(\langle w_\ell, x\rangle_2 + b_\ell)\right| > t\sqrt{m}\right) \le 2\mathcal{M}_\gamma\exp\left(-c\frac{t^2}{D^2K^4}\right).$$

Thus by setting $t = CDK^2\sqrt{\log(c\mathcal{M}_\gamma/\delta)}$ assuming $t \leq \sqrt{m}DK^2$ we have with probability at least $1-\delta$

$$\max_{x\in\Delta}\left|\sum_{\ell=1}^{m}\frac{a_\ell}{\sqrt{m}}\sigma(\langle w_\ell, x\rangle_2 + b_\ell)\right| \leq t. \tag{2.4}$$

On the other hand by Lemma 2.5.15 our prior definition of t is large enough (up to a redefinition of the constants c, C) to ensure that with probability at least $1-\delta$

$$|b_0| \leq t. \tag{2.5}$$

When (2.4) and (2.5) hold simultaneously we have that $\max_{x'\in\Delta}|f(x',\theta_0)| \leq 2t$. By a union bound we have with probability at least $1-3\delta$ that (2.3), (2.4), (2.5) hold simultaneously. Well then for any $x \in B_M$ we may choose $x' \in \Delta$ so that $\|x-x'\|_2 \leq \gamma$. Then

$$|f(x;\theta_0)| \leq |f(x';\theta_0)| + |f(x;\theta_0) - f(x';\theta_0)| \leq 2t + L\gamma.$$

Therefore

$$\sup_{x\in B_M}|f(x;\theta_0)| \leq 2t + L\gamma$$

and this argument runs through for any $\gamma > 0$. We will set $\gamma = 1/L$. Note that for this choice of γ the hypothesis on m ensures that $t \leq \sqrt{m}DK^2$. Thus the preceding argument goes through in this case. Thus by replacing δ with $\delta/3$ in the previous argument we get the desired conclusion up to a redefinition of c, C. □

We quickly introduce the following lemma.

Lemma 2.5.22.

$$\|\hat{r}(0)\|_{\mathbb{R}^n} \leq \|f(\bullet;\theta_0)\|_\infty + \|y\|_{\mathbb{R}^n}.$$

Proof. Let $\hat{y} \in \mathbb{R}^n$ be the vector whose ith entry is equal to $f(x_i;\theta_0)$. Well then note that $\|\hat{y}\|_{\mathbb{R}^n} \leq \|f(\bullet;\theta_0)\|_\infty$. Therefore

$$\|\hat{r}(0)\|_{\mathbb{R}^n} = \|\hat{y}-y\|_{\mathbb{R}^n} \leq \|\hat{y}\|_{\mathbb{R}^n} + \|y\|_{\mathbb{R}^n} \leq \|f(\bullet;\theta_0)\|_\infty + \|y\|_{\mathbb{R}^n}.$$

□

Finally we prove one last lemma that will be useful later.

Lemma 2.5.23. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\|1\|_{\psi_2}, \|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Let $\Gamma > 1$, $D = 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$,*

$$L(m, \sigma, d, K, \delta) := \sqrt{m}\|\sigma'\|_\infty \left\{1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2,$$

$$\rho := CDK^2\sqrt{d\log(CML) + \log(c/\delta)} = \tilde{O}(\sqrt{d}).$$

Suppose

$$m \geq \frac{4D^2\|y\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2} \text{ and } m \geq \max\left\{\frac{4D^2\rho^2T^2}{[\log(\Gamma)]^2}, \left(\frac{\rho}{DK^2}\right)^2\right\}.$$

Then with probability at least $1 - \delta$

$$\max_{t\leq T}\xi(t) \leq \Gamma\xi(0) \quad \max_{t\leq T}\tilde{\xi}(t) \leq \Gamma\tilde{\xi}(0),$$

where $\xi(t)$ and $\tilde{\xi}(t)$ are defined as in Lemmas 2.5.2 and 2.5.3. If one instead does the doubling trick the second hypothesis on m can be removed and the conclusion holds with probability 1.

Proof. First assume we are not doing the doubling trick. Well from the condition $m \geq \left(\frac{\rho}{DK^2}\right)^2$ we have by Lemma 2.5.21 that with probability at least $1 - \delta$

$$\sup_{x\in B_M}|f(x;\theta_0)| \leq CDK^2\sqrt{d\log(CML) + \log(c/\delta)} =: \rho.$$

Also by Lemma 2.5.22 and the above bound we have

$$\|\hat{r}(0)\|_{\mathbb{R}^n} \leq \|y\|_{\mathbb{R}^n} + \rho.$$

Well in this case

$$\frac{D\|\hat{r}(0)\|_{\mathbb{R}^n}t}{\sqrt{m}} \leq \frac{D[\|y\|_{\mathbb{R}^n} + \rho]t}{\sqrt{m}} \leq \frac{2D\max\{\|y\|_{\mathbb{R}^n}, \rho\}t}{\sqrt{m}} \leq \log(\Gamma),$$

where we have used the hypothesis on m in the last inequality. Therefore by Lemmas 2.5.2, 2.5.3 we have in this case that

$$\max_{t\leq T}\xi(t) \leq \Gamma\xi(0), \quad \max_{t\leq T}\tilde{\xi}(t) \leq \Gamma\tilde{\xi}(0).$$

Now assume we are performing the doubling trick so that $f(\bullet;\theta_0) \equiv 0$. Then ρ in the previous argument can simply be replaced with zero and the same argument runs through, except using Lemma 2.5.21 is no longer necessary (and thus the second hypothesis on m is not needed). Without using Lemma 2.5.21 the whole argument is deterministic so that the conclusion holds with probability 1. □

2.5.2.6 *NTK* Time Deviations Bounds

In this section we bound the deviations of the NTK throughout time. This section runs through the following steps

- Bound $\sup_{(x,y)\in B_M\times B_M} |\partial_t K_t(x,y)|$
- Bound $\sup_{(x,y)\in B_M\times B_M} |K_t(x,y) - K_0(x,y)|$

In the following lemma we will provide an upper bound on the NTK derivative

$$\sup_{x,y\in B_M\times B_M} |\partial_t K_t(x,y)|.$$

Lemma 2.5.24. *Let*

$$\xi(t) = \max\{\frac{1}{\sqrt{m}} \|W(t)\|_{op}, \frac{1}{\sqrt{m}} \|b(t)\|_2, \frac{1}{\sqrt{m}} \|a(t)\|_2, 1\},$$

$$\tilde{\xi}(t) = \max\{\max_{\ell\in[m]} \|w_\ell(t)\|_2, \|a(t)\|_\infty, \|b(t)\|_\infty, 1\},$$

$$D = 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\},$$

$$D' := \left[\max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\}^2[M^2+1] + D\|\sigma'\|_\infty\right]\max\{1, M\}.$$

Then for any initial conditions $W(0)$, $b(0)$, $a(0)$ we have for all t

$$\sup_{x,y\in B_M\times B_M} |\partial_t K_t(x,y)| \le \frac{CDD'}{\sqrt{m}}\xi(t)^2\tilde{\xi}(t)\|\hat{r}(t)\|_{\mathbb{R}^n}.$$

Proof. We need to bound the following time derivatives

$$\partial_t \partial_a f(x;\theta) = \partial_t x^{(1)} = -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \left[\frac{1}{\sqrt{m}}\sigma_1'(x)\sigma_1'(x_i)\frac{a}{\sqrt{m}}\right][\langle x, x_i\rangle_2 + 1],$$

$$\partial_t \partial_W f(x;\theta) = \partial_t \frac{1}{\sqrt{m}}\sigma_1'(x) a \otimes x$$
$$= \frac{1}{\sqrt{m}}[(\partial_t \sigma_1'(x))a + \sigma_1'(x)(\partial_t a)] \otimes x,$$

$$\partial_t \partial_b f(x;\theta) = \partial_t \frac{1}{\sqrt{m}}\sigma_1'(x) a = \frac{1}{\sqrt{m}}\left([\partial_t \sigma_1'(x)]a + \sigma_1'(x)\partial_t a\right).$$

Note that

$$\left\| \frac{1}{\sqrt{m}}\sigma_1'(x)\sigma_1'(x_i)\frac{a}{\sqrt{m}}[\langle x, x_i\rangle_2 + 1] \right\|_2 \leq \frac{1}{\sqrt{m}} \left\|\sigma'\right\|_\infty^2 \frac{1}{\sqrt{m}} \left\|a\right\|_2 [M^2 + 1].$$

Thus by Lemma 2.5.1

$$\left\|\partial_t \partial_a f(x;\theta)\right\|_2 \leq \frac{1}{\sqrt{m}} \left\|\sigma'\right\|_\infty^2 \frac{1}{\sqrt{m}} \left\|a\right\|_2 [M^2 + 1] \left\|\hat{r}(t)\right\|_{\mathbb{R}^n}$$
$$\leq \frac{\left\|\sigma'\right\|_\infty^2 [M^2 + 1]]}{\sqrt{m}} \xi(t) \left\|\hat{r}(t)\right\|_{\mathbb{R}^n}.$$

On the other hand

$$[\partial_t \sigma_1'(x)]a = -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \frac{1}{\sqrt{m}}\sigma_1''(x)\sigma_1'(x_i) diag(a) a[\langle x, x_i\rangle_2 + 1].$$

Well

$$\frac{1}{\sqrt{m}} \left\|\sigma_1''(x)\sigma_1'(x_i) diag(a) a[\langle x, x_i\rangle_2 + 1\right\| \leq \frac{1}{\sqrt{m}} \left\|\sigma''\right\|_\infty \left\|\sigma'\right\|_\infty \left\|a\right\|_\infty \left\|a\right\|_2 [M^2 + 1]$$
$$\leq \left\|\sigma''\right\|_\infty \left\|\sigma'\right\|_\infty [M^2 + 1] \xi(t) \tilde{\xi}(t).$$

Thus by Lemma 2.5.1 we have that

$$\left\|[\partial_t \sigma_1'(x)]a\right\| \leq \left\|\sigma''\right\|_\infty \left\|\sigma'\right\|_\infty [M^2 + 1] \xi(t) \tilde{\xi}(t) \left\|\hat{r}(t)\right\|_{\mathbb{R}^n}.$$

Finally we have

$$\sigma_1'(x)\partial_t a = -\frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \sigma_1'(x) x_i^{(1)}.$$

Well

$$\begin{aligned}
\left\|\sigma_1'(x)x_i^{(1)}\right\| &\leq \|\sigma'\|_\infty \left\|x_i^{(1)}\right\|_2 = \|\sigma'\|_\infty \frac{1}{\sqrt{m}} \|\sigma(Wx_i+b)\|_2 \\
&\leq \|\sigma'\|_\infty [|\sigma(0)| + \frac{1}{\sqrt{m}} \|\sigma'\|_\infty (\|W\|_{op} M + \|b\|_2)] \\
&\leq \|\sigma'\|_\infty [|\sigma(0)| + M\|\sigma'\|_\infty + \|\sigma'\|_\infty] \xi(t) \leq \|\sigma'\|_\infty D\xi(t).
\end{aligned}$$

Thus we finally by Lemma 2.5.1 again we get that

$$\|\sigma_1'(x)\partial_t a\| \leq \|\sigma'\|_\infty D\xi(t) \|\hat{r}(t)\|_{\mathbb{R}^n}.$$

It follows that

$$\begin{aligned}
&\|\partial_t \partial_b f(x;\theta)\| \\
&\leq \frac{1}{\sqrt{m}} \left[\|\sigma''\|_\infty \|\sigma'\|_\infty [M^2+1] + \|\sigma'\|_\infty D\right] \xi(t)\tilde{\xi}(t) \|\hat{r}(t)\|_{\mathbb{R}^n},
\end{aligned}$$

and similarly

$$\begin{aligned}
&\|\partial_t \partial_W f(x;\theta)\| \\
&\leq \frac{M}{\sqrt{m}} \left[\|\sigma''\|_\infty \|\sigma'\|_\infty [M^2+1] + \|\sigma'\|_\infty D\right] \xi(t)\tilde{\xi}(t) \|\hat{r}(t)\|_{\mathbb{R}^n}.
\end{aligned}$$

Thus in total we can say

$$\|\partial_t \partial_a f(x;\theta)\|_2, \|\partial_t \partial_b f(x;\theta)\|_2, \|\partial_t \partial_w f(x;\theta)\|_F \leq \frac{D'}{\sqrt{m}} \xi(t)\tilde{\xi}(t) \|\hat{r}(t)\|_{\mathbb{R}^n}.$$

It thus follows by the chain rule and Lemma 2.5.5 that

$$\sup_{(x,y)\in B_M \times B_M} |\partial_t K_t(x,y)| \leq \frac{CDD'}{\sqrt{m}} \xi(t)^2 \tilde{\xi}(t) \|\hat{r}(t)\|_{\mathbb{R}^n}.$$

□

Using the previous lemma we can now bound the deviations of the NTK.

Theorem 2.5.25. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Let $\Gamma > 1$ and $T > 0$ be positive constants,*

$$D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\},$$

$$D' := \left[\max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\}^2[M^2+1] + D\|\sigma'\|_\infty\right]\max\{1, M\},$$

and assume

$$m \geq \frac{4D^2\|y\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2} \text{ and } m \geq \max\left\{\frac{4D^2 O(\log(c/\delta)+\tilde{O}(d))T^2}{[\log(\Gamma)]^2}, O(\log(c/\delta)+\tilde{O}(d))\right\}.$$

Then with probability at least $1-\delta$

$$\sup_{(x,y)\in B_M\times B_M} |K_t(x,y) - K_0(x,y)| \leq$$
$$t\Gamma^3 \frac{CDD'}{\sqrt{m}}\|\hat{r}(0)\|_{\mathbb{R}^n}\left\{1 + C\frac{\sqrt{d}+K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2\left\{\sqrt{d}+CK^2\left[\sqrt{\log(c/\delta)}+\sqrt{\log m}\right]\right\}.$$

If one instead does the doubling trick then the second condition on m can be removed from the hypothesis and the same conclusion holds.

Proof. First assume we are not doing the doubling trick. By Lemmas 2.5.17, 2.5.18 and a union bound we have with probability at least $1-\delta$

$$\xi(0)^2\tilde{\xi}(0) \leq \left\{1 + C\frac{\sqrt{d}+K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2\left\{\sqrt{d}+CK^2\left[\sqrt{\log(c/\delta)}+\sqrt{\log m}\right]\right\}. \quad (2.6)$$

Note that ρ as defined in Lemma 2.5.23 satisfies $\rho^2 = O(\log(c/\delta)+\tilde{O}(d))$. Thus the hypothesis on m is strong enough to apply Lemma 2.5.23, therefore by applying this lemma we have with probability at least $1-\delta$

$$\max_{t\leq T}\xi(t) \leq \Gamma\xi(0), \quad \max_{t\leq T}\tilde{\xi}(t) \leq \Gamma\tilde{\xi}(0). \quad (2.7)$$

Thus by replacing δ with $\delta/2$ and taking a union bound we have that with probability at least $1-\delta$ (2.6) and (2.7) hold simultaneously. Then using Lemma 2.5.24 and the fact that $\|\hat{r}(t)\|_{\mathbb{R}^n} \leq \|\hat{r}(0)\|_{\mathbb{R}^n}$ we have for $t \leq T$

$$\sup_{(x,y)\in B_M\times B_M} |\partial_t K_t(x,y)| \leq \Gamma^3 \frac{CDD'}{\sqrt{m}} \xi(0)^2 \tilde{\xi}(0) \|\hat{r}(0)\|_{\mathbb{R}^n}.$$

Therefore by the fundamental theorem of calculus for $t \leq T$

$$\sup_{(x,y)\in B_M\times B_M} |K_t(x,y) - K_0(x,y)| \leq t\Gamma^3 \frac{CDD'}{\sqrt{m}} \xi(0)^2 \tilde{\xi}(0) \|\hat{r}(0)\|_{\mathbb{R}^n}$$

$$\leq t\Gamma^3 \frac{CDD'}{\sqrt{m}} \|\hat{r}(0)\|_{\mathbb{R}^n} \left\{1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2 \left\{\sqrt{d} + CK^2\left[\sqrt{\log(c/\delta)} + \sqrt{\log m}\right]\right\}.$$

Now consider if one instead does the doubling trick where one does the following swaps $W(0) \to \begin{bmatrix} W(0) \\ W(0) \end{bmatrix}$, $b(0) \to \begin{bmatrix} b(0) \\ b(0) \end{bmatrix}$, $a(0) \to \begin{bmatrix} a(0) \\ -a(0) \end{bmatrix}$ and $m \to 2m$ where $W(0)$, $b(0)$, and $a(0)$ are initialized as before. Then $\xi(0)$ and $\tilde{\xi}(0)$ do not change. We can then run through the same exact proof as before except when we apply Lemma 2.5.23 the second hypothesis on m is no longer needed. □

Theorem 2.5.26. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Let $\Gamma > 1$ and $T > 0$ be positive constants and let $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$. Assume*

$$m \geq \frac{4D^2\|y\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2} \text{ and } m \geq \max\left\{\frac{4D^2 O(\log(c/\delta) + \tilde{O}(d))T^2}{[\log(\Gamma)]^2}, O(\log(c/\delta) + \tilde{O}(d))\right\}.$$

Then with probability at least $1-\delta$ we have for $t \leq T$

$$\sup_{(x,y)\in B_M\times B_M} |K_t(x,y) - K^\infty(x,y)| = \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3\|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right).$$

If one instead does the doubling trick then one can remove the assumption

$$m \geq \frac{4D^2 O(\log(c/\delta) + \tilde{O}(d))T^2}{[\log(\Gamma)]^2}$$

and have the same conclusion hold.

Proof. The condition $m \geq O(\log(c/\delta) + \tilde{O}(d))$ is sufficient to satisfy the hypothesis of Theorem 2.5.19. The condition on m also immediately satisfies the hypothesis of Theorem 2.5.25. The desired result then follows from a union bound. □

Theorem 2.5.27. *Under the same assumptions as Theorem 2.5.26 we have that with probability at least $1-\delta$ for all $t \leq T$*

$$\|H^\infty - H_t\|_{op} \leq n\tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 \|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right),$$

$$\sup_{s \leq T} \|G^\infty - G_t\|_{op} \leq \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 \|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right).$$

Proof. Recall that for a matrix $A \in \mathbb{R}^{m\times n}$ $\|A\|_{op} \leq \sqrt{mn}\max_{i,j}|A_{i,j}|$. Thus by Theorem 2.5.26 with probability at least $1-\delta$

$$\begin{aligned}\|H^\infty - H_t\|_{op} &\leq n\max_{i,j}|H^\infty_{i,j} - (H_t)_{i,j}| \leq n\sup_{(x,y)\in B_M\times B_M}|K_t(x,y) - K^\infty(x,y)| \\ &= n\tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3\|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right).\end{aligned}$$

The second bound follows from $G_s = \frac{1}{n}H_s$ and $G^\infty = \frac{1}{n}H^\infty$. □

2.5.2.7 NTK Deviations for ReLU Approximations

The NTK deviation bounds given in the previous subsections assumed $\|\sigma''\|_\infty < \infty$. For ReLU this assumption is not satisfied. It is natural to ask to what extent we might expect the results to hold when the activation function is $\sigma(x) = ReLU(x) = \max\{0, x\}$. The closest we can get to ReLU without modifying the proofs is to use the Softmax approximation to ReLU, namely $\sigma(x) = \frac{1}{\alpha}\ln(1+\exp(\alpha x))$, and consider what happens as $\alpha \to \infty$. For this choice of σ we have that $\|\sigma''\|_\infty = O(\alpha)$. In Subsection 2.5.2.6 where you will pay the biggest penalty is in Theorem 2.5.25 via the constant $D' = O(\|\sigma''\|^2_\infty) = O(\alpha^2)$. Since the final bound depends on the ratio $\frac{D'}{\sqrt{m}}$ you will have that m will grow like $O(\alpha^4)$. This is no

moderate penalty, although we might expect the results to hold for wide ReLU networks if a finite α provides a reasonable approximation. In particular Softmax $\ln(1+\exp(x))$ leads to a fixed constant for D'.

2.5.3 Underparameterized Regime

In this section we build the tools to study the implicit bias in the underparameterized case. Our ultimate goal is prove Theorem 2.3.5.

Outline of this section

- Review operator theory
- Prove damped deviations equation
- Bound $\|(T_{K^\infty} - T_n^s)r_t\|_2$
  - Bound $\|(T_n - T_n^s)r_t\|_2$ using NTK deviation results (comparatively easy)
  - Bound $\|(T_{K^\infty} - T_n)r_t\|_2$
    - Derive covering number for a class of functions $\mathcal{C}$
    - Use covering number to bound $\sup_{g\in\mathcal{C}} \|(T_{K^\infty} - T_n)g\|$
    - Show that r_t is in class $\mathcal{C}$
- Prove Theorem 2.3.5

2.5.3.1 RKHS and Mercer's Theorem

We recall some facts about Reproducing Kernel Hilbert Spaces (RKHS) and Mercer's Theorem. For additional background we suggest [BT04]. Let $X \subset \mathbb{R}^d$ be a compact space equipped with a strictly positive (regular Borel) probability measure ρ. Let $K : X \times X \to \mathbb{R}$ be a continuous, symmetric, positive definite function. We define the integral operator

$T_K : L^2_\rho(X) \to L^2_\rho(X)$

$$T_K f(x) := \int_X K(x,s) f(s) d\rho(s).$$

In this setting T_K is a compact, positive, self-adjoint operator. By the spectral theorem there is a countable nonincreasing sequence of nonnegative values $\{\sigma_i\}_{i=1}^\infty$ and an orthonormal set $\{\phi_i\}_{i=1}^\infty$ in L^2 such that $T_K \phi_i = \sigma_i \phi_i$. We will assume that T_K is strictly positive, i.e. $\langle f, T_K f \rangle_2 > 0$ for $f \neq 0$, so that we have further that $\{\phi_i\}_{i=1}^\infty$ is an orthonormal basis of L^2 and $\sigma_i > 0$ for all i. Moreover since K is continuous we may select the ϕ_i so that they are continuous functions, i.e. $\phi_i \in C(X)$ for each i. Then by Mercer's theorem we can decompose

$$K(x,y) = \sum_{i=1}^\infty \sigma_i \phi_i(x) \phi_i(y),$$

where the convergence is uniform. Furthermore the RKHS $\mathcal{H}$ associated with K is given by the set of functions

$$\mathcal{H} = \left\{ f \in L^2 : \sum_{i=1}^\infty \frac{|\langle f, \phi_i \rangle_2|^2}{\sigma_i} < \infty \right\},$$

where the inner product on $\mathcal{H}$ is given by

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^\infty \frac{\langle f, \phi_i \rangle_2 \langle g, \phi_i \rangle_2}{\sigma_i}.$$

Note that in this setting $\{\sqrt{\sigma_i}\phi_i\}_{i=1}^\infty$ is an orthonormal basis of $\mathcal{H}$. Define $K_x := K(\bullet, x)$. Recall the RKHS has the defining properties

$$K_x \in \mathcal{H} \quad \forall x \in X,$$

$$h(x) = \langle h, K_x \rangle_{\mathcal{H}} \quad \forall (x,h) \in X \times \mathcal{H}.$$

We will let $\kappa := \sup_{x \in X} K(x,x) < \infty$. From this we will have the useful inequality: for $h \in \mathcal{H}$

$$|h(x)| = |\langle h, K_x \rangle_{\mathcal{H}}| \leq \|h\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}} = \|h\|_{\mathcal{H}} \sqrt{\langle K_x, K_x \rangle_{\mathcal{H}}} = \|h\|_{\mathcal{H}} \sqrt{K(x,x)}$$
$$\leq \kappa^{1/2} \|h\|_{\mathcal{H}}.$$

Furthermore the elements of $\mathcal{H}$ are bounded continuous functions and $\mathcal{H}$ is seperable.

2.5.3.2 Hilbert-Schmidt and Trace Class Operators

We will recall some definitions from [RBV10]. A bounded operator on a separable Hilbert space with associated norm $\|\bullet\|$ is called *Hilbert-Schmidt* if

$$\sum_{i=1}^{\infty} \|Ae_i\|^2 < \infty$$

for some (any) orthonormal basis $\{e_i\}_i$. For such an operator we define its Hilbert-Schmidt norm $\|A\|_{HS}$ to be the square root of the above sum. The Hilbert-Schmidt norm is the analog of the Frobenius norm for matrices. It is useful to note that every Hilbert-Schmidt operator is compact. The space of Hilbert-Schmidt operators is a Hilbert space with respect to the inner product

$$\langle A, B\rangle = \sum_j \langle Ae_j, Be_j\rangle.$$

A stronger notion is that of a *trace class* operator. We say a bounded operator on a separable Hilbert space is *trace class* if

$$\sum_{i=1}^{\infty} \langle \sqrt{A^*A} e_i, e_i\rangle < \infty$$

for some (any) orthonormal bases $\{e_i\}_i$. For such an operator we may define

$$Tr(A) := \sum_{i=1}^{\infty} \langle Ae_i, e_i\rangle.$$

By Lidskii's theorem the above sum is also equal to the sum of the eigenvalues of A repeated by multiplicity. The space of trace class operators is a Banach space with the norm $\|A\|_{TC} = Tr(\sqrt{A^*A})$. The following inequalities will be useful

$$\|A\| \leq \|A\|_{HS} \leq \|A\|_{TC}.$$

Furthermore if A is Hilbert-Schmidt and B is bounded we have

$$\|BA\|_{HS}, \|AB\|_{HS} \leq \|A\|_{HS} \|B\|.$$

Note that in our setting we have

$$\kappa \geq \int_X K(x,x)d\rho(x) = \int_X \sum_{i=1}^{\infty} \sigma_i |\phi_i(x)|^2 d\rho(x) = \sum_{i=1}^{\infty} \sigma_i \int_X |\phi_i(x)|^2 d\rho(x)$$
$$= \sum_{i=1}^{\infty} \sigma_i = Tr(T_K),$$

where the interchange of integration and summation is justified by the monotone convergence theorem. Thus T_K is a trace class operator and we have the inequality

$$\kappa \geq \sum_{i=1}^{\infty} \sigma_i$$

which will prove useful later.

2.5.3.3 Damped Deviations

Let $x \mapsto g_s(x) \in L^2$ for each $s \in [0,t]$ such that $s \mapsto \langle \phi_i, g_s \rangle_2$ is measureable for each i and $\int_0^t \|g_s\|_2^2 < \infty$. Then we define the integral

$$\int_0^t g_s ds$$

coordinatewise, meaning that $\int_0^t g_s ds$ is the L^2 function h such that

$$\langle h, \phi_i \rangle_2 = \int_0^t \langle g_s, \phi_i \rangle_2 ds.$$

Using this definition, we can now prove the "Damped Deviations" lemma.

Lemma 2.2.3. *Let $K(x,x')$ be an arbitrary continuous, symmetric, positive-definite kernel. Let $[T_K h](\bullet) = \int_X K(\bullet, s)h(s)d\rho(s)$ be the integral operator associated with K and let $[T_n^s h](\bullet) = \frac{1}{n}\sum_{i=1}^n K_s(\bullet, x_i)h(x_i)$ denote the operator associated with the time-dependent NTK K_s. Then*

$$r_t = \exp(-T_K t)r_0 + \int_0^t \exp(-T_K(t-s))(T_K - T_n^s)r_s ds,$$

where the equality is in the L^2 sense.

Proof. We have that

$$\partial_s r_s(x) = -\frac{1}{n}\sum_{i=1}^{n} K_s(x, x_i) r_s(x_i) = -[T_n^s r_s](x),$$

where the equality is pointwise over x. $\partial_s r_s(x)$ is a continuous function of x since K_s is continuous and is thus in L^2. Therefore we can consider

$$\langle \partial_s r_s, \phi_i \rangle_2 = \langle -T_n^s r_s, \phi_i \rangle_2.$$

By the continuity of $s \mapsto \theta_s$ we have the parameters are locally bounded in time and thus by Lemma 2.5.5 we have that $\|K_s\|_\infty$ is also locally bounded therefore for any $\delta > 0, s_0$: $\sup_{|s-s_0|\leq\delta} \|K_s\|_\infty < \infty$. Note then that

$$|\partial_s r_s(x)| \leq \frac{1}{n}\sum_{i=1}^{n} |K_s(x, x_i)||r_s(x_i)| \leq \|K_s\|_\infty \|\hat{r}_s\|_{\mathbb{R}^n} \leq \|K_s\|_\infty \|\hat{r}_0\|_{\mathbb{R}^n}.$$

It follows that $\|\partial_s r_s\|_\infty$ is bounded locally uniformly in s. Therefore the following differentiation under the integral sign is justified

$$\frac{d}{ds}\langle r_s, \phi_i \rangle_2 = \langle \partial_s r_s, \phi_i \rangle_2.$$

Thus combined with our previous equality we get

$$\frac{d}{ds}\langle r_s, \phi_i \rangle_2 = \langle -T_n^s r_s, \phi_i \rangle_2 = \langle -T_K r_s, \phi_i \rangle_2 + \langle (T_K - T_n^s) r_s, \phi_i \rangle_2$$
$$= \langle r_s, -T_K \phi_i \rangle_2 + \langle (T_K - T_n^s) r_s, \phi_i \rangle_2 = -\sigma_i \langle r_s, \phi_i \rangle_2 + \langle (T_K - T_n^s) r_s, \phi_i \rangle_2,$$

where we have used that T_K is self-adjoint. Therefore

$$\frac{d}{ds}\langle r_s, \phi_i \rangle_2 + \sigma_i \langle r_s, \phi_i \rangle_2 = \langle (T_K - T_n^s) r_s, \phi_i \rangle_2.$$

Multiplying by the integrating factor $\exp(\sigma_i s)$ we get

$$\frac{d}{ds}[\exp(\sigma_i s)\langle r_s, \phi_i \rangle_2] = \exp(\sigma_i s)\langle (T_K - T_n^s) r_s, \phi_i \rangle_2.$$

Therefore applying the fundamental theorem of calculus after rearrangement we get

$$\langle r_t, \phi_i \rangle_2 = \exp(-\sigma_i t)\langle r_0, \phi_i \rangle_2 + \int_0^t \exp(-\sigma_i (t-s))\langle (T_K - T_n^s) r_s, \phi_i \rangle_2 ds,$$

which is just the coordinatewise version of the desired result.

$$r_t = \exp(-T_K t) r_0 + \int_0^t \exp(-T_K (t-s))(T_K - T_n^s) r_s ds.$$

□

2.5.3.4 Covering Number of Class

We will now estimate the covering number of the class of shallow networks with bounds on their parameter norms. This lemma is slightly more general than what we will use but we will particularize it latter as it's general formulation presents no additional difficulty.

Lemma 2.5.28. *Let*

$$\mathcal{C} = \{ \frac{a^T}{\sqrt{m}} \sigma(Wx + b) + b_0 : \|a - a'\|_2 \leq \rho_1, \|W - W'\|_F \leq \rho_2,$$
$$\|b - b'\|_2 \leq \rho_3, |b_0 - b_0'| \leq \rho_4,$$
$$\frac{1}{\sqrt{m}} \|a\|_2 \leq \rho_1', \frac{1}{\sqrt{m}} \|W\|_{op} \leq \rho_2', \frac{1}{\sqrt{m}} \|b\|_2 \leq \rho_3' \},$$

and

$$\gamma' := |\sigma(0)| + \|\sigma'\|_\infty [\rho_2' M + \rho_3'].$$

Then the (proper) covering number of $\mathcal{C}$ in the uniform norm satisfies

$$\mathcal{N}(\mathcal{C}, \epsilon, \|\|_\infty) \leq \left(\frac{C'}{\epsilon} \right)^p$$

where $p = md + 2m + 1$ is the total number of parameters and C' equals

$$C' = C \max \{ \rho_1 \gamma', \rho_2 M \|\sigma'\|_\infty \rho_1', \rho_3 \|\sigma'\|_\infty \rho_1', \rho_4 \}$$

where $C > 0$ is an absolute constant.

Proof. We will bound the pertubation of the function when changing the weights, specifically we will bound

$$\sup_{x\in B_M}\left|\frac{a^T}{\sqrt{m}}\sigma(Wx+b)-\frac{\tilde{a}^T}{\sqrt{m}}\sigma(\tilde{W}x+\tilde{b})\right|.$$

Let $x^{(1)}=\frac{1}{\sqrt{m}}\sigma(Wx+b)$ and $\tilde{x}^{(1)}=\frac{1}{\sqrt{m}}\sigma(\tilde{W}x+\tilde{b})$. Then note that we have

$$\begin{aligned}\left\|x^{(1)}-\tilde{x}^{(1)}\right\|_2 &\le \frac{1}{\sqrt{m}}\left\|\sigma'\right\|_\infty\left\|(W-\tilde{W})x+b-\tilde{b}\right\|_2\\ &\le \frac{1}{\sqrt{m}}\left\|\sigma'\right\|_\infty\left[\left\|W-\tilde{W}\right\|_{op}\left\|x\right\|_2+\left\|b-\tilde{b}\right\|_2\right]\\ &\le \frac{1}{\sqrt{m}}\left\|\sigma'\right\|_\infty\left[\left\|W-\tilde{W}\right\|_F M+\left\|b-\tilde{b}\right\|_2\right]=:\gamma.\end{aligned}$$

Well then

$$\begin{aligned}|a^Tx^{(1)}-\tilde{a}^T\tilde{x}^{(1)}| &\le |a^T(x^{(1)}-\tilde{x}^{(1)})|+|(a-\tilde{a})^T\tilde{x}^{(1)}|\\ &\le \left\|a\right\|_2\gamma+\left\|a-\tilde{a}\right\|_2\left\|\tilde{x}^{(1)}\right\|_2.\end{aligned}$$

Finally

$$\begin{aligned}\left\|\tilde{x}^{(1)}\right\|_2=\left\|\frac{1}{\sqrt{m}}\sigma(\tilde{W}x+\tilde{b})\right\|_2 &\le |\sigma(0)|+\frac{1}{\sqrt{m}}\left\|\sigma'\right\|_\infty\left\|\tilde{W}x+\tilde{b}\right\|_2\\ &\le |\sigma(0)|+\frac{1}{\sqrt{m}}\left\|\sigma'\right\|_\infty\left[\left\|\tilde{W}\right\|_{op}\left\|x\right\|_2+\left\|\tilde{b}\right\|_2\right]\\ &\le |\sigma(0)|+\frac{1}{\sqrt{m}}\left\|\sigma'\right\|_\infty\left[\left\|\tilde{W}\right\|_{op}M+\left\|\tilde{b}\right\|_2\right]\\ &\le |\sigma(0)|+\left\|\sigma'\right\|_\infty[\rho_2'M+\rho_3']=:\gamma'.\end{aligned}$$

Therefore

$$|a^Tx^{(1)}-\tilde{a}^T\tilde{x}^{(1)}|\le\left\|a\right\|_2\gamma+\left\|a-\tilde{a}\right\|_2\gamma'.$$

Thus if we have

$$\left\|a-\tilde{a}\right\|_2\le\frac{\epsilon}{4\gamma'}=:\epsilon_1,\quad \left\|W-\tilde{W}\right\|_F\le\frac{\epsilon}{8M\left\|\sigma'\right\|_\infty\rho_1'}=:\epsilon_2,\quad \left\|b-\tilde{b}\right\|_2\le\frac{\epsilon}{8\left\|\sigma'\right\|_\infty\rho_1'}=:\epsilon_3,$$

then

$$\left\|a\right\|_2\gamma\le\frac{\epsilon\left\|a\right\|_2}{4\rho_1'\sqrt{m}}\le\frac{\epsilon}{4}.$$

Therefore

$$|a^T x^{(1)} - \tilde{a}^T \tilde{x}^{(1)}| \leq \epsilon/2$$

and this bound holds for any $x \in B_M$. If add biases b_0 and $\tilde{b_0}$ such that $|b_0 - \tilde{b_0}| \leq \epsilon/2$ we simply get by the triangle inequality

$$|a^T x^{(1)} + b_0 - (\tilde{a}^T \tilde{x}^{(1)} + \tilde{b_0})| \leq \epsilon.$$

Thus to get a cover we can simply cover the sets

$$\{a : \|a - a'\|_2 \leq \rho_1\}, \quad \{W : \|W - W'\|_F \leq \rho_2\},$$

$$\{b : \|b - b'\|_2 \leq \rho_3\}, \quad \{b_0 : |b_0 - b_0'| \leq \rho_4\},$$

in the Euclidean norm and multiply the covering numbers. Recall that the ϵ covering number for a Euclidean ball of radius R in $\mathbb{R}^s$, say $\mathcal{N}_\epsilon$, using the Euclidean norm satisfies

$$\left(\frac{cR}{\epsilon}\right)^s \leq \mathcal{N}_\epsilon \leq \left(\frac{CR}{\epsilon}\right)^s$$

for two absolute constants $c, C > 0$. Therefore we get that

$$\mathcal{N}(\mathcal{C}, \epsilon, \|\|_\infty) \leq \left(\frac{C\rho_1}{\epsilon_1}\right)^m \left(\frac{C\rho_2}{\epsilon_2}\right)^{md} \left(\frac{C\rho_3}{\epsilon_3}\right)^m \left(\frac{2C\rho_4}{\epsilon}\right).$$

The desired result follows from

$$\max\left\{\frac{\rho_1}{\epsilon_1}, \frac{\rho_2}{\epsilon_2}, \frac{\rho_3}{\epsilon_3}, \frac{2\rho_4}{\epsilon}\right\} \leq \frac{C}{\epsilon} \max\left\{\rho_1\gamma', \rho_2 M \|\sigma'\|_\infty \rho_1', \rho_3 \|\sigma'\|_\infty \rho_1', \rho_4\right\}.$$

□

We can now prove the following corollary which is the version of the previous lemma that we will actually use for our neural network.

Corollary 2.5.29. *Let*

$$\mathcal{C} = \{ \frac{a^T}{\sqrt{m}}\sigma(Wx + b) + b_0 : \frac{1}{\sqrt{m}}\|a\|_2, \frac{1}{\sqrt{m}}\|W\|_{op}, \frac{1}{\sqrt{m}}\|b\|_2 \leq A, |b_0| \leq B \}$$

and

$$\gamma' := |\sigma(0)| + \|\sigma'\|_\infty [AM + A]$$

and assume $m \geq d$. Then the (proper) covering number of $\mathcal{C}$ in the uniform norm satisfies

$$\mathcal{N}(\mathcal{C}, \epsilon, \|\|_\infty) \leq \left(\frac{\Psi(m,d)}{\epsilon}\right)^p,$$

where

$$\begin{aligned}\Psi(m,d) &= C \max\{\sqrt{m}A\gamma', \sqrt{md}A^2 \|\sigma'\|_\infty M, \sqrt{m}A^2 \|\sigma'\|_\infty, B\} \\ &= \sqrt{md}O\left(\max\left\{A^2, \frac{B}{\sqrt{md}}\right\}\right).\end{aligned}$$

Proof. The idea is to apply Lemma 2.5.28 with $a' = 0$, $W' = 0$, $b' = 0$, and $b'_0 = 0$. Note that $\|W\|_F \leq \sqrt{d}\|W\|_{op} \leq \sqrt{md}A$. The result then follows by applying lemma with $\rho_1 = \sqrt{m}A$, $\rho_2 = \sqrt{md}A$, $\rho_3 = \sqrt{m}A$ and $\rho_4 = B$ and $\rho'_1 = \rho'_2 = \rho'_3 = A$. □

2.5.3.5 Uniform Convergence over the Class

We now show that $\|(T_n - T_{K^\infty})g\|_2$ is uniformly small for all g in a suitable class of functions $\mathcal{C}'$. Ultimately we will show that $r_t \in \mathcal{C}'$ and thus this result is towards proving that $\|(T_n - T_{K^\infty})r_t\|_2$ is small.

Lemma 2.5.30. *Let $K(x,x')$ by a continuous, symmetric, positive-definite kernel and let $\kappa = \max_{x\in X} K(x,x) < \infty$. Let $T_K h(\bullet) = \int_X K(\bullet, s)h(s)d\rho(s)$ and $T_n h(\bullet) = \frac{1}{n}\sum_{i=1}^n K(\bullet, x_i)h(x_i)$ be the associated operators. Let σ_1 denote the largest eigenvalue of T_K. Let $\mathcal{C}$ and $\Psi(m,d)$ be defined as in Corollary 2.5.29. We let $\mathcal{C}' = \{g - f^* : g \in \mathcal{C}\} \cap \{g : \|g\|_\infty \leq S\}$ be the set where $\mathcal{C}$ is translated by the target function f^* then intersected with the L^∞ ball of radius $S > 0$. Then with probability at least $1 - \delta$ over the sampling of $x_1, \ldots, x_n$*

$$\begin{aligned}\sup_{g\in\mathcal{C}'} \|(T_n - T_K)g\|_2 &\leq \frac{2S\sqrt{\sigma_1\kappa}\sqrt{2\log(c/\delta) + 2p\log(\|K\|_\infty \Psi(m,d)\sqrt{n})}}{\sqrt{n}} + \frac{2}{\sqrt{n}} \\ &= \frac{2\left[1 + S\sqrt{\sigma_1\kappa}\sqrt{2\log(c/\delta) + \tilde{O}(p)}\right]}{\sqrt{n}}.\end{aligned}$$

Proof. Let $g \in \mathcal{C}'$. We introduce the random variables $Y_i := K_{x_i} g(x_i) - \mathbb{E}_{x\sim\rho}[K_x g(x)]$ taking values in the Hilbert space $\mathcal{H}$ for $i \in [n]$ where $\mathcal{H}$ is the RKHS associated with K. Note that for any x

$$\|K_x g(x)\|_{\mathcal{H}} = |g(x)|\sqrt{\langle K_x, K_x\rangle_{\mathcal{H}}} \leq S\sqrt{K(x,x)} \leq S\kappa^{1/2}.$$

Thus $\|Y_i\|_{\mathcal{H}} \leq 2S\kappa^{1/2}$ a.s. Thus by Hoeffding's inequality for random variables taking values in a separable Hilbert space (see Section 2.4 [RBV10]) we have

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\|_{\mathcal{H}} > t\right) \leq 2\exp\left(-nt^2/2[2S\kappa^{1/2}]^2\right).$$

Note that by basic properties of the covering number we have that

$$\mathcal{N}(\mathcal{C}', \epsilon, \|\|_\infty) \leq \mathcal{N}(\mathcal{C}, \epsilon/2, \|\|_\infty),$$

thus by Corollary 2.5.29 the covering number of $\mathcal{C}'$ satisfies (up to a redefinition of C)

$$\mathcal{N}(\mathcal{C}', \epsilon, \|\|_\infty) \leq \left(\frac{\Psi(m,d)}{\epsilon}\right)^p.$$

Let Δ be an ϵ net of $\mathcal{C}'$ in the uniform norm. Note that $\frac{1}{n}\sum_{i=1}^n Y_i = (T_n - T_K)g$. Thus by taking a union bound we have

$$\mathbb{P}\left(\max_{g\in\Delta} \|(T_n - T_K)g\|_{\mathcal{H}} \geq t\right) \leq \left(\frac{\Psi(m,d)}{\epsilon}\right)^p 2\exp\left(-nt^2/2[2S\kappa^{1/2}]^2\right).$$

Note that for any probability measure ν and $h \in L^\infty$

$$\left|\int_X K(x,s)h(s)d\nu(s)\right| \leq \int_X |K(x,s)||h(s)|d\nu(s) \leq \|K\|_\infty \|h\|_\infty.$$

It follows that for any $h \in L^\infty$

$$\|(T_K - T_n)h\|_\infty \leq 2\|K\|_\infty \|h\|_\infty.$$

Note for any $g \in \mathcal{C}'$ we can pick $\hat{g}$ in Δ such that $\|g - \hat{g}\|_\infty \leq \epsilon$. Then

$$\begin{aligned}
\|(T_n - T_K)g\|_2 &\leq \|(T_n - T_K)\hat{g}\|_2 + \|(T_n - T_K)(g - \hat{g})\|_2 \\
&\leq \sqrt{\sigma_1}\|(T_n - T_K)\hat{g}\|_{\mathcal{H}} + \|(T_n - T_K)(g - \hat{g})\|_\infty \\
&\leq \sqrt{\sigma_1}t + 2\|K\|_\infty \|g - \hat{g}\|_\infty \\
&\leq \sqrt{\sigma_1}t + 2\|K\|_\infty \epsilon,
\end{aligned}$$

where we have used the fact that $\|\bullet\|_2 \leq \sqrt{\sigma_1}\|\bullet\|_{\mathcal{H}}$ and $\|\bullet\|_2 \leq \|\bullet\|_\infty$ in the second inequality. Thus by setting

$$t = \frac{2S\kappa^{1/2}\sqrt{2\log(c/\delta) + 2p\log(\Psi(m,d)/\epsilon)}}{\sqrt{n}}$$

we have with probability at least $1-\delta$

$$\sup_{g\in\mathcal{C}'} \|(T_n - T_K)g\|_2 \leq$$
$$\sqrt{\sigma_1}\frac{2S\kappa^{1/2}\sqrt{2\log(c/\delta) + 2p\log(\Psi(m,d)/\epsilon)}}{\sqrt{n}} + 2\|K\|_\infty \epsilon.$$

This argument runs through for any $\epsilon > 0$. Thus by setting $\epsilon = \frac{1}{\|K\|_\infty \sqrt{n}}$ we get the desired result. □

2.5.3.6 Neural Network is in the Class

In this section we demonstrate that the neural network in such a class as $\mathcal{C}$ as defined in Lemma 2.5.28. Once we have this we can use Lemma 2.5.30 to show that $\|(T_{K^\infty} - T_n)r_t\|_2$ is uniformly small. The first step is to bound the parameter norms, hence the following lemma.

Lemma 2.5.31. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\|1\|_{\psi_2}, \|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Let $\Gamma > 1$, $T > 0$, $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$, and*

$$\xi(t) = \max\{\frac{1}{\sqrt{m}}\|W\|_{op}, \frac{1}{\sqrt{m}}\|b\|_2, \frac{1}{\sqrt{m}}\|a\|_2, 1\}.$$

Furthermore assume

$$m \geq \frac{4D^2\|y\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2} \text{ and } m \geq \max\left\{\frac{4D^2 O(\log(c/\delta) + \tilde{O}(d))T^2}{[\log(\Gamma)]^2}, O(\log(c/\delta) + \tilde{O}(d))\right\}.$$

Then with probability at least $1-\delta$

$$\max_{t\in[0,T]} \xi(t) \leq \Gamma\left[1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right].$$

If one instead does the doubling trick then the second condition on m can be removed from the hypothesis and the same conclusion holds.

Proof. First assume we are not doing the doubling trick. Note that the hypothesis on m is strong enough to satisfy the hypothesis of Lemma 2.5.23, therefore we have with probability at least $1-\delta$

$$\max_{t\leq T}\xi(t)\leq\Gamma\xi(0).$$

Well then separately by Lemma 2.5.17 with probability at least $1-\delta$

$$\xi(0)\leq 1+C\frac{\sqrt{d}+K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}.$$

Thus by replacing δ with $\delta/2$ in the previous statements and taking a union bound we have with probability at least $1-\delta$

$$\max_{t\in[0,T]}\xi(t)\leq\Gamma\left[1+C\frac{\sqrt{d}+K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right]$$

which is the desired result. Now suppose instead one does the doubling trick. We recall that the doubling trick does not change $\xi(0)$. Thus we can run through the exact same argument as before except when we apply Lemma 2.5.23 we can remove the second condition on m from the hypothesis. □

The following lemma bounds the bias term.

Lemma 2.5.32. *For any initial conditions we have*

$$|b_0(t)|\leq|b_0(0)|+t\left\|\hat{r}(0)\right\|_{\mathbb{R}^n}.$$

Proof. Note that

$$|\partial_t b_0(t)|=\left|\frac{1}{n}\sum_{i=1}^{n}\hat{r}(t)_i\right|\leq\left\|\hat{r}(t)\right\|_{\mathbb{R}^n}\leq\left\|\hat{r}(0)\right\|_{\mathbb{R}^n}.$$

Thus by the fundamental theorem of calculus

$$|b_0(t)|\leq|b_0(0)|+t\left\|\hat{r}(0)\right\|_{\mathbb{R}^n}.$$

□

The following lemma demonstrates that the residual $r_t = f_t - f^*$ is bounded.

Lemma 2.5.33. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\|1\|_{\psi_2}, \|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Let $\Gamma > 1$, $T > 0$, $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$, and assume*

$$m \geq \frac{4D^2\|y\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2} \text{ and } m \geq \max\left\{\frac{4D^2 O(\log(c/\delta) + \tilde{O}(d))T^2}{[\log(\Gamma)]^2}, O(\log(c/\delta) + \tilde{O}(d))\right\}.$$

Then with probability at least $1-\delta$ for $t \leq T$

$$\|f_t - f^*\|_\infty \leq \|f_0 - f^*\|_\infty + t\|\hat{r}(0)\|_{\mathbb{R}^n} CD^2\Gamma^2\left[1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right]^2.$$

If one instead does the doubling trick then the second condition on m can be removed from the hypothesis and the same conclusion holds.

Proof. Recall that

$$\partial_t(f_t(x) - f^*(x)) = -\frac{1}{n}\sum_{i=1}^n K_t(x, x_i)(f_t(x_i) - f^*(x_i)) = -\frac{1}{n}\sum_{i=1}^n K_t(x, x_i)\hat{r}(t)_i.$$

Thus

$$|\partial_t(f_t(x) - f^*(x))| \leq \frac{1}{n}\sum_{i=1}^n |K_t(x, x_i)||\hat{r}(t)_i| \leq \|K_t\|_\infty \|\hat{r}(t)\|_{\mathbb{R}^n} \leq \|K_t\|_\infty \|\hat{r}(0)\|_{\mathbb{R}^n}.$$

Well by Lemma 2.5.5 we have that $\|K_t\|_\infty \leq CD^2\xi^2(t)$ where

$$\xi(t) = \max\{\frac{1}{\sqrt{m}}\|W\|_{op}, \frac{1}{\sqrt{m}}\|b\|_2, \frac{1}{\sqrt{m}}\|a\|_2, 1\}.$$

Well by Lemma 2.5.31 we have that with probability at least $1-\delta$

$$\max_{t\in[0,T]} \xi(t) \leq \Gamma\left[1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right].$$

Thus by the fundamental theorem of calculus for $t \leq T$

$$|f_t(x) - f^*(x)| \leq |f_0(x) - f^*(x)| + t\|\hat{r}(0)\|_{\mathbb{R}^n} CD^2\Gamma^2\left[1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right]^2$$

Thus by taking the supremum over x we get

$$\|f_t - f^*\|_\infty \leq \|f_0 - f^*\|_\infty + t\,\|\hat{r}(0)\|_{\mathbb{R}^n} CD^2\Gamma^2 \left[1 + C\frac{\sqrt{d} + K^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right]^2,$$

which is the desired conclusion. □

We can now finally prove that $\|(T_{K^\infty} - T_n)r_t\|_2$ is uniformly small.

Lemma 2.5.34. *Let $K(x, x')$ by a continuous, symmetric, positive-definite kernel and let $\kappa = \max_x K(x, x) < \infty$. Let $T_K h(\bullet) = \int_X K(\bullet, s)h(s)d\rho(s)$ and $T_n h(\bullet) = \frac{1}{n}\sum_{i=1}^n K(\bullet, x_i)h(x_i)$ be the associated operators. Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume*

$$\|1\|_{\psi_2}, \|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2}, \|b_0\|_{\psi_2} \leq K'$$

for each $\ell \in [m]$ where $K' \geq 1$. Let $\Gamma > 1$, $T > 0$, $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$, and assume

$$m \geq \frac{4D^2\|y\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2} \text{ and } m \geq \max\left\{\frac{4D^2 O(\log(c/\delta) + \tilde{O}(d))T^2}{[\log(\Gamma)]^2}, O(\log(c/\delta) + \tilde{O}(d))\right\}.$$

If we are doing the doubling trick set $S' = 0$ and otherwise set

$$S' = CD(K')^2\sqrt{d\log(CM\tilde{O}(\sqrt{m})) + \log(c/\delta)} = \tilde{O}(\sqrt{d}).$$

Then with probability at least $1 - \delta$

$$\sup_{t \leq T}\|(T_n - T_K)r_t\|_2 = \tilde{O}\left(\frac{(\|f^*\|_\infty + S')(1 + T\Gamma^2)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right)$$

and

$$\|r_0\|_\infty \leq \|f^*\|_\infty + S'.$$

If we are performing the doubling trick the second condition on m can be removed and the same conclusion holds.

Proof. By Lemma 2.5.31 we have with probability at least $1-\delta$

$$\max_{t\in[0,T]} \xi(t) \leq \Gamma\left[1 + C\frac{\sqrt{d} + (K')^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right] =: A. \tag{2.8}$$

Also by Lemma 2.5.32

$$|b_0(t)| \leq |b_0(0)| + t\left\|\hat{r}(0)\right\|_{\mathbb{R}^n}.$$

If we are doing the doubling trick then $b_0(0) = 0$. Otherwise by Lemma 2.5.15 we have with probability at least $1-\delta$

$$|b_0(0)| \leq CK'\sqrt{\log(c/\delta)}.$$

Furthermore by Lemma 2.5.22 we have

$$\left\|\hat{r}(0)\right\|_{\mathbb{R}^n} \leq \|y\|_{\mathbb{R}^n} + \left\|f(\bullet;\theta_0)\right\|_{\infty}.$$

Let L be defined as in Lemma 2.5.21, i.e.

$$L(m,\sigma,d,K',\delta) := \sqrt{m}\left\|\sigma'\right\|_{\infty}\left\{1 + C\frac{\sqrt{d} + (K')^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right\}^2 = \tilde{O}(\sqrt{m}).$$

If we are not performing the doubling trick set

$$S' = CD(K')^2\sqrt{d\log(CML) + \log(c/\delta)}.$$

Otherwise if we are performing the doubling trick set $S' = 0$. In either case by Lemma 2.5.21 we have with probability at least $1-\delta$

$$\left\|f(\bullet;\theta_0)\right\|_{\infty} \leq S'. \tag{2.9}$$

In particular by Lemma 2.5.22 we have

$$\left\|\hat{r}(0)\right\|_{\mathbb{R}^n} \leq \|y\|_{\mathbb{R}^n} + \left\|f(\bullet;\theta_0)\right\|_{\infty} \leq \|y\|_{\mathbb{R}^n} + S'.$$

Thus we can say

$$|b_0(t)| \leq |b_0(0)| + t\left\|\hat{r}(0)\right\|_{\mathbb{R}^n} \leq CK'\sqrt{\log(c/\delta)} + T(\|y\|_{\mathbb{R}^n} + S') =: B$$

and this holds whether or not we are performing the doubling trick. Thus up until time T the neural network is in class $\mathcal{C}$ as defined in Corollary 2.5.29 with parameters A and B as defined above. Moreover by Lemma 2.5.33 separate from the randomness before we have that with probability at least $1-\delta$

$$\|r_t\|_\infty \leq \|r_0\|_\infty + t\,\|\hat{r}(0)\|_{\mathbb{R}^n} CD^2\Gamma^2 \left[1 + C\frac{\sqrt{d} + (K')^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right]^2.$$

Well note that when (2.9) holds we have

$$\|\hat{r}(0)\|_{\mathbb{R}^n} \leq \|r_0\|_\infty \leq \|f^*\|_\infty + \|f(\bullet;\theta_0)\|_\infty \leq \|f^*\|_\infty + S'.$$

Thus

$$\|r_t\|_\infty \leq (\|f^*\|_\infty + S')\left\{1 + TCD^2\Gamma^2\left[1 + C\frac{\sqrt{d} + (K')^2\sqrt{\log(c/\delta)}}{\sqrt{m}}\right]^2\right\} =: S.$$

Thus by taking a union bound and redefining δ we have by an application of Lemma 2.5.30 with S as defined in the hypothesis of the current theorem that with probability at least $1-\delta$

$$\begin{aligned}\sup_{t\leq T}\|(T_n - T_K)r_t\|_2 &\leq \frac{2\left[1 + S\sqrt{\sigma_1\kappa}\sqrt{2\log(c/\delta) + \tilde{O}(p)}\right]}{\sqrt{n}} \\ &= \tilde{O}\left(\frac{(\|f^*\|_\infty + S')(1 + T\Gamma^2)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right)\end{aligned}$$

where we have used that $S = \tilde{O}([\|f^*\|_\infty + S'][1 + T\Gamma^2])$. □

2.5.3.7 Proof of Theorem 2.3.5

We are almost ready to prove Theorem 2.3.5. However first we must introduce a couple lemmas. The following lemma uses the damped deviations equation to bound the difference between r_t and $\exp(-T_K t)r_0$.

Lemma 2.5.35. *Let $K(x, x')$ be a continuous, symmetric, positive-definite kernel with associated operator $T_K h(\bullet) = \int_X K(\bullet, s)h(s)d\rho(s)$. Let $T_n^s h(\bullet) = \frac{1}{n}\sum_{i=1}^n K_s(\bullet, x_i)h(x_i)$ denote the operator associated with the time-dependent NTK. Then*

$$\|P_k(r_t - \exp(-T_K t)r_0)\|_2 \leq \frac{1-\exp(-\sigma_k t)}{\sigma_k} \sup_{s\leq t} \|(T_K - T_n^s)r_s\|_2 ,$$

and

$$\|r_t - \exp(-T_K t)r_0\|_2 \leq t \cdot \sup_{s\leq t} \|(T_K - T_n^s)r_s\|_2 .$$

Proof. From Lemma 2.2.3 we have

$$r_t = \exp(-T_K t)r_0 + \int_0^t \exp(-T_K(t-s))(T_K - T_n^s)r_s ds.$$

Thus for any $k \in \mathbb{N}$

$$\begin{aligned} P_k(r_t - \exp(-T_K t)r_0) &= P_k \int_0^t \exp(-T_K(t-s))(T_K - T_n^s)r_s ds \\ &= \int_0^t P_k \exp(-T_K(t-s))(T_K - T_n^s)r_s ds. \end{aligned}$$

Therefore

$$\begin{aligned} \|P_k(r_t - \exp(-T_K t)r_0)\|_2 &= \left\| \int_0^t P_k \exp(-T_K(t-s))(T_K - T_n^s)r_s ds \right\|_2 \\ &\leq \int_0^t \|P_k \exp(-T_K(t-s))(T_K - T_n^s)r_s\|_2 \, ds \\ &\leq \int_0^t \|P_k \exp(-T_K(t-s))\| \, \|(T_K - T_n^s)r_s\|_2 \, ds \\ &\leq \int_0^t \exp(-\sigma_k(t-s)) \, \|(T_K - T_n^s)r_s\|_2 \, ds \\ &\leq \frac{1-\exp(-\sigma_k t)}{\sigma_k} \sup_{s\leq t} \|(T_K - T_n^s)r_s\|_2 . \end{aligned}$$

Similarly

$$\begin{aligned}\|r_t - \exp(-T_K t)r_0\|_2 &= \left\| \int_0^t \exp(-T_K(t-s))(T_K - T_n^s) r_s ds \right\|_2 \\ &\leq \int_0^t \|\exp(-T_K(t-s))(T_K - T_n^s) r_s\|_2 \, ds \\ &\leq \int_0^t \|\exp(-T_K(t-s))\| \, \|(T_K - T_n^s) r_s\|_2 \, ds \\ &\leq \int_0^t \|(T_K - T_n^s) r_s\|_2 \, ds \leq t \cdot \sup_{s \leq t} \|(T_K - T_n^s) r_s\|_2 .\end{aligned}$$

□

In light of the previous lemma we would like to have a bound for $\|(T_K - T_n^s) r_s\|_2$. This is accomplished by the following lemma.

Lemma 2.5.36. *Let $K(x, x')$ by a continuous, symmetric, positive-definite kernel. Let $T_K h(\bullet) = \int_X K(\bullet, s) h(s) d\rho(s)$ and $T_n h(\bullet) = \frac{1}{n} \sum_{i=1}^n K(\bullet, x_i) h(x_i)$ be the associated operators. Let $T_n^s h(\bullet) = \frac{1}{n} \sum_{i=1}^n K_s(\bullet, x_i) h(x_i)$ denote the operator associated with the time-dependent NTK. Then*

$$\sup_{s \leq T} \|(T_K - T_n^s) r_s\|_2 \leq \sup_{s \leq T} \|(T_K - T_n) r_s\|_2 + \sup_{s \leq T} \|K - K_s\|_\infty \|\hat{r}(0)\|_{\mathbb{R}^n} .$$

Proof. We have that

$$\|(T_K - T_n^s) r_s\|_2 \leq \|(T_K - T_n) r_s\|_2 + \|(T_n - T_n^s) r_s\|_2 .$$

Now observe that

$$\begin{aligned}|(T_n - T_n^s) r_s(x)| &= \left| \frac{1}{n} \sum_{i=1}^n [K(x, x_i) - K_s(x, x_i)] r_s(x_i) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n |K(x, x_i) - K_s(x, x_i)| |r_s(x_i)| \\ &\leq \|K - K_s\|_\infty \|\hat{r}(s)\|_{\mathbb{R}^n} \leq \|K - K_s\|_\infty \|\hat{r}(0)\|_{\mathbb{R}^n} .\end{aligned}$$

Therefore

$$\|(T_n - T_n^s) r_s\|_2 \leq \|(T_n - T_n^s) r_s\|_\infty \leq \|K - K_s\|_\infty \|\hat{r}(0)\|_{\mathbb{R}^n} .$$

Thus

$$\sup_{s\leq T}\|(T_K - T_n^s)r_s\|_2 \leq \sup_{s\leq T}\|(T_K - T_n)r_s\|_2 + \sup_{s\leq T}\|K - K_s\|_\infty \|\hat{r}(0)\|_{\mathbb{R}^n}.$$

□

We are almost ready to finally prove Theorem 2.3.5. We must prove one final lemma that combines Lemma 2.5.34 with the NTK deviation bounds in Theorem 2.5.26 to show that $\|(T_{K^\infty} - T_n^s)r_s\|_2$ is uniformly small.

Lemma 2.5.37. *Assume that $W_{i,j} \sim \mathcal{W}$, $b_\ell \sim \mathcal{B}$, $a_\ell \sim \mathcal{A}$ are all i.i.d zero-mean, subgaussian random variables with unit variance. Furthermore assume $\|1\|_{\psi_2}, \|w_\ell\|_{\psi_2}, \|a_\ell\|_{\psi_2}, \|b_\ell\|_{\psi_2} \leq K$ for each $\ell \in [m]$ where $K \geq 1$. Let $\Gamma > 1$, $T > 0$, $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$, and assume*

$$m \geq \frac{4D^2\|y\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2} \text{ and } m \geq \max\left\{\frac{4D^2O(\log(c/\delta)+\tilde{O}(d))T^2}{[\log(\Gamma)]^2}, O(\log(c/\delta)+\tilde{O}(d))\right\}.$$

If we are doing the doubling trick set $S' = 0$ and otherwise set

$$S' = CDK^2\sqrt{d\log(CM\tilde{O}(\sqrt{m}))+\log(c/\delta)} = \tilde{O}(\sqrt{d}), S = S' + \|f^*\|_\infty.$$

Then with probability at least $1-\delta$

$$\sup_{s\leq t}\|(T_{K^\infty} - T_n^s)r_s\|_2 = \tilde{O}\left(S\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 S\right] + \frac{S(1+T\Gamma^2)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right).$$

If we are performing the doubling trick the condition $m \geq \frac{4D^2O(\log(c/\delta)+\tilde{O}(d))T^2}{[\log(\Gamma)]^2}$ can be removed and the same conclusion holds.

Proof. Note by Lemma 2.5.36 we have

$$\sup_{s\leq T}\|(T_{K^\infty} - T_n^s)r_s\|_2 \leq \sup_{s\leq T}\|(T_{K^\infty} - T_n)r_s\|_2 + \sup_{s\leq T}\|K^\infty - K_s\|_\infty \|\hat{r}(0)\|_{\mathbb{R}^n}.$$

Well then by Theorem 2.5.26 we have with probability at least $1-\delta$ that

$$\sup_{t\leq T}\|K_t - K^\infty\|_\infty = \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3\|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right).$$

Separately by Lemma 2.5.34 we have with probability at least $1-\delta$

$$\sup_{s\leq T}\|(T_{K^\infty}-T_n)r_s\|_2=\tilde{O}\left(\frac{(\|f^*\|_\infty+S')(1+T\Gamma^2)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right)=\tilde{O}\left(\frac{S(1+T\Gamma^2)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right),$$

and

$$\|\hat{r}(0)\|_{\mathbb{R}^n}\leq\|r_0\|_\infty\leq S.$$

The result follows then from taking a union bound and replacing δ with $\delta/2$. □

We now proceed to prove the main theorem of this paper.

Theorem 2.3.5. *Assume that Assumptions 2.3.3 and 2.3.4 hold. Let P_k be the orthogonal projection in L^2 onto span$\{\phi_1,\ldots,\phi_k\}$ and let $D:=3\max\{|\sigma(0)|,M\|\sigma'\|_\infty,\|\sigma'\|_\infty,1\}$. If we are doing the doubling trick set $S'=0$ and otherwise set $S'=O\left(\sqrt{\tilde{O}(d)+\log(c/\delta)}\right)$, $S=\|f^*\|_\infty+S'$. Also let $T>0$. Assume $m\geq D^2\|y\|_{\mathbb{R}^n}^2T^2$, and*

$$m\geq O(\log(c/\delta)+\tilde{O}(d))\max\left\{T^2,1\right\}.$$

Then with probability at least $1-\delta$ we have that for all $t\leq T$ and $k\in\mathbb{N}$

$$\|P_k(r_t-\exp(-T_{K^\infty}t)r_0)\|_2\leq\frac{1-\exp(-\sigma_kt)}{\sigma_k}\tilde{O}\left(S\left[1+tS\right]\frac{\sqrt{d}}{\sqrt{m}}+S(1+T)\frac{\sqrt{p}}{\sqrt{n}}\right),$$

and

$$\|r_t-\exp(-T_{K^\infty}t)r_0\|_2\leq t\tilde{O}\left(S\left[1+tS\right]\frac{\sqrt{d}}{\sqrt{m}}+S(1+T)\frac{\sqrt{p}}{\sqrt{n}}\right).$$

Proof. By Lemma 2.5.35 we have for any $k\in\mathbb{N}$

$$\|P_k(r_t-\exp(-T_{K^\infty}t)r_0)\|_2\leq\frac{1-\exp(-\sigma_kt)}{\sigma_k}\sup_{s\leq t}\|(T_{K^\infty}-T_n^s)r_s\|_2,$$

and furthermore

$$\|r_t-\exp(-T_{K^\infty}t)r_0\|\leq t\sup_{s\leq t}\|(T_{K^\infty}-T_n^s)r_s\|_2.$$

Well the conditions on m in the hypothesis suffice to apply Lemma 2.5.37 with $\Gamma=e^2$ ensure that with probability at least $1-\delta$

$$\sup_{s\leq t}\|(T_{K^\infty}-T_n^s)r_s\|_2=\tilde{O}\left(S\frac{\sqrt{d}}{\sqrt{m}}\left[1+tS\right]+\frac{S(1+T)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right).$$

Since κ and σ_1 only depend on K^∞ which is fixed we will treat them as constants for simplicity of presentation of the main result (note that they were tracked in all previous results for anyone interested in the specific constants). The desired result follows from plugging in the above expression into the previous bounds after setting σ_1 and κ as constants. $\square$

Theorem 2.3.5 is strong enough to get a bound on the test error, which is demonstrated by the following corollary.

Corollary 2.3.6. *Assume Assumptions 2.3.3 and 2.3.4 hold. Suppose that $f^* = O(1)$ and assume we are performing the doubling trick where $f_0 \equiv 0$ so that $r_0 = -f^*$. Let $k \in \mathbb{N}$ and let P_k be the orthogonal projection onto span$\{\phi_1, \dots, \phi_k\}$. Set $t = \frac{\log(\sqrt{2}\|P_k f^*\|_2/\epsilon^{1/2})}{\sigma_k}$ Then we have that $m = \tilde{\Omega}(\frac{d}{\epsilon\sigma_k^4})$ and $n = \tilde{\Omega}\left(\frac{p}{\sigma_k^4\epsilon}\right)$ suffices to ensure with probability at least $1-\delta$*

$$\frac{1}{2}\|r_t\|_2^2 \le 2\epsilon + 2\|(I-P_k)f^*\|_2^2.$$

Proof. Set $t = \frac{\log(\sqrt{2}\|P_k f^*\|_2/\epsilon^{1/2})}{\sigma_k}$. Note that

$$\begin{aligned}
\frac{1}{2}\|r_t\|_2^2 &\le \frac{1}{2}\left[\|\exp(-T_{K^\infty}t)r_0\|_2 + \|r_t - \exp(-T_{K^\infty}t)r_0\|_2\right]^2 \\
&\le 2\max\{\|\exp(-T_{K^\infty}t)r_0\|_2, \|r_t - \exp(-T_{K^\infty}t)r_0\|_2\}^2 \\
&\le 2\left[\|\exp(-T_{K^\infty}t)r_0\|_2^2 + \|r_t - \exp(-T_{K^\infty}t)r_0\|_2^2\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\|\exp(-T_{K^\infty}t)r_0\|_2^2 = \|\exp(-T_{K^\infty}t)f^*\|_2^2 &= \sum_{i=1}^{\infty}\exp(-2\sigma_i t)|\langle f^*, \phi_i\rangle_2|^2 \\
&\le \exp(-2\sigma_k t)\sum_{i=1}^{k}|\langle f^*, \phi_i\rangle_2|^2 + \sum_{i=k+1}^{\infty}|\langle f^*, \phi_i\rangle_2|^2 \\
&= \frac{\epsilon}{2} + \|(I-P_k)f^*\|_2^2.
\end{aligned}$$

We want to apply Theorem 2.3.5 with $T = t$. We need

$$m \ge D^2\|y\|_{\mathbb{R}^n}^2 t^2 \text{ and } m \ge O(\log(c/\delta) + \tilde{O}(d))\max\{t^2, 1\}.$$

Note that since $f^* = O(1)$ we have that $\|\hat{r}(0)\|_{\mathbb{R}^n} = \|y\|_{\mathbb{R}^n} = O(1)$. Then

$$D^2 \|y\|_{\mathbb{R}^n}^2 t^2 = \tilde{O}\left(t^2\right) = \tilde{O}\left(\frac{1}{\sigma_k^2}\right).$$

thus our condition on m is strong enough to satisfy the first condition. Also $O(\log(c/\delta) + \tilde{O}(d)) \max\{t^2, 1\} = \tilde{O}(dt^2)$ which is satisfied by our condition on m. Thus by an application of Theorem 2.3.5 with $T = t$ we have with probability at least $1 - \delta$

$$\|r_t - \exp(-T_{K^\infty} t) r_0\|_2 \leq t\tilde{O}\left(\|f^*\|_\infty \left[1 + t\|f^*\|_\infty\right] \frac{\sqrt{d}}{\sqrt{m}} + \|f^*\|_\infty (1+t) \frac{\sqrt{p}}{\sqrt{n}}\right).$$

Recall that $f^* = O(1)$. Thus the first term above is

$$\tilde{O}\left(t^2 \frac{\sqrt{d}}{\sqrt{m}}\right) = \tilde{O}\left(\frac{\sqrt{d}}{\sigma_k^2 \sqrt{m}}\right).$$

Thus setting $m = \tilde{\Omega}(\frac{d}{\epsilon \sigma_k^4})$ suffices to ensure the first term is bounded by $\epsilon^{1/2}/(2\sqrt{2})$. Similarly the second term is

$$\tilde{O}\left(\frac{t^2\sqrt{p}}{\sqrt{n}}\right) = \tilde{O}\left(\frac{\sqrt{p}}{\sigma_k^2 \sqrt{n}}\right).$$

Thus setting $n = \tilde{\Omega}\left(\frac{p}{\sigma_k^4 \epsilon}\right)$ suffices to ensure that the second term bounded by $\epsilon^{1/2}/(2\sqrt{2})$. Thus in this case we have

$$\|r_t - \exp(-T_{K^\infty} t) r_0\|_2 \leq \frac{\epsilon^{1/2}}{\sqrt{2}}.$$

Thus we have

$$\frac{1}{2}\|r_t\|_2^2 \leq 2\left[\|\exp(-T_{K^\infty} t) r_0\|_2^2 + \|r_t - \exp(-T_{K^\infty} t) r_0\|_2^2\right] \leq 2\epsilon + 2\|(I - P_k) f^*\|_2^2.$$

□

2.5.3.8 Deterministic Initialization

In this section we will prove a version of Theorem 2.3.5 where instead of θ_0 being chosen randomly we take θ_0 to be some deterministic value. θ_0 could represent the parameters given by the output of some pretraining procedure that is independent of the training data, or selected with a priori knowledge.

Lemma 2.5.38. *Let θ_0 be a fixed parameter initialization. Let $\Gamma > 1$, $T > 0$, $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$,*

$$\xi(t) = \max\{\frac{1}{\sqrt{m}}\|W(t)\|_{op}, \frac{1}{\sqrt{m}}\|b(t)\|_2, \frac{1}{\sqrt{m}}\|a(t)\|_2, 1\},$$

$$\tilde{\xi}(t) = \max\{\max_{\ell\in[m]}\|w_\ell(t)\|_2, \|a(t)\|_\infty, \|b(t)\|_\infty, 1\}.$$

Furthermore assume

$$m \geq \frac{D^2\|\hat{r}(0)\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2}.$$

Then

$$\max_{t\in[0,T]}\xi(t) \leq \Gamma\xi(0), \quad \max_{t\in[0,T]}\tilde{\xi}(t) \leq \Gamma\tilde{\xi}(0).$$

Proof. By the hypothesis on m we have that for $t \leq T$

$$\frac{D\|\hat{r}(0)\|_{\mathbb{R}^n} t}{\sqrt{m}} \leq \log\Gamma.$$

Therefore by Lemmas 2.5.2 and 2.5.3 the desired result holds. □

Lemma 2.5.39. *Let θ_0 be a fixed parameter initialization. Let $\Gamma > 1$, $T > 0$, $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$,*

$$\xi(t) = \max\{\frac{1}{\sqrt{m}}\|W(t)\|_{op}, \frac{1}{\sqrt{m}}\|b(t)\|_2, \frac{1}{\sqrt{m}}\|a(t)\|_2, 1\},$$

and assume

$$m \geq \frac{D^2\|\hat{r}(0)\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2}.$$

Then for $t \leq T$

$$\|f_t - f^*\|_\infty \leq \|f_0 - f^*\|_\infty + t\|\hat{r}(0)\|_{\mathbb{R}^n} CD^2\Gamma^2\xi(0)^2.$$

Proof. Recall that

$$\partial_t(f_t(x) - f^*(x)) = -\frac{1}{n}\sum_{i=1}^n K_t(x, x_i)(f_t(x_i) - f^*(x_i)) = -\frac{1}{n}\sum_{i=1}^n K_t(x, x_i)\hat{r}(t)_i.$$

Thus

$$|\partial_t(f_t(x) - f^*(x))| \leq \frac{1}{n}\sum_{i=1}^{n} |K_t(x, x_i)||\hat{r}(t)_i| \leq \|K_t\|_\infty \|\hat{r}(t)\|_{\mathbb{R}^n} \leq \|K_t\|_\infty \|\hat{r}(0)\|_{\mathbb{R}^n}.$$

Well by Lemma 2.5.5 we have that $\|K_t\|_\infty \leq CD^2\xi^2(t)$. Also by Lemma 2.5.38 we have that

$$\max_{t\in[0,T]} \xi(t) \leq \Gamma\xi(0).$$

Thus by the fundamental theorem of calculus for $t \leq T$

$$|f_t(x) - f^*(x)| \leq |f_0(x) - f^*(x)| + t\|\hat{r}(0)\|_{\mathbb{R}^n} CD^2\Gamma^2\xi(0)^2.$$

Thus by taking the supremum over x we get

$$\|f_t - f^*\|_\infty \leq \|f_0 - f^*\|_\infty + t\|\hat{r}(0)\|_{\mathbb{R}^n} CD^2\Gamma^2\xi(0)^2$$

which is the desired conclusion. □

Lemma 2.5.40. *Let θ_0 be a fixed parameter initialization. Let K_0 denote the time-dependent NTK at initialization θ_0. Let*

$$T_{K_0}h(\bullet) = \int_X K_0(\bullet, s)h(s)d\rho(s)$$

and

$$T_n h(\bullet) = \frac{1}{n}\sum_{i=1}^{n} K_0(\bullet, x_i)h(x_i)$$

be the associated operators. Let $\kappa = \max_x K_0(x, x)$ and let σ_1 denote the largest eigenvalue of T_{K_0}. Let $\Gamma > 1$, $T > 0$, $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$,

$$\xi(0) = \max\{\frac{1}{\sqrt{m}}\|W(0)\|_{op}, \frac{1}{\sqrt{m}}\|b(0)\|_2, \frac{1}{\sqrt{m}}\|a(0)\|_2, 1\},$$

and assume

$$m \geq \frac{D^2\left[\|f^*\|_\infty + \|f_0\|_\infty\right]^2 T^2}{[\log(\Gamma)]^2}.$$

Then with probability at least $1 - \delta$ over the sampling of $x_1, \ldots, x_n$ we have that

$$\sup_{t\leq T} \|(T_n - T_{K_0})r_t\|_2 = \tilde{O}\left(\frac{(\|f^*\|_\infty + \|f_0\|_\infty)(1 + T\Gamma^2\xi(0)^2)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right).$$

Proof. First note that

$$\|\hat{r}(0)\|_{\mathbb{R}^n} \leq \|r_0\|_\infty \leq \|f^*\|_\infty + \|f_0\|_\infty . \tag{2.10}$$

Thus our hypothesis on m is strong enough to apply Lemma 2.5.38 so that we have

$$\max_{t\in[0,T]} \xi(t) \leq \Gamma\xi(0) =: A. \tag{2.11}$$

Also by Lemma 2.5.32

$$|b_0(t)| \leq |b_0(0)| + t\,\|\hat{r}(0)\|_{\mathbb{R}^n} ,$$

therefore

$$\max_{t\leq T} |b_0(t)| \leq |b_0(0)| + T\,\|\hat{r}(0)\|_{\mathbb{R}^n} := B.$$

Thus up until time T the neural network is in class $\mathcal{C}$ as defined in Corollary 2.5.29 with parameters A and B as defined above. Furthermore by Lemma 2.5.39 we have

$$\|f_t - f^*\|_\infty \leq \|f_0 - f^*\|_\infty + t\,\|\hat{r}(0)\|_{\mathbb{R}^n} CD^2\Gamma^2\xi(0)^2.$$

Well then by (2.10) and the above we have

$$\|r_t\|_\infty \leq (\|f^*\|_\infty + \|f_0\|_\infty)\left\{1 + TCD^2\Gamma^2\xi(0)^2\right\} =: S.$$

Thus by an application of Lemma 2.5.30 with $K = K_0$ we have with probability at least $1-\delta$ over the sampling of $x_1, \ldots, x_n$ that

$$\begin{aligned}\sup_{t\leq T} \|(T_n - T_{K_0})r_t\|_2 &\leq \frac{2\left[1 + S\sqrt{\sigma_1\kappa}\sqrt{2\log(c/\delta) + \tilde{O}(p)}\right]}{\sqrt{n}} \\ &= \tilde{O}\left(\frac{(\|f^*\|_\infty + \|f_0\|_\infty)(1 + T\Gamma^2\xi(0)^2)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right)\end{aligned}$$

where we have used that $S = \tilde{O}([\|f^*\|_\infty + \|f_0\|_\infty][1 + T\Gamma^2\xi(0)^2])$. □

Lemma 2.5.41. *Let θ_0 be a fixed parameter initialization. Let $\Gamma > 1$, $T > 0$,*

$$D := 3\max\{|\sigma(0)|, M\,\|\sigma'\|_\infty , \|\sigma'\|_\infty , 1\},$$

$$D' := \left[\max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\}^2[M^2+1] + D\|\sigma'\|_\infty\right]\max\{1, M\},$$
$$\xi(t) = \max\{\frac{1}{\sqrt{m}}\|W(t)\|_{op}, \frac{1}{\sqrt{m}}\|b(t)\|_2, \frac{1}{\sqrt{m}}\|a(t)\|_2, 1\},$$
$$\tilde{\xi}(t) = \max\{\max_{\ell\in[m]}\|w_\ell(t)\|_2, \|a(t)\|_\infty, \|b(t)\|_\infty, 1\}.$$

Furthermore assume

$$m \geq \frac{D^2\|\hat{r}(0)\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2}.$$

Then for $t \leq T$

$$\|K_0 - K_t\|_\infty \leq t\frac{CDD'}{\sqrt{m}}\Gamma^3\xi(0)^2\tilde{\xi}(0)\|\hat{r}(0)\|_{\mathbb{R}^n}.$$

Proof. Note by Lemma 2.5.24 we have that

$$\sup_{x,y\in B_M\times B_M}|\partial_t K_t(x,y)| \leq \frac{CDD'}{\sqrt{m}}\xi(t)^2\tilde{\xi}(t)\|\hat{r}(t)\|_{\mathbb{R}^n}.$$

Now applying Lemma 2.5.38 and the fact that $\|\hat{r}(t)\|_{\mathbb{R}^n} \leq \|\hat{r}(0)\|_{\mathbb{R}^n}$ from the above we get that for $t \leq T$

$$\sup_{x,y\in B_M\times B_M}|\partial_t K_t(x,y)| \leq \frac{CDD'}{\sqrt{m}}\Gamma^3\xi(0)^2\tilde{\xi}(0)\|\hat{r}(0)\|_{\mathbb{R}^n}.$$

Thus by the fundamental theorem of calculus we have that for $t \leq T$

$$\|K_0 - K_t\|_\infty \leq t\frac{CDD'}{\sqrt{m}}\Gamma^3\xi(0)^2\tilde{\xi}(0)\|\hat{r}(0)\|_{\mathbb{R}^n}.$$

□

Theorem 2.5.42. *Let* θ_0 *be a fixed parameter initialization. Assume that Assumption 2.3.3 holds. Let* $\{\phi_i\}_i$ *denote the eigenfunctions of* T_{K_0} *corresponding to the nonzero eigenvalues, which we enumerate* $\sigma_1 \geq \sigma_2 \geq \cdots$. *Let* P_k *be the orthogonal projection in* L^2 *onto* $span\{\phi_1, \ldots, \phi_k\}$ *and let* $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$. *Also let* $T > 0$ *and set*

$$\xi(0) = \max\{\frac{1}{\sqrt{m}}\|W(0)\|_{op}, \frac{1}{\sqrt{m}}\|b(0)\|_2, \frac{1}{\sqrt{m}}\|a(0)\|_2, 1\},$$
$$\tilde{\xi}(0) = \max\{\max_{\ell\in[m]}\|w_\ell(0)\|_2, \|a(0)\|_\infty, \|b(0)\|_\infty, 1\},$$

$$S := \tilde{O}([\|f^*\|_\infty + \|f_0\|_\infty][1 + T\xi(0)^2]).$$

Assume

$$m \geq D^2 \left[\|f^*\|_\infty + \|f_0\|_\infty\right]^2 T^2.$$

Then with probability at least $1-\delta$ *over the sampling of* $x_1, \ldots, x_n$ *we have that for all* $t \leq T$ *and* $k \in \mathbb{N}$

$$\|P_k(r_t - \exp(-T_{K_0}t)r_0)\|_2 \leq \frac{1-\exp(-\sigma_k t)}{\sigma_k}\tilde{O}\left(\frac{t}{\sqrt{m}}\xi(0)^2\tilde{\xi}(0)\,\|\hat{r}(0)\|_{\mathbb{R}^n}^2 + \frac{S\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right),$$

and

$$\|r_t - \exp(-T_{K_0}t)r_0\|_2 \leq t\tilde{O}\left(\frac{t}{\sqrt{m}}\xi(0)^2\tilde{\xi}(0)\,\|\hat{r}(0)\|_{\mathbb{R}^n}^2 + \frac{S\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right).$$

Proof. By Lemma 2.5.35 we have for any $k \in \mathbb{N}$

$$\|P_k(r_t - \exp(-T_{K_0}t)r_0)\|_2 \leq \frac{1-\exp(-\sigma_k t)}{\sigma_k}\sup_{s\leq t}\|(T_{K_0} - T_n^s)r_s\|_2,$$

and furthermore

$$\|r_t - \exp(-T_{K_0}t)r_0\| \leq t\sup_{s\leq t}\|(T_{K_0} - T_n^s)r_s\|_2.$$

Let $T_n h(\bullet) = \frac{1}{n}\sum_{i=1}^n K_0(\bullet, x_i)h(x_i)$ be the discretization of T_{K_0}. Thus by Lemma 2.5.36 we have

$$\sup_{s\leq t}\|(T_{K_0} - T_n^s)r_s\|_2 \leq \sup_{s\leq t}\|(T_{K_0} - T_n)r_s\|_2 + \sup_{s\leq t}\|K_0 - K_s\|_\infty \|\hat{r}(0)\|_{\mathbb{R}^n}.$$

Note from the inequality

$$\|\hat{r}(0)\|_{\mathbb{R}^n} \leq \|r_0\|_\infty \leq \|f^*\|_\infty + \|f_0\|_\infty$$

the hypothesis on m is strong enough to apply Lemma 2.5.41 with $\Gamma = e$. Well then by an application of Lemma 2.5.41 with $\Gamma = e$ we have that

$$\sup_{s\leq t}\|K_s - K_0\|_\infty = \tilde{O}\left(\frac{t}{\sqrt{m}}\xi(0)^2\tilde{\xi}(0)\,\|\hat{r}(0)\|_{\mathbb{R}^n}\right).$$

Separately by Lemma 2.5.40 we have with probability at least $1-\delta$

$$\sup_{s\leq t}\|(T_{K_0} - T_n)r_s\|_2 = \tilde{O}\left(\frac{(\|f^*\|_\infty + \|f_0\|_\infty)(1+T\xi(0)^2)\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right).$$

Combining these results we get that

$$\sup_{s\leq t} \|(T_{K_0} - T_n^s)r_s\|_2 \leq \tilde{O}\left(\frac{t}{\sqrt{m}}\xi(0)^2\tilde{\xi}(0)\,\|\hat{r}(0)\|_{\mathbb{R}^n}^2 + \frac{S\sqrt{\sigma_1\kappa p}}{\sqrt{n}}\right).$$

The desired result follows from plugging in the above expression into the previous bounds. □

2.5.4 Damped Deviations on the Training Set

The damped deviations lemma for the training set is incredibly simple to prove and yet is incredibly powerful as we will see later. Here is the proof.

Lemma 2.2.1. *Let $G \in \mathbb{R}^{n\times n}$ be an arbitrary positive semidefinite matrix and let G_s be the time dependent NTK matrix at time s. Then*

$$\hat{r}_t = \exp(-Gt)\hat{r}_0 + \int_0^t \exp(-G(t-s))(G - G_s)\hat{r}_s ds.$$

Proof. Note that we have the equation

$$\partial_t \hat{r}_t = -G_t\hat{r}_t = -G\hat{r}_t + (G - G_t)\hat{r}_t.$$

Thus by multiplying by the integrating factor $\exp(Gt)$ and using the fact that $\exp(Gt)$ and G commute we have that

$$\partial_t \exp(Gt)\hat{r}_t = \exp(Gt)(G - G_t)\hat{r}_t.$$

Therefore by the fundamental theorem of calculus

$$\exp(Gt)\hat{r}_t - \hat{r}_0 = \int_0^t \exp(Gs)(G - G_s)\hat{r}_s ds,$$

which after rearrangement gives

$$\hat{r}_t = \exp(-Gt)\hat{r}_0 + \int_0^t \exp(-G(t-s))(G - G_s)\hat{r}_s ds.$$

□

Throughout we will let $u_1, \ldots, u_n$ denote the eigenvectors of G^∞ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$, normalized to have unit norm in $\|\bullet\|_{\mathbb{R}^n}$, i.e. $\|u_i\|_{\mathbb{R}^n} = 1$. The following corollary demonstrates that if one is only interested in approximating the top eigenvectors, then the deviations of the NTK only need to be small relative to the cutoff eigenvalue λ_i that you care about.

Corollary 2.5.43. *Let P_k be the orthogonal projection onto span$\{u_1, \ldots, u_k\}$. Then for any $k \in [n]$*

$$\begin{aligned}\|P_k(\hat{r}_t - \exp(-G^\infty t)\hat{r}_0)\|_{\mathbb{R}^n} &\leq \sup_{s \leq t} \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} \frac{1 - \exp(-\lambda_k t)}{\lambda_k} \\ &\leq \sup_{s \leq t} \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} t.\end{aligned}$$

In particular

$$\begin{aligned}\|\hat{r}_t - \exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} &\leq \sup_{s \leq t} \|G^\infty - G_s\|_{\mathbb{R}^n} \|\hat{r}_0\|_{\mathbb{R}^n} \frac{1 - \exp(-\lambda_n t)}{\lambda_n} \\ &\leq \sup_{s \leq t} \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} t.\end{aligned}$$

Proof. Note by Lemma 2.2.1 we have that

$$\hat{r}_t - \exp(-G^\infty t)\hat{r}_0 = \int_0^t \exp(-G^\infty(t-s))(G^\infty - G_s)\hat{r}_s ds.$$

Therefore for any $k \in [n]$

$$\begin{aligned}P_k(\hat{r}_t - \exp(-G^\infty t)\hat{r}_0) &= P_k \int_0^t \exp(-G^\infty(t-s))(G^\infty - G_s)\hat{r}_s ds \\ &= \int_0^t P_k \exp(-G^\infty(t-s))(G^\infty - G_s)\hat{r}_s ds.\end{aligned}$$

Thus

$$\begin{aligned}
\|P_k(\hat{r}_t - \exp(-G^\infty t)\hat{r}_0)\|_{\mathbb{R}^n} &= \left\| \int_0^t P_k \exp(-G^\infty(t-s))(G^\infty - G_s)\hat{r}_s ds \right\|_{\mathbb{R}^n} \\
&\leq \int_0^t \|P_k \exp(-G^\infty(t-s))(G^\infty - G_s)\hat{r}_s\|_{\mathbb{R}^n} \, ds \\
&\leq \int_0^t \|P_k \exp(-G^\infty(t-s))\|_{op} \|G^\infty - G_s\|_{op} \|\hat{r}_s\|_{\mathbb{R}^n} \, ds \\
&\leq \int_0^t \exp(-\lambda_k(t-s)) \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} \, ds \\
&\leq \sup_{s \leq t} \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} \int_0^t \exp(-\lambda_k(t-s)) ds \\
&\leq \sup_{s \leq t} \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} \frac{1 - \exp(-\lambda_k t)}{\lambda_k} \\
&\leq \sup_{s \leq t} \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} \, t
\end{aligned}$$

where we have used the inequality $1 + x \leq \exp(x)$ in the last inequality. By specializing to the case $k = n$ since $\text{span}\{u_1, \ldots, u_n\} = \mathbb{R}^n$ we have

$$\begin{aligned}
\|\hat{r}_t - \exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} &\leq \sup_{s \leq t} \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} \frac{1 - \exp(-\lambda_n t)}{\lambda_n} \\
&\leq \sup_{s \leq t} \|G^\infty - G_s\|_{op} \|\hat{r}_0\|_{\mathbb{R}^n} \, t.
\end{aligned}$$

This completes the proof. □

Theorem 2.3.5 uses the concept of damped deviations to compare r_t with $\exp(-T_{K^\infty} t)r_0$. We can also prove the analogous statement on the training set that compares $\hat{r}_t$ to $\exp(-G^\infty t)\hat{r}_0$. The following is the analog of Theorem 2.3.5 on the training set.

Theorem 2.5.44. *Let* $D := 3\max\{|\sigma(0)|, M\|\sigma'\|_\infty, \|\sigma'\|_\infty, 1\}$. *Also let* $\Gamma > 1, T > 0$. *Furthermore assume*

$$m \geq \frac{4D^2 \|y\|_{\mathbb{R}^n}^2 T^2}{[\log(\Gamma)]^2} \text{ and } m \geq \max\left\{ \frac{4D^2 O(\log(c/\delta) + \tilde{O}(d))T^2}{[\log(\Gamma)]^2}, O(\log(c/\delta) + \tilde{O}(d)) \right\}.$$

Let P_k be the orthogonal projection onto span$\{u_1, \dots, u_k\}$. Then with probability at least $1-\delta$ we have for any $k \in [n]$ and $t \leq T$

$$\begin{aligned}\|P_k(\hat{r}_t - \exp(-G^\infty t)\hat{r}_0)\|_{\mathbb{R}^n} &\leq \frac{1-\exp(-\lambda_k t)}{\lambda_k} \|\hat{r}_0\|_{\mathbb{R}^n} \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 \|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right) \\ &\leq t \|\hat{r}_0\|_{\mathbb{R}^n} \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 \|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right),\end{aligned}$$

in particular

$$\begin{aligned}\|\hat{r}_t - \exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} &\leq \frac{1-\exp(-\lambda_n t)}{\lambda_n} \|\hat{r}_0\|_{\mathbb{R}^n} \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 \|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right) \\ &\leq t \|\hat{r}_0\|_{\mathbb{R}^n} \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 \|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right).\end{aligned}$$

If one instead does the doubling trick the term $\frac{4D^2 O(\log(c/\delta)+\tilde{O}(d))T^2}{[\log(\Gamma)]^2}$ can be removed from the hypothesis on m and the same conclusion holds.

Proof. By Corollary 2.5.43 we have

$$\begin{aligned}\|P_k(\hat{r}_t - \exp(-G^\infty t)\hat{r}_0)\|_{\mathbb{R}^n} &\leq \sup_{s \leq t} \|G^\infty - G_s\|_{\mathbb{R}^n} \|\hat{r}_0\|_{\mathbb{R}^n} \frac{1-\exp(-\lambda_k t)}{\lambda_k} \\ &\leq \sup_{s \leq t} \|G^\infty - G_s\|_{\mathbb{R}^n} \|\hat{r}_0\|_{\mathbb{R}^n} t.\end{aligned}$$

Well by Theorem 2.5.27 we have with probability at least $1-\delta$

$$\sup_{s \leq t} \|G^\infty - G_t\|_{op} \leq \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3 \|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right).$$

The desired result follows from plugging this in to the previous bounds. □

2.5.5 Proof of Theorem 2.3.7

We can now quickly prove our analog of Theorem 4.1 from [ADH19a].

Theorem 2.3.7. *Assume $m = \tilde{\Omega}(dn^5\epsilon^{-2}\lambda_n(H^\infty)^{-4})$ and $m \geq O(\log(c/\delta) + \tilde{O}(d))$ and $f^* = O(1)$. Assume we are performing the doubling trick so that $\hat{r}_0 = -y$. Let $v_1, \dots, v_n$ denote*

the eigenvectors of G^∞ normalized to have unit L2 norm $\|v_i\|_2 = 1$. Then with probability at least $1-\delta$

$$\hat{r}_t = \exp(-G^\infty t)(-y) + \delta(t),$$

where $\sup_{t\geq 0}\|\delta(t)\|_2 \leq \epsilon$. In particular

$$\|\hat{r}_t\|_2 = \sqrt{\sum_{i=1}^{n}\exp(-2\lambda_i t)|\langle y, v_i\rangle_2|^2} \pm \epsilon.$$

Proof. Set $T = \log(\|\hat{r}(0)\|_{\mathbb{R}^n}\sqrt{n}/\epsilon)/\lambda_n$. Note that since $f^* = O(1)$ and we are performing the doubling trick we have that $\|\hat{r}_0\|_{\mathbb{R}^n} = \|y\|_{\mathbb{R}^n} = O(1)$. Recall that $\lambda_n := \frac{1}{n}\lambda_n(H^\infty)$ therefore $m = \tilde{\Omega}(dn^5\epsilon^{-2}\lambda_n(H^\infty)^{-4}) = \tilde{\Omega}(dn\epsilon^{-2}\lambda_n^{-4})$ is strong enough to ensure that

$$m \geq \frac{4D^2\|y\|_{\mathbb{R}^n}^2 T^2}{[\log(2)]^2} = \tilde{O}(\lambda_n^{-2}), \quad m \geq O(\log(c/\delta) + \tilde{O}(d)) = \tilde{O}(d).$$

Then by an application of Theorem 2.5.44 with $\Gamma = 2$ we have with probability at least $1-\delta$ that for all $t \leq T$

$$\begin{aligned}\|\hat{r}_t - \exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} &\leq \frac{1-\exp(-\lambda_n t)}{\lambda_n}\|\hat{r}_0\|_{\mathbb{R}^n}\tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + t\Gamma^3\|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right)\\ &\leq \frac{\|\hat{r}_0\|_{\mathbb{R}^n}}{\lambda_n}\tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + T\Gamma^3\|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right).\end{aligned}$$

Since $f^* = O(1)$ we have that $\|\hat{r}_0\|_{\mathbb{R}^n} = \|y\|_{\mathbb{R}^n} = O(1)$ therefore the above bound is

$$\tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\frac{T}{\lambda_n}\right) = \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\frac{1}{\lambda_n^2}\right).$$

Thus $m = \tilde{\Omega}(dn^5\epsilon^{-2}\lambda_n(H^\infty)^{-4}) = \tilde{\Omega}(dn\epsilon^{-2}\lambda_n^{-4})$ suffices to make the above term bounded by $\epsilon/\sqrt{n}$. Thus in this case

$$\sup_{t\leq T}\|\hat{r}_t - \exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} \leq \epsilon/\sqrt{n}.$$

Let $\delta(t) = \hat{r}_t - \exp(-G^\infty t)\hat{r}_0$. We have just shown that $\sup_{t\leq T}\|\delta(t)\|_{\mathbb{R}^n} \leq \epsilon/\sqrt{n}$. We will now bound $\delta(t)$ for $t \geq T$. Note that for $t \geq T$

$$\|\exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} \leq \exp(-\lambda_n t)\|\hat{r}_0\|_{\mathbb{R}^n} \leq \exp(-\lambda_n T)\|\hat{r}_0\|_{\mathbb{R}^n} \leq \epsilon/\sqrt{n}.$$

Also for $t \geq T$

$$\|\hat{r}_t\|_{\mathbb{R}^n} \leq \|\hat{r}_T\|_{\mathbb{R}^n} \leq \|\exp(-G^\infty T)\hat{r}_0\|_{\mathbb{R}^n} + \|\delta(T)\|_{\mathbb{R}^n} \leq 2\epsilon/\sqrt{n}$$

where we have used that $\|\hat{r}_t\|_{\mathbb{R}^n}$ is nonincreasing for gradient flow. Therefore for $t \geq T$

$$\|\delta(t)\|_{\mathbb{R}^n} \leq \|\hat{r}_t\|_{\mathbb{R}^n} + \|\exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} \leq 3\epsilon/\sqrt{n}.$$

Thus we have shown

$$\sup_{t\geq 0} \|\delta(t)\|_{\mathbb{R}^n} \leq 3\epsilon/\sqrt{n}.$$

The desired result follows from replacing ϵ with $\epsilon/3$ in the previous argument and using the fact that $\|\bullet\|_2 = \sqrt{n}\,\|\bullet\|_{\mathbb{R}^n}$ and $\hat{r}_0 = -y$. □

2.5.6 Proof of Theorem 2.3.8

Using some lemmas that we leave to the following section, we can prove Theorem 2.3.8 quite quickly using the damped deviations equation and the NTK deviation bounds.

2.5.6.1 Main Theorem

Theorem 2.3.8. *Assume Assumptions 2.3.3 and 2.3.4 hold. Furthermore assume* $m = \tilde{\Omega}\left(\epsilon^{-2} dT^2 \|f^*\|_\infty^2 (1 + T\|f^*\|_\infty)^2\right)$ *where* $T > 0$ *is a time parameter and* $m \geq O(\log(c/\delta) + \tilde{O}(d))$ *and* $n \geq \frac{128\kappa^2 \log(2/\delta)}{(\sigma_k - \sigma_{k+1})^2}$. *Also assume* $f^* \in L^\infty(X) \subset L^2(X)$ *and let* $P^{T_{K^\infty}}$ *be the orthogonal projection onto the eigenspaces of* T_{K^∞} *corresponding to the eigenvalue* $\alpha \in \sigma(T_{K^\infty})$ *and higher. Assume that* $\left\|(I - P^{T_{K^\infty}})f^*\right\|_\infty \leq \epsilon'$ *for some* $\epsilon' \geq 0$. *Pick* k *so that* $\sigma_k = \alpha$ *and* $\sigma_{k+1} < \alpha$, *i.e.* k *is the index of the last repeated eigenvalue corresponding to* α *in the ordered sequence* $\{\sigma_i\}_i$. *Also assume we are performing the doubling trick so that* $\hat{r}(0) = -y$. *Then we have with probability at least* $1 - 3\delta$ *over the sampling of* $x_1, \ldots, x_n$ *and* θ_0 *that for* $t \leq T$

$$\|\hat{r}_t\|_{\mathbb{R}^n} \leq \exp(-\lambda_k t)\|y\|_{\mathbb{R}^n} + \frac{4\kappa\|f^*\|_2\sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}} + 2\epsilon' + \epsilon.$$

Proof. Recall that we have $\|\hat{r}(0)\|_{\mathbb{R}^n} = \|-y\|_{\mathbb{R}^n} \leq \|f^*\|_\infty$. Note that

$$m = \tilde{\Omega}\left(\epsilon^{-2} dT^2 \|f^*\|_\infty^2 (1 + T\|f^*\|_\infty)^2\right)$$

and $m \geq O(\log(c/\delta) + \tilde{O}(d))$ are strong enough to ensure the hypothesis of Theorem 2.5.44 is satisfied with $\Gamma = 2$. From now on $\Gamma = 2 = O(1)$ and will be treated as a constant. Then by Theorem 2.5.44 with probability at least $1 - \delta$ we have for $t \leq T$

$$\|\hat{r}_t - \exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} \leq T\|\hat{r}_0\|_{\mathbb{R}^n} \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\left[1 + T\Gamma^3\|\hat{r}(0)\|_{\mathbb{R}^n}\right]\right).$$

Thus using the fact from the doubling trick that $\|\hat{r}(0)\|_{\mathbb{R}^n} = \|y\|_{\mathbb{R}^n} \leq \|f^*\|_\infty$ setting $m = \tilde{\Omega}\left(\epsilon^{-2} dT^2 \|f^*\|_\infty^2 (1 + T\|f^*\|_\infty)^2\right)$ suffices to ensure that $\|\hat{r}_t - \exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} \leq \epsilon$ for $t \leq T$. Let P_k be the orthogonal projection onto $\text{span}\{u_1, \ldots, u_k\}$. Well then for $t \leq T$

$$\begin{aligned}\|\hat{r}_t\|_{\mathbb{R}^n} &\leq \|\exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} + \epsilon \leq \|P_k \exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} + \|(I - P_k)\exp(-G^\infty t)\hat{r}_0\|_{\mathbb{R}^n} + \epsilon \\ &\leq \exp(-\lambda_k t)\|\hat{r}_0\|_{\mathbb{R}^n} + \|(I - P_k)\hat{r}_0\|_{\mathbb{R}^n} + \epsilon.\end{aligned}$$

By Theorem 2.5.50 we have with probability at least $1 - 2\delta$ over the sampling of $x_1, \ldots, x_n$ that

$$\|(I - P_k)y\|_{\mathbb{R}^n} \leq 2\epsilon' + \frac{4\kappa\|f^*\|_2\sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}}.$$

Since we are using the doubling trick we have $\hat{r}_0 = -y$. Thus we have

$$\|(I - P_k)\hat{r}_0\|_{\mathbb{R}^n} \leq 2\epsilon' + \frac{4\kappa\|f^*\|_2\sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}}.$$

Thus by taking a union bound we have with probability at least $1 - 3\delta$ for all $t \leq T$

$$\|\hat{r}_t\|_{\mathbb{R}^n} \leq \exp(-\lambda_k t)\|\hat{r}_0\|_{\mathbb{R}^n} + \frac{4\kappa\|f^*\|_2\sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}} + +2\epsilon' + \epsilon.$$

The desired result follows from $\hat{r}_0 = -y$. □

2.5.6.2 Control of Initial Residual

We will use some of the notation and operator theory from Section 2.5.3.1 and Section 2.5.3.2 in this section, thus it is recommended to have read those sections first. Let $u_1, \ldots, u_n$ denote

the eigenvectors of G^∞ normalized to have unit norm in $\|\bullet\|_{\mathbb{R}^n}$, i.e. $\|u_i\|_{\mathbb{R}^n} = 1$. Let P_k be the orthogonal projection onto $\text{span}\{u_1, \ldots, u_k\}$. The goal of this section is to upper bound the extent to which the labels y participate in the bottom eigendirections of G^∞, i.e. to show that $\|(I - P_k)y\|_{\mathbb{R}^n}$ is small. Let $P^{T_{K^\infty}}$ be some projection onto the top eigenspaces of T_{K^∞}. The idea is to show that if $\left\|(I - P^{T_{K^\infty}})f^*\right\|_2$ is small then by picking P_k so that $\text{rank}(P_k) = \text{rank}(P^{T_{K^\infty}})$ then $\|(I - P_k)y\|_{\mathbb{R}^n}$ is also small with high probability. The results in this section essentially all appear in the proofs in [SY19]. We repeat the arguments here for completeness and due to differences in notation and constants.

We use some of the same machinery in [RBV10]. We define operators $L_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ and $T_n : \mathcal{H} \to \mathcal{H}$ by

$$T_{\mathcal{H}} f := \int_X \langle f, K_s \rangle_{\mathcal{H}} K_s d\rho(s),$$
$$T_n f := \frac{1}{n} \sum_{i=1}^n \langle f, K_{x_i} \rangle_{\mathcal{H}} K_{x_i}.$$

Note that $T_{\mathcal{H}}$ is equal to T_{K^∞} on $\mathcal{H}$ and T_n is simply the operator you get if you replace ρ in the defintion of $T_{\mathcal{H}}$ with the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. We define the "restriction" operator $R_n : \mathcal{H} \to \mathbb{R}^n$ by

$$R_n f = [f(x_1), f(x_2), \ldots, f(x_n)]^T.$$

Note here the domain of R_n is $\mathcal{H}$ but in other parts of this paper we will allow R_n to take more general functions as input. Define $R_n^* : \mathbb{R}^n \to \mathcal{H}$ by

$$R_n^*(v_1, \ldots, v_n) = \frac{1}{n} \sum_{i=1}^n v_i K_{x_i}.$$

It can be seen that

$$\langle R_n^* v, f \rangle_{\mathcal{H}} = \langle v, R_n f \rangle_{\mathbb{R}^n},$$

and thus R_n^* is the adjoint of R_n. Using these operators we may write $T_n = R_n^* R_n$ and $G^\infty = R_n R_n^*$. It will follow that T_n and G^∞ have the same eigenvalues (up to some zero eigenvalues) and their eigenvectors are related. We recall the following result from [RBV10] (Proposition 9):

Theorem 2.5.45. *[RBV10] The following hold*

- *The operator T_n is finite rank, self-adjoint and positive, and the matrix G^∞ is symmetric and semi-positive definite. In particular the spectrum $\sigma(T_n)$ of T_n has finitely many non-zero elements and they are contained in $[0, \kappa]$.*

- *The spectrum of T_n and G^∞ are the same up to zero, specifically $\sigma(G^\infty) \setminus \{0\} = \sigma(T_n) \setminus \{0\}$. Moreover if λ_i is a nonzero eigenvalue and u_i and v_i are the corresponding eigenvector and eigenfunction for G^∞ and T_n respectively (normalized to norm 1 in $\|\bullet\|_{\mathbb{R}^n}$ and $\|\bullet\|_{\mathcal{H}}$ respectively), then*

$$u_i = \frac{1}{\lambda_i^{1/2}} R_n v_i,$$

$$v_i = \frac{1}{\lambda_i^{1/2}} R_n^* u_i = \frac{1}{\lambda_i^{1/2}} \frac{1}{n} \sum_{j=1}^{n} K_{x_j} (u_i)_j,$$

  where $(u_i)_j$ is the jth component of the vector u_i.

- *The following decompositions hold*

$$G^\infty w = \sum_{j=1}^{k} \lambda_j \langle w, u_j \rangle_{\mathbb{R}^n} u_j,$$

$$T_n f = \sum_{j=1}^{k} \lambda_j \langle f, v_j \rangle_{\mathcal{H}} v_j,$$

  where $k = rank(G^\infty) = rank(T_n)$ and both sums run over the positive eigenvalues. $\{u_i\}_{i=1}^k$ is an orthonormal basis for $ker(G^\infty)^\perp$ and $\{v_i\}_{i=1}^k$ is an orthonormal basis for $ker(T_n)^\perp$.

We will make use of the following lemma from (Proposition 6 [RBV10]):

Lemma 2.5.46. *[RBV10] Let $\alpha_1 > \alpha_2 > \ldots > \alpha_N > \alpha_{N+1}$ be the top $N+1$ distinct eigenvalues of $T_\mathcal{H}$. Let $P^{T_\mathcal{H}}$ be the orthogonal projection onto the eigenfunctions of $T_\mathcal{H}$ corresponding to eigenvalues α_N and above. Let P^{T_n} be the projection onto the top k eigenvectors*

of T_n so that $k = dim(range(T_n)) = dim(range(T_{\mathcal{H}}))$. Assume further that

$$\|T_{\mathcal{H}} - T_n\|_{HS} \leq \frac{\alpha_N - \alpha_{N+1}}{4}.$$

Then

$$\left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS} \leq \frac{2}{\alpha_N - \alpha_{N+1}} \|T_{\mathcal{H}} - T_n\|_{HS}.$$

The following lemma will be useful.

Lemma 2.5.47. *Let $f^* \in L^2$ and let $P^{T_{\mathcal{H}}}$ and P^{T_n} be defined as in Lemma 2.5.46. Then*

$$\sum_{i=k+1}^{n} |\langle R_n P^{T_{K^\infty}} f^*), u_i\rangle_{\mathbb{R}^n}|^2 \leq \frac{\|f^*\|_2^2 \lambda_{k+1}}{\sigma_k} \left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS}^2.$$

Proof. We repeat the same proof as in [SY19] for completeness and to remove confusion that may arise from differences in notation. The proof was originally given in [RBV10] albeit with a minor error involving missing multiplicative factors. Note that

$$P^{T_{K^\infty}} f^* = \sum_{j=1}^{k} \langle f^*, \phi_j\rangle_2 \phi_j.$$

Therefore

$$\langle R_n P^{T_{K^\infty}} f^*, u_i\rangle_{\mathbb{R}^n} = \sum_{j=1}^{k} \langle f^*, \phi_j\rangle_2 \langle R_n\phi_j, u_i\rangle_{\mathbb{R}^n}.$$

Applying Cauchy-Schwarz we get

$$\begin{aligned} |\langle R_n P^{T_{K^\infty}} f^*, u_i\rangle_{\mathbb{R}^n}|^2 &\leq \left[\sum_{j=1}^{k} |\langle f^*, \phi_j\rangle_2|^2\right] \left[\sum_{j=1}^{k} |\langle R_n\phi_j, u_i\rangle_{\mathbb{R}^n}|^2\right] \\ &\leq \|f^*\|_2^2 \sum_{j=1}^{k} |\langle R_n\phi_j, u_i\rangle_{\mathbb{R}^n}|^2. \end{aligned}$$

Well then note that

$$\sum_{j=1}^{k} |\langle R_n\phi_j, u_i\rangle_{\mathbb{R}^n}|^2 = \sum_{j=1}^{k} |\langle \phi_j, R_n^* u_i\rangle_{\mathcal{H}}|^2 = \sum_{j=1}^{k} \lambda_i |\langle \phi_j, v_i\rangle_{\mathcal{H}}|^2.$$

Therefore

$$\sum_{i=k+1}^{n} |\langle R_n P^{T_{K^\infty}} f^*), u_i\rangle_{\mathbb{R}^n}|^2 \leq \|f^*\|_2^2 \sum_{i=k+1}^{n}\sum_{j=1}^{k} \lambda_i |\langle \phi_j, v_i\rangle_{\mathcal{H}}|^2$$
$$\leq \|f^*\|_2^2 \, \lambda_{k+1} \sum_{i=k+1}^{n}\sum_{j=1}^{k} |\langle \phi_j, v_i\rangle_{\mathcal{H}}|^2. \tag{2.12}$$

On the other hand

$$\left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS}^2 \geq \sum_{j=1}^{k} \left\|(P^{T_{\mathcal{H}}} - P^{T_n})\sqrt{\sigma_j}\phi_j\right\|_{\mathcal{H}}^2$$
$$\geq \sum_{j=1}^{k}\sum_{i=k+1}^{n} |\langle (P^{T_{\mathcal{H}}} - P^{T_n})\sqrt{\sigma_j}\phi_j, v_i\rangle_{\mathcal{H}}|^2.$$

Note that for $1 \leq j \leq k$ and $k+1 \leq i \leq n$ we have

$$\langle (P^{T_{\mathcal{H}}} - P^{T_n})\sqrt{\sigma_j}\phi_j, v_i\rangle_{\mathcal{H}} = \langle P^{T_{\mathcal{H}}}\sqrt{\sigma_j}\phi_j, v_i\rangle_{\mathcal{H}} - \langle P^{T_n}\sqrt{\sigma_j}\phi_j, v_i\rangle_{\mathcal{H}}$$
$$= \langle \sqrt{\sigma_j}\phi_j, v_i\rangle_{\mathcal{H}} - \langle \sqrt{\sigma_j}\phi_j, P^{T_n} v_i\rangle_{\mathcal{H}} = \langle \sqrt{\sigma_j}\phi_j, v_i\rangle_{\mathcal{H}}.$$

So

$$\sum_{j=1}^{k}\sum_{i=k+1}^{n} |\langle (P^{T_{\mathcal{H}}} - P^{T_n})\sqrt{\sigma_j}\phi_j, v_i\rangle_{\mathcal{H}}|^2 = \sum_{j=1}^{k}\sum_{i=k+1}^{n} |\langle \sqrt{\sigma_j}\phi_j, v_i\rangle_{\mathcal{H}}|^2$$
$$\geq \sigma_k \sum_{j=1}^{k}\sum_{i=k+1}^{n} |\langle \phi_j, v_i\rangle_{\mathcal{H}}|^2.$$

To summarize we have shown

$$\frac{1}{\sigma_k}\left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS}^2 \geq \sum_{j=1}^{k}\sum_{i=k+1}^{n} |\langle \phi_j, v_i\rangle_{\mathcal{H}}|^2.$$

Combining this with (2.12) we get the final result

$$\sum_{i=k+1}^{n} |\langle R_n P^{T_{K^\infty}} f^*), u_i\rangle_{\mathbb{R}^n}|^2 \leq \frac{\|f^*\|_2^2 \, \lambda_{k+1}}{\sigma_k}\left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS}^2.$$

□

We can use Lemma 2.5.47 to produce the following bound.

Lemma 2.5.48. *Let $f^* \in L^2$ and let $P^{T_{\mathcal{H}}}$ and P^{T_n} be defined as in Lemma 2.5.46. Then*

$$\sum_{i=k+1}^{n} |\langle R_n f^*, u_i \rangle_{\mathbb{R}^n}|^2 \leq \frac{2}{n} \sum_{i=1}^{n} |(I - P^{T_{K^\infty}}) f^*(x_i)|^2 + 2 \frac{\|f^*\|_2^2 \lambda_{k+1}}{\sigma_k} \left\| P^{T_{\mathcal{H}}} - P^{T_n} \right\|_{HS}^2.$$

Proof. We have that

$$\langle R_n f^*, u_i \rangle_{\mathbb{R}^n} = \langle R_n (I - P^{T_{K^\infty}}) f^*, u_i \rangle_{\mathbb{R}^n} + \langle R_n P^{T_{K^\infty}} f^*, u_i \rangle_{\mathbb{R}^n}.$$

Thus from the inequality $(a+b)^2 \leq 2(a^2+b^2)$ we get

$$\sum_{i=k+1}^{n} |\langle R_n f^*, u_i \rangle_{\mathbb{R}^n}|^2 \leq 2 \sum_{i=k+1}^{n} |\langle R_n (I - P^{T_{K^\infty}}) f^*, u_i \rangle_{\mathbb{R}^n}|^2 + 2 \sum_{i=k+1}^{n} |\langle R_n P^{T_{K^\infty}} f^*, u_i \rangle_{\mathbb{R}^n}|^2.$$

To control the first term we have

$$\sum_{i=k+1}^{n} |\langle R_n (I - P^{T_{K^\infty}}) f^*, u_i \rangle_{\mathbb{R}^n}|^2 \leq \left\| (I - P^{T_{K^\infty}}) f^* \right\|_{\mathbb{R}^n}^2 = \frac{1}{n} \sum_{i=1}^{n} |(I - P^{T_{K^\infty}}) f^*(x_i)|^2.$$

Then by applying Lemma 2.5.47 to the second term we get the desired result. □

We recall the following lemma from (Theorem 7 [RBV10]):

Lemma 2.5.49. *[RBV10] With probability at least $1 - \delta$ over the sampling of $x_1, \ldots, x_n$*

$$\|T_{\mathcal{H}} - T_n\|_{HS} \leq \frac{2\kappa \sqrt{2 \log(2/\delta)}}{\sqrt{n}}.$$

Now finally we can provide a bound on the labels participation in the bottom eigendirections.

Theorem 2.5.50. *Assume $f^* \in L^2(X)$ and let $P^{T_{K^\infty}}$ be the orthogonal projection onto the eigenspaces of T_{K^∞} corresponding to the eigenvalue $\alpha \in \sigma(T_{K^\infty})$ and higher. Assume that*

$$\left\| (I - P^{T_{K^\infty}}) f^* \right\|_\infty \leq \epsilon'$$

for some $\epsilon' \geq 0$. Pick k so that $\sigma_k = \alpha$ and $\sigma_{k+1} < \alpha$, i.e. k is the index of the last repeated eigenvalue corresponding to α in the ordered sequence $\{\sigma_i\}_i$. Let P_k denote the orthogonal projection onto span$\{u_1, \ldots, u_k\}$. Finally assume

$$n \geq \frac{128 \kappa^2 \log(2/\delta)}{(\sigma_k - \sigma_{k+1})^2}.$$

Then we have with probability at least $1-2\delta$ over the sampling of $x_1, \ldots, x_n$ that

$$\|(I-P_k)R_n f^*\|_{\mathbb{R}^n} = \|(I-P_k)y\|_{\mathbb{R}^n} \leq 2\epsilon' + \frac{4\kappa \|f^*\|_2 \sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}}.$$

Proof. From Lemma 2.5.48 we have

$$\sum_{i=k+1}^{n} |\langle R_n f^*, u_i\rangle_{\mathbb{R}^n}|^2 \leq \frac{2}{n}\sum_{i=1}^{n} |(I - P^{T_{K^\infty}})f^*(x_i)|^2 + 2\frac{\|f^*\|_2^2 \lambda_{k+1}}{\sigma_k} \left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS}^2.$$

By assumption we have that the first term is bounded by $2(\epsilon')^2$. Now we must control the term

$$2\frac{\|f^*\|_2^2 \lambda_{k+1}}{\sigma_k} \left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS}^2.$$

By Lemma 2.5.49 we have with probability at least $1-\delta$

$$\|T_{\mathcal{H}} - T_n\|_{HS} \leq \frac{2\kappa\sqrt{2\log(2/\delta)}}{\sqrt{n}}.$$

Then

$$n \geq \frac{128\kappa^2 \log(2/\delta)}{(\sigma_k - \sigma_{k+1})^2}$$

suffices so that the right hand side above is less than or equal to $\frac{\sigma_k - \sigma_{k+1}}{4}$. Thus by Lemma 2.5.46 we have that

$$\left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS} \leq \frac{2}{\sigma_k - \sigma_{k+1}} \|T_{\mathcal{H}} - T_n\|_{HS} \leq \frac{2}{\sigma_k - \sigma_{k+1}} \frac{2\kappa\sqrt{2\log(2/\delta)}}{\sqrt{n}}.$$

Thus from the above inequality we get that

$$2\frac{\|f^*\|_2^2 \lambda_{k+1}}{\sigma_k} \left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS}^2 \leq \frac{64\kappa^2 \|f^*\|_2^2 \lambda_{k+1} \log(2/\delta)}{\sigma_k(\sigma_k - \sigma_{k+1})^2 \cdot n}.$$

By Proposition 10 in [RBV10] we have separately with probability at least $1-\delta$

$$\lambda_{k+1} \leq \sigma_{k+1} + \frac{2\kappa\sqrt{2\log(2/\delta)}}{\sqrt{n}}.$$

Note that

$$n \geq \frac{128\kappa^2 \log(2/\delta)}{(\sigma_k - \sigma_{k+1})^2}$$

implies that

$$\frac{1}{\sqrt{n}} \leq \frac{\sigma_k - \sigma_{k+1}}{8\kappa\sqrt{2\log(2/\delta)}},$$

therefore

$$\lambda_{k+1} \leq \sigma_{k+1} + \frac{2\kappa\sqrt{2\log(2/\delta)}}{\sqrt{n}} \leq \sigma_k + \frac{1}{4}(\sigma_k - \sigma_{k+1}) \leq \frac{5}{4}\sigma_k.$$

Thus

$$2\frac{\|f^*\|_2^2 \lambda_{k+1}}{\sigma_k} \left\|P^{T_{\mathcal{H}}} - P^{T_n}\right\|_{HS}^2 \leq \frac{64\kappa^2 \|f^*\|_2^2 \lambda_{k+1}\log(2/\delta)}{\sigma_k(\sigma_k - \sigma_{k+1})^2 n} \leq \frac{80\kappa^2 \|f^*\|_2^2 \log(2/\delta)}{(\sigma_k - \sigma_{k+1})^2 n}.$$

Thus combined with our previous results we finally get that

$$\|(I - P_k)R_n f^*\|_{\mathbb{R}^n}^2 = \sum_{i=k+1}^{n} |\langle R_n f^*, u_i\rangle_{\mathbb{R}^n}|^2 \leq 2(\epsilon')^2 + \frac{80\kappa^2 \|f^*\|_2^2 \log(2/\delta)}{(\sigma_k - \sigma_{k+1})^2 n}.$$

Thus from the inequality $\sqrt{a+b} \leq \sqrt{2}(\sqrt{a} + \sqrt{b})$ which holds for all $a, b \geq 0$ we have

$$\|(I - P_k)R_n f^*\|_{\mathbb{R}^n} \leq 2\epsilon' + \frac{4\kappa \|f^*\|_2 \sqrt{10\log(2/\delta)}}{(\sigma_k - \sigma_{k+1})\sqrt{n}}.$$

Since $y = R_n f^*$ this provides the desired conclusion. □

2.5.7 NTK Integral Operator is Strictly Positive

Note that

$$K^{\infty}(x, x') = \mathbb{E}[\sigma(\langle w, x\rangle_2 + b)\sigma(\langle w, x'\rangle_2 + b)] + \mathbb{E}[a^2\sigma'(\langle w, x\rangle_2 + b)\sigma'(\langle w, x'\rangle_2 + b)][\langle x, x'\rangle_2 + 1] + 1$$

where the expectation is taken with respect to the parameter initialization. It suffices to show that the kernel corresponding to the first term above

$$K_a(x, x') := \mathbb{E}[\sigma(\langle w, x\rangle_2 + b)\sigma(\langle w, x'\rangle_2 + b)]$$

induces a strictly positive operator $T_{K_a} f(x) = \int_X K_a(x, s) f(s) d\rho(s)$. From the discussion in Section 2.5.3.1 it suffices to show that the RKHS corresponding to K_a is dense in L^2. In Proposition 4.1 in [RR08b] they showed that the RKHS associated with K_a has dense subset

$$\mathcal{F} := \left\{ x \mapsto \int_{\Theta} a(w, b)\sigma(\langle w, x\rangle_2 + b) d\mu(w, b) : \int_{\Theta} |a(w, b)|^2 d\mu(w, b) < \infty \right\}$$

where μ is the measure for the parameter initialization, i.e. $(w, b) \sim \mu$. Since $C(X)$ is dense in $L^2(X)$ it suffices to show that $\mathcal{F}$ is dense in $C(X)$ which is provided by the following theorem:

Theorem 2.5.51. *Let σ be L-Lipschitz and not a polynomial. Assume that μ is a strictly positive measure supported on all of $\mathbb{R}^{d+1}$. Also assume that*

$$\int_{\mathbb{R}^{d+1}} [\|w\|_2^2 + \|b\|_2^2] d\mu(w, b) < \infty.$$

Then $\mathcal{F}$ is dense in $C(X)$ under the uniform norm.

Proof. We first show that $\mathcal{F} \subset C(X)$. Suppose we have $f \in \mathcal{F}$ and write

$$f(x) = \int_{\mathbb{R}^{d+1}} a(w, b)\sigma(\langle w, x\rangle_2 + b) d\mu(w, b).$$

Well then

$$\begin{aligned} |f(x) - f(x')| &= \left| \int_{\mathbb{R}^{d+1}} a(w, b)[\sigma(\langle w, x\rangle_2 + b) - \sigma(\langle w, x'\rangle_2 + b)] d\mu(w, b) \right| \\ &\leq \int_{\mathbb{R}^{d+1}} |a(w, b)||\sigma(\langle w, x\rangle_2 + b) - \sigma(\langle w, x'\rangle_2 + b)| d\mu(w, b) \\ \leq \int_{\mathbb{R}^{d+1}} |a(w, b)| L |\langle w, x - x'\rangle| d\mu(w, b) &\leq \int_{\mathbb{R}^{d+1}} |a(w, b)| L \|w\|_2 \|x - x'\|_2 \, d\mu(w, b) \\ &\leq L \|x - x'\|_2 \left[\int_{\mathbb{R}^{d+1}} |a(w, b)|^2 d\mu(w, b) \right]^{1/2} \left[\int_{\mathbb{R}^{d+1}} \|w\|_2^2 \, d\mu(w, b) \right]^{1/2}. \end{aligned}$$

Thus f is Lipschitz and thus continuous. Now suppose that $\mathcal{F}$ is not dense in $C(X)$. Then by the Riesz representation theorem there exists a nonzero signed measure $\nu(x)$ with finite total variation such that $\int_X f(x) d\nu(x) = 0$ for all $f \in \mathcal{F}$. Well then writing $f(x) = \int_{\mathbb{R}^{d+1}} a(w, b)\sigma(\langle w, x\rangle_2 + b) d\mu(w, b)$ as before we have

$$\int_X \int_{\mathbb{R}^{d+1}} a(w, b)\sigma(\langle w, x\rangle_2 + b) d\mu(w, b) d\nu(x) = 0. \tag{2.13}$$

Note that

$$\begin{aligned}
&\int_{\mathbb{R}^{d+1}} |a(w,b)||\sigma(\langle w,x\rangle_2 + b)|d\mu(w,b) \\
&\leq \int_{\mathbb{R}^{d+1}} |a(w,b)|[|\sigma(0)| + L|\langle w,x\rangle_2 + b|]d\mu(w,b) \\
&\leq \int_{\mathbb{R}^{d+1}} |a(w,b)|[|\sigma(0)| + L(\|w\|_2 M + \|b\|_2)]d\mu(w,b) < \infty,
\end{aligned}$$

where we have used Cauchy-Schwarz and the hypothesis on the integrability of $\|w\|_2^2, \|b\|_2^2$ in the last step. Thus the integrand in (2.13) is $\mu \times \nu$ integrable thus by Fubini's theorem we may interchange the order of integration. To get that

$$\int_{\mathbb{R}^{d+1}} a(w,b) \int_X \sigma(\langle w,x\rangle_2 + b)d\nu(x)d\mu(w,b)$$

and the above holds for any $a \in L^2(\mathbb{R}^{d+1}, \mu)$. Thus $\int_X \sigma(\langle w,x\rangle_2 + b)d\nu(x) = 0$ for μ-almost every w, b. However by essentially the same proof as when we showed $\mathcal{F} \subset C(X)$ we may show that $\int_X \sigma(\langle w,x\rangle_2 + b)d\nu(x) = 0$ is a continuous function of (w,b). Thus since μ is a strictly positive measure on $\mathbb{R}^{d+1}$ this implies that $\int_X \sigma(\langle w,x\rangle_2 + b)d\nu(x) = 0$ for every $(w,b) \in \mathbb{R}^{d+1}$. However by Theorem 1 in [LLP93] we have that $\text{span}\{\sigma(\langle w,x\rangle_2 + b) : (w,b) \in \mathbb{R}^{d+1}\}$ is dense in $C(X)$. However by our previous conclusion and linearity we have that $\int g(x)d\nu(x) = 0$ for any g in $\text{span}\{\sigma(\langle w,x\rangle_2 + b) : (w,b) \in \mathbb{R}^{d+1}\}$, which implies then that ν must equal 0. Thus $\mathcal{F}$ is dense in $C(X)$. □

Since Gaussians are supported on all of $\mathbb{R}^{d+1}$ we have the following corollary:

Corollary 2.5.52. *If $(w,b) \sim N(0, I_{d+1})$ then K^∞ is strictly positive.*

CHAPTER 3

Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime

3.1 Introduction

Training heavily overparameterized networks via gradient based optimization has become standard operating procedure in deep learning. Overparameterized networks are able to interpolate arbitrary labels both in principle and in practice [ZBH17], rendering classical PAC learning theory insufficient to explain the generalization of networks within this modality. The high capacity of modern networks ensures that there are both good and bad empirical risk minimizers. Miraculously the network preferentially chooses the good solutions and sidesteps those that are unfavorable, posing a challenge and opportunity to today's researchers.

The success of overparameterized networks has prompted the theoretical community to search for more subtle forms of capacity control [NTS15, NTS17, GWB17]. The contemporary point-of-view is that the data distribution, model parameterization, and optimization algorithm are all relevant in limiting complexity. This has led to a variety of efforts to characterize the properties that networks and related models are biased towards when optimized via gradient descent. Examples include max-margin bias for classification problems [SHN18, JT19, NLG19, GLS18], minimum nuclear norm bias for matrix factorization [GWB17, LMZ18, GLS18], and minimum RKHS norm bias in the kernel regime [ZXL20].

Empirically it is known that neural networks tend to learn low Fourier frequencies first and add higher frequencies only later in training [RBA19, XZX19, YAA22], the phenomenon that has been titled "Spectral Bias" or the "Frequency Principle". Theoretical justifications of this have been proposed by studying networks in the kernel regime. For shallow univariate ReLU networks [RJK19, BGG20] demonstrate that the dominant eigenfunctions of the Neural Tangent Kernel (NTK) [JGH18] correspond to the low Fourier frequencies for the uniform distribution and more generally to smoother components for nonuniform distributions. This echos the results by [WTP19] and [JM21] that show that univariate ReLU networks in the kernel regime are biased towards smooth interpolants. Abstracting away from Fourier frequencies, "Spectral Bias" can be interpreted more broadly to mean bias towards learning the top eigenfunctions of the Neural Tangent Kernel. By looking at empirical approximations to the eigenfunctions, spectral bias was demonstrated to hold on the training set by [ADH19a], [BGG20], and [CFW21]. A recent work by [BM22a] was able to demonstrate that spectral bias holds off the training set for shallow feedforward networks when the network is underparameterized. In the present chapter we exploit the low-effective-rank property of the Fisher Information Matrix and are able to demonstrate that spectral bias holds outside the training set without the underparameterization requirement. In fact the number of samples can be on the same order as the width of the network. Furthermore, by leveraging a recent work by [LZB20b] bounding the Hessian of wide networks, our result permits deep networks with fully connected, convolutional, and residual layers. Consequently we are able to conclude that spectral bias holds for more realistic sample complexities and diverse architectures.

3.1.1 Our Contributions

- We provide quantitative bounds measuring the L^2 difference in function space between the trajectory of a finite-width network trained on finitely many samples from the idealized kernel dynamics of infinite width and infinite data (see Theorem 3.3.5 and Corollary 3.3.7).

- As an implication of these bounds, eigenfunctions of the NTK integral operator (not just their empirical approximations) are learned at rates corresponding to their eigenvalues (see Corollary 3.3.7 and Observation 3.3.8).
- We demonstrate that the network will inherit the bias of the kernel at the beginning of training even when the width only grows linearly with the number of samples (see Observation 3.3.9).

3.1.2 Related Work

NTK Convergence Results The NTK was introduced by [JGH18] while almost concurrently [DZP19] used it implicitly to prove a global convergence guarantee for gradient descent applied to a shallow ReLU network. These two highly charismatic works led to a flurry of subsequent works, of which we can only hope to provide a partial list. Global convergence for arbitrary labels was addressed in a series of works [DZP19, DLL19, OS20, ALS19a, NM20, Ngu21, ZCZ20, ZG19]. For arbitrary labels to our knowledge all works require the network width to either grow polynomially with the number of samples n or the inverse desired accuracy ϵ^{-1}. If one assumes the target function aligns with the NTK model, for shallow networks this can be reduced to polylogarithmic width for the logistic loss [JT20] or linear width for the squared loss [EMW20, SY19, BM22a].

Spectrum of the NTK/Hessian and Generalization The fact that the NTK tends to have a small number of large outlier eigenvalues has been observed in many works (e.g. [ADH19a, OFL19, LSO20]). [Pap20] demonstrated that for classification problems the logit gradients cluster within classes, which produces outliers in the spectra of the NTK and the Hessian of the loss. There have been a series of works analyzing the NTK/Hessian spectrum theoretically using random matrix theory and other tools (e.g. [KAA21, PW18, PB17, FW20, YS19]). Recently the spectrum of the NTK integral operator for ReLU networks has been shown to asymptotically follow a power law [VY21]. [ADH19a] provided a generalization

bound that is effective when the labels align with the top eigenvectors of the NTK. [OFL19] were able to use the low effective rank of the NTK to obtain generalization bounds, and [LSO20] used the same property to demonstrate robustness to label noise. The low effective rank of the Hessian has also been incorporated into PAC-Bayes bounds, most recently by [YMC22]. Interestingly, the notion of the effective dimension they define is essentially the same quantity we use to bound the model complexity of the network's linearization.

NTK Eigenvector and Eigenfunction Convergence Rates [LMX22] explicitly tracked the dynamics of the infinite-width shallow model in the Fourier domain. [ADH19a] demonstrated that when training the hidden layer of a shallow ReLU network, the residual error on the training set projected along eigenvectors of the NTK Gram matrix decays linearly at rates corresponding to the eigenvalues. [CFW21] proved a similar statement for training both layers, and [BGG20] proved the analogous statement for a deep fully connected ReLU network where the first and last layer are fixed. Our result can be viewed as the corresponding statement for the test residual instead of the empirical residual: projections of the test residual along eigen*functions* of the NTK *integral operator* are learned at rates corresponding to their eigenvalues. This was shown in a recent work [BM22a] for shallow fully connected networks that are underparameterized. By contrast our result does not require the network to be underparameterized, and holds for deep networks with fully connected, convolutional, and residual layers. We view our fundamental contribution as demonstrating that spectral bias holds with more realistic sample complexities and in considerable generality with respect to model architecture.

3.2 Preliminaries

3.2.1 Notation

Vectors $v \in \mathbb{R}^k$ will be column vectors by default. We will let $\langle \bullet, \bullet \rangle$ and $\|\bullet\|_2$ denote the Euclidean inner product and norm. We define $\langle \bullet, \bullet \rangle_{\mathbb{R}^n} = \frac{1}{n}\langle \bullet, \bullet \rangle$ and $\|\bullet\|_{\mathbb{R}^n} := \sqrt{\langle \bullet, \bullet \rangle_{\mathbb{R}^n}}$ to be the normalized Euclidean inner product and norm. The notation $\overline{B}(v, r) := \{w : \|w - v\|_2 \leq r\}$ will denote the *closed* Euclidean ball centered at v of radius r. $\|A\|_{op} := \sup_{\|v\|_2=1} \|Av\|_2$ will denote the operator norm for matrices. For a symmetric matrix $A \in \mathbb{R}^{k \times k}$, $\lambda_i(A)$ denotes its i-th largest eigenvalue, i.e. $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_k(A)$. For a set A we will let $|A|$ denote its cardinality. For a natural number $k \geq 1$, we will let $[k] := \{1, \ldots, k\}$. We will let $L^p(X, \nu)$ denote the L^p space over domain X with measure ν. We will denote the inner product associated with $L^2(X, \nu)$ as $\langle \bullet, \bullet \rangle_\nu$. We will use the standard big O and Ω notation with $\tilde{O}$ and $\tilde{\Omega}$ hiding logarithmic terms.

3.2.2 NTK Dynamics

Let $f(x; \theta)$ be our scalar-valued neural network model taking inputs $x \in X \subset \mathbb{R}^d$ parameterized by $\theta \in \mathbb{R}^p$. For now we will not specify a specific architecture. Our training data will be n input-label pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ where we assume that the labels y_i are generated from a fixed scalar-valued target function f^*, namely $f^*(x_i) = y_i$. We will let $y \in \mathbb{R}^n$ denote the label vector $y = (y_1, \ldots, y_n)^T$. Let $\hat{r}(\theta) \in \mathbb{R}^n$ denote the vector that measures the residual error on the training set, whose i-th entry is $\hat{r}(\theta)_i := f(x_i; \theta) - y_i$. We will optimize the squared loss

$$\Phi(\theta) := \frac{1}{2n} \|\hat{r}(\theta)\|_2^2 = \frac{1}{2} \|\hat{r}(\theta)\|_{\mathbb{R}^n}^2$$

via gradient flow

$$\partial_t \theta_t = -\partial_\theta \Phi(\theta),$$

which is the continuous time analog of gradient descent. For conciseness we will denote $\hat{r}(\theta_t)$ by $\hat{r}_t$ and let $r_t(x) := f(x;\theta_t) - f^*(x)$ denote the residual for an arbitrary input x not necessarily in the training set. We may also write $r(x;\theta) := f(x;\theta) - f^*(x)$ for the residual for an arbitrary θ.

We recall some key definitions and facts about the NTK. For a comprehensive introduction we refer the reader to [JGH18]. We recall the definition of the analytical NTK

$$K^\infty(x,x') := \lim_{m\to\infty} \langle \nabla_\theta f(x;\theta_0), \nabla_\theta f(x';\theta_0)\rangle,$$

where m is the width of the network and the convergence is in probability over the parameter initialization $\theta_0 \sim \mu$. The kernel K^∞ induces an integral operator $T_{K^\infty} : L^2(X,\rho) \to L^2(X,\rho)$

$$T_{K^\infty} g(x) := \int_X K^\infty(x,s) g(s) d\rho(s), \tag{3.1}$$

where X is our input space and ρ is the input distribution. We assume our training inputs $x_1,\dots,x_n$ are i.i.d. samples from ρ. More generally, for a continuous kernel $K(x,x')$ we define $T_K : L^2(X,\rho) \to L^2(X,\rho)$

$$T_K g(x) := \int_X K(x,s) g(s) d\rho(s). \tag{3.2}$$

Returning back to K^∞, by Mercer's theorem we have the decomposition

$$K^\infty(x,x') = \sum_{i=1}^\infty \sigma_i \phi_i(x) \phi_i(x'),$$

where $\{\phi_i\}$ is an orthonormal basis for $L^2(X,\rho)$ and $\{\sigma_i\}$ is a nonincreasing sequence of positive values. We will see that the bias at the beginning of training within our framework can be described entirely through the operator T_{K^∞} and its eigenfunctions. We note that T_{K^∞} depends only on the model architecture, parameter initialization distribution μ, and input distribution ρ. The training data sample $x_1,\dots,x_n$ introduces a discretization of the operator T_{K^∞}

$$T_n g(x) := \frac{1}{n}\sum_{i=1}^n K^\infty(x,x_i) g(x_i) = \int_X K^\infty(x,s) g(s) d\widehat{\rho}(s), \tag{3.3}$$

where $\widehat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ is the empirical measure. We now introduce the time-dependent NTK

$$K_t(x, x') := \langle \nabla_\theta f(x; \theta_t), \nabla_\theta f(x'; \theta_t) \rangle$$

with the associated time-dependent operator T_n^t

$$T_n^t g(x) := \frac{1}{n}\sum_{i=1}^{n} K_t(x, x_i) g(x_i) = \int_X K_t(x, s) g(s) d\widehat{\rho}(s). \tag{3.4}$$

The update rule for the residual r_t under gradient flow is given by

$$\partial_t r_t(x) = -\frac{1}{n}\sum_{i=1}^{n} K_t(x, x_i) r_t(x_i) = -T_n^t r_t.$$

Speaking loosely, as the network width tends to infinity the time-dependent NTK $K_t(x, x')$ becomes constant so that $K_t(x, x') = K^\infty(x, x')$ uniformly in t. If $K_t = K^\infty$ then we have the operator equality $T_n^t = T_n$. Similarly, heuristically as $n \to \infty$ we have $T_n \to T_{K^\infty}$. Thus in the idealized infinite-width, infinite-data limit the update rule becomes

$$\partial_t r_t = -T_{K^\infty} r_t,$$

which has the solution $r_t = \exp(-T_{K^\infty} t) r_0$ which is defined via its projections

$$\langle r_t, \phi_i \rangle_\rho = \exp(-\sigma_i t) \langle r_0, \phi_i \rangle_\rho.$$

Thus in this idealized setting the network learns eigenfunctions ϕ_i at rates determined by their eigenvalues σ_i. The dependence of the convergence rate on the magnitude of σ_i is particularly relevant as the NTK tends to have a very skewed spectrum. We can estimate the spectrum of K^∞ by randomly initializing a network and computing the Gram matrix $(G_0)_{i,j} := K_0(x_i, x_j)$. In Figure 3.1 we plot the spectrum of the NTK Gram Matrix $(G_0)_{i,j} := K_0(x_i, x_j)$ at initialization. We observe a small number of outlier eigenvalues of large magnitude followed by a long tail of small eigenvalues. This phenomenon has appeared in many works (e.g. [ADH19a, OFL19, LSO20]). For ReLU networks the spectrum is known to asymptotically follow a power law $\sigma_i \sim \Lambda i^{-\nu}$ [VY21]. The goal of this chapter is to quantify the extent to which a finite-width network trained on finitely many samples behaves like the idealized kernel dynamics $r_t = \exp(-T_{K^\infty} t) r_0$ corresponding to infinite width and infinite data.

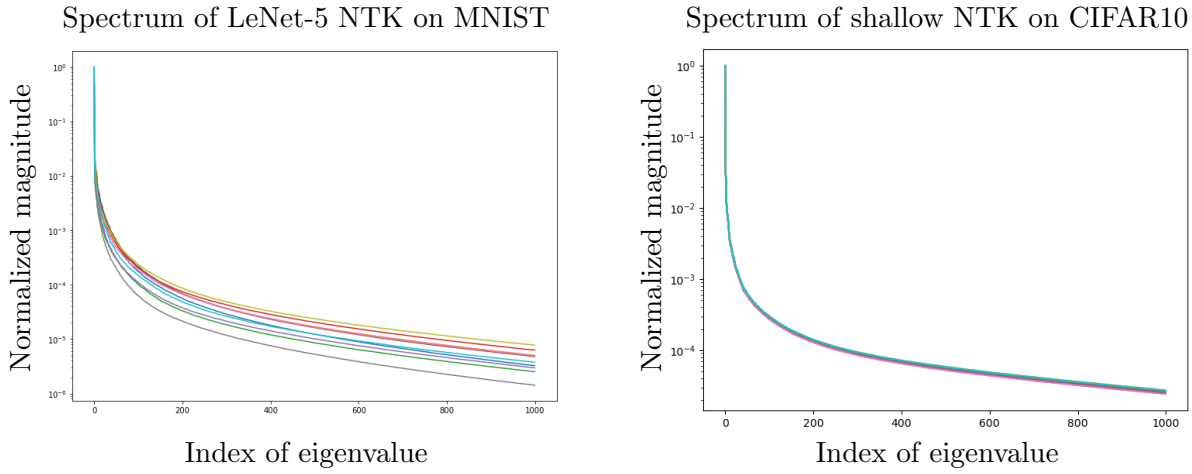

Figure 3.1: **NTK Spectrum on MNIST and CIFAR10** We plot the NTK spectrum on MNIST and CIFAR10 for two networks using 10 random parameter initializations and data batches. In both plots the x-axis represents the eigenvalue index k (linear scale) and the y-axis the normalized eigenvalue λ_k/λ_1 magnitude (log scale). To avoid numerical issues, we compute the NTK on a batch of size 2000 and plot the first 1000 eigenvalues. The left plot computed the NTK corresponding to the logit of class 0 for LeNet-5 on MNIST. The right plot is for a shallow fully-connected softplus network with 4000 hidden units on CIFAR10.

3.2.3 Applicable Architectures

We now specify an architecture for our model $f(x;\theta)$. We consider deep networks of the form

$$\begin{aligned} \alpha^{(0)} &:= x, \\ \alpha^{(l)} &:= \psi_l(\theta^{(l)}, \alpha^{(l-1)}), \quad l \in [L], \\ f(x;\theta) &:= \frac{1}{\sqrt{m_L}} v^T \alpha^{(L)}, \end{aligned}$$

where each $\psi_l(\theta^{(l)}, \bullet) : \mathbb{R}^{m_{l-1}} \to \mathbb{R}^{m_l}$ is a vector-valued function parameterized by $\theta^{(l)} \in \mathbb{R}^{p_l}$ and $v \in \mathbb{R}^{m_L}$. We define $\theta^{(L+1)} := v$ and set $\theta := ((\theta^{(1)})^T, \ldots, (\theta^{(L+1)})^T)^T$ to be the collection of all parameters. We assume each layer mapping ψ_l has one of the following forms:

$$\begin{aligned} \text{Fully Connected} : \psi_l(\theta^{(l)}, \alpha^{(l-1)}) &= \omega\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right) \\ \text{Convolutional} : \psi_l(\theta^{(l)}, \alpha^{(l-1)}) &= \omega\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} * \alpha^{(l-1)}\right) \\ \text{Residual} : \psi_l(\theta^{(l)}, \alpha^{(l-1)}) &= \omega\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right) + \alpha^{(l-1)} \end{aligned}$$

Here $\theta^{(l)} = vec(W^{(l)})$ and ω is a twice continuously differentiable function such that ω and ω' are Lipschitz. All parameters of the network will be trained as in practice. For feedforward and residual layers $W^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$ is a matrix. For the case of convolutional layers $W^{(l)} \in \mathbb{R}^{K \times m_l \times m_{l-1}}$ is an order-3 tensor with filter size K. The precise definition of the convolution $*$ is offered in Section 3.6.2.2. We will let $m = \min_l m_l$ denote the minimum width of the network. We will assume that $\max_l \frac{m_l}{m} = O(1)$. The input dimension $d := m_0$, the depth L, and the filter sizes K of convolutional layers will be treated as constant. The depth L being constant is essential for NTK convergence; see [HN20] for an explanation of failure modes whenever depth is nonconstant.

We will now discuss our initialization scheme. We will perform the antisymmetric initialization trick introduced by [ZXL20] so that the model is identically zero at initialization $f(\bullet;\theta_0) \equiv 0$. Let $f(x;\theta)$ be any neural network of the form described above. Then let

$\tilde{\theta} = \left[\begin{smallmatrix} \theta \\ \theta' \end{smallmatrix}\right]$ where $\theta, \theta' \in \mathbb{R}^p$. We then define

$$f_{ASI}(x; \tilde{\theta}) := \frac{1}{\sqrt{2}} f(x; \theta) - \frac{1}{\sqrt{2}} f(x; \theta')$$

which takes the difference of two rescaled copies of our original model $f(x; \theta)$ with parameters θ and θ' that are optimized freely. The antisymmetric initialization trick initializes $\theta_0 \sim N(0, I)$ then sets $\tilde{\theta}_0 = \left[\begin{smallmatrix} \theta_0 \\ \theta_0 \end{smallmatrix}\right]$. We then optimize the model f_{ASI} starting from the initialization $\tilde{\theta}_0$. This trick simultaneously ensures that the model is identically zero at initialization without changing the NTK at initialization [ZXL20]. For ease of notation we will simply assume from now on that $f(x; \theta) = f_{ASI}(x; \theta)$ and not write the subscript ASI.

3.3 Main Results

Before stating our main result, we enumerate our key assumptions for the sake of clarity, assumed to hold throughout. Detailed proofs are deferred to Section 3.6.

Assumption 3.3.1. *The activation ω is twice continuously differentiable and ω and ω' are Lipschitz.*

Assumption 3.3.2. *The input domain X is compact with strictly positive Borel measure ρ.*

Assumption 3.3.3. *The target function f^* satisfies $\|f^*\|_{L^\infty(X,\rho)} = O(1)$.*

Assumption 3.3.4. *We use the antisymmetric initilization trick so that $f(\bullet; \theta_0) \equiv 0$.*

Most activation functions except for ReLU satisfy Assumption 3.3.1, such as Softplus $\omega(x) = \ln(1 + e^x)$, Sigmoid $\omega(x) = \frac{1}{1+e^{-x}}$, and Tanh $\omega(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Assumption 3.3.2 is a sufficient condition for Mercer's Theorem to hold. While Mercer's theorem is often assumed to hold implicitly, we prefer to make this assumption explicit. Assumption 3.3.3 simply means the target function is bounded. We believe the antisymmetric initialization specified in Assumption 3.3.4 is not strictly necessary but it greatly simplifies the proofs and associated

bounds. To sidestep 3.3.4 one would utilize high probability bounds on the magnitude $|f(x;\theta_0)|$ at initialization. In the following results $f(x;\theta)$ will be any of the architectures discussed in Section 3.2.3. We are now ready to introduce the main result.

Theorem 3.3.5. *Let $T \geq 1, \epsilon > 0$. Let $K(x,x')$ be a fixed continuous, symmetric, positive definite kernel. For $k \in \mathbb{N}$ let $P_k : L^2(X,\rho) \to L^2(X,\rho)$ denote the orthogonal projection onto the span of the top k eigenfunctions of the operator T_K defined in Equation (3.2). Let $\sigma_k > 0$ denote the k-th eigenvalue of T_K. Then $m = \tilde{\Omega}(T^4/\epsilon^2)$ and $n = \tilde{\Omega}(T^2/\epsilon^2)$ suffices to ensure with probability at least $1 - O(mn)\exp(-\Omega(\log^2(m))$ over the parameter initilization θ_0 and the training samples $x_1, \ldots, x_n$ that for all $t \leq T$ and $k \in \mathbb{N}$*

$$\|P_k(r_t - \exp(-T_K t)r_0)\|^2_{L^2(X,\rho)} \leq \left[\frac{1-\exp(-\sigma_k t)}{\sigma_k}\right]^2 \cdot \left[4\|f^*\|^2_\infty \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right]$$

and

$$\|r_t - \exp(-T_K t)r_0\|^2_{L^2(X,\rho)} \leq t^2 \cdot \left[4\|f^*\|^2_\infty \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right].$$

3.3.1 Interpretation and Consequences

Theorem 3.3.5 compares the dynamics of the residual $r_t(x) := f(x;\theta_t) - f^*(x)$ of our finite-width model trained on finitely many samples to the idealized dynamics of a kernel method $\exp(-T_K t)r_0$ with infinite data. We recall that if ϕ_i is an eigenfunction of T_K with eigenvalue σ_i then $\langle \exp(-T_K t)r_0, \phi_i\rangle_\rho = \exp(-\sigma_i t)\langle r_0, \phi_i\rangle_\rho$. Thus the term $\exp(-T_K t)r_0$ learns the projection along eigenfunction ϕ_i linearly at rate σ_i. Whenever the NTK at initialization K_0 concentrates around K, the residual r_t will inherit this bias of the kernel dynamics $\exp(-T_K t)r_0$. Furthermore, the bound for the projected difference $\|P_k(r_t - \exp(-T_K t)r_0)\|^2_{L^2(X,\rho)}$ is smaller whenever σ_k is large. Therefore the bias appears more pronounced along eigendirections with large eigenvalues.

Consequences for the special case $K = K^\infty$ In the infinite-width limit, we have that K_0 approaches K^∞ for general architectures [Yan20]. For fixed x, x', by concentration results

the typical rate of convergence is $|K_0(x,x') - K^\infty(x,x')| = \tilde{O}(1/\sqrt{m})$ with high probability [DZP19, DLL19, HY20]. Bounds that hold uniformly over x, x' of the same rate were provided by [BM22a] and [BGW21]. A more pessimistic estimate of $1/m^{1/4}$ is provided by [ADH19b]. Even if the rate is $1/m^{1/4}$, we have that $m = \tilde{\Omega}(\epsilon^{-2})$ is strong enough to ensure that $|K_0(x,x') - K^\infty(x,x')| \leq \epsilon^{1/2}$. Given these results, it is reasonable to make the following assumption for the architectures we consider (see Section 3.6.5).

Assumption 3.3.6. *$m = \tilde{\Omega}(\epsilon^{-2})$ suffices to ensure that $\|K_0 - K^\infty\|^2_{L^2(X\times X, \rho\otimes\rho)} \leq \epsilon$ holds with high probability $1 - \delta(m)$ over the initialization θ_0 where $\delta(m) = o(1)$.*

Under this assumption, by setting $K = K^\infty$ in Theorem 3.3.5 we get the following corollary.

Corollary 3.3.7. *Let $\delta(m)$ be defined as in Assumption 3.3.6 which we assume to hold. Let $T \geq 1$ and $\epsilon > 0$. For $k \in \mathbb{N}$ let $P_k : L^2(X,\rho) \to L^2(X,\rho)$ denote the orthogonal projection onto the span of the top k eigenfunctions of the operator T_{K^∞} defined in Equation (3.1). Let $\sigma_k > 0$ denote the k-th eigenvalue of T_{K^∞}. Then $m = \tilde{\Omega}(T^4/\epsilon^2)$ and $n = \tilde{\Omega}(T^2/\epsilon^2)$ suffices to ensure with probability at least $1 - O(mn)\exp(-\Omega(\log^2(m)) - \delta(m)$ that for all $t \leq T$ and $k \in \mathbb{N}$*

$$\|P_k(r_t - \exp(-T_{K^\infty}t)r_0)\|^2_{L^2(X,\rho)} \leq \left[\frac{1-\exp(-\sigma_k t)}{\sigma_k}\right]^2 \cdot \epsilon$$

and

$$\|r_t - \exp(-T_{K^\infty}t)r_0\|^2_{L^2(X,\rho)} \leq t^2 \cdot \epsilon.$$

Informally Corollary 3.3.7 states that up to the stopping time T, we have that

$$r_t \approx \exp(-T_{K^\infty}t)r_0.$$

As discussed before, the term $\exp(-T_{K^\infty}t)r_0$ projected along the i-th eigenfunction of K^∞ decays linearly, $\langle \exp(-T_{K^\infty}t)r_0, \phi_i\rangle_\rho = \exp(-\sigma_i t)\langle r_0, \phi_i\rangle_\rho$. Given that K^∞ tends to have a highly skewed spectrum (see, e.g. Figure 3.1), the effect the magnitude of σ_i has on the

convergence rate is particularly relevant. Furthermore the bound on the projected difference $\|P_k(r_t - \exp(-T_{K^\infty}t)r_0)\|_{L^2(X,\rho)}$ is smaller whenever σ_k is large due to the dependence of the bound on the inverse eigenvalue σ_k^{-1}. Thus we have that the bias along the top eigenfunctions is particularly pronounced. Hence we make the following important observation.

Observation 3.3.8. *At the beginning of training the network learns projections along eigenfunctions of the Neural Tangent Kernel integral operator T_{K^∞} at rates corresponding to their eigenvalues. This is particularly true for the eigenfunctions with large eigenvalues.*

Scaling with respect to width and number of training data samples Now let us interpret how the width m and number of training samples n in the theorem scale. We note that as long as $n \leq m^\alpha$ for some $\alpha > 0$ the failure probability $O(mn)\exp(-\Omega(\log^2(m)))$ goes to zero as $m \to \infty$. Thus once m and n are sufficiently large relative to the stopping time T and precision ϵ, they can both tend to infinity at just about any rate to achieve a high probability bound. We also observe that m and n both have the same scaling with respect to ϵ, namely $m, n = \tilde{\Omega}(\epsilon^{-2})$. Thus for a fixed stopping time T we can send m and n to infinity at the same rate $m \sim n$ to send the error $\epsilon \to 0$. This is significant as typical NTK analysis requires $m = \Omega(poly(n))$. We reach following important conclusion.

Observation 3.3.9. *The network will inherit the bias of the kernel at the beginning of training even when the width m only grows linearly with the number of samples n.*

Scaling with respect to stopping time We will now address the scaling with respect to the stopping time T. The relevant question is how quickly the terms $P_k \exp(-T_{K^\infty}t)r_0$ and $\exp(-T_{K^\infty}t)r_0$ converge to zero. We observe that

$$\|P_k \exp(-T_{K^\infty}t)r_0\|_{L^2(X,\rho)} \leq \exp(-\sigma_k t)\,\|r_0\|_{L^2(X,\rho)} \leq \exp(-\sigma_k t)\,\|f^*\|_{L^\infty(X,\rho)}\,,$$

where we have used the antisymmetric initialization $r_0 = f(\bullet;\theta_0) - f^* = 0 - f^* = -f^*$ and the basic inequality $\|\bullet\|_{L^2(X,\rho)} \leq \|\bullet\|_{L^\infty(X,\rho)}$. Based on this we have that

$$t \geq \log(\|f^*\|_{L^\infty(X,\rho)}/\epsilon)/\sigma_k$$

suffices to ensure $\|P_k \exp(-T_{K^\infty} t) r_0\|_{L^2(X,\rho)} \leq \epsilon$. Using this fact we get the following corollary.

Corollary 3.3.10. *Let $\delta(m)$ be defined as in Assumption 3.3.6 which is assumed to hold. Let $T = \tilde{\Omega}(1/\sigma_k)$ and $\epsilon > 0$. For $k \in \mathbb{N}$ let $P_k : L^2(X,\rho) \to L^2(X,\rho)$ denote the orthogonal projection onto the span of the top k eigenfunctions of the operator T_{K^∞} defined in Equation (3.1). Let $\sigma_k > 0$ denote the k-th eigenvalue of T_{K^∞}. Then $m = \tilde{\Omega}({\sigma_k}^{-8}/\epsilon^2)$ and $n = \tilde{\Omega}(\sigma_k^{-6}/\epsilon^2)$ suffices to ensure that with probability at least $1 - O(mn)\exp(-\Omega(\log^2(m)) - \delta(m)$*

$$\|P_k r_T\|^2_{L^2(X,\rho)} \leq \epsilon$$

and in particular

$$\frac{1}{2}\|r_T\|^2_{L^2(X,\rho)} \leq \tilde{O}(\epsilon) + \|(I - P_k) r_0\|^2_{L^2(X,\rho)}.$$

The interpretation of the Corollary 3.3.10 is that the stopping time $T = \tilde{\Omega}(1/\sigma_k)$ is long enough to ensure that the network has learned the top k eigenfunctions to ϵ accuracy provided that $m = \tilde{\Omega}(\sigma_k^{-8}\epsilon^{-2})$ and $n = \tilde{\Omega}(\sigma_k^{-6}\epsilon^{-2})$. We note that the second conclusion of Corollary 3.3.10 is a bound on the test error $\frac{1}{2}\|r_t\|^2_{L^2(X,\rho)}$. From the antisymmetric initialization $r_0 = -f^*$ so that $\|(I - P_k) r_0\|^2_{L^2(X,\rho)} = \|(I - P_k) f^*\|^2_{L^2(X,\rho)}$. For a general target f^*, this quantity can decay arbitrary slowly with respect to k. Our goal with Theorem 3.3.5 was not to get a learning guarantee, but to describe how the bias of the kernel K^∞ is inherited by the finite-width network at the beginning of training even for general target functions. Nevertheless we will briefly sketch how it is possible to get a learning guarantee from Corollary 3.3.7 when f^* is in the RKHS of K^∞. In this case one can show that $\|\exp(-T_{K^\infty} t) r_0\|^2_{L^2(X,\rho)} = O\left(\frac{\|f^*\|^2_{\mathcal{H}}}{t}\right)$ where $\|\bullet\|_{\mathcal{H}}$ is the RKHS norm. Then treating $\|f^*\|_{\mathcal{H}}$ as a constant one can choose the stopping time $T \sim \epsilon^{-1}$ to bring the test error to ϵ provided

that $m, n = \tilde{\Omega}(poly(\epsilon^{-1}))$. More generally [VY21] derive sufficient conditions for the power law $\|\exp(-T_{K^\infty}t)r_0\|^2_{L^2(X,\rho)} \sim Ct^{-\xi}$ to hold. Using a similar argument in this case one can choose the stopping time $T \sim \epsilon^{-1/\xi}$ and get a learning guarantee for $m, n = \tilde{\Omega}(poly(\epsilon^{-1}))$.

3.3.2 Technical Comparison to Prior Work

[LXS19, ADH19b] compared the network $f(x;\theta)$ to its linearization

$$f_{lin}(x;\theta) := \langle \nabla_\theta f(x;\theta_0), \theta - \theta_0 \rangle + f(x;\theta_0)$$

in the regime where $m = \Omega(poly(n))$. When $m = \Omega(poly(n))$ one can show the loss converges to zero and the parameter changes $\|\theta_t - \theta_0\|_2$ are bounded. By contrast we avoid the condition $m = \Omega(poly(n))$ by employing a stopping time. [ADH19a, CFW21, BGG20] proved statements similar to Theorem 3.3.5 and Corollary 3.3.7 that roughly correspond to replacing T_{K^∞} with its Gram matrix induced by the training data $(G^\infty)_{i,j} = K^\infty(x_i, x_j)$ and replacing ρ with the empirical measure $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. [ADH19a, BGG20] operate in the regime where $m = \Omega(poly(n))$ and as a benefit do not need to employ a stopping time. [CFW21] instead of requiring $m = \Omega(poly(n))$ requires that the width m satisfies at least $m = \Omega(\max\{\sigma_k^{-14}, \epsilon^{-6}\})$ where σ_k is the cutoff eigenvalue. The most similar work is [BM22a], which demonstrated a version of Corollary 3.3.7 for a shallow feedforward network that is underparameterized. If p is the total number of parameters, they require $m = \tilde{\Omega}(\epsilon^{-1}T^2)$ and $n = \tilde{\Omega}(\epsilon^{-1}pT^2)$. This requires the network to be greatly underparameterized $n \gg p$. Our result was able to remove the dependence of n on p and demonstrate the result for general deep architectures at the expense of slightly worse scaling with respect to T and ϵ.

3.4 Proof Sketch

For simplicity we will go through the case where $K = K^\infty$. At a high level the proof revolves around bounding the difference between the operators T_{K^∞} and T_n^t defined in Equations

(3.1) and (3.4).

Bounding Operator Deviations [BM22a] demonstrated

$$r_t = \exp(-T_{K^\infty} t) r_0 + \int_0^t \exp(-T_{K^\infty}(t-s))(T_{K^\infty} - T_n^s) r_s ds.$$

This exhibits the residual r_t as a sum of $\exp(-T_{K^\infty} t) r_0$ and a correction term. The proof of Theorem 3.3.5 revolves around bounding the correction term which involves bounding

$$\|(T_{K^\infty} - T_n^s) r_s\|_{L^2(X,\rho)} \leq \|(T_{K^\infty} - T_n) r_s\|_{L^2(X,\rho)} + \|(T_n - T_n^s) r_s\|_{L^2(X,\rho)}.$$

At a high level $\|(T_n - T_n^s) r_s\|_{L^2(X,\rho)}$ will be small whenever the kernel deviations $K_0 - K_s$ are small. On the other hand by metric entropy based arguments we have that

$$\|(T_{K^\infty} - T_n) r_s\|_{L^2(X,\rho)}$$

will be small whenever n is large enough relative to the complexity of the residual functions r_s.

Comparison with Linearization Let $H(x;\theta) := \nabla_\theta^2 f(x;\theta)$ denote the Hessian of our network with respect to the parameters θ for a fixed input x. It turns out that if $\|H(x,\theta)\|_{op}$ was uniformly small over x and θ then the kernel deviations $K_0 - K_s$ would be bounded and the complexity of our model $f(x;\theta)$ would be controlled by the complexity of the linearized model $f_{lin}(x;\theta) := \langle \nabla_\theta f(x;\theta_0), \theta - \theta_0 \rangle$. The caveat to this approach is we do not in fact have a way to bound the Hessian $H(x,\theta)$ uniformly. However [LZB20b] demonstrated that for *fixed* x and $R > 0$ we have with high probability over the initialization θ_0

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|H(x,\theta)\|_{op} = \tilde{O}\left(\frac{R}{\sqrt{m}} poly(R/\sqrt{m})\right). \tag{3.5}$$

Using a priori parameter norm deviation bounds we have that $\|\theta_t - \theta_0\|_2 = O(\sqrt{t})$ and thus we can set $R = O(\sqrt{T})$. The difficulty then arises to get bounds that only depend on the Hessian $H(x;\theta)$ evaluated only on finitely many inputs x. We overcome this difficulty by

showing for fixed θ_0 one has high probability bounds over the sampling of the training data $x_1, \ldots, x_n$ that only require the Hessian evaluated on a finite point set. This requires some elaborate calculations involving Rademacher complexity. We then use the Fubini-Tonelli theorem and the Hessian bound (3.5) to get a bound over the simultaneous sampling of θ_0 and $x_1, \ldots, x_n$.

Covering Number of the Linearized Model The complexity of the residual functions r_s up to the stopping time T can be controlled by bounding the complexity of the function class $\mathcal{C} = \{f_{lin}(x;\theta) : \theta \in \overline{B}(\theta_0, R)\}$. In Section 3.6.1 we show that the $L^2(X,\rho)$ metric entropy of the linearized model $\mathcal{C} = \{f_{lin}(x;\theta) : \theta \in \overline{B}(\theta_0, R)\}$ is determined by the spectrum of the Fisher Information Matrix

$$F := \int_X \nabla_\theta f(x;\theta_0) \nabla_\theta f(x;\theta_0)^T d\rho(x). \tag{3.6}$$

Let $\lambda_1^{1/2} \geq \lambda_2^{1/2} \geq \cdots \geq 0$ denote the eigenvalues of $F^{1/2}$. We define the effective rank of $F^{1/2}$ at scale ϵ as

$$\tilde{p}(F^{1/2}, \epsilon) = |\{i : \lambda_i^{1/2} > \epsilon\}|.$$

This measures the number of dimensions within the unit ball whose image under $F^{1/2}$ can be larger than ϵ in Euclidean norm. In Section 3.6.1 we demonstrate that the ϵ covering number of $\mathcal{C}$ in $L^2(X,\rho)$, denoted $\mathcal{N}(\mathcal{C}, \|\bullet\|_{L^2(X,\rho)}, \epsilon)$, has the bound

$$\log \mathcal{N}(\mathcal{C}, \|\bullet\|_{L^2(X,\rho)}, \epsilon) = \tilde{O}(\tilde{p}(F^{1/2}, 0.75\epsilon/R)).$$

It turns out that for $\|(T_{K^\infty} - T_n) r_s\|_{L^2(X,\rho)}$ to be on the order of ϵ we merely need n to be large relative to $\tilde{p}(F^{1/2}, 0.75\epsilon/R)$. By contrast [BM22a] required that the network was underparameterized so that n was large relative to the total number of parameters p. Since $\tilde{p} \ll p$, this is what lets us relax the sample complexity dramatically. In fact for fixed R and ϵ we have that $\tilde{p} = \tilde{O}(1)$ with high probability as the width grows to infinity whereas $p \to \infty$. Interestingly, the quantity $\tilde{p}$ for the loss Hessian at convergence was used recently to derive

analytical PAC-Bayes bounds [YMC22]. Note for the squared loss the (empirical) FIM[1] can be taken as an approximation to the Hessian, and at a minimizer this approximation becomes exact. Thus these two notions are closely related.

3.5 Conclusion and Future Directions

We provided quantitative bounds measuring the L^2 difference in function space between a finite-width network trained on finitely many samples and the corresponding kernel method with infinite width and infinite data. As a consequence, the network will inherit the bias of the kernel at the beginning of training even when the width scales linearly with the number of samples. This bias is not only over the training data but over the entire input space. The key property that allows this is the low-effective-rank property of the Fisher Information Matrix (FIM) at initialization which controls the capacity of the model at the beginning of training. An interesting avenue for future work is to investigate if flat minima manifesting a FIM of low effective rank at the end of training can be related to the behavior of the network on out-of-sample data after training. One limitation of the results we present is that our framework can only characterize the network's bias up to a stopping time. There is compelling evidence that the kernel adapts to the target function later in training [BGL21, ABP22], and this falls outside our framework. Accounting for adaptations in the kernel is an important problem that is still being addressed by the theoretical community.

3.6 Appendix

The section is organized as follows.

• In Section 3.6.1 we bound the $L^2(X, \rho)$ metric entropy of the linearized model. This is

[1]Note that we define F as an expectation over the true input distribution ρ. To approximate the Hessian of the empirical loss one must replace ρ with the empirical measure $\hat{\rho}$.

necessary to bound the operator deviation $T_K - T_n$.

- In Section 3.6.2 we bound the Hessian of the network and introduce some technical lemmas. This is necessary in order to relate the network to the linearized model.
- In Section 3.6.3 we bound the quantity $\|(T_K - T_n^t)r_t\|_{L^2(X,\rho)}$. This section contains the bulk of the proof for the main result Theorem 3.3.5.
- In Section 3.6.4 we put the aforementioned results together to prove Theorem 3.3.5.
- In Section 3.6.5 we explain the merit of Assumption 3.3.6.
- In Section 3.6.6 we describe the details of our experiments with a link to the relevant code.

3.6.1 Covering Number for the Linearized Model

Our approach to generalization will be based on metric entropy(see e.g. [Wai19]), a fundamental tool in learning theory. We recall some basic definitions.

Definition 3.6.1. *Let V be a vector space with seminorm $\|\bullet\|$. For a subset $A \subset V$ we say that B is a proper ϵ-covering of A if $B \subset A$ and for all $a \in A$ there exists $b \in B$ such that $\|a - b\| \leq \epsilon$.*

Since we will concern ourselves solely with proper coverings we may remove the adjective "proper" when discussing ϵ-coverings. A closely related notion is the ϵ-covering number.

Definition 3.6.2. *Let V be a vector space with seminorm $\|\bullet\|$ and let $A \subset V$. For $\epsilon > 0$ we define the proper ϵ-covering number of A, denoted $\mathcal{N}(A, \|\bullet\|, \epsilon)$, by*

$$\mathcal{N}(A, \|\bullet\|, \epsilon) = \min_{N \,:\, N \text{ is proper } \epsilon\text{-covering of } A} |N|.$$

It is also useful to define the covering number of a set K with respect to another set L.

Definition 3.6.3. *Let K and L be two subsets of a vector space V. We define $\mathcal{N}(K, L)$ as the smallest $n \in \mathbb{N}$ such that there exists $v^{(1)}, \ldots, v^{(n)} \in K$ satisfying*

$$K \subset \bigcup_{i=1}^{n} (v^{(i)} + L).$$

Now consider a model $f_{lin}(x; \theta)$ that is potentially nonlinear in x but affine in θ. The motivating example is the following NTK model

$$f_{lin}(x; \theta) = f(x; \theta_0) + \langle \nabla_\theta f(x; \theta_0), \theta - \theta_0 \rangle.$$

We will be interested in deriving covering numbers for such classes of functions. Since translation by a fixed function does not change the covering number we will for convenience assume the model is linear in θ. Thus we will consider models of the form

$$f_{lin}(x; \theta) = \langle g(x), \theta \rangle.$$

The function g can be nonlinear and thus $x \mapsto f_{lin}(x; \theta)$ is typically nonlinear. For the NTK model we have $g(x) = \nabla_\theta f(x; \theta_0)$. Let X be our input space and let ν be some measure on X. We consider $L^2(X, \nu)$ where

$$\|h\|^2_{L^2(X,\nu)} = \int_X |h(x)|^2 d\nu(x).$$

Throughout we will assume that $\|g\|_2 \in L^2(X, \nu)$ i.e. $\int_X \|g\|_2^2 \, d\nu < \infty$. We will be interested in deriving covering numbers for classes of functions

$$\mathcal{C}_A := \{f_{lin}(x; \theta) : \theta \in A\}$$

where $A \subset \Theta$ is some subset of parameter space Θ. For now we will assume that $\Theta = \mathbb{R}^p$. We observe that

$$\|f_{lin}(\bullet; \theta_1) - f_{lin}(\bullet; \theta_2)\|^2_{L^2(X,\nu)} = \int_X |\langle g(x), \theta_1 - \theta_2 \rangle|^2 d\nu(x)$$

$$= \int_X (\theta_1 - \theta_2)^T g(x) g(x)^T (\theta_1 - \theta_2) d\nu(x) = (\theta_1 - \theta_2)^T \left[\int_X g(x) g(x)^T d\nu(x) \right] (\theta_1 - \theta_2).$$

Thus of primary importance is the symmetric positive semidefinite matrix

$$M := \int_X g(x)g(x)^T d\nu.$$

When ν is a probability measure and $f_{lin}(x;\theta)$ is the NTK model we have that

$$M = \mathbb{E}_{x\sim\nu}\left[\nabla_\theta f(x;\theta_0)\nabla_\theta f(x;\theta_0)^T\right]$$

is the (uncentered) gradient covariance matrix, which can be interpreted as the Fisher Information Matrix (FIM) for the squared loss. The two most interesting cases are when ν is the true input distribution or $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical distribution arising from the training samples. In the former case M is the true (uncentered) gradient covariance matrix and in the latter case M is the (uncentered) empirical covariance. For neural networks the FIM tends to have a very skewed spectrum (is approximately low rank), and thus the relations between the spectrum of M and the covering number will be particularly relevant. We will define the seminorm $\|\bullet\|_M$ as

$$\|v\|_M := \sqrt{v^T M v}.$$

The following lemma relates the covering number $\mathcal{N}(\mathcal{C}_A, \|\bullet\|_{L^2(X,\nu)}, \epsilon)$ to $\mathcal{N}(A, \|\bullet\|_M, \epsilon)$.

Lemma 3.6.4. *Let $N \subset A \subset \mathbb{R}^p$. Then N is a proper ϵ-covering of A with respect to the seminorm $\|\bullet\|_M$ if and only if $\mathcal{C}_N$ is a proper ϵ-covering of $\mathcal{C}_A$ with respect to the $L^2(X,\nu)$ norm.*

Proof. As we argued before we have that

$$\begin{aligned}\|f_{lin}(\bullet;\theta_1) - f_{lin}(\bullet;\theta_2)\|^2_{L^2(X,\nu)} &= (\theta_1-\theta_2)^T\left[\int_X g(x)g(x)^T d\nu(x)\right](\theta_1-\theta_2)\\ &= (\theta_1-\theta_2)^T M(\theta_1-\theta_2) = \|\theta_1-\theta_2\|^2_M.\end{aligned}$$

For each function in $h \in \mathcal{C}_A$ pick a representative parameter $\hat{\theta}(h) \in A$ so that $h = f_{lin}(\bullet;\hat{\theta}(h))$ (if M is strictly positive definite $\hat{\theta}(h)$ is unique). We can choose the mapping $h \mapsto \hat{\theta}(h)$ so

that the image of $\mathcal{C}_N$ under this mapping is N. Suppose N is an ϵ-covering for A with respect to $\|\bullet\|_M$. Then for each $\theta \in A$ we can choose θ' such that $\|\theta - \theta'\|_M \leq \epsilon$. Well then for any $h \in \mathcal{C}_A$ we can consider $\hat{\theta}(h)$ and choose $\theta' \in N$ such that $\epsilon \geq \left\|\hat{\theta}(h) - \theta'\right\|_M = \left\|f_{lin}(\bullet;\hat{\theta}(h)) - f_{lin}(\bullet;\theta')\right\|_{L^2(X,\nu)}$. It follows that $\mathcal{C}_N$ is an ϵ-covering of $\mathcal{C}_A$. Conversely suppose now that $\mathcal{C}_N$ is an ϵ-covering of $\mathcal{C}_A$ with respect to $\|\bullet\|_{L^2(X,\nu)}$. Well then for any $\theta \in A$ we can consider $f_{lin}(x;\theta)$ and take $h \in \mathcal{C}_N$ such that $\|f_{lin}(\bullet;\theta) - h(\bullet)\|_{L^2(X,\nu)} \leq \epsilon$. However since $h(\bullet) = f_{lin}(\bullet;\hat{\theta}(h))$ we have that $\epsilon \geq \left\|f_{lin}(\bullet;\theta) - f_{lin}(\bullet;\hat{\theta}(h))\right\|_2 = \left\|\theta - \hat{\theta}(h)\right\|_M$. Thus $\hat{\theta}(\mathcal{C}_N) = N$ is an ϵ-covering for A. $\square$

Thus covering the space $\mathcal{C}_A$ in $L^2(X,\nu)$ reduces to covering a subset of Euclidean space under the seminorm $\|\bullet\|_M$. By a change of coordinates we will assume without loss of generality that M is diagonal. Let $M^{1/2}$ be the square root of M and let $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ be the eigenvalues of $M^{1/2}$. We note that

$$\{v \in \mathbb{R}^p : \|v\|_M \leq 1\} = \left\{v \in \mathbb{R}^p : \sum_{i=1}^{p} \sigma_i^2 v_i^2 \leq 1\right\}.$$

Thus the unit ball in $\mathbb{R}^p$ determined by $\|\bullet\|_M$ is the ellipsoid with half-axis lengths σ_i^{-1} (if $\sigma_i = 0$ we consider the ellipsoid as being infinite along that dimension). For a general vector $a \in \mathbb{R}^p$ with nonnegative entries we define the ellipse

$$E_a := \left\{v \in \mathbb{R}^p : \sum_{i=1}^{p} \frac{v_i^2}{a_i^2} \leq 1\right\}$$

where in the sum if $a_i = 0$ we interpret $\frac{v_i^2}{a_i^2}$ as 0 if $v_i = 0$ and infinity otherwise. E_a is the ellipse with half-axis lengths $a_1, a_2, \ldots, a_n$. We will also let $B_r^k \subset \mathbb{R}^k$ denote the closed Euclidean ball in dimension k of radius r, specifically

$$B_r^k := \{v \in \mathbb{R}^k : \sum_{i=1}^{k} v_i^2 \leq r\}.$$

Our main study will be bounding $\mathcal{N}(A, \|\bullet\|_M, \epsilon)$ when $A = \{\theta \in \mathbb{R}^p : \|\theta\|_2 \leq R\} = B_R^p$. This amounts to covering a Euclidean ball with ellipsoids determined by $\|\bullet\|_M$. Fortunately, there

are well established results for coverings involving ellipsoids. Let $\sigma = (\sigma_1, \ldots, \sigma_p)^T$ denote the spectrum of $M^{1/2}$ and let $M^{-1/2}$ denote the pseudo-inverse of $M^{1/2}$. Let L denote the closed unit ball in $\mathbb{R}^p$ under the seminorm $\|\bullet\|_M$. In geometric terms $\mathcal{N}(B_R^p, \|\bullet\|_M, \epsilon) = \mathcal{N}(B_R^p, \epsilon L)$. We claim that up to an application of $M^{1/2}$ or $M^{-1/2}$, covering B_R^p with translates of ϵL is equivalent to covering $E_{\frac{R}{\epsilon}\sigma}$ with translates of B_1^p. This is formalized in the following lemma.

Lemma 3.6.5. *Let $M \in \mathbb{R}^{p\times p}$ be a symmetric positive semidefinite matrix and let $\sigma = (\sigma_1, \ldots, \sigma_p)^T \in \mathbb{R}^p$ denote the eigenvalues of $M^{1/2}$. Then $\mathcal{N}(B_R^p, \|\bullet\|_M, \epsilon) = \mathcal{N}(E_{\frac{R}{\epsilon}\sigma}, B_1^p)$.*

Proof. By a change of basis we can assume without loss of generality that M is diagonal. Let L denote the closed unit ball of $\mathbb{R}^p$ under $\|\bullet\|_M$. We note that in geometric terms $\mathcal{N}(B_R^p, \|\bullet\|_M, \epsilon) = \mathcal{N}(B_R^p, \epsilon L)$. Since we can dilate by $1/\epsilon$ we can replace R with R/ϵ and ϵ with 1. Thus for convenience we will assume for now that $\epsilon = 1$. We note that if $v^{(1)}, \ldots, v^{(n)}$ form an L covering of B_R^p as in

$$B_R^p \subset \bigcup_{i=1}^{n} (v^{(i)} + L),$$

then

$$E_{R\sigma} = M^{1/2}(B_R^p) \subset \bigcup_{i=1}^{n} (M^{1/2}v^{(i)} + M^{1/2}(L)) \subset \bigcup_{i=1}^{n} (M^{1/2}v^{(i)} + B_1^p).$$

Thus $M^{1/2}v^{(1)}, \ldots, M^{1/2}v^{(n)}$ forms a B_1^p covering of $E_{R\sigma}$. Conversely suppose $v^{(1)}, \ldots, v^{(n)}$ satisfy

$$E_{R\sigma} \subset \bigcup_{i=1}^{n} (v^{(i)} + B_1^p)$$

and let P be the projection onto $\text{span}\{e_i : \sigma_i \neq 0\}$ where e_i denotes the ith standard basis vector. Then

$$P(B_R^p) = M^{-1/2}(E_{R\sigma}) \subset \bigcup_{i=1}^{n} (M^{-1/2}v^{(i)} + M^{-1/2}(B_1^p)) = \bigcup_{i=1}^{n} (M^{-1/2}v^{(i)} + P(L)).$$

However L is infinitely long along the dimensions outside $im(P)$, and thus

$$B_R^p \subset \bigcup_{i=1}^{n} (M^{-1/2}v^{(i)} + L).$$

Thus $M^{-1/2}v^{(1)}, \ldots, M^{-1/2}v^{(n)}$ form an L covering of B_R^p. We conclude that $\mathcal{N}(B_R^p, L) = \mathcal{N}(E_{R\sigma}, B_1^p)$. Thus for general $\epsilon > 0$ we have that

$$\mathcal{N}(B_R^p, \|\bullet\|_M, \epsilon) = \mathcal{N}(B_R^p, \epsilon L) = \mathcal{N}(B_{R/\epsilon}^p, L) = \mathcal{N}(E_{\frac{R}{\epsilon}\sigma}, B_1^p).$$

□

We will let $vol(\bullet)$ denote volume in the standard Lebesgue sense. If $a \in \mathbb{R}^p$ is a vector with positive entries we recall that the volume of an ellipsoid E_a is given by the formula

$$vol(E_a) = vol(B_1^p) \prod_{i=1}^{p} a_i.$$

When most of the a_i are very small we have that E_a is very thin and has small volume and thus we expect the covering number to be small. Coverings for ellipsoids are well established with roots in geometric functional analysis. The following lemma is phrased the same as Theorems 1 and 2 in [Dum06]. The result dates back to classic results in geometric functional analysis. Specifically a similar result for more general convex bodies is sketched at the end of Chapter 5 in [Pis89] which also appeared in [GKS87, Proposition 1.7]. We don't need the additional generality for our purposes. We will offer the simplest proof needed for our purposes for completeness and clarity.

Lemma 3.6.6 ([Dum06, Pis89, GKS87])**.** *Let $a \in \mathbb{R}^p$ be a vector with nonnegative entries. Let $J = \{i : a_i > 1\}$, $K = \sum_{i \in J} \log(a_i)$, $\gamma \in (0, 1/2)$, and $\mu_\gamma = |\{i : a_i^2 > (1-\gamma)^2\}|$. Then the proper covering number $\mathcal{N}(E_a, B_1^p)$ satisfies*

$$K \leq \log \mathcal{N}(E_a, B_1^p) \leq K + \mu_\gamma \log\left(\frac{3}{\gamma}\right).$$

Proof. We first prove the lower bound. Let $J = \{i : a_i > 1\}$, $m = |J|$, and let P be the orthogonal projection onto $\text{span}\{e_i : i \in J\}$ where e_i denotes the standard basis. Suppose $v^{(1)}, \ldots, v^{(n)}$ are the centers of a B_1^p covering of E_a, specifically

$$E_a \subset \bigcup_{i=1}^{n} (v^{(i)} + B_1^p).$$

Well then

$$P(E_a) \subset \bigcup_{i=1}^{n} P(v^{(i)} + B_1^p) = \bigcup_{i=1}^{n} (Pv^{(i)} + B_1^m).$$

Well then by the standard volume estimate we get that

$$n \cdot vol(B_1^m) \geq vol\left(\bigcup_{i=1}^{n} (Pv^{(i)} + B_1^m)\right) \geq vol(P(E_a))$$

and thus

$$n \geq \frac{vol(P(E_a))}{vol(B_1^m)} = \prod_{i \in J} a_i.$$

Now we prove the upper bound. Let $\gamma \in (0, 1/2)$ and let $J_\gamma = \{i : a_i^2 > (1-\gamma)^2\}$, $\mu_\gamma = |J_\gamma|$, and let P be the orthogonal projection onto $\text{span}\{e_i : i \in J_\gamma\}$. We first notice that if $v \in E_a$ we have that $\|(I-P)v\|_2 \leq 1-\gamma$, indeed because for $v \in E_a$

$$\sum_{i \notin J_\gamma} \frac{v_i^2}{(1-\gamma)^2} \leq \sum_{i \notin J_\gamma} \frac{v_i^2}{a_i^2} \leq 1.$$

Thus if $v^{(1)}, \ldots, v^{(n)}$ are the centers of a proper $B_\gamma^{\mu_\gamma}$ covering of $P(E_a)$ then by the triangle inequality they also induce a proper B_1^p covering of E_a. Thus let $v^{(1)}, \ldots, v^{(n)}$ be a maximal subset of $P(E_a)$ such that for $i \neq j$ $\left\|v^{(i)} - v^{(j)}\right\|_2 > \gamma$. By maximality $v^{(1)}, \ldots, v^{(n)}$ form a $B_\gamma^{\mu_\gamma}$ covering of $P(E_a)$. Well then the balls $v^{(i)} + B_{\gamma/2}^{\mu_\gamma}$ are all disjoint and contained in $P(E_a) + B_{\gamma/2}^{\mu_\gamma}$. Thus by the volume estimates

$$n \cdot vol(B_{\gamma/2}^{\mu_\gamma}) = vol\left(\bigcup_{i=1}^{n} (v^{(i)} + B_{\gamma/2}^{\mu_\gamma})\right) \leq vol\left(P(E_a) + B_{\gamma/2}^{\mu_\gamma}\right).$$

Thus

$$n \leq \frac{vol\left(P(E_a) + B_{\gamma/2}^{\mu_\gamma}\right)}{vol(B_{\gamma/2}^{\mu_\gamma})}.$$

Note that $B_{1-\gamma}^{\mu_\gamma} \subset P(E_a)$ and thus $B_{\gamma/2}^{\mu_\gamma} \subset \frac{\gamma}{2(1-\gamma)} P(E_a)$. Now let $\|\bullet\|_{P(E_a)}$ be the norm on $\mathbb{R}^{\mu_\gamma}$ such that $P(E_a)$ is the unit ball. Then note for v, w such that $v \in P(E_a)$ and $w \in B_{\gamma/2}^{\mu_\gamma}$ we have that

$$\|v + w\|_{P(E_a)} \leq \|v\|_{P(E_a)} + \|w\|_{P(E_a)} \leq 1 + \frac{\gamma}{2(1-\gamma)}.$$

We conclude that $P(E_a) + B_{\gamma/2}^{\mu_\gamma} \subset \left(1 + \frac{\gamma}{2(1-\gamma)}\right) P(E_a)$. Therefore

$$n \leq \frac{vol\left(P(E_a) + B_{\gamma/2}^{\mu_\gamma}\right)}{vol(B_{\gamma/2}^{\mu_\gamma})} \leq \frac{vol\left[\left(1 + \frac{\gamma}{2(1-\gamma)}\right) P(E_a)\right]}{vol(B_{\gamma/2}^{\mu_\gamma})} = \left(\frac{2}{\gamma} + \frac{1}{1-\gamma}\right)^{\mu_\gamma} \prod_{i \in J_\gamma} a_i.$$

Note that since $\gamma < 1/2$ we have that $\frac{1}{1-\gamma} < \frac{1}{\gamma}$. Therefore $\frac{2}{\gamma} + \frac{1}{1-\gamma} \leq \frac{3}{\gamma}$. Moreover $\prod_{i \in J_\gamma} a_i \leq \prod_{i \in J} a_i$. Thus

$$n \leq \left(\frac{2}{\gamma} + \frac{1}{1-\gamma}\right)^{\mu_\gamma} \prod_{i \in J_\gamma} a_i \leq \left(\frac{3}{\gamma}\right)^{\mu_\gamma} \prod_{i \in J} a_i.$$

After taking logarithms we get the desired result. □

From the Lemmas 3.6.5 and 3.6.6 we see that the covering number $\mathcal{N}(B_R^p, \|\bullet\|_M, \epsilon)$ will depend on how many eigenvalues of M lie above a certain threshold. Let $A \in \mathbb{R}^p$ be a symmetric positive semidefinite square matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. We define the effective rank of A at scale ϵ as

$$\tilde{p}(A, \epsilon) = |\{i : \lambda_i > \epsilon\}|.$$

This measures the number of dimensions within B_1 whose image under A can be larger than ϵ in Euclidean norm. We will also define

$$|A|_{>c} = \prod_{i:\lambda_i > c} \lambda_i,$$

which can be thought of the determinant of A after removing some eigenvalues. We then have our main result.

Theorem 3.6.7. *Let $g : X \to \mathbb{R}^p$ such that $\|g\|_2 \in L^2(X, \nu)$. Let $\mathcal{C} = \{x \mapsto \langle g(x), \theta \rangle : \|\theta\|_2 \leq R\}$, $\gamma \in (0, 1/2)$. Define $M \in \mathbb{R}^{p \times p}$ by*

$$M = \int_X g(x) g(x)^T d\nu(x).$$

Then the proper covering number $\mathcal{N}(\mathcal{C}, \|\bullet\|_{L^2(X,\nu)}, \epsilon)$ satisfies

$$\log \left| \frac{R}{\epsilon} M^{1/2} \right|_{>1} \leq \log \mathcal{N}(\mathcal{C}, \|\bullet\|_{L^2(X,\nu)}, \epsilon) \leq \log \left| \frac{R}{\epsilon} M^{1/2} \right|_{>1} + \tilde{p}\left(\frac{R}{\epsilon} M^{1/2}, (1-\gamma)\right) \log\left(\frac{3}{\gamma}\right).$$

Proof. We have by Lemmas 3.6.4 and 3.6.5 that $\mathcal{N}(\mathcal{C}, \|\bullet\|_{L^2(X,\nu)}, \epsilon) = \mathcal{N}(B_R^p, \|\bullet\|_M, \epsilon) = \mathcal{N}(E_{\frac{R}{\epsilon}\sigma}, B_1^p)$ where $\sigma = (\sigma_1, \ldots, \sigma_p)^T \in \mathbb{R}^p$ is the vector of eigenvalues of $M^{1/2}$. Well then by applying Lemma 3.6.6 with $a = \frac{R}{\epsilon}\sigma$ we have that

$$\log \left| \frac{R}{\epsilon} M^{1/2} \right|_{>1} \leq \log \mathcal{N}(E_{\frac{R}{\epsilon}\sigma}, B_1) \leq \log \left| \frac{R}{\epsilon} M^{1/2} \right|_{>1} + \tilde{p}\left(\frac{R}{\epsilon} M^{1/2}, (1-\gamma) \right) \log \left(\frac{3}{\gamma} \right).$$

The desired result thus follows. □

Corollary 3.6.8. *Let $g : X \to \mathbb{R}^p$ such that $\|g\|_2 \in L^2(X, \nu)$. Let $\mathcal{C} = \{x \mapsto \langle g(x), \theta \rangle : \|\theta\|_2 \leq R\}$, $\gamma \in (0, 1/2)$. Define $M \in \mathbb{R}^{p \times p}$ by*

$$M = \int_X g(x) g(x)^T d\nu(x).$$

Then the proper covering number $\mathcal{N}(\mathcal{C}, \|\bullet\|_{L^2(X,\nu)}, \epsilon)$ satisfies

$$\log \mathcal{N}(\mathcal{C}, \|\bullet\|_{L^2(X,\nu)}, \epsilon) = \tilde{O}\left(\tilde{p}\left(M^{1/2}, \frac{3\epsilon}{4R} \right) \right).$$

Proof. This follows from setting $\gamma = 1/4$ and the fact that

$$\begin{aligned} \log \left| \frac{R}{\epsilon} M^{1/2} \right| &= \log \left(\prod_{\sigma_i > \epsilon/R} \frac{R}{\epsilon} \sigma_i \right) \\ &\leq \tilde{p}(M^{1/2}, \epsilon/R) \log \left(\frac{R\sigma_1}{\epsilon} \right) \\ &\leq \tilde{p}\left(M^{1/2}, \frac{3\epsilon}{4R} \right) \log \left(\frac{R\sigma_1}{\epsilon} \right). \end{aligned}$$

□

3.6.2 Bounding the Network Hessian and other Technical Items

3.6.2.1 Main Hessian Bound

For a fixed input x we will let $H(x, \theta) := \nabla_\theta^2 f(x; \theta)$ denote the Hessian of the network with respect to the parameters. We will use the following result, which follows from the proof of a result by [LZB20a, Theorem 3.3], which we state here explicitly for reference.

Theorem 3.6.9 (Reformulation of [LZB20a, Theorem 3.3]). *Let $f(x;\theta)$ be a general neural network of the form specified in Section 3.2.3 which can be a fully connected network, CNN, ResNet or a mixture of these types. Let m be the minimum of the hidden layer widths and assume $\max_l \frac{m_l}{m} = O(1)$. Given any fixed $R \geq 1$ and $x \in X$ then with probability at least $1 - Cme^{-c\log^2(m)}$*

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|H(x,\theta)\|_{op} = \tilde{O}\left(\frac{R}{\sqrt{m}}\left[\max\left\{1, \frac{R}{\sqrt{m}}\right\}\right]^{O(L)}\right).$$

In particular if $\sqrt{m} \geq R$ then

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|H(x,\theta)\|_{op} = \tilde{O}\left(\frac{R}{\sqrt{m}}\right).$$

The constants $c, C > 0$ depend on the architecture but are independent of the width.

Discussion of the statement of Theorem 3.6.9 We note that our statement of Theorem 3.6.9 is not exactly the same as the result of [LZB20a, Theorem 3.3]. [LZB20a] do not explicitly write the failure probability and the dependence of the Hessian bound on R in the statement of the theorem. In Theorem 3.6.9 we write the failure probability and dependence on the radius R according to the proof[2] provided by the authors [LZB20a]. We also add the assumption $\max_l \frac{m_l}{m} = O(1)$ to the hypothesis. This assumption is so that the initial weight matrices satisfy $\frac{1}{\sqrt{m}}\left\|W_0^{(l)}\right\|_{op} = O(1)$ with high probability (see Lemma 3.6.10). This condition on the initial weight matrices appears in the proof by [LZB20a]. The authors [LZB20a] do not need to explicitly add this assumption because they perform the proof for the case where all the layers have equal width for simplicity of presentation, while stating that the proof generalizes to the case where the layers do not have equal width.

Exponential dependence on depth We note that under the $\tilde{O}$ notation in Theorem 3.6.9 there are constants that depend exponentially on the network depth L. For this reason it

[2]We communicated with the authors to better understand the dependence of the bound on the quantity R. Nevertheless we accept full liability for any misinterpretation of their proof.

is essential that the depth L be treated as constant. We will now briefly explain how the term $\max\{1, R/\sqrt{m}\}^{O(L)}$ arises in the bound in Theorem 3.6.9. For simplicity assume the network is fully connected at each layer (the same form of argument holds for the other cases). Let $\xi(\theta) = \max_l \frac{1}{\sqrt{m}} \left\| W^{(l)} \right\|_{op}$. With high probability over the initialization we have that $\xi(\theta_0) = O(1)$ (see Lemma 3.6.10). Furthermore for θ such that $\|\theta - \theta_0\|_2 \leq R$ we have that $\xi(\theta) \leq \xi(\theta_0) + \frac{R}{\sqrt{m}} = O(\max\{1, R/\sqrt{m}\})$. It turns out that the features $\alpha^{(l)}$ at each layer l satisfy $\frac{1}{\sqrt{m}} \left\| \alpha^{(l)} \right\|_2 = O(\xi^{O(L)})$. Well for θ such that $\|\theta - \theta_0\|_2 \leq R$ as stated before we have that $\xi(\theta) = O(\max\{1, R/\sqrt{m}\})$. Consequently for such θ we get that $\frac{1}{\sqrt{m}} \left\| \alpha^{(l)} \right\|_2 = O(\xi^{O(L)}) = O(\max\{1, R/\sqrt{m}\}^{O(L)})$. The Hessian bound inherits dependence on the quantity $O(\max\{1, R/\sqrt{m}\}^{O(L)})$ from its dependence the normalized feature $\frac{1}{\sqrt{m}} \left\| \alpha^{(l)} \right\|_2$ norms.

Antisymmetric initialization and the Hessian We will now explain how the antisymmetric initialization trick will not hinder us from bounding the Hessian via Theorem 3.6.9. Let $f(x; \theta)$ denote any model of the form specified in Section 3.2.3 where $\theta \in \mathbb{R}^p$. Let $\tilde{\theta} = \left[\begin{smallmatrix} \theta \\ \theta' \end{smallmatrix} \right]$ where $\theta, \theta' \in \mathbb{R}^p$. Recall the antisymmetric initialization trick defines the model

$$f_{ASI}(x; \tilde{\theta}) := \frac{1}{\sqrt{2}} f(x; \theta) - \frac{1}{\sqrt{2}} f(x; \theta')$$

which takes the difference of two rescaled copies of the model $f(x; \bullet)$ with parameters θ and θ' that are optimized freely. We then note that the Hessian of f_{ASI} has the block diagonal structure

$$\nabla^2_{\tilde{\theta}} f_{ASI}(x; \tilde{\theta}) = \frac{1}{\sqrt{2}} \begin{bmatrix} \nabla^2_{\theta} f(x; \theta) & 0 \\ 0 & -\nabla^2_{\theta'} f(x; \theta') \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} H(x, \theta) & 0 \\ 0 & -H(x, \theta') \end{bmatrix}.$$

Well then it is not too hard to show that

$$\left\| \nabla^2_{\tilde{\theta}} f_{ASI}(x; \tilde{\theta}) \right\|_{op} \leq \max \left[\|H(x, \theta)\|_{op}, \|H(x, \theta')\|_{op} \right].$$

Now recall that the antisymmetric initialization trick initializes $\theta_0 \sim N(0, I)$ then sets $\tilde{\theta}_0 = \left[\begin{smallmatrix} \theta_0 \\ \theta_0 \end{smallmatrix} \right]$. Furthermore note that if $\left\| \tilde{\theta} - \tilde{\theta_0} \right\|_2 \leq R$ then $\|\theta - \theta_0\|_2 \leq R$ and $\|\theta' - \theta_0\|_2 \leq R$. Thus

if θ_0 is an initialization such that the conclusion of Theorem 3.6.9 holds for the model $f(x;\theta)$ then the same conclusion holds for $f_{ASI}(x;\tilde{\theta})$ with initialization $\tilde{\theta}_0$.

3.6.2.2 Definition of the Convolution Operation

In this subsection we will formally define the convolution operation $*$ introduced in Section 3.2.3. We use the same convention for the convolution operation as [LZB20a]. A convolutional layer of the network has the form

$$\alpha^{(l)} = \psi_l(\theta^{(l)}, \alpha^{(l-1)}) = \omega\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} * \alpha^{(l-1)}\right).$$

Here $W^{(l)} \in \mathbb{R}^{K \times m_l \times m_{l-1}}$ is an order-3 tensor where K denotes the filter size, m_l is the number of output channels, and m_{l-1} is the number of input channels. The input $\alpha^{(l-1)} \in \mathbb{R}^{m_{l-1} \times Q}$ is a matrix with m_{l-1} rows as channels and Q columns as pixels. The output of the layer ψ_l is of size $\mathbb{R}^{m_l \times Q}$. From now on we will drop the superscripts and just denote $W = W^{(l)}$ and $\alpha = \alpha^{(l)}$. The convolution operation is defined as

$$(W * \alpha)_{i,q} = \sum_{k=1}^{K} \sum_{j=1}^{m_{l-1}} W_{k,i,j} \alpha_{j,q+k-\frac{K+1}{2}}.$$

This can be reformulated as follows. For each $k \in [K]$ define the matrices $W^{[k]} := W_{k,i,j}$ and $(\alpha^{[k]})_{j,q} := \alpha_{j,q+k-\frac{K+1}{2}}$. Then the convolution operation can be rewritten as

$$(W * \alpha) = \sum_{k=1}^{K} W^{[k]} \alpha^{[k]}.$$

Under this reformulation the convolutional layer can be rewritten as

$$\psi(W, \alpha) = \omega\left(\sum_{k=1}^{K} \frac{1}{\sqrt{m_{l-1}}} W^{[k]} \alpha^{[k]}\right).$$

By treating each $W^{[k]}$ as if it were a weight matrix in a fully connected layer, the convolutional layers can be treated similarly to fully connected layers. Thus when we refer to weight matrices in the context of a convolutional layer we are referring to the matrices $W^{[k]}$.

3.6.2.3 Technical Lemmas

This section will cover some miscellaneous technical lemmas that will be of significance later. The following lemma bounds the operator norm of the weight matrices at initialization.

Lemma 3.6.10. *Let $f(x;\theta)$ be a neural network of the form specified in Section 3.2.3. Assume $m \geq d$ and $\max_l \frac{m_l}{m} \leq A$. Then with probability at least $1 - C\exp(-cm)$ over the initialization θ_0 each weight matrix W_0 at initialization satisfies*

$$\frac{1}{\sqrt{m}} \|W_0\| \leq 2\sqrt{A} + 1.$$

The constant $C > 0$ depends on the architecture but is independent of the width m.

Proof. Fix a weight matrix $W \in \mathbb{R}^{m_l \times m_{l-1}}$ in the model. Following [Ver12, Corollary 5.35] we have with probability at least $1 - 2\exp(-t^2/2)$ over the initialization

$$\|W_0\|_{op} \leq \sqrt{m_l} + \sqrt{m_{l-1}} + t$$

and thus

$$\frac{1}{\sqrt{m}} \left\|W_0^{(l)}\right\|_{op} \leq \frac{\sqrt{m_l}}{\sqrt{m}} + \frac{\sqrt{m_{l-1}}}{\sqrt{m}} + \frac{t}{\sqrt{m}} \leq 2\sqrt{A} + \frac{t}{\sqrt{m}}.$$

Thus by setting $t = \sqrt{m}$ and taking the union bound over all weight matrices in the model (which depends on the architecture) we get the desired result. □

We now state for reference the following lemma which follows from the proof in [LZB20a].

Lemma 3.6.11. *Let $R \geq 1$ and let $f(x;\theta)$ be a neural network of the form specified in Section 3.2.3. If θ_0 is an initialization such that each weight matrix W_0 satisfies $\frac{1}{\sqrt{m}}\left\|W_0^{(l)}\right\|_2 = O(1)$ then*

$$\sup_{x \in X} \sup_{\theta \in \overline{B}(\theta_0, R)} \|\nabla_\theta f(x;\theta)\|_2 = O\left(\max\left\{1, \frac{R}{\sqrt{m}}\right\}^{O(L)}\right).$$

In particular if $\sqrt{m} \geq R$ then

$$\sup_{x \in X} \sup_{\theta \in \overline{B}(\theta_0, R)} \|\nabla_\theta f(x;\theta)\|_2 = O(1).$$

As a consequence of the previous lemma we get the following high probability bound on the gradients norm $\|\nabla_\theta f(x;\theta)\|_2$.

Lemma 3.6.12. *Let $R \geq 1$ and let $f(x;\theta)$ be a neural network of the form specified in Section 3.2.3. Assume that $m \geq d$, $\max_l \frac{m_l}{m} = O(1)$, and $\sqrt{m} \geq R$. Then with probability at least $1 - C\exp(-cm)$ over the initialization θ_0 we have that*

$$\sup_{x\in X} \sup_{\theta\in\overline{B}(\theta_0,R)} \|\nabla_\theta f(x;\theta)\|_2 = O(1).$$

The constant $C > 0$ depends on the architecture but is independent of the width m

Proof. This follows immediately from Lemma 3.6.10 and Lemma 3.6.11. □

The following lemma bounds the kernel deviations $K^\theta - K^{\theta_0}$ in terms of the network Hessian.

Lemma 3.6.13. *Let $S = \{z_1, \ldots, z_k\} \subset X$. Let $B = \sup_{x\in X} \sup_{\theta\in\overline{B}(\theta_0,R)} \|\nabla_\theta f(x;\theta)\|$ and let $H_{max} = \max_{z\in S} \sup_{\theta\in\overline{B}(\theta_0,R)} \|H(z,\theta)\|_{op}$. Then for $\theta \in \overline{B}(\theta_0, R)$*

$$\max_{i,j\in[k]} |K^\theta(z_i,z_j) - K^{\theta_0}(z_i,z_j)| \leq 2BH_{max}R.$$

Proof. We have that

$$\begin{aligned}
&|K^\theta(z_i,z_j) - K^{\theta_0}(z_i,z_j)| \\
&\leq \|\nabla_\theta f(z_i;\theta)\| \|\nabla_\theta f(z_j;\theta) - \nabla_\theta f(z_j;\theta_0)\| + \|\nabla_\theta f(z_i;\theta) - \nabla_\theta f(z_i;\theta_0)\| \|\nabla_\theta f(z_j;\theta_0)\| \\
&\leq 2BH_{max}R.
\end{aligned}$$

Here we have used the fact that

$$\begin{aligned}
\|\nabla_\theta f(z_i;\theta) - \nabla_\theta f(z_i;\theta_0)\|_2 &= \left\| \int_0^1 H(z_i, s\theta + (1-s)\theta_0)(\theta - \theta_0) ds \right\|_2 \\
&\leq \int_0^1 \|H(z_i, s\theta + (1-s)\theta_0)\|_{op} \|\theta - \theta_0\|_2 \leq H_{max}R.
\end{aligned}$$

□

The following lemma provides a trivial bound on $\|\theta_t - \theta_0\|_2$.

Lemma 3.6.14.

$$\|\theta_t - \theta_0\|_2 \leq \frac{\sqrt{t}}{\sqrt{2}} \|\hat{r}_0\|_{\mathbb{R}^n} \leq \frac{\sqrt{t}}{\sqrt{2}} \|f^*\|_{L^\infty(X,\rho)}.$$

Proof.

$$\|\theta_t - \theta_0\|_2 \leq \int_0^t \|\partial_s \theta_s\|_2 \, ds = \int_0^t \|\partial_\theta L(\theta_s)\|_2 \, ds \leq \sqrt{t} \left[\int_0^t \|\partial_\theta L(\theta_s)\|_2^2 \, ds \right]^{1/2}$$
$$= \sqrt{t} \left[\int_0^t -\partial_s L(\theta_s) ds \right]^{1/2} = \sqrt{t} \left[L(\theta_0) - L(\theta_t) \right]^{1/2} \leq \sqrt{t} \left[L(\theta_0) \right]^{1/2} = \frac{\sqrt{t}}{\sqrt{2}} \|\hat{r}_0\|_{\mathbb{R}^n}$$
$$\leq \frac{\sqrt{t}}{\sqrt{2}} \|f^*\|_{L^\infty(X,\rho)}$$

where the second inequality above follows from the Cauchy-Schwarz inequality and the final inequality follows from the fact that $\|\hat{r}_0\|_{\mathbb{R}^n} = \|y\|_{\mathbb{R}^n} \leq \|f^*\|_{L^\infty(X,\rho)}$ from the antisymmetric initialization. □

3.6.3 Convergence of the Operators

Throughout this section $K(x, x')$ will be a fixed continuous, symmetric, positive definite kernel. We will let $\kappa := \max_{x \in X} K(x, x)$. We note that since K is continuous and X is compact we have that $\kappa < \infty$. We will thus treat κ as a constant. We also note that since K is a kernel for any $x, x' \in X$ we have the inequality $K(x, x') \leq \sqrt{K(x, x)}\sqrt{K(x', x')} \leq \kappa$.

We will let $K^\theta(x, x') = \langle \nabla_\theta f(x; \theta), \nabla_\theta f(x'; \theta) \rangle$ denote the NTK for a specific parameter θ. In this section θ_0 will be treated as fixed. We will show that for fixed θ_0 we have bounds on $\|(T_K - T_n^s) r_s\|_{L^2(X,\rho)}$ that hold with high probability over the sampling of $S = (x_1, \ldots, x_n)$. By the Fubini-Tonelli theorem this suffices to get bounds that hold with high probability over the parameter initialization $\theta_0 \sim \mu$ and data sampling $S \sim \rho^{\otimes n}$ as long as one makes sure that the appropriate events are measureable on the product space. Fortunately, due to the continuity of $K^\theta(x, x')$ and $H(x, \theta)$ with respect to x, x' and θ we can avoid such issues and we thus will not address measureability line-by-line.

In this section we will bound $\|(T_K - T_n^s)r_s\|_{L^2(X,\rho)}$ for all s such that $\|\theta_s - \theta_0\|_2 \leq R$. This will be done by bounding $\|(T_K - T_n)r_s\|_{L^2(X,\rho)}$ and $\|(T_n - T_n^s)r_s\|_{L^2(X,\rho)}$ separately. At a high level $\|(T_n - T_n^s)r_s\|_{L^2(X,\rho)}$ will be small whenever $K_0 - K_s$ is small. On the other hand $\|(T_K - T_n)r_s\|_{L^2(X,\rho)}$ will be small whenever n is large enough relative to the complexity of the function class $\{f(x;\theta) : \theta \in \overline{B}(\theta_0, R)\}$. If $\sup_{\theta \in \overline{B}(\theta_0,R)} \|H(x,\theta)\|_2$ was uniformly small over x then the kernel deviations $K_0 - K_s$ would be bounded and the complexity of $\{f(x;\theta) : \theta \in \overline{B}(\theta_0, R)\}$ would be controlled by the complexity of the linearized model $f_{lin}(x;\theta) = \langle \nabla_\theta f(x;\theta_0), \theta - \theta_0 \rangle$. However, Theorem 3.6.9 only gives us the ability to bound $\|H(x,\theta)\|$ for finitely many values of x. For this reason we will need to do somewhat elaborate gymnastics using Rademacher complexity to form estimates that only require the evaluation of $\sup_{\theta \in \overline{B}(\theta_0,R)} \|H(x,\theta)\|$ over finitely many values of x.

Let $\mathcal{F}$ denote some family of real valued functions and let $S = (z_1, \ldots, z_k)$ be a finite point set. We define

$$\mathcal{F}_{|S} = \{(g(z_1), \ldots, g(z_k)) : g \in \mathcal{F}\}$$

to be the set of all vectors in $\mathbb{R}^k$ formed by restricting a function in $\mathcal{F}$ to the point set S. Now let $\epsilon \in \mathbb{R}^k$ be a vector with entries that are i.i.d. Rademacher random variables, i.e. $\epsilon_i \sim \text{Unif}\{+1, -1\}$. We define the (unnormalized) Rademacher complexity of $\mathcal{F}_{|S}$.

$$URad(\mathcal{F}_{|S}) := \mathbb{E}_\epsilon \sup_{v \in \mathcal{F}_{|S}} \langle v, \epsilon \rangle = \mathbb{E} \sup_{g \in \mathcal{F}} \sum_{i=1}^k \epsilon_i g(x_i).$$

We will use the following classic result, see e.g. [Tel21, Theorem 13.1]

Theorem 3.6.15. *Let $\mathcal{F}$ be given with $g(z) \in [a,b]$ a.s. for all $g \in \mathcal{F}$. Then with probability at least $1 - \delta$ over the sampling of $z_1, \ldots, z_n$*

$$\sup_{g \in \mathcal{F}} \left[\mathbb{E}[g(Z)] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right] \leq \frac{2}{n} URad(\mathcal{F}_{|S}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

We will also make use of the following lemma which is also classic, see e.g. [Tel21, Lemma 13.3]

Lemma 3.6.16. *Let $\ell : \mathbb{R}^n \to \mathbb{R}^n$ be a vector of univariate C-lipschitz functions. Then $URad((\ell \circ \mathcal{F})_{|S}) \leq C \cdot URad(\mathcal{F}_{|S})$.*

Using this we will now prove the following technical lemma. For the purpose of this lemma $x_1, \ldots, x_n$ will be treated as fixed and the randomness will be over a ghost sample $S' = (x'_1, \ldots, x'_n)$.

Lemma 3.6.17. *Let $R \geq 1$ and $B = \sup_{x \in X} \sup_{\theta \in \overline{B}(\theta_0, R)} \|\nabla_\theta f(x, \theta)\|_2$. Consider $x_1, \ldots, x_n \in X$ to be fixed. Then let*

$$\mathcal{F} = \{x \mapsto \frac{1}{n} \sum_{i=1}^{n} |K^\theta(x, x_i) - K^{\theta_0}(x, x_i)|^2 : \theta \in \overline{B}(\theta_0, R)\}.$$

Let $x'_1, \ldots, x'_n$ be sampled i.i.d. from ρ. Let $S = (x_1, \ldots, x_n)$ and $S' = \{x'_1, \ldots, x'_n\}$ and define

$$H_{max} := \max_{z \in S \cup S'} \sup_{\theta \in \overline{B}(\theta_0, R)} \|H(z, \theta)\|_{op}$$

Then with probability at least $1 - \delta$ over the sampling of $x'_1, \ldots, x'_n$ we have that every $g \in \mathcal{F}$ satisfies

$$\mathbb{E}_{x \sim \rho}[g(x)] \leq 12B^2 H_{max}^2 R^2 + 12B^4 \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. We note that for $\theta \in \overline{B}(\theta_0, R)$

$$|K^\theta(x, x_i) - K^{\theta_0}(x, x_i)|^2 \leq [|K^\theta(x, x_i)| + |K^{\theta_0}(x, x_i)|]^2 \leq [2B^2]^2 = 4B^4.$$

Therefore for all $g \in \mathcal{F}$ we have that $g(x) \in [0, 4B^4]$ a.s. Then by Theorem 3.6.15 we have with probability at least $1 - \delta$ over the sampling of $S' = \{x'_1, \ldots, x'_n\}$

$$\sup_{g \in \mathcal{F}} \left[\mathbb{E}_{x \sim \rho}[g(x)] - \frac{1}{n} \sum_{i=1}^{n} g(x'_i) \right] \leq \frac{2}{n} URad(\mathcal{F}_{|S'}) + 12B^4 \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Then we note that for any $z, z' \in S \cup S'$ we by Lemma 3.6.13 that $\theta \in \overline{B}(\theta_0, R)$ implies

$$|K^\theta(z, z') - K^{\theta_0}(z, z')| \leq 2BH_{max}R.$$

It follows that for any member of $\mathcal{F}_{|S\cup S'}$ is bounded in infinity norm by $4B^2H_{max}^2R^2$. Thus for any $g \in \mathcal{F}$ we have that

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i') \leq 4B^2H_{max}^2R^2$$

and

$$\frac{1}{n}URad(\mathcal{F}_{|S'}) \leq 4B^2H_{max}^2R^2.$$

Therefore for any $g \in \mathcal{F}$ we have that

$$\begin{aligned}\mathbb{E}_{x\sim\rho}[g(x)] &\leq \frac{1}{n}\sum_{i=1}^{n} g(x_i') + \frac{2}{n}URad(\mathcal{F}_{|S'}) + 12B^4\sqrt{\frac{\log(2/\delta)}{2n}} \\ &\leq 12B^2H_{max}^2R^2 + 12B^4\sqrt{\frac{\log(2/\delta)}{2n}}.\end{aligned}$$

□

Using the previous lemma we can now bound $\|(T_n - T_n^t)r_t\|_{L^2(X,\rho)}$.

Lemma 3.6.18. *Let $R \geq 1$ and $B = \sup_{x\in X}\sup_{\theta\in\overline{B}(\theta_0,R)}\|\nabla_\theta f(x,\theta)\|_2$. Let $S = (x_1,\ldots,x_n)$ and $S' = (x_1',\ldots,x_n')$ be two independent sequences of i.i.d. samples from ρ. Define*

$$H_{max} := \max_{z\in S\cup S'}\sup_{\theta\in\overline{B}(\theta_0,R)}\|H(z,\theta)\|_{op}.$$

Then with probability at least $1-\delta$ over the sampling of S and S' we have that for any θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$,

$$\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 2\|f^*\|_{L^\infty(X,\rho)}^2\left[\|K - K_0\|_{L^2(X^2,\rho\otimes\rho)}^2 + 12B^2H_{max}^2R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right)\right].$$

Proof. We note that

$$\begin{aligned}\left|(T_n - T_n^t)r_t(x)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}[K(x,x_i) - K_t(x,x_i)]r_t(x_i)\right| \\ \leq \|\hat{r}_t\|_{\mathbb{R}^n}\left[\frac{1}{n}\sum_{i=1}^{n}|K(x,x_i) - K_t(x,x_i)|^2\right]^{1/2} &\leq \|\hat{r}_0\|_{\mathbb{R}^n}\left[\frac{1}{n}\sum_{i=1}^{n}|K(x,x_i) - K_t(x,x_i)|^2\right]^{1/2}\end{aligned}$$

where we have used the property $\|\hat{r}_t\|_{\mathbb{R}^n} \leq \|\hat{r}_0\|_{\mathbb{R}^n}$ from gradient flow. Well from the inequality $(a+b)^2 \leq 2(a^2+b^2)$ we have that

$$\frac{1}{n}\sum_{i=1}^{n} |K(x,x_i) - K_t(x,x_i)|^2$$
$$\leq \frac{2}{n}\sum_{i=1}^{n} |K(x,x_i) - K_0(x,x_i)|^2 + \frac{2}{n}\sum_{i=1}^{n} |K_0(x,x_i) - K_t(x,x_i)|^2.$$

For conciseness let

$$h_1(x) := \frac{1}{n}\sum_{i=1}^{n} |K(x,x_i) - K_0(x,x_i)|^2$$
$$h_2^t(x) := \frac{1}{n}\sum_{i=1}^{n} |K_0(x,x_i) - K_t(x,x_i)|^2.$$

Then by the above we have that

$$\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 2\left\|\hat{r}_0\right\|_{\mathbb{R}^n}^2 \left[\mathbb{E}_{x\sim\rho}[h_1(x)] + \mathbb{E}_{x\sim\rho}[h_2^t(x)]\right].$$

Well we note that $|K(x,x')| \leq \kappa$ and $|K_0(x,x')| \leq B^2$ uniformly over x, x'. Now consider the random variables $Z_i := \|K(\bullet, x_i) - K_0(\bullet, x_i)\|_{L^2(X,\rho)}^2$ where the randomness is over the sampling of x_i. Then we have that $|Z_i| \leq [\kappa + B^2]^2$ a.s. Thus by Hoeffding's inequality we have that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}_{x_1\sim\rho}[Z_1] > s\right) \leq \exp\left(\frac{-ns^2}{2[\kappa + B^2]^4}\right).$$

Thus with probability at least $1-\delta$ over the sampling of $x_1, \ldots, x_n$

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \leq \mathbb{E}_{x_1\sim\rho}[Z_1] + \frac{\sqrt{2}[\kappa + B^2]^2\sqrt{\log(1/\delta)}}{\sqrt{n}}. \tag{3.7}$$

Now note that

$$\frac{1}{n}\sum_{i=1}^{n} Z_i = \mathbb{E}_{x\sim\rho}[h_1(x)] \qquad \mathbb{E}_{x_1\sim\rho}[Z_1] = \|K - K_0\|_{L^2(X^2,\rho\otimes\rho)}^2.$$

Thus whenever (3.7) holds we have that

$$\mathbb{E}_{x\sim\rho}[h_1(x)] \leq \|K - K_0\|_{L^2(X^2,\rho\otimes\rho)}^2 + \frac{\sqrt{2}[\kappa + B^2]^2\sqrt{\log(1/\delta)}}{\sqrt{n}}$$
$$= \|K - K_0\|_{L^2(X^2,\rho\otimes\rho)}^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right).$$

On the other hand we have by Lemma 3.6.17 for any fixed $x_1, \ldots, x_n$ that with probability $1-\delta$ over the sampling of $x'_1, \ldots, x'_n$ i.i.d. from ρ we have that for all $\theta \in \overline{B}(\theta_0, R)$

$$\mathbb{E}_{x\sim\rho}\left[\frac{1}{n}\sum_{i=1}^{n}|K^{\theta}(x,x_i)-K^{\theta_0}(x,x_i)|^2\right] \leq 12B^2H_{max}^2R^2 + 12B^4\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{3.8}$$

Whenever the above holds we have that for any θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$ we have that

$$\mathbb{E}_{x\sim\rho}[h_2^t(x)] \leq 12B^2H_{max}^2R^2 + 12B^4\sqrt{\frac{\log(2/\delta)}{2n}} = 12B^2H_{max}^2R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right).$$

Thus combining these together we have with probability at least $(1-\delta)^2 \geq 1-2\delta$ over the sampling of $x_1, \ldots, x_n, x'_1, \ldots, x'_n$ that Equations (3.7) and (3.8) hold simultaneously for all $\theta \in \overline{B}(\theta_0, R)$. In such a case we have that for all θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$ that

$$\mathbb{E}_{x\sim\rho}[h_1(x)] + \mathbb{E}_{x\sim\rho}[h_2^t(x)] \leq \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + 12B^2H_{max}^2R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right).$$

Well then

$$\begin{aligned}
\left\|(T_n - T_n^t)r_t\right\|^2_{L^2(X,\rho)} &\leq 2\|\hat{r}_0\|^2_{\mathbb{R}^n}\left[\mathbb{E}_{x\sim\rho}[h_1(x)] + \mathbb{E}_{x\sim\rho}[h_2^t(x)]\right] \\
&\leq 2\|\hat{r}_0\|^2_{\mathbb{R}^n}\left[\|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + 12B^2H_{max}^2R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right)\right] \\
&\leq 2\|f^*\|^2_{L^\infty(X,\rho)}\left[\|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + 12B^2H_{max}^2R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right)\right].
\end{aligned}$$

In the last line above we have used the fact that $\|\hat{r}_0\|_{\mathbb{R}^n} = \|y\|_{\mathbb{R}^n} \leq \|f^*\|_{L^\infty(X,\rho)}$ from the antisymmetric initialization. The desired result follows after replacing δ with $\delta/2$ in the previous argument. □

From Lemma 3.6.18 we get the following corollary.

Corollary 3.6.19. *Let $R \geq 1$, $B = \sup_{x\in X}\sup_{\theta\in\overline{B}(\theta_0,R)}\|\nabla_\theta f(x,\theta)\|_2$. Let $S = (x_1, \ldots, x_n)$ and $S' = (x'_1, \ldots, x'_n)$ be two independent sequences of i.i.d. samples from ρ. Define*

$$H_{max} := \max_{z\in S\cup S'}\sup_{\theta\in\overline{B}(\theta_0,R)}\|H(z,\theta)\|_{op}.$$

Then with probability at least $1-\delta$ over the sampling of S and S' we have that for any θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$

$$\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 2\left\|f^*\right\|_{L^\infty(X,\rho)}^2 \left\|K - K_0\right\|_{L^2(X^2,\rho\otimes\rho)}^2 + \epsilon$$

provided that $B = \tilde{O}(1)$, $H_{max} = \tilde{O}(\epsilon^{1/2}/R)$ and $n = \tilde{\Omega}(\epsilon^{-2})$.

Proof. We have by Lemma 3.6.18 with probability at least $1-\delta$ over the sampling of S, S'

$$\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 2\left\|f^*\right\|_{L^\infty(X,\rho)}^2 \left[\left\|K - K_0\right\|_{L^2(X^2,\rho\otimes\rho)}^2 + 12B^2H_{max}^2R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right)\right].$$

Thus if $B = \tilde{O}(1)$ then $H_{max} = \tilde{O}(\epsilon^{1/2}/R)$ and $n = \tilde{\Omega}(\epsilon^{-2})$ is sufficient to ensure that

$$\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 2\left\|f^*\right\|_{L^\infty(X,\rho)}^2 \left\|K - K_0\right\|_{L^2(X^2,\rho\otimes\rho)}^2 + \epsilon.$$

□

Now we will begin the work to bound $\|(T_K - T_n)r_s\|_{L^2(X,\rho)}$. The following technical lemma bounds the Rademacher complexity of the difference between the network $f(x;\theta)$ and the linearization $f_{lin}(x;\theta) = \langle \nabla_\theta f(x;\theta_0), \theta - \theta_0\rangle$ in terms of the Hessian norm for finitely many values $z \in X$.

Lemma 3.6.20. *Let $R \geq 1$, $\mathcal{F} = \{x \mapsto f(x;\theta) - f_{lin}(x;\theta) : \theta \in \overline{B}(\theta_0, R)\}$, $B = \sup_{x\in X}\sup_{\theta\in\overline{B}(\theta_0,R)}\|\nabla_\theta f(x;\theta)\|$, and let $S = (z_1 \ldots, z_n) \subset X$. Furthermore let*

$$H_{max} := \max_{z\in S}\sup_{\theta\in\overline{B}(\theta_0,R)}\left\|H(z,\theta)\right\|_{op}.$$

Then

$$\sup_{g\in\mathcal{F}}\|g\|_{L^\infty(X,\rho)} \leq 2BR$$

and

$$\sup_{g\in\mathcal{F}}\max_{z\in S}|g(z)| \leq \frac{1}{2}R^2H_{max}.$$

In particular

$$\frac{1}{n}URad((\mathcal{F}\cup -\mathcal{F})_{|S}) \leq \frac{1}{2}R^2H_{max}.$$

Proof. We note that

$$|f(x;\theta) - f_{lin}(x;\theta)| \leq |f(x;\theta)| + |f_{lin}(x;\theta)|.$$

Well then using the fact that $f(\bullet;\theta_0) = 0$ from the antisymmetric initialization we get

$$|f(x;\theta)| = |f(x;\theta) - f(x;\theta_0)| = \left| \int_0^1 \langle \nabla_\theta f(x;\theta s + (1-s)\theta_0), \theta - \theta_0 \rangle ds \right|$$
$$\leq \int_0^1 |\langle \nabla_\theta f(x;\theta s + (1-s)\theta_0), \theta - \theta_0 \rangle| \leq B \|\theta - \theta_0\| \leq BR.$$

On the other hand

$$|f_{lin}(x;\theta)| = |\langle \nabla_\theta f(x;\theta_0), \theta - \theta_0 \rangle| \leq \|\nabla_\theta f(x;\theta_0)\|_2 \|\theta - \theta_0\|_2 \leq BR.$$

Thus

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|f(\bullet;\theta) - f_{lin}(\bullet;\theta)\|_{L^\infty(X,\rho)} \leq 2BR$$

and the first conclusion follows. Furthermore by the Lagrange form of the remainder in Taylor's theorem we have for $z \in S$

$$|f(z;\theta) - f_{lin}(z;\theta)| = \left| (\theta - \theta_0)^T \frac{H(z,\xi)}{2} (\theta - \theta_0) \right| \leq \frac{1}{2} \|\theta - \theta_0\|_2^2 \|H(z,\xi)\|_{op}$$

where ξ is some point on the line between θ and θ_0. Thus if we set

$$H_{max} := \max_{z \in S} \sup_{\theta \in \overline{B}(\theta_0, R)} \|H(z,\theta)\|_{op}$$

we have that

$$|f(z;\theta) - f_{lin}(z;\theta)| \leq \frac{1}{2} R^2 H_{max}$$

for all $\theta \in \overline{B}(\theta_0, R)$. Therefore $\frac{1}{n} URad((\mathcal{F} \cup -\mathcal{F})_{|S}) \leq \frac{1}{2} R^2 H_{max}$ and the desired result follows. □

We now introduce another technical lemma that provides Rademacher complexity and L^∞ norm bounds for the linear model $x \mapsto \langle \nabla_\theta f(x;\theta_0), \theta \rangle$.

Lemma 3.6.21. *Let $R \geq 1$, $\mathcal{F} = \{x \mapsto \langle \nabla_\theta f(x;\theta_0), \theta\rangle : \|\theta\|_2 \leq 2R\}$. Let*

$$B = \sup_{x\in X} \sup_{\theta \in \overline{B}(\theta_0,R)} \|\nabla_\theta f(x;\theta)\|.$$

Then

$$\sup_{g\in\mathcal{F}} \|g\|_{L^\infty(X,\rho)} \leq 2BR$$

and

$$\frac{1}{n} URad(\mathcal{F}_{|S}) \leq \frac{2BR}{\sqrt{n}}.$$

Proof. By Cauchy-Schwarz

$$|\langle \nabla_\theta f(x;\theta_0), \theta\rangle| \leq 2BR$$

and thus $\|g\|_{L^\infty(X,\rho)} \leq 2BR$ for all $g \in \mathcal{F}$. Now let $\epsilon \in \mathbb{R}^n$ be a vector with i.i.d Rademacher entries $\epsilon_i \sim \text{Unif}\{+1,-1\}$. Then as was shown by [BM03, Lemma 22]

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{\theta\in\overline{B}(\theta_0,2R)} \sum_{i=1}^n \epsilon_i \langle \nabla_\theta f(x_i,\theta_0), \theta\rangle \right] &= 2R\mathbb{E}_\epsilon \left\| \sum_{i=1}^n \epsilon_i \nabla_\theta f(x_i;\theta_0) \right\|_2 \\
&\leq 2R \left[\mathbb{E}_\epsilon \left\| \sum_{i=1}^n \epsilon_i \nabla_\theta f(x_i;\theta_0) \right\|_2^2 \right]^{1/2} \\
&= 2R \left[\mathbb{E}_\epsilon \left[\sum_{1\leq i,j\leq n} \epsilon_i\epsilon_j \langle \nabla_\theta f(x_i;\theta_0), \nabla_\theta f(x_j;\theta_0)\rangle \right] \right]^{1/2} \\
&= 2R \sqrt{\sum_{i=1}^n K^{\theta_0}(x_i,x_i)} \\
&\leq 2RB\sqrt{n}.
\end{aligned}$$

where the first inequality above is an application of Jensen's inequality. The Rademacher complexity bound then follows from the bound above. □

The following lemma compares the $L^2(X,\rho)$ norm to that of its empirical counterpart $L^2(X,\widehat{\rho})$ for the function classes discussed in Lemmas 3.6.20 and 3.6.21.

Lemma 3.6.22. *Let $R \geq 1$, $\mathcal{F}_1 = \{x \mapsto f(x;\theta) - f_{lin}(x;\theta) : \theta \in \overline{B}(\theta_0, R)\}$, $\mathcal{F}_2 = \{x \mapsto \langle \nabla_\theta f(x;\theta_0), \theta\rangle : \|\theta\|_2 \leq 2R\}$, and $B = \sup_{x\in X} \sup_{\theta\in\overline{B}(\theta_0,R)} \|\nabla_\theta f(x;\theta)\|$. Then with probability at least $1-\delta$ over the sampling of $S = (x_1, \ldots, x_n)$*

$$\sup_{g\in\mathcal{F}_1\cup\mathcal{F}_2} \left| \|g\|^2_{L^2(X,\rho)} - \|g\|^2_{L^2(X,\widehat{\rho})} \right| \leq 4BR^3 H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right).$$

where $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical measure induced by $x_1, \ldots, x_n$ and

$$H_{max} := \max_{z\in S} \sup_{\theta\in\overline{B}(\theta_0,R)} \|H(z,\theta)\|_{op}.$$

Proof. Let $\mathcal{F} = \{|g|^2 : g \in \mathcal{F}_1 \cup \mathcal{F}_2\}$. Note that by Lemmas 3.6.20 and 3.6.21 we have that for $g \in \mathcal{F}_1 \cup \mathcal{F}_2$ that $\|g\|_{L^\infty(X,\rho)} \leq 2BR$. Thus every $g \in \mathcal{F}$ satisfies $g(x) \in [0, 4B^2R^2]$ a.s. Well then by Theorem 3.6.15 we have with probability at least $1-\delta$ over the sampling of $S = (x_1, \ldots, x_n)$ that

$$\sup_{g\in\mathcal{F}} \left[\mathbb{E}_{x\sim\rho}[g(x)] - \frac{1}{n}\sum_{i=1}^n g(x_i) \right] \leq \frac{2}{n} URad(\mathcal{F}_{|S}) + 12B^2R^2\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Well note that x^2 is $4BR$ Lipschitz on the interval $[-2BR, 2BR]$. Then by Lemma 3.6.16 we have that

$$URad(\mathcal{F}_{|S}) \leq 4BR \cdot URad((\mathcal{F}_1 \cup \mathcal{F}_2)_{|S}).$$

Well then we have that

$$URad((\mathcal{F}_1 \cup \mathcal{F}_2)_{|S}) \leq URad((\mathcal{F}_1 \cup -\mathcal{F}_1 \cup \mathcal{F}_2)_{|S}) \leq URad((\mathcal{F}_1 \cup -\mathcal{F}_1)_{|S}) + URad((\mathcal{F}_2)_{|S})$$

where we have used the property that if A, A' are vector classes such that $\sup_{u\in A}\langle \epsilon, u\rangle \geq 0$ and $\sup_{u\in A'}\langle \epsilon, u\rangle \geq 0$ for all $\epsilon \in \{1,-1\}^n$ then $URad(A \cup A') \leq URad(A) + URad(A')$. Well by Lemma 3.6.20 we have that

$$\frac{1}{n} URad((\mathcal{F}_1 \cup -\mathcal{F}_1)_{|S}) \leq \frac{1}{2} R^2 H_{max}.$$

On the other hand by Lemma 3.6.21 we have that

$$\frac{1}{n} URad((\mathcal{F}_2)_{|S}) \leq \frac{2BR}{\sqrt{n}}.$$

Therefore combining these two bounds we get that

$$\frac{1}{n}URad((\mathcal{F}_1 \cup \mathcal{F}_2)_{|S}) \leq \frac{1}{2}R^2 H_{max} + \frac{2BR}{\sqrt{n}}$$

and thus

$$\frac{1}{n}URad(\mathcal{F}_{|S}) \leq \frac{4BR}{n} \cdot URad((\mathcal{F}_1 \cup \mathcal{F}_2)_{|S}) \leq 4BR\left[\frac{1}{2}R^2 H_{max} + \frac{2BR}{\sqrt{n}}\right].$$

Therefore by putting everything together we have that

$$\begin{aligned}\sup_{g\in\mathcal{F}}\left[\mathbb{E}_{x\sim\rho}[g(x)] - \frac{1}{n}\sum_{i=1}^{n} g(x_i)\right] &\leq 8BR\left[\frac{1}{2}R^2 H_{max} + \frac{2BR}{\sqrt{n}}\right] + 12B^2R^2\sqrt{\frac{\log(2/\delta)}{2n}} \\ &= 4BR^3 H_{max} + \frac{16B^2R^2}{\sqrt{n}} + 12B^2R^2\sqrt{\frac{\log(2/\delta)}{2n}}.\end{aligned}$$

By repeating the same argument for the class $-\mathcal{F}$ and taking a union bound we have with probability at least $1 - 2\delta$ that

$$\sup_{g\in\mathcal{F}}\left|\mathbb{E}_{x\sim\rho}[g(x)] - \frac{1}{n}\sum_{i=1}^{n} g(x_i)\right| \leq 4BR^3 H_{max} + \frac{16B^2R^2}{\sqrt{n}} + 12B^2R^2\sqrt{\frac{\log(2/\delta)}{2n}}.$$

The above can be reinterpreted as

$$\begin{aligned}\sup_{g\in\mathcal{F}_1\cup\mathcal{F}_2}\left|\|g\|^2_{L^2(X,\rho)} - \|g\|^2_{L^2(X,\widehat{\rho})}\right| &\leq 4BR^3 H_{max} + \frac{16B^2R^2}{\sqrt{n}} + 12B^2R^2\sqrt{\frac{\log(2/\delta)}{2n}} \\ &= 4BR^3 H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right).\end{aligned}$$

The desired result then follows from replacing δ with $\delta/2$ in the previous argument. □

Now we are ready to provide a bound on the quantity $\|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)}$ for θ satisfying $\|\theta - \theta_0\|_2 \leq R$.

Lemma 3.6.23. *Let $R \geq 1$ and let B and H_{max} be defined as in Lemma 3.6.22. Let $\mathcal{C} = \{x \mapsto f_{lin}(x;\theta) - f^*(x) : \theta \in \overline{B}(\theta_0, R)\}$. Then there are quantities Γ and Φ such that*

$$\Gamma = \tilde{O}\left(\frac{BR\sqrt{\log(\mathcal{N}(\mathcal{C}, L^2(X,\rho), \epsilon))}}{\sqrt{n}}\right)$$

and

$$\Phi = 4BR^3H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right)$$

such that with probability at least $1-\delta$ *over the sampling of* $x_1, \ldots, x_n$

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)} \leq \Gamma + \kappa\left[\sqrt{R^4H_{max}^2 + 2\Phi} + \sqrt{4\epsilon^2 + 2\Phi}\right].$$

Proof. We will define $r_{lin}(x;\theta) = f_{lin}(x;\theta) - f^*(x)$. Well then we have that

$$\begin{aligned}&\|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)}\\ &\leq \|(T_K - T_n)r_{lin}(\bullet;\theta)\|_{L^2(X,\rho)} + \|(T_K - T_n)(f - f_{lin})(\bullet;\theta)\|_{L^2(X,\rho)}.\end{aligned}$$

Now let E be a proper ϵ-covering of $\mathcal{C} = \{r_{lin}(x;\theta) : \theta \in \overline{B}(\theta_0, R)\}$ with respect to $L^2(X,\rho)$. Furthermore assume E is of minimal cardinality so that $|E| = \mathcal{N}(\mathcal{C}, L^2(X,\rho), \epsilon)$. Then for any $r_{lin}(\bullet;\theta)$ we can choose $\hat{\theta} \in \overline{B}(\theta_0, R)$ so that $r_{lin}(\bullet;\hat{\theta}) \in E$ and

$$\left\|r_{lin}(\bullet;\theta) - r_{lin}(\bullet;\hat{\theta})\right\|_{L^2(X,\rho)} \leq \epsilon.$$

Well then

$$\begin{aligned}&\|(T_K - T_n)r_{lin}(\bullet;\theta)\|_{L^2(X,\rho)}\\ &\leq \left\|(T_K - T_n)r_{lin}(\bullet;\hat{\theta})\right\|_{L^2(X,\rho)} + \left\|(T_K - T_n)(r_{lin}(\bullet;\theta) - r_{lin}(\bullet;\hat{\theta}))\right\|_{L^2(X,\rho)}.\end{aligned}$$

We note that for any $r_{lin}(x;\theta) \in \mathcal{C}$ that

$$\begin{aligned}|r_{lin}(x;\theta)| &\leq |f_{lin}(x;\theta)| + |f^*(x)| = |\langle \nabla_\theta f(x;\theta_0), \theta - \theta_0\rangle| + |f^*(x)|\\ &\leq BR + \|f^*\|_{L^\infty(X,\rho)} =: S.\end{aligned}$$

To handle the term $\left\|(T_K - T_n)r_{lin}(\bullet;\hat{\theta})\right\|_{L^2(X,\rho)}$, for $g \in E$ we define the random variables $Z_i := g(x_i)K_{x_i} - \mathbb{E}_{x\sim\rho}[g(x)K_x]$ taking values in the separable Hilbert space $\mathcal{H}$ where $\mathcal{H}$ is the RKHS associated with K. We note that $(T_n - T_K)g$ is equal to $\frac{1}{n}\sum_{i=1}^n Z_i$. Well then note that $\|g(x)K_x\|_{\mathcal{H}} = |g(x)|\,\|K_x\|_{\mathcal{H}} \leq \|g\|_{L^\infty(X,\rho)}\sqrt{K(x,x)} \leq S\kappa^{1/2}$ a.s. Well then

$$\begin{aligned}\|Z_i\|_{\mathcal{H}} &\leq \|g(x_i)K_{x_i}\|_{\mathcal{H}} + \|\mathbb{E}_{x\sim\rho}[g(x)K_x]\|_{\mathcal{H}}\\ &\leq S\kappa^{1/2} + \mathbb{E}_{x\sim\rho}\|g(x)K_x\|_{\mathcal{H}} \leq 2S\kappa^{1/2}.\end{aligned}$$

Then using Hoeffding's inequality for random variables taking values in a separable Hilbert space (see [RBV10, Section 2.4]) we have

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\|_{\mathcal{H}} > s\right) \leq 2\exp\left(-ns^2/2[2S\kappa^{1/2}]^2\right).$$

Thus by the union bound and the fact that $\frac{1}{n}\sum_{i=1}^{n} Z_i = (T_n - T_K)g$ we have that

$$\mathbb{P}\left(\max_{g\in E}\|(T_n - T_K)g\|_{\mathcal{H}} > s\right) \leq 2|E|\exp\left(-ns^2/2[2S\kappa^{1/2}]^2\right).$$

By setting

$$s = \frac{2\sqrt{2}\cdot S\kappa^{1/2}\sqrt{\log\left(\frac{2|E|}{\delta}\right)}}{\sqrt{n}} = \tilde{O}\left(\frac{BR\sqrt{\log(\mathcal{N}(\mathcal{C}, L^2(X,\rho),\epsilon))}}{\sqrt{n}}\right)$$

we get that with probability at least $1-\delta$ over the sampling of $x_1,\ldots,x_n$

$$\max_{g\in E}\|(T_n - T_K)g\|_{\mathcal{H}} \leq s$$

and thus from the inequality $\|\bullet\|_{L^2(X,\rho)} \leq \sqrt{\sigma_1}\,\|\bullet\|_{\mathcal{H}}$ we get

$$\max_{g\in E}\|(T_n - T_K)g\|_{L^2(X,\rho)} \leq s\sqrt{\sigma_1} \leq s\sqrt{\kappa}. \tag{3.9}$$

On the other hand we must bound

$$\left\|(T_K - T_n)(r_{lin}(\bullet;\theta) - r_{lin}(\bullet;\hat{\theta}))\right\|_{L^2(X,\rho)}$$

and

$$\|(T_K - T_n)(f - f_{lin})(\bullet;\theta)\|_{L^2(X,\rho)}.$$

Well note since $K(\bullet,\bullet) \leq \kappa$ pointwise it follows by Cauchy-Schwarz that for any h

$$|T_K h(x)| = \left|\int K(x,s)h(s)d\rho(s)\right| \leq \kappa\,\|h\|_{L^2(X,\rho)}$$

and similarly

$$|T_n h(x)| = \left|\int K(x,s)h(s)d\widehat{\rho}(s)\right| \leq \kappa\,\|h\|_{L^2(X,\widehat{\rho})}.$$

Therefore

$$\|(T_K - T_n)h\|_{L^2(X,\rho)} \leq \|(T_K - T_n)h\|_{L^\infty(X,\rho)} \leq \|T_K h\|_{L^\infty(X,\rho)} + \|T_n h\|_{L^\infty(X,\rho)}$$
$$\leq \kappa[\|h\|_{L^2(X,\rho)} + \|h\|_{L^2(X,\widehat{\rho})}].$$

Thus we will bound $r_{lin}(\bullet;\theta) - r_{lin}(\bullet;\hat{\theta})$ and $(f - f_{lin})(\bullet;\theta)$ in $L^2(X,\rho)$ and $L^2(X,\widehat{\rho})$. Well since $\theta \in \overline{B}(\theta_0, R)$ we have that $(f - f_{lin})(\bullet;\theta) \in \mathcal{F}_1$ where $\mathcal{F}_1$ is defined as in Lemma 3.6.22. On the other hand we note that $r_{lin}(x;\theta) - r_{lin}(x;\hat{\theta}) = \langle \nabla_\theta f(x;\theta_0), \theta - \hat{\theta} \rangle$. Note that since $\theta, \hat{\theta} \in \overline{B}(\theta_0, R)$ we have that $\left\|\theta - \hat{\theta}\right\|_2 \leq 2R$. Thus $r_{lin}(\bullet;\theta) - r_{lin}(\bullet;\hat{\theta}) \in \mathcal{F}_2$ where $\mathcal{F}_2$ is defined as in Lemma 3.6.22. Thus by Lemma 3.6.22 separate from the randomness before we have with probability at least $1 - \delta$ over the sampling of $x_1, \ldots, x_n$

$$\sup_{g \in \mathcal{F}_1 \cup \mathcal{F}_2} \left| \|g\|^2_{L^2(X,\rho)} - \|g\|^2_{L^2(X,\widehat{\rho})} \right| \leq 4BR^3 H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right) := \Phi. \tag{3.10}$$

Well note that by Lemma 3.6.20 we have that for each $i \in [n]$

$$|f(x_i;\theta) - f_{lin}(x_i;\theta)| \leq \frac{1}{2}R^2 H_{max}$$

and consequently

$$\|f(\bullet;\theta) - f_{lin}(\bullet;\theta)\|_{L^2(X,\widehat{\rho})} \leq \frac{1}{2}R^2 H_{max}.$$

On the other hand we had by the selection of $\hat{\theta}$ that

$$\left\| r_{lin}(\bullet;\theta) - r_{lin}(\bullet;\hat{\theta}) \right\|_{L^2(X,\rho)} \leq \epsilon.$$

Now for conciseness let $h_1 = f(\bullet;\theta) - f_{lin}(\bullet;\theta)$ and $h_2 = r_{lin}(\bullet;\theta) - r_{lin}(\bullet;\hat{\theta})$. Then by (3.10) we have

$$\|h_1\|^2_{L^2(X,\rho)} \leq \|h_1\|^2_{L^2(X,\widehat{\rho})} + \Phi \leq \frac{1}{4}R^4 H^2_{max} + \Phi$$

and

$$\|h_2\|^2_{L^2(X,\widehat{\rho})} \leq \|h_2\|^2_{L^2(X,\rho)} + \Phi \leq \epsilon^2 + \Phi.$$

This implies

$$\|h_1\|^2_{L^2(X,\rho)} + \|h_1\|^2_{L^2(X,\widehat{\rho})} \leq \frac{1}{2}R^4 H^2_{max} + \Phi$$
$$\|h_2\|^2_{L^2(X,\rho)} + \|h_2\|^2_{L^2(X,\widehat{\rho})} \leq 2\epsilon^2 + \Phi.$$

Thus using the inequality $a+b \leq \sqrt{2}(a^2+b^2)^{1/2}$ for $a, b \geq 0$ combined with the previous estimates we have

$$\|h_1\|_{L^2(X,\rho)} + \|h_1\|_{L^2(X,\widehat{\rho})} \leq \sqrt{2}\sqrt{\frac{1}{2}R^4 H_{max}^2 + \Phi} = \sqrt{R^4 H_{max}^2 + 2\Phi}$$

and

$$\|h_2\|_{L^2(X,\rho)} + \|h_2\|_{L^2(X,\widehat{\rho})} \leq \sqrt{2}\sqrt{2\epsilon^2 + \Phi} = \sqrt{4\epsilon^2 + 2\Phi}.$$

Thus we have just shown that assuming (3.10) holds that

$$\|(T_K - T_n)h_1\|_{L^2(X,\rho)} \leq \kappa[\|h_1\|_{L^2(X,\rho)} + \|h_1\|_{L^2(X,\widehat{\rho})}] \leq \kappa\sqrt{R^4 H_{max}^2 + 2\Phi}$$

and

$$\|(T_K - T_n)h_2\|_{L^2(X,\rho)} \leq \kappa[\|h_2\|_{L^2(X,\rho)} + \|h_2\|_{L^2(X,\widehat{\rho})}] \leq \kappa\sqrt{4\epsilon^2 + 2\Phi}.$$

Then by taking a union bound we can assume with probability at least $1-2\delta$ that (3.9) and (3.10) hold simultaneously. In which case our previous estimates combine to give us the bound

$$\begin{aligned}
&\|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)} \\
\leq &\left\|(T_K - T_n)r_{lin}(\bullet;\hat{\theta})\right\|_{L^2(X,\rho)} + \|(T_K - T_n)h_1\|_{L^2(X,\rho)} + \|(T_K - T_n)h_2\|_{L^2(X,\rho)} \\
&\leq s\sqrt{\kappa} + \kappa\left[\sqrt{R^4 H_{max}^2 + 2\Phi} + \sqrt{4\epsilon^2 + 2\Phi}\right].
\end{aligned}$$

We now note that as long as (3.9) and (3.10) hold the same argument runs through for any $\theta \in \overline{B}(\theta_0, R)$. Thus with probability at least $1-2\delta$

$$\sup_{\theta \in \overline{B}(\theta_0,R)} \|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)} \leq s\sqrt{\kappa} + \kappa\left[\sqrt{R^4 H_{max}^2 + 2\Phi} + \sqrt{4\epsilon^2 + 2\Phi}\right].$$

The desired conclusion follows by setting $\Gamma = s\sqrt{\kappa}$ and replacing δ with $\delta/2$ in the previous argument. □

From Lemma 3.6.23 we get the following corollary.

Corollary 3.6.24. *Let $R \geq 1$ and*

$$B = \sup_{x \in X} \sup_{\theta \in \overline{B}(\theta_0, R)} \|\nabla_\theta f(x, \theta)\|_2 .$$

Then with probability at least $1 - \delta$ over the sampling of $x_1, \ldots, x_n$ we have that

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|(T_K - T_n) r(\bullet; \theta)\|^2_{L^2(X,\rho)} \leq \epsilon$$

provided that $B = \tilde{O}(1)$, $H_{max} = \tilde{O}(\epsilon / R^3)$ and $n = \tilde{\Omega}(R^4/\epsilon^2)$ where the expressions under the $\tilde{O}$ and $\tilde{\Omega}$ notation do not depend on the values $x_1, \ldots, x_n$.

Proof. After substituting $\epsilon^{1/2}$ for ϵ in Lemma 3.6.23 we have that with probability at least $1 - \delta$ over the sampling of $x_1, \ldots, x_n$

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|(T_K - T_n) r(\bullet; \theta)\|_{L^2(X,\rho)} \leq \Gamma + \kappa \left[\sqrt{R^4 H^2_{max} + 2\Phi} + \sqrt{4\epsilon + 2\Phi} \right]$$

where

$$\Gamma = \tilde{O}\left(\frac{BR \sqrt{\log(\mathcal{N}(\mathcal{C}, L^2(X, \rho), \epsilon^{1/2}))}}{\sqrt{n}} \right),$$

$$\Phi = 4BR^3 H_{max} + \tilde{O}\left(\frac{B^2 R^2}{\sqrt{n}} \right),$$

and

$$\mathcal{C} = \{x \mapsto f_{lin}(x; \theta) - f^*(x) : \theta \in \overline{B}(\theta_0, R)\}.$$

Now define

$$F := \int_X \nabla_\theta f(x; \theta_0) \nabla_\theta f(x; \theta_0)^T d\rho(x).$$

Since translation by a fixed function does not change the covering number we have by Corollary 3.6.8 that

$$\log \mathcal{N}(\mathcal{C}, L^2(X, \rho), \epsilon^{1/2}) = \tilde{O}\left(\tilde{p}\left(F^{1/2} \frac{3\epsilon^{1/2}}{4R} \right) \right) = \tilde{O}\left(\tilde{p}\left(F, \frac{9\epsilon}{16R^2} \right) \right).$$

Well using the fact that $\tilde{p}(A, \epsilon) \leq \frac{Tr(A)}{\epsilon}$ we have that

$$\tilde{p}\left(F, \frac{9\epsilon}{16R^2} \right) \leq \frac{16R^2 Tr(F)}{9\epsilon}.$$

Well we note that

$$Tr(F) = Tr(\mathbb{E}_{x\sim\rho}[\nabla_\theta f(x;\theta_0)\nabla_\theta f(x;\theta_0)^T]) = \mathbb{E}_{x\sim\rho} Tr(\nabla_\theta f(x;\theta_0)\nabla_\theta f(x;\theta_0)^T)$$
$$= \mathbb{E}_{x\sim\rho} \left\|\nabla_\theta f(x;\theta_0)\right\|^2 \leq B^2.$$

Therefore assuming $B = \tilde{O}(1)$ we have that

$$\Gamma = \tilde{O}\left(\frac{R\sqrt{\log \mathcal{N}(\mathcal{C}, L^2(X,\rho), \epsilon^{1/2})}}{\sqrt{n}}\right) = \tilde{O}\left(\frac{R^2}{\epsilon^{1/2}\sqrt{n}}\right).$$

Thus $n = \tilde{\Omega}(R^4/\epsilon^2)$ suffices to ensure that $\Gamma = O(\epsilon^{1/2})$. Now we must bound

$$\Phi = 4BR^3 H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right).$$

We note that whenever $B = \tilde{O}(1)$ we have that $H_{max} = \tilde{O}(\epsilon/R^3)$ and $n = \tilde{\Omega}(R^4/\epsilon^2)$ guarantees that $\Phi = O(\epsilon)$. Finally we have that $H_{max} = \tilde{O}(\epsilon/R^3) \subset \tilde{O}(\epsilon^{1/2}/R^2)$ suffices to ensure that $R^4 H_{max}^2 = O(\epsilon)$. Thus given all these conditions are met we have that

$$\Gamma + \kappa\left[\sqrt{R^4H_{max}^2 + 2\Phi} + \sqrt{4\epsilon + 2\Phi}\right] = O(\epsilon^{1/2}).$$

The desired result then follows from setting the constants under the $\tilde{O}$ and $\tilde{\Omega}$ notation appropriately. □

The following lemma combines the results in this section to get the ultimate bound on the operator deviations $T_K - T_n^t$.

Lemma 3.6.25. *Let $R \geq 1$ and $\epsilon \in (0, R)$. Let $S = (x_1, \ldots, x_n)$ and $S' = (x_1', \ldots, x_n')$ be two separate i.i.d. samples from ρ and denote*

$$H_{max} := \max_{z\in S\cup S'} \sup_{\theta\in\overline{B}(\theta_0,R)} \|H(z,\theta)\|_{op}$$
$$B := \sup_{x\in X} \sup_{\theta\in\overline{B}(\theta_0,R)} \|\nabla_\theta f(x,\theta)\|_2 .$$

Then with probability at least $1-\delta$ over the sampling of S, S' we have that for any t such that $\|\theta_t - \theta_0\|_2 \leq R$ that

$$\left\|(T_K - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 4\left\|f^*\right\|_{L^\infty(X,\rho)}^2 \left\|K - K_0\right\|_{L^2(X^2,\rho\otimes\rho)}^2 + \epsilon$$

provided that $B = \tilde{O}(1)$, $H_{max} = \tilde{O}(\epsilon/R^3)$ *and* $n = \tilde{\Omega}(R^4/\epsilon^2)$ *where the expressions under the* $\tilde{O}$ *and* $\tilde{\Omega}$ *notation do not depend on* S *and* S'.

Proof. We note that for θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$ that

$$\begin{aligned}\left\|(T_K - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 &\leq [\|(T_K - T_n)r_t\|_{L^2(X,\rho)} + \left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}]^2 \\ &\leq 2\left\|(T_K - T_n)r_t\right\|_{L^2(X,\rho)}^2 + 2\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \\ &\leq 2\sup_{\theta\in\overline{B}(\theta_0,R)}\|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)}^2 + 2\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2.\end{aligned}$$

Well by Corollary 3.6.24 we have with probability at least $1-\delta$ over the sampling of $x_1, \ldots, x_n$

$$\sup_{\theta\in\overline{B}(\theta_0,R)}\|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)}^2 \leq \epsilon$$

provided that $B = \tilde{O}(1)$, $H_{max} = \tilde{O}(\epsilon/R^3)$ and $n = \tilde{\Omega}(R^4/\epsilon^2)$. This result also does not depend in any way on S'. On the other hand by Corollary 3.6.19 separate from the randomness before we have with probability at least $1-\delta$ over the sampling of S and S' that for any θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$

$$\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 2\|f^*\|_{L^\infty(X,\rho)}^2\|K - K_0\|_{L^2(X^2,\rho\otimes\rho)}^2 + \epsilon.$$

provided that $B = \tilde{O}(1)$, $H_{max} = \tilde{O}(\epsilon^{1/2}/R)$ and $n = \tilde{\Omega}(\epsilon^{-2})$. The desired result then follows from taking a union bound and replacing δ with $\delta/2$ and ϵ with $\epsilon/4$. □

3.6.4 Main Result

3.6.4.1 Damped Deviations

In this subsection we will recall some definitions and results from [BM22a]. The main theorems in [BM22a] assume that the network architecture is shallow, however the results we recall in this section do not depend on the architecture. Let $K(x, x')$ be a continuous, symmetric, positive-definite kernel. Recall that K defines the integral operator

$$T_K g(x) := \int_X K(x, s)g(s)d\rho(s).$$

Then by Mercer's theorem

$$K(x,x') = \sum_{i=1}^{\infty} \sigma_i \phi_i(x)\phi_i(x')$$

where $\{\phi_i\}_i$ is an orthonormal basis of $L^2(X,\rho)$ and $\{\sigma_i\}_i$ is a nonincreasing sequence of positive values. Each ϕ_i is an eigenfunction of T_K with eigenvalue σ_i, i.e. $T_K\phi_i = \sigma_i\phi_i$. Let $x \mapsto g_s(x)$ be a $L^2(X,\rho)$ function for each $s \in [0,t]$. Assume $s \mapsto \langle \phi_i, g_s\rangle_\rho$ is measureable for each i and $\int_0^t \|g_s\|^2_{L^2(X,\rho)}\, ds < \infty$. Then we write

$$\int_0^t g_s ds$$

to denote the coordinate-wise integral, meaning that $\int_0^t g_s ds$ is the $L^2(X,\rho)$ function h such that

$$\langle h, \phi_i\rangle_\rho = \int_0^t \langle g_s, \phi_i\rangle_\rho ds.$$

With this definition in hand we now recall the following "Damped Deviations" lemma given by [BM22a, Lemma 2.4].

Lemma 3.6.26. *Let $K(x,x')$ be a continuous, symmetric, positive-definite kernel. Let $[T_K h](\bullet) = \int_X K(\bullet,s)h(s)d\rho(s)$ be the integral operator associated with K and let $[T_n^s h](\bullet) = \frac{1}{n}\sum_{i=1}^n K_s(\bullet,x_i)h(x_i)$ denote the operator associated with the time-dependent NTK K_s. Then*

$$r_t = \exp(-T_K t)r_0 + \int_0^t \exp(-T_K(t-s))(T_K - T_n^s)r_s ds,$$

where the equality is in the $L^2(X,\rho)$ sense.

Furthermore we have the following lemma [BM22a, Lemma C.8]

Lemma 3.6.27. *Let $K(x,x')$ be a continuous, symmetric, positive-definite kernel with associated operator $T_K h(\bullet) = \int_X K(\bullet,s)h(s)d\rho(s)$. Let $T_n^s h(\bullet) = \frac{1}{n}\sum_{i=1}^n K_s(\bullet,x_i)h(x_i)$ denote the operator associated with the time-dependent NTK. Then*

$$\|P_k(r_t - \exp(-T_K t)r_0)\|_{L^2(X,\rho)} \le \frac{1-\exp(-\sigma_k t)}{\sigma_k} \sup_{s\le t} \|(T_K - T_n^s)r_s\|_{L^2(X,\rho)}.$$

and

$$\|r_t - \exp(-T_K t)r_0\|_{L^2(X,\rho)} \le t \cdot \sup_{s\le t} \|(T_K - T_n^s)r_s\|_{L^2(X,\rho)}.$$

3.6.4.2 Proof of Theorem 3.3.5

We are now ready to prove the main result of this paper.

Theorem 3.3.5. *Let $T \geq 1, \epsilon > 0$. Let $K(x, x')$ be a fixed continuous, symmetric, positive definite kernel. For $k \in \mathbb{N}$ let $P_k : L^2(X, \rho) \to L^2(X, \rho)$ denote the orthogonal projection onto the span of the top k eigenfunctions of the operator T_K defined in Equation (3.2). Let $\sigma_k > 0$ denote the k-th eigenvalue of T_K. Then $m = \tilde{\Omega}(T^4/\epsilon^2)$ and $n = \tilde{\Omega}(T^2/\epsilon^2)$ suffices to ensure with probability at least $1 - O(mn)\exp(-\Omega(\log^2(m))$ over the parameter initilization θ_0 and the training samples $x_1, \ldots, x_n$ that for all $t \leq T$ and $k \in \mathbb{N}$*

$$\|P_k(r_t - \exp(-T_K t)r_0)\|^2_{L^2(X,\rho)} \leq \left[\frac{1 - \exp(-\sigma_k t)}{\sigma_k}\right]^2 \cdot \left[4\|f^*\|^2_\infty \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right]$$

and

$$\|r_t - \exp(-T_K t)r_0\|^2_{L^2(X,\rho)} \leq t^2 \cdot \left[4\|f^*\|^2_\infty \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right].$$

Proof. Let θ_0 be the parameter initialization and let $S = (x_1, \ldots, x_n)$ and $S' = (x'_1, \ldots, x'_n)$ be two i.i.d. samples from ρ. Furthermore let $1 \leq R \leq \sqrt{m}$. Let $E_1 \subset \mathbb{R}^p \times X^{2n}$ be the set of values (θ_0, S, S') so that the conclusion of Lemma 3.6.25 holds. Similarly let E_2 be the set of values (θ_0, S, S') satisfying

$$B := \max_{x\in X} \sup_{\theta\in\overline{B}(\theta_0,R)} \|\nabla_\theta f(x;\theta)\|_2 = O(1)$$

and

$$H_{max} := \max_{z\in S\cup S'} \sup_{\theta\in\overline{B}(\theta_0,R)} \|H(z,\theta)\|_{op} = \tilde{O}(\epsilon/R^3)$$

where the expression $O(1)$ above is the bound on B given by Lemma 3.6.12 and the expression $\tilde{O}(\epsilon/R^3)$ is precisely the condition on H_{max} in the conclusion of Lemma 3.6.25. By Lemma 3.6.25 for any fixed θ_0 we have that the conclusion holds with probability at least $1 - \delta$ over the sampling of S, S'. Thus for any θ_0 we have that

$$\mathbb{E}_{S,S'}[\mathbb{I}\{(\theta_0, S, S') \in E_1\}] \geq 1 - \delta.$$

It follows then by the Fubini-Tonelli theorem that

$$\mathbb{P}(E_1) = \mathbb{E}_{\theta_0}\mathbb{E}_{S,S'}[\mathbb{I}\{(\theta_0, S, S') \in E_1\}] \geq 1 - \delta.$$

On the other hand by Theorem 3.6.9 and Lemma 3.6.12 combined with a union bound we have that for any fixed S, S' then with probability at least $1 - 2Cmn\exp(-c\log^2(m)) - C\exp(-cm)$ that $H_{max} = \tilde{O}(R/\sqrt{m})$ and $B = O(1)$. Thus if $m = \tilde{\Omega}(R^8/\epsilon^2)$ we ensure that $H_{max} = \tilde{O}(\epsilon/R^3)$. Then by the same Fubini-Tonelli argument as before we get that

$$\mathbb{P}(E_2) = \mathbb{E}_{S,S'}\mathbb{E}_{\theta_0}\mathbb{I}\{(\theta_0, S, S') \in E_2\} \geq 1 - 2Cmn\exp(-c\log^2(m)) - C\exp(-cm).$$

Thus by taking a union bound we have with probability at least

$$1 - \delta - O(mn)\exp(-\Omega(\log^2(m))$$

that the events E_1 and E_2 both hold simultaneously. This holds for any δ so we may as well set $\delta = O(mn)\exp(-\Omega(\log^2(m)))$ and absorb it into the other term. Whenever E_1 and E_2 hold simultaneously we have by Lemma 3.6.25 that for any θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$

$$\left\|(T_K - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 4\left\|f^*\right\|_{L^\infty(X,\rho)}^2 \left\|K - K_0\right\|_{L^2(X^2,\rho\otimes\rho)}^2 + \epsilon. \tag{3.11}$$

Well by Lemma 3.6.14 we have that $\|\theta_t - \theta_0\| \leq \frac{\sqrt{t}}{\sqrt{2}}\|f^*\|_{L^\infty(X,\rho)}$. Thus for $t \leq \frac{2R^2}{\|f^*\|_{L^\infty(X,\rho)}^2}$ we have that $\|\theta_t - \theta_0\| \leq R$. Well then by Lemma 3.6.27 and the inequality (3.11) we have that

$$\begin{aligned}&\|P_k(r_t - \exp(-T_K t)r_0)\|_{L^2(X,\rho)}^2 \\ &\leq \left[\frac{1 - \exp(-\sigma_k t)}{\sigma_k}\right]^2 \cdot \left[4\left\|f^*\right\|_{L^\infty(X,\rho)}^2 \left\|K - K_0\right\|_{L^2(X^2,\rho\otimes\rho)}^2 + \epsilon\right]\end{aligned}$$

and

$$\|r_t - \exp(-T_K t)r_0\|_{L^2(X,\rho)}^2 \leq t^2 \cdot \left[4\left\|f^*\right\|_{L^\infty(X,\rho)}^2 \left\|K - K_0\right\|_{L^2(X^2,\rho\otimes\rho)}^2 + \epsilon\right].$$

The desired result then follows by setting $T = \frac{2R^2}{\|f^*\|_{L^\infty(X,\rho)}^2}$. □

3.6.5 Discussion of Assumption 3.3.6

We will discuss why it is reasonable to assume that $m = \tilde{\Omega}(\epsilon^{-2})$ suffices to ensure that $\|K_0 - K^\infty\|^2_{L^2(X\times X, \rho\otimes\rho)} \leq \epsilon$ holds with high probability over the initialization. We note that for fixed θ_0, K_0 and K^∞ are bounded and thus by Hoeffding's inequality we have that with high probability

$$\begin{aligned}&\|K_0 - K^\infty\|^2_{L^2(X\times X, \rho\otimes\rho)}\\ &\leq \frac{1}{N}\sum_{i=1}^N |K_0(x_i, x_i') - K^\infty(x_i, x_i')|^2 + \tilde{O}\left(\frac{\|K_0 - K^\infty\|^2_{L^\infty(X\times X, \rho\times\rho)}}{\sqrt{N}}\right),\end{aligned}$$

where $(x_1, x_1'), \ldots, (x_N, x_N')$ is an i.i.d. sample from $\rho \otimes \rho$. Furthermore we have by Lemma 3.6.12 that $\|K_0 - K^\infty\|^2_{L^\infty(X\times X, \rho\times\rho)} = \tilde{O}(1)$ with high probability over the initialization of θ_0. Thus if we set $N = \tilde{\Omega}(\epsilon^{-2})$ we have that Assumption 3.3.6 holds provided that

$$\frac{1}{N}\sum_{i=1}^N |K_0(x_i, x_i') - K^\infty(x_i, x_i')|^2 = O(\epsilon)$$

with high probability over the simultaneous sampling of θ_0 and $(x_1, x_1'), \ldots, (x_N, x_N')$.

It is been shown in many settings that the pointwise deviations satisfy

$$|K_0(x, x') - K^\infty(x, x')| = \tilde{O}(1/\sqrt{m})$$

with high probability over θ_0. The earliest was [DZP19] who demonstrate that for a shallow ReLU network for fixed x, x' we have with probability at least $1 - \delta$ over the initialization

$$|K_0(x, x') - K^\infty(x, x')| \leq O\left(\frac{\log(1/\delta)}{\sqrt{m}}\right).$$

Analyzing the portion of the Neural Tangent Kernel corresponding to the last hidden layer, [DLL19] get an analogous bound for deep fully-connected, ResNet, and convolutional networks with smooth activations. This is substantiated by the results of [HY20] for deep fully-connected networks with smooth activations. In their work they demonstrate that for a fixed training set $x_1, \ldots, x_n$

$$\max_{i,j} |K_0(x_i, x_j) - K^\infty(x_i, x_j)| = \tilde{O}(1/\sqrt{m})$$

with high probability over the initialization. In their result there are constants that depend on how well dispersed $x_1, \ldots, x_n$ are. [BM22a] demonstrated that for shallow fully-connected networks with smooth activations

$$\sup_{(x,x')\in X\times X} |K_0(x,x') - K^\infty(x,x')| = \tilde{O}(1/\sqrt{m})$$

with high probability over the initialization. For deep fully-connected ReLU networks [ADH19b] demonstrate that for fixed x, x' if $m = \Omega(L^6 \log(L/\delta)/\epsilon^4)$ then with probability at least $1-\delta$

$$|K_0(x,x') - K^\infty(x,x')| \leq (L+1)\epsilon.$$

In terms of the width m this translates to $|K_0(x,x') - K^\infty(x,x')| = \tilde{O}(1/m^{1/4})$ with high probability. This was improved in a recent work by [BGW21] that demonstrated that if $\mathcal{M}$ is a Riemannian submanifold of the unit sphere then with high probability over the initialization

$$\sup_{x,x'\in\mathcal{M}\times\mathcal{M}} |K_0(x,x') - K^\infty(x,x')| = \tilde{O}(1/\sqrt{m}).$$

Furthermore as stated by [BGW21] their analysis should be amenable to other architectures.

Now note that $\max_{i\in[N]} |K_0(x_i,x_i') - K^\infty(x_i,x_i')| = O(\epsilon^{1/2})$ suffices to ensure that

$$\frac{1}{N}\sum_{i=1}^{N} |K_0(x_i,x_i') - K^\infty(x_i,x_i')|^2 = O(\epsilon).$$

Based on the previous discussion, we expect that with high probability

$$\max_{i\in[N]} |K_0(x_i,x_i') - K^\infty(x_i,x_i')| = \tilde{O}(1/\sqrt{m}).$$

Thus if $m = \tilde{\Omega}(1/\epsilon^2)$ then we would have that $\max_{i\in[N]} |K_0(x_i,x_i') - K^\infty(x_i,x_i')| = \tilde{O}(\epsilon)$ which is stronger than what we need. In fact $\max_{i\in[N]} |K_0(x_i,x_i') - K^\infty(x_i,x_i')| = \tilde{O}(1/m^{1/4})$ is sufficient. For these reasons, we view Assumption 3.3.6 as quite reasonable. Nevertheless, we are not aware of an out-of-the box result that simultaneously addresses all the cases we consider and thus we must add this as an external assumption. However, if desired one can bypass Assumption 3.3.6 by citing the aforementioned results to get statements for the cases in which they apply to.

3.6.6 Experimental Details

Architecture and Parameterization The code to produce Figure 3.1 is available at `https://github.com/bbowman223/deepspec` The NTK Gram matrix

$$(G_0)_{i,j} := K^{\theta_0}(x_i, x_j) = \langle \nabla_\theta f(x_i; \theta_0), \nabla_\theta f(x_j; \theta_0) \rangle$$

was computed for two separate networks. The first network corresponds to LeNet-5 [LBB98] where the output is the logit corresponding to class 0. The second network is a feedforward network with one hidden layer with the Softplus activation $\omega(x) = \log(1 + \exp(x))$. For LeNet-5 we compute the NTK using PyTorch [PGM19] using the default PyTorch initialization and parameterization. For the shallow network we implement the network directly and use the Neural Tangent Kernel parameterization:

$$f(x; \theta) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \omega(\langle w_i, x \rangle + b_i) + b_0,$$

where there is an explicit $1/\sqrt{m}$ factor. All parameters for the shallow network are initialized as i.i.d. standard Gaussian random variables $N(0, 1)$.

Details of Computation For each network we compute the NTK Gram matrix G_0 for 10 separate pairs of (θ_0, S) where θ_0 is the parameter initialization and $S = (x_1, \ldots, x_n)$ is the data batch. Each line in the plots of Figure 3.1 corresponds to a different pair (θ_0, S). We simultaneously sample the parameter initialization θ_0 and a random batch of 2000 training samples $x_1, \ldots, x_{2000}$. We load the batches using "DataLoader" in PyTorch with the "shuffle" parameter set to True. This means the batches will be sampled sequentially from a random permutation of the training data and thus are sampled without replacement. We then compute the NTK Gram matrix $(G_0)_{i,j} := K^{\theta_0}(x_i, x_j) = \langle \nabla_\theta f(x_i; \theta_0), \nabla_\theta f(x_j; \theta_0) \rangle$. Once we compute G_0 we compute its spectrum and plot the first 1000 eigenvalues. Note that the number of eigenvalues that we plot is half the batch size. We observe that if one plots all n eigenvalues (the number of eigenvalues equals the number of samples) one gets a sharp drop

in log scale magnitude starting near the bottom 5-10% of eigenvalues. We observed this to occur even as one varies n. We suspect this is due to numerical errors and thus we only plot the first half of the spectrum.

Data The dataset used for LeNet-5 is MNIST [LBB98] and the dataset for the shallow model is CIFAR-10 [Kri09]. MNIST is made available through the Creative Commons Attribution-Share Alike 3.0 license. CIFAR-10 does not specify a license. Neither of these datasets have personally identifiable information nor offensive content.

Computational Resources and Runtime The experiments were run on a 2016 Macbook Pro with a 2.6 Ghz Quad-Core Intel Core i7 processor and 16GB of RAM. The experiment took less than an hour in wall-clock time.

Software Licenses and Attribution Our experiments were implemented in Python with the aid of the following software libraries/tools: PyTorch [PGM19], NumPy [HMW20], SciPy [VGO20], Matplotlib [Hun07], Jupyter Notebook [KRP16], IPython [PG07], and autograd-hacks `https://github.com/cybertronai/autograd-hacks`. PyTorch, Numpy, and SciPy are available under the BSD license. Jupyter and IPython are available under the new/modified BSD license. Matplotlib uses only BSD compatible code and is available under the PSF license. The code for autograd-hacks belongs to the public domain as specified by the public-domain-equivalent-license "Unlicense" `https://unlicense.org/`.

CHAPTER 4

Characterizing the Spectrum of the NTK via a Power Series Expansion

4.1 Introduction

Neural networks currently dominate modern artificial intelligence, however, despite their empirical success establishing a principled theoretical foundation for them remains an active challenge. The key difficulties are that neural networks induce nonconvex optimization objectives [SS89] and typically operate in an overparameterized regime which precludes classical statistical learning theory [AB02]. The persistent success of overparameterized models tuned via non-convex optimization suggests that the relationship between the parameterization, optimization, and generalization is more sophisticated than that which can be addressed using classical theory.

A recent breakthrough on understanding the success of overparameterized networks was established through the Neural Tangent Kernel (NTK) [JGH18]. In the infinite-width limit the optimization dynamics are described entirely by the NTK and the parameterization behaves like a linear model [LXS19]. In this regime explicit guarantees for the optimization and generalization can be obtained [DLL19, DZP19, ADH19a, ALS19a, ZCZ20]. While one must be judicious when extrapolating insights from the NTK to finite-width networks [LSP20], the NTK remains one of the most promising avenues for understanding deep learning on a principled basis.

The spectrum of the NTK is fundamental to both the optimization and generaliza-

tion of wide networks. In particular, bounding the smallest eigenvalue of the NTK Gram matrix is a staple technique for establishing convergence guarantees for the optimization [DLL19, DZP19, OS20]. Furthermore, the full spectrum of the NTK Gram matrix governs the dynamics of the empirical risk [ADH19b], and the eigenvalues of the associated integral operator characterize the dynamics of the generalization error outside the training set [BM22b, BM22a]. Moreover, the decay rate of the generalization error for Gaussian process regression using the NTK can be characterized by the decay rate of the spectrum [CD07, CLK21, JBM22].

The importance of the spectrum of the NTK has led to a variety of efforts to characterize its structure via random matrix theory and other tools [YS19, FW20]. There is a broader body of work studying the closely related Conjugate Kernel, Fisher Information Matrix, and Hessian [PLR16, PW17, PW18, LLC18, KAA20]. These results often require complex random matrix theory or operate in a regime where the input dimension is sent to infinity. By contrast, using a just a power series expansion we are able to characterize a variety of attributes of the spectrum for fixed input dimension and recover key results from prior work.

4.1.1 Contributions

In Theorem 4.3.2 we derive coefficients for the power series expansion of the NTK under unit variance initialization, see Assumption 4.3.1. Consequently we are able to derive insights into the NTK spectrum, notably concerning the outlier eigenvalues as well as the asymptotic decay.

- In Theorem 4.4.1 and Observation 4.4.2 we demonstrate that the largest eigenvalue $\lambda_1(\mathbf{K})$ of the NTK takes up an $\Omega(1)$ proportion of the trace and that there are $O(1)$ outlier eigenvalues of the same order as $\lambda_1(\mathbf{K})$.
- In Theorem 4.4.3 and Theorem 4.4.5 we show that the effective rank $Tr(\mathbf{K})/\lambda_1(\mathbf{K})$ of the NTK is upper bounded by a constant multiple of the effective rank $Tr(\mathbf{X}\mathbf{X}^T)/\lambda_1(\mathbf{X}\mathbf{X}^T)$

of the input data Gram matrix for both infinite and finite-width networks.

- In Theorem 4.4.6 and Theorem 4.4.8 we characterize the asymptotic behavior of the NTK spectrum for both uniform and nonuniform data distributions on the sphere.

4.1.2 Related Work

Neural Tangent Kernel (NTK): the NTK was introduced by [JGH18], who demonstrated that in the infinite-width limit neural network optimization is described via a kernel gradient descent. As a consequence, when the network is polynomially wide in the number of samples, global convergence guarantees for gradient descent can be obtained [DLL19, DZP19, ALS19a, ZG19, LXS19, ZCZ20, OS20, NM20, Ngu21]. Furthermore, the connection between infinite-width networks and Gaussian processes, which traces back to [Nea96], has been reinvigorated in light of the NTK. Recent investigations include [LBN18, dHR18, NXB19].

Analysis of NTK Spectrum: theoretical analysis of the NTK spectrum via random matrix theory was investigated by [YS19, FW20] in the high dimensional limit. [VY21] demonstrated that for ReLU networks the spectrum of the NTK integral operator asymptotically follows a power law, which is consistent with our results for the uniform data distribution. [BJK19] calculated the NTK spectrum for shallow ReLU networks under the uniform distribution, which was then expanded to the nonuniform case by [BGG20]. [GGJ22] analyzed the spectrum of the conjugate kernel and NTK for convolutional networks with ReLU activations whose pixels are uniformly distributed on the sphere. [GYK20, BB21, CX21] analyzed the reproducing kernel Hilbert spaces of the NTK for ReLU networks and the Laplace kernel via the decay rate of the spectrum of the kernel. In contrast to previous works, we are able to address the spectrum in the finite dimensional setting and characterize the impact of different activation functions on it.

Hermite Expansion: [DFS16] used Hermite expansion to the study the expressivity of the Conjugate Kernel. [SAD22] used this technique to demonstrate that any dot product

kernel can be realized by the NTK or Conjugate Kernel of a shallow, zero bias network. [OS20] use Hermite expansion to study the NTK and establish a quantitative bound on the smallest eigenvalue for shallow networks. This approach was incorporated by [NM20] to handle convergence for deep networks, with sharp bounds on the smallest NTK eigenvalue for deep ReLU networks provided by [NMM21]. The Hermite approach was utilized by [PSG20] to analyze the smallest NTK eigenvalue of shallow networks under various activations. Finally, in a concurrent work [HZL22] use Hermite expansions to develop a principled and efficient polynomial based approximation algorithm for the NTK and CNTK. In contrast to the aforementioned works, here we employ the Hermite expansion to characterize both the outlier and asymptotic portions of the spectrum for both shallow and deep networks under general activations.

4.2 Preliminaries

For our notation, lower case letters, e.g., x, y, denote scalars, lower case bold characters, e.g., $\mathbf{x}, \mathbf{y}$ are for vectors, and upper case bold characters, e.g., $\mathbf{X}, \mathbf{Y}$, are for matrices. For natural numbers $k_1, k_2 \in \mathbb{N}$ we let $[k_1] = \{1, \ldots, k_1\}$ and $[k_2, k_1] = \{k_2, \ldots, k_1\}$. If $k_2 > k_1$ then $[k_2, k_1]$ is the empty set. We use $\|\cdot\|_p$ to denote the p-norm of the matrix or vector in question and as default use $\|\cdot\|$ as the operator or 2-norm respectively. We use $\mathbf{1}_{m\times n} \in \mathbb{R}^{m\times n}$ to denote the matrix with all entries equal to one. We define $\delta_{p=c}$ to take the value 1 if $p = c$ and be zero otherwise. We will frequently overload scalar functions $\phi : \mathbb{R} \to \mathbb{R}$ by applying them elementwise to vectors and matrices. The entry in the ith row and jth column of a matrix we access using the notation $[\mathbf{X}]_{ij}$. The Hadamard or entrywise product of two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$ we denote $\mathbf{X} \odot \mathbf{Y}$ as is standard. The pth Hadamard power we denote $\mathbf{X}^{\odot p}$ and define it as the Hadamard product of $\mathbf{X}$ with itself p times,

$$\mathbf{X}^{\odot p} := \mathbf{X} \odot \mathbf{X} \odot \cdots \odot \mathbf{X}.$$

Given a Hermitian or symmetric matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$, we adopt the convention that $\lambda_i(\mathbf{X})$ denotes the ith largest eigenvalue,

$$\lambda_1(\mathbf{X}) \geq \lambda_2(\mathbf{X}) \geq \cdots \geq \lambda_n(\mathbf{X}).$$

Finally, for a square matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ we let $Tr(\mathbf{X}) = \sum_{i=1}^{n} [\mathbf{X}]_{ii}$ denote the trace.

4.2.1 Hermite Expansion

We say that a function $f \colon \mathbb{R} \to \mathbb{R}$ is square integrable with respect to the standard Gaussian measure $\gamma(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ if $\mathbb{E}_{X \sim \mathcal{N}(0,1)}[f(X)^2] < \infty$. We denote by $L^2(\mathbb{R}, \gamma)$ the space of all such functions. The normalized probabilist's Hermite polynomials are defined as

$$h_k(x) = \frac{(-1)^k e^{x^2/2}}{\sqrt{k!}} \frac{d^k}{dx^k} e^{-x^2/2}, \quad k = 0, 1, \ldots$$

and form a complete orthonormal basis in $L^2(\mathbb{R}, \gamma)$ [OD14, §11]. The Hermite expansion of a function $\phi \in L^2(\mathbb{R}, \gamma)$ is given by $\phi(x) = \sum_{k=0}^{\infty} \mu_k(\phi) h_k(x)$, where $\mu_k(\phi) = \mathbb{E}_{X \sim \mathcal{N}(0,1)}[\phi(X) h_k(X)]$ is the kth normalized probabilist's Hermite coefficient of ϕ.

4.2.2 NTK Parameterization

In what follows, for $n, d \in \mathbb{N}$ let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote a matrix which stores n points in $\mathbb{R}^d$ row-wise. Unless otherwise stated, we assume $d \leq n$ and denote the ith row of $\mathbf{X}_n$ as $\mathbf{x}_i$. In this work we consider fully-connected neural networks of the form $f^{(L+1)} \colon \mathbb{R}^d \to \mathbb{R}$ with $L \in \mathbb{N}$ hidden layers and a linear output layer. For a given input vector $\mathbf{x} \in \mathbb{R}^d$, the activation $f^{(l)}$ and preactivation $g^{(l)}$ at each layer $l \in [L+1]$ are defined via the following recurrence relations,

$$\begin{aligned}
&g^{(1)}(\mathbf{x}) = \gamma_w \mathbf{W}^{(1)} \mathbf{x} + \gamma_b \mathbf{b}^{(1)}, \; f^{(1)}(\mathbf{x}) = \phi\left(g^{(1)}(\mathbf{x})\right), \\
&g^{(l)}(\mathbf{x}) = \frac{\sigma_w}{\sqrt{m_{l-1}}} \mathbf{W}^{(l)} f^{(l-1)}(\mathbf{x}) + \sigma_b \mathbf{b}^{(l)}, \; f^{(l)}(\mathbf{x}) = \phi\left(g^{(l)}(\mathbf{x})\right), \; \forall l \in [2, L], \\
&g^{(L+1)}(\mathbf{x}) = \frac{\sigma_w}{\sqrt{m_L}} \mathbf{W}^{(L+1)} f^{(L)}(\mathbf{x}), \; f^{(L+1)}(\mathbf{x}) = g^{(L+1)}(\mathbf{x}).
\end{aligned} \tag{4.1}$$

The parameters $\mathbf{W}^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^{m_l}$ are the weight matrix and bias vector at the lth layer respectively, $m_0 = d$, $m_{L+1} = 1$, and $\phi\colon \mathbb{R} \to \mathbb{R}$ is the activation function applied elementwise. The variables $\gamma_w, \sigma_w \in \mathbb{R}_{>0}$ and $\gamma_b, \sigma_b \in \mathbb{R}_{\geq 0}$ correspond to weight and bias hyperparameters respectively. Let $\theta_l \in \mathbb{R}^p$ denote a vector storing the network parameters $(\mathbf{W}^{(h)}, \mathbf{b}^{(h)})_{h=1}^{l}$ up to and including the lth layer. The Neural Tangent Kernel [JGH18] $\tilde{\Theta}^{(l)}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ associated with $f^{(l)}$ at layer $l \in [L+1]$ is defined as

$$\tilde{\Theta}^{(l)}(\mathbf{x}, \mathbf{y}) := \langle \nabla_{\theta_l} f^{(l)}(\mathbf{x}), \nabla_{\theta_l} f^{(l)}(\mathbf{y}) \rangle. \tag{4.2}$$

We will mostly study the NTK under the following standard assumptions.

Assumption 4.2.1. *NTK initialization.*

1. *At initialization all network parameters are distributed as $\mathcal{N}(0,1)$ and are mutually independent.*
2. *The activation function satisfies $\phi \in L^2(\mathbb{R}, \gamma)$, is differentiable almost everywhere and its derivative, which we denote ϕ', also satisfies $\phi' \in L^2(\mathbb{R}, \gamma)$.*
3. *The widths are sent to infinity in sequence, $m_1 \to \infty, m_2 \to \infty, \ldots, m_L \to \infty$. We refer to this regime as the sequential infinite-width limit.*

Under Assumption 4.2.1, for any $l \in [L+1]$, $\tilde{\Theta}^{(l)}(\mathbf{x}, \mathbf{y})$ converges in probability to a deterministic limit $\Theta^{(l)}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ [JGH18] and the network behaves like a kernelized linear predictor during training; see, e.g., [ADH19b, LXS19, WGL20]. Given access to the rows $(\mathbf{x}_i)_{i=1}^n$ of $\mathbf{X}$ the NTK matrix at layer $l \in [L+1]$, which we denote $\mathbf{K}_l$, is the $n \times n$ matrix with entries defined as

$$[\mathbf{K}_l]_{ij} = \frac{1}{n}\Theta^{(l)}(\mathbf{x}_i, \mathbf{x}_j), \ \forall (i,j) \in [n] \times [n]. \tag{4.3}$$

4.3 Expressing the NTK as a Power Series

The following assumption allows us to study a power series for the NTK of deep network and with general activation functions. We remark that power series for the NTK of deep networks with positive homogeneous activation functions, namely ReLU, have been studied in prior works [HZL22, CX21, BB21, GGJ22]. We further remark that while these works focus on the asymptotics of the NTK spectrum we also study the large eigenvalues.

Assumption 4.3.1. *The hyperparameters of the network satisfy*

$$\gamma_w^2 + \gamma_b^2 = 1, \quad \sigma_w^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi(Z)^2] \leq 1, \quad \sigma_b^2 = 1 - \sigma_w^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi(Z)^2].$$

Furthermore, the data is normalized so that $\|\mathbf{x}_i\| = 1$ *for all* $i \in [n]$.

Recall under Assumption 4.2.1 that the preactivations of the network are centered Gaussian processes [Nea96, LBN18]. Assumption 4.3.1 ensures the preactivation of each neuron has unit variance and thus is reminiscent of the [LBO12], [GB10] and [HZR15] initializations, which are designed to avoid vanishing and exploding gradients. We refer the reader to Section 4.6.1.3 for a thorough discussion. Under Assumption 4.3.1 we will show it is possible to write the NTK not only as a dot-product kernel but also as an analytic power series on $[-1, 1]$ and derive expressions for the coefficients. In order to state this result recall, given a function $f \in L^2(\mathbb{R}, \gamma)$, that the pth normalized probabilist's Hermite coefficient of f is denoted $\mu_p(f)$, we refer the reader to Section 4.6.1.4 for an overview of the Hermite polynomials and their properties. Furthermore, letting $\bar{a} = (a_j)_{j=0}^{\infty}$ denote a sequence of real numbers, then for any $p, k \in \mathbb{Z}_{\geq 0}$ we define

$$F(p, k, \bar{a}) = \begin{cases} 1, & k = 0 \text{ and } p = 0, \\ 0, & k = 0 \text{ and } p \geq 1, \\ \sum_{(j_i)\in\mathcal{J}(p,k)} \prod_{i=1}^{k} a_{j_i}, & k \geq 1 \text{ and } p \geq 0, \end{cases} \tag{4.4}$$

where

$$\mathcal{J}(p,k) := \left\{(j_i)_{i\in[k]} \;:\; j_i \geq 0 \;\forall i \in [k],\; \sum_{i=1}^{k} j_i = p\right\} \quad \text{for all } p \in \mathbb{Z}_{\geq 0},\; k \in \mathbb{N}.$$

Here $\mathcal{J}(p,k)$ is the set of all k-tuples of nonnegative integers which sum to p and $F(p,k,\bar{a})$ is therefore the sum of all ordered products of k elements of $\bar{a}$ whose indices sum to p. We are now ready to state the key result of this section, Theorem 4.3.2, whose proof is provided in Section 4.6.2.1.

Theorem 4.3.2. *Under Assumptions 4.2.1 and 4.3.1, for all* $l \in [L+1]$

$$n\mathbf{K}_l = \sum_{p=0}^{\infty} \kappa_{p,l} \left(\mathbf{X}\mathbf{X}^T\right)^{\odot p}. \tag{4.5}$$

The series for each entry $n[\mathbf{K}_l]_{ij}$ *converges absolutely and the coefficients* $\kappa_{p,l}$ *are nonnegative and can be evaluated using the recurrence relationships*

$$\kappa_{p,l} = \begin{cases} \delta_{p=0}\gamma_b^2 + \delta_{p=1}\gamma_w^2, & l = 1, \\ \alpha_{p,l} + \sum_{q=0}^{p} \kappa_{q,l-1}\upsilon_{p-q,l}, & l \in [2, L+1], \end{cases} \tag{4.6}$$

where

$$\alpha_{p,l} = \begin{cases} \sigma_w^2 \mu_p^2(\phi) + \delta_{p=0}\sigma_b^2, & l = 2, \\ \sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{4.7}$$

and

$$\upsilon_{p,l} = \begin{cases} \sigma_w^2 \mu_p^2(\phi'), & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{4.8}$$

are likewise nonnegative for all $p \in \mathbb{Z}_{\geq 0}$ *and* $l \in [2, L+1]$.

As already remarked, power series for the NTK have been studied in previous works, however, to the best of our knowledge Theorem 4.3.2 is the first to explicitly express the coefficients at a layer in terms of the coefficients of previous layers. To compute the coefficients of the NTK as per Theorem 4.3.2, the Hermite coefficients of both ϕ and ϕ' are required.

Under Assumption 4.3.3 below, which has minimal impact on the generality of our results, this calculation can be simplified. In short, under Assumption 4.3.3 $\upsilon_{p,2} = (p+1)\alpha_{p+1,2}$ and therefore only the Hermite coefficients of ϕ are required. We refer the reader to Lemma 4.6.7 in Section 4.6.2.2 for further details.

Assumption 4.3.3. *The activation function $\phi\colon \mathbb{R} \to \mathbb{R}$ is absolutely continuous on $[-a, a]$ for all $a > 0$, differentiable almost everywhere, and is polynomially bounded, i.e., $|\phi(x)| = \mathcal{O}(|x|^{\beta})$ for some $\beta > 0$. Further, the derivative $\phi'\colon \mathbb{R} \to \mathbb{R}$ satisfies $\phi' \in L^2(\mathbb{R}, \gamma)$.*

We remark that ReLU, Tanh, Sigmoid, Softplus and many other commonly used activation functions satisfy Assumption 4.3.3. In order to understand the relationship between the Hermite coefficients of the activation function and the coefficients of the NTK, we first consider the simple two-layer case with $L = 1$ hidden layers. From Theorem 4.3.2

$$\kappa_{p,2} = \sigma_w^2(1 + \gamma_w^2 p)\mu_p^2(\phi) + \sigma_w^2\gamma_b^2(1+p)\mu_{p+1}^2(\phi) + \delta_{p=0}\sigma_b^2. \tag{4.9}$$

As per Table 4.1, a general trend we observe across all activation functions is that the first few coefficients account for the large majority of the total NTK coefficient series.

Table 4.1: **Dominance of the Early Coefficients** Percentage of $\sum_{p=0}^{\infty} \kappa_{p,2}$ accounted for by the first $T+1$ NTK coefficients assuming $\gamma_w^2 = 1$, $\gamma_b^2 = 0$, $\sigma_w^2 = 1$ and $\sigma_b^2 = 1 - \mathbb{E}[\phi(Z)^2]$.

$T =$	0	1	2	3	4	5
ReLU	43.944	77.277	93.192	93.192	95.403	95.403
Tanh	41.362	91.468	91.468	97.487	97.487	99.090
Sigmoid	91.557	99.729	99.729	99.977	99.977	99.997
Gaussian	95.834	95.834	98.729	98.729	99.634	99.634

However, the asymptotic rate of decay of the NTK coefficients varies significantly by activation function, due to the varying behavior of their tails. In Lemma 4.3.4 we choose ReLU, Tanh and Gaussian as prototypical examples of activations functions with growing,

constant, and decaying tails respectively, and analyze the corresponding NTK coefficients in the two layer setting. For typographical ease we denote the zero mean Gaussian density function with variance σ^2 as $\omega_\sigma(z) := (1/\sqrt{2\pi\sigma^2})\exp(-z^2/(2\sigma^2))$.

Lemma 4.3.4. *Under Assumptions 4.2.1 and 4.3.1,*

1. *if* $\phi(z) = ReLU(z)$, *then* $\kappa_{p,2} = \delta_{(\gamma_b>0)\cup(p\ even)}\Theta(p^{-3/2})$,
2. *if* $\phi(z) = Tanh(z)$, *then* $\kappa_{p,2} = \mathcal{O}\left(\exp\left(-\frac{\pi\sqrt{p-1}}{2}\right)\right)$,
3. *if* $\phi(z) = \omega_\sigma(z)$, *then* $\kappa_{p,2} = \delta_{(\gamma_b>0)\cup(p\ even)}\Theta(p^{1/2}(\sigma^2+1)^{-p})$.

The trend we observe from Lemma 4.3.4 is that activation functions whose Hermite coefficients decay quickly, such as ω_σ, result in a faster decay of the NTK coefficients. We remark that analyzing the rates of decay in the deep setting is challenging due to the calculation of $F(p,k,\bar{\alpha}_{l-1})$ (4.4) and therefore leave this study to future work.

Finally, we briefly pause here to highlight the potential for using a truncation of (4.5) in order to perform efficient numerical approximation of the infinite-width NTK. We remark that this idea is also addressed in a concurrent work by [HZL22], albeit under a somewhat different set of assumptions [1]. As per our observations thus far that the coefficients of the NTK power series (4.5) typically decay quite rapidly, one might consider approximating $\Theta^{(l)}$ by computing just the first few terms in each series of (4.5). Furthermore Figure 4.2 in Section 4.6.2.3 shows the absolute error between the truncated ReLU NTK and the analytical expression for the ReLU NTK, which is also defined in Section 4.6.2.3. Let ρ denote the input correlation. The key takeaway is that while for $|\rho|$ close to one the approximation is poor, for $|\rho| < 0.5$, which is arguably more realistic for real-world data, with just 50 coefficients machine level precision can be achieved. We refer the interested reader to Section 4.6.2.3 for a proper discussion.

[1] In particular, in [HZL22] the authors focus on homogeneous activation functions and allow the data to lie off the sphere. By contrast, we require the data to lie on the sphere but can handle non-homogeneous activation functions in the deep setting.

4.4 Analyzing the Spectrum of the NTK via its Power Series

In this section, we consider a general kernel matrix power series of the form

$$n\mathbf{K} = \sum_{p=0}^{\infty} c_p(\mathbf{X}\mathbf{X}^T)^{\odot p}$$

where $\{c_p\}_{p=0}^{\infty}$ are coefficients and $\mathbf{X}$ is the data matrix. According to Theorem 4.3.2, the coefficients of the NTK power series (4.5) are always nonnegative, thus we only consider the case where c_p are nonnegative. We will also consider the kernel function power series, which we denote as $K(x_1, x_2) = \sum_{p=0}^{\infty} c_p \langle x_1, x_2 \rangle^p$. Later on we will analyze the spectrum of kernel matrix $\mathbf{K}$ and kernel function K.

4.4.1 Analysis of the Upper Spectrum and Effective Rank

In this section we analyze the upper part of the spectrum of the NTK, corresponding to the large eigenvalues, using the power series given in Theorem 4.3.2. Our first result concerns the *effective rank* [HHV22] of the NTK. Given a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ we define the effective rank of $\mathbf{A}$ to be

$$\text{eff}(\mathbf{A}) = \frac{Tr(\mathbf{A})}{\lambda_1(\mathbf{A})}.$$

The effective rank quantifies how many eigenvalues are on the order of the largest eigenvalue. This follows from the Markov-like inequality

$$|\{p : \lambda_p(\mathbf{A}) \geq c\lambda_1(\mathbf{A})\}| \leq c^{-1}\text{eff}(\mathbf{A}) \tag{4.10}$$

and the eigenvalue bound

$$\frac{\lambda_p(\mathbf{A})}{\lambda_1(\mathbf{A})} \leq \frac{\text{eff}(\mathbf{A})}{p}.$$

Our first result is that the effective rank of the NTK can be bounded in terms of a ratio involving the power series coefficients. As we are assuming the data is normalized so that

$\|\mathbf{x}_i\| = 1$ for all $i \in [n]$, then observe by the linearity of the trace

$$Tr(n\mathbf{K}) = \sum_{p=0}^{\infty} c_p Tr((\mathbf{X}\mathbf{X}^T)^{\odot p}) = n\sum_{p=0}^{\infty} c_p,$$

where we have used the fact that $Tr((\mathbf{X}\mathbf{X}^T)^{\odot p}) = n$ for all $p \in \mathbb{N}$. On the other hand,

$$\lambda_1(n\mathbf{K}) \geq \lambda_1(c_0(\mathbf{X}\mathbf{X}^T)^0) = \lambda_1(c_0\mathbf{1}_{n\times n}) = nc_0.$$

Combining these two results we get the following theorem.

Theorem 4.4.1. *Assume that we have a kernel Gram matrix* $\mathbf{K}$ *of the form*

$$n\mathbf{K} = \sum_{p=0}^{\infty} c_p(\mathbf{X}\mathbf{X}^T)^{\odot p}$$

where $c_0 \neq 0$*. Furthermore, assume the input data* $\mathbf{x}_i$ *are normalized so that* $\|\mathbf{x}_i\| = 1$ *for all* $i \in [n]$*. Then*

$$\text{eff}(\mathbf{K}) \leq \frac{\sum_{p=0}^{\infty} c_p}{c_0}.$$

By Theorem 4.3.2 $c_0 \neq 0$ provided the network has biases or the activation function has nonzero Gaussian expectation (i.e., $\mu_0(\phi) \neq 0$). Thus we have that the effective rank of $\mathbf{K}$ is bounded by an $O(1)$ quantity. In the case of ReLU for example, as evidenced by Table 4.1, the effective rank will be roughly 2.3 for a shallow network. By contrast, a well-conditioned matrix would have an effective rank that is $\Omega(n)$. Combining Theorem 4.4.1 and the Markov-type bound (4.10) we make the following important observation.

Observation 4.4.2. *The largest eigenvalue* $\lambda_1(\mathbf{K})$ *of the NTK takes up an* $\Omega(1)$ *fraction of the entire trace and there are* $O(1)$ *eigenvalues on the same order of magnitude as* $\lambda_1(\mathbf{K})$*, where the* $O(1)$ *and* $\Omega(1)$ *notation are with respect to the parameter* n*.*

While the constant term $c_0\mathbf{1}_{n\times n}$ in the kernel leads to a significant outlier in the spectrum of $\mathbf{K}$, it is rather uninformative beyond this. What interests us is how the structure of the data $\mathbf{X}$ manifests in the spectrum of the kernel matrix $\mathbf{K}$. For this reason we will examine the centered kernel matrix $\widetilde{\mathbf{K}} := \mathbf{K} - \frac{c_0}{n}\mathbf{1}_{n\times n}$. By a very similar argument as before we get the following result.

Theorem 4.4.3. *Assume that we have a kernel Gram matrix* $\mathbf{K}$ *of the form*

$$n\mathbf{K} = \sum_{p=0}^{\infty} c_p (\mathbf{X}\mathbf{X}^T)^{\odot p}$$

where $c_1 \neq 0$*. Furthermore, assume the input data* $\mathbf{x}_i$ *are normalized so that* $\|\mathbf{x}_i\| = 1$ *for all* $i \in [n]$*. Then the centered kernel* $\widetilde{\mathbf{K}} := \mathbf{K} - \frac{c_0}{n}\mathbf{1}_{n\times n}$ *satisfies*

$$\text{eff}(\widetilde{\mathbf{K}}) \leq \text{eff}(\mathbf{X}\mathbf{X}^T)\frac{\sum_{p=1}^{\infty} c_p}{c_1}.$$

Thus we have that the effective rank of the centered kernel $\widetilde{\mathbf{K}}$ is upper bounded by a constant multiple of the effective rank of the input data Gram $\mathbf{X}\mathbf{X}^T$. Furthermore, we can take the ratio $\frac{\sum_{p=1}^{\infty} c_p}{c_1}$ as a measure of how much the NTK inherits the behavior of the linear kernel $\mathbf{X}\mathbf{X}^T$: in particular, if the input data gram has low effective rank and this ratio is moderate then we may conclude that the centered NTK must also have low effective rank. Again from Table 4.1, in the shallow setting we see that this ratio tends to be small for many of the common activations, for example, for ReLU it is roughly 1.3. To summarize then from Theorem 4.4.3 we make the important observation.

Observation 4.4.4. *Whenever the input data are approximately low rank, the centered kernel matrix* $\widetilde{\mathbf{K}} = \mathbf{K} - \frac{c_0}{n}\mathbf{1}_{n\times n}$ *is also approximately low rank.*

It turns out that this phenomenon also holds for finite-width networks at initialization. Consider the shallow model

$$\sum_{\ell=1}^{m} a_\ell \phi(\langle \mathbf{w}_\ell, \mathbf{x}\rangle),$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{w}_\ell \in \mathbb{R}^d$, $a_\ell \in \mathbb{R}$ for all $\ell \in [m]$. The following theorem demonstrates that when the width m is linear in the number of samples n then $\text{eff}(\mathbf{K})$ is upper bounded by a constant multiple of $\text{eff}(\mathbf{X}\mathbf{X}^T)$.

Theorem 4.4.5. *Assume* $\phi(x) = ReLU(x)$ *and* $n \geq d$*. Fix* $\epsilon > 0$ *small. Suppose that* $\mathbf{w}_1, \ldots, \mathbf{w}_m \sim N(0, \nu_1^2 I_d)$ *i.i.d. and* $a_1, \ldots, a_m \sim N(0, \nu_2^2)$*. Set* $M = \max_{i\in[n]} \|\mathbf{x}_i\|_2$*, and let*

$$\Sigma := \mathbb{E}_{\mathbf{w}\sim N(0,\nu_1^2 I)}[\phi(\mathbf{X}\mathbf{w})\phi(\mathbf{w}^T\mathbf{X}^T)].$$

Then

$$m = \Omega\left(\max(\lambda_1(\Sigma)^{-2}, 1)\max(n, \log(1/\epsilon))\right), \quad \nu_1 = O(1/M\sqrt{m})$$

suffices to ensure that, with probability at least $1 - \epsilon$ *over the sampling of the parameter initialization,*

$$\text{eff}(\mathbf{K}) \leq C \cdot \text{eff}(\mathbf{X}\mathbf{X}^T),$$

where $C > 0$ *is an absolute constant.*

Many works consider the model where the outer layer weights are fixed and have constant magnitude and only the inner layer weights are trained. This is the setting considered by [XLS17], [ADH19a], [DZP19], [OFL19], [LSO20], and [OS20]. In this setting we can reduce the dependence on the width m to only be logarithmic in the number of samples n, and we have an accompanying lower bound. See Theorem 4.6.13 in the Section 4.6.4.3 for details.

In Figure 4.1 we empirically validate our theory by computing the spectrum of the NTK on both Caltech101 [LAR22] and isotropic Gaussian data for feedforward networks. We use the `functorch`[2] module in PyTorch [PGM19] using an algorithmic approach inspired by [NSS22]. As per Theorem 4.1 and Observation 4.2, we observe all network architectures exhibit a dominant outlier eigenvalue due to the nonzero constant coefficient in the power series. Furthermore, this dominant outlier becomes more pronounced with depth, as can be observed if one carries out the calculations described in Theorem 4.3.2. Additionally, this outlier is most pronounced for ReLU, as the combination of its Gaussian mean plus bias term is the largest out of the activations considered here. As predicted by Theorem 4.3, Observation 4.4 and Theorem 4.5, we observe real-world data, which has a skewed spectrum and hence a low effective rank, results in the spectrum of the NTK being skewed. By contrast, isotropic Gaussian data has a flat spectrum, and as a result beyond the outlier the decay of eigenvalues of the NTK is more gradual. These observations support the claim that the NTK inherits its spectral structure from the data. We also observe that the spectrum

[2] `https://pytorch.org/functorch/stable/notebooks/neural_tangent_kernels.html`

for Tanh is closer to the linear activation relative to ReLU: intuitively this should not be surprising as close to the origin Tanh is well approximated by the identity. Our theory provides a formal explanation for this observation, indeed, the power series coefficients for Tanh networks decay quickly relative to ReLU. We provide further experimental results in Section 4.6.5, including for CNNs where we observe the same trends. We note that the effective rank has implications for the generalization error. The Rademacher complexity of a kernel method (and hence the NTK model) within a parameter ball is determined by its its trace [BM03]. Since for the NTK $\lambda_1(\mathbf{K}) = O(1)$, lower effective rank implies smaller trace and hence limited complexity.

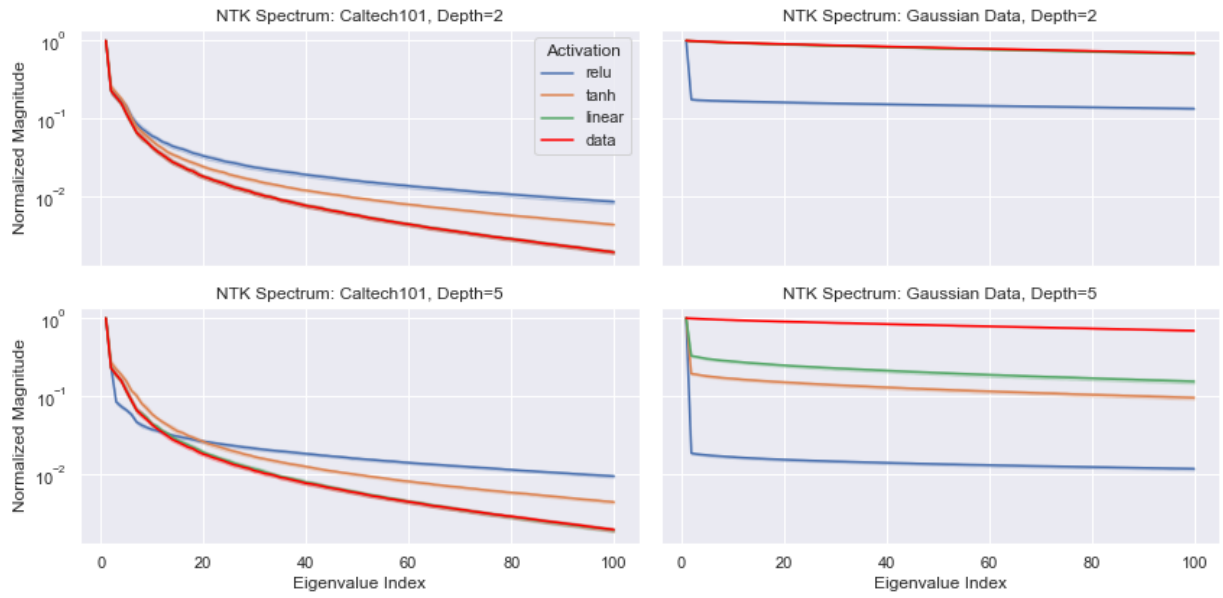

Figure 4.1: **Feedforward NTK Spectrum** We plot the normalized eigenvalues λ_p/λ_1 of the NTK Gram matrix $\mathbf{K}$ and the data Gram matrix $\mathbf{X}\mathbf{X}^T$ for Caltech101 and isotropic Gaussian datasets. To compute the NTK we randomly initialize feedforward networks of depths 2 and 5 with width 500. We use the standard parameterization and Pytorch's default Kaiming uniform initialization in order to better connect our results with what is used in practice. We consider a batch size of $n = 200$ and plot the first 100 eigenvalues. The thick part of each curve corresponds to the mean across 10 trials, while the transparent part corresponds to the 95% confidence interval

4.4.2 Analysis of the Lower Spectrum

In this section, we analyze the lower part of the spectrum using the power series. We first analyze the kernel function K which we recall is a dot-product kernel of the form $K(x_1, x_2) = \sum_{p=0}^{\infty} c_p \langle x_1, x_2 \rangle^p$. Assuming the training data is uniformly distributed on a hypersphere it was shown by [BJK19, BM19] that the eigenfunctions of K are the spherical harmonics. [AM15] gave the eigenvalues of the kernel K in terms of the power series coefficients.

Theorem 4.4.6. *[AM15] Suppose that the training data are uniformly sampled from the unit hypersphere $\mathbb{S}^d$, $d \geq 2$. If the dot-product kernel function has the expansion $K(x_1, x_2) = \sum_{p=0}^{\infty} c_p \langle x_1, x_2 \rangle^p$ where $c_p \geq 0$, then the eigenvalue of every spherical harmonic of frequency k is given by*

$$\overline{\lambda_k} = \frac{\pi^{d/2}}{2^{k-1}} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} c_p \frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)},$$

where Γ is the gamma function.

A proof of Theorem 4.4.6 is provided in Section 4.6.6 for the reader's convenience. This theorem connects the coefficients c_p of the kernel power series with the eigenvalues $\overline{\lambda_k}$ of the kernel. In particular, given a specific decay rate for the coefficients c_p one may derive the decay rate of $\overline{\lambda_k}$: for example, [SH21] examined the decay rate of $\overline{\lambda_k}$ if c_p admits a polynomial decay or exponential decay. The following Corollary summarizes the decay rates of $\overline{\lambda_k}$ corresponding to two layer networks with different activations.

Corollary 4.4.7. *Under the same setting as in Theorem 4.4.6,*

1. *if $c_p = \Theta(p^{-a})$ where $a \geq 1$, then $\overline{\lambda_k} = \Theta(k^{-d-2a+2})$,*
2. *if $c_p = \delta_{(p \text{ even})}\Theta(p^{-a})$, then $\overline{\lambda_k} = \delta_{(k \text{ even})}\Theta(k^{-d-2a+2})$,*
3. *if $c_p = \mathcal{O}\left(\exp\left(-a\sqrt{p}\right)\right)$, then $\overline{\lambda_k} = \mathcal{O}\left(k^{-d+1/2}\exp\left(-a\sqrt{k}\right)\right)$,*
4. *if $c_p = \Theta(p^{1/2}a^{-p})$, then $\overline{\lambda_k} = \mathcal{O}\left(k^{-d+1}a^{-k}\right)$ and $\overline{\lambda_k} = \Omega\left(k^{-d/2+1}2^{-k}a^{-k}\right)$.*

In addition to recovering existing results for ReLU networks [BJK19, VY21, GYK20, BB21], Corollary 4.4.7 also provides the decay rates for two-layer networks with Tanh and Gaussian activations. As faster eigenvalue decay implies a smaller RKHS Corollary 4.4.7 shows using ReLU results in a larger RKHS relative to Tanh or Gaussian activations. Numerics for Corollary 4.4.7 are provided in Figure 4.4 in Section 4.6.5. Finally, in Theorem 4.4.8 we relate a kernel's power series to its spectral decay for arbitrary data distributions.

Theorem 4.4.8 (Informal)**.** *Let the rows of* $\mathbf{X} \in \mathbb{R}^{n \times d}$ *be arbitrary points on the unit sphere. Consider the kernel matrix* $n\mathbf{K} = \sum_{p=0}^{\infty} c_p \left(\mathbf{X}\mathbf{X}^T\right)^{\odot p}$ *and let* $r(n) \leq d$ *denote the rank of* $\mathbf{X}\mathbf{X}^T$*. Then*

1. *if* $c_p = \mathcal{O}(p^{-\alpha})$ *with* $\alpha > r(n) + 1$ *for all* $n \in \mathbb{Z}_{\geq 0}$ *then* $\lambda_n(\mathbf{K}) = \mathcal{O}\left(n^{-\frac{\alpha-1}{r(n)}}\right)$,
2. *if* $c_p = \mathcal{O}(e^{-\alpha\sqrt{p}})$ *then* $\lambda_n(\mathbf{K}) = \mathcal{O}\left(n^{\frac{1}{2r(n)}} \exp\left(-\alpha' n^{\frac{1}{2r(n)}}\right)\right)$ *for any* $\alpha' < \alpha 2^{-1/2r(n)}$,
3. *if* $c_p = \mathcal{O}(e^{-\alpha p})$ *then* $\lambda_n(\mathbf{K}) = \mathcal{O}\left(\exp\left(-\alpha' n^{\frac{1}{r(n)}}\right)\right)$ *for any* $\alpha' < \alpha 2^{-1/2r(n)}$.

Although the presence of the factor $1/r(n)$ in the exponents of n in these bounds is a weakness, Theorem 4.4.8 still illustrates how, in a highly general setting, the asymptotic decay of the coefficients of the power series ensures a certain asymptotic decay in the eigenvalues of the kernel matrix. A formal version of this result is provided in Section 4.6.7 along with further discussion.

4.5 Conclusion

Using a power series expansion we derived a number of insights into both the outliers as well as the asymptotic decay of the spectrum of the NTK, in particular highlighting the role of the activation function. We performed our analysis without recourse to a high dimensional limit or the use of random matrix theory. Interesting avenues for future work include better analyzing the role of depth as well as characterizing the outlier eigenvalues and spectrum as a whole for networks with convolutional, residual or transformer layers.

Reproducibility Statement To ensure reproducibility, we make the code public at `https://github.com/bbowman223/data_ntk`

4.6 Appendix

This section is organized as follows.

- Section 4.6.1 gives background material on Gaussan kernels, NTK, unit variance intitialization, and Hermite polynomial expansions.
- Section 4.6.2 derives the power series expansion for the NTK.
- Section 4.6.3 analyzes the effective rank of power series kernels.
- Section 4.6.4 analyzes the effective rank of finite-width networks.
- Section 4.6.5 empirically validates the theoretical results on the effective rank of the NTK and the asymptotic decay of its spectrum.
- Section 4.6.6 analyzes the asymptotic decay of the spectrum for data uniformly distributed on the sphere.
- Section 4.6.7 analyzes the asymptotic decay of the spectrum for nonuniform distributions.

4.6.1 Background Material

4.6.1.1 Gaussian Kernel

Observe by construction that the flattened collection of preactivations at the first layer $(g^{(1)}(\mathbf{x}_i))_{i=1}^n$ form a centered Gaussian process, with the covariance between the αth and βth neuron being described by

$$\Sigma^{(1)}_{\alpha_\beta}(\mathbf{x}_i, \mathbf{x}_j) := \mathbb{E}[g^{(1)}_\alpha(\mathbf{x}_i) g^{(1)}_\beta(\mathbf{x}_j)] = \delta_{\alpha=\beta} \left(\gamma_w^2 \mathbf{x}_i^T \mathbf{x}_j + \gamma_b^2\right).$$

Under the Assumption 4.2.1, the preactivations at each layer $l \in [L+1]$ converge also in distribution to centered Gaussian processes [Nea96, LBN18]. We remark that the sequential width limit condition of Assumption 4.2.1 is not necessary for this behavior, for example the

same result can be derived in the setting where the widths of the network are sent to infinity simultaneously under certain conditions on the activation function [dHR18]. However, as our interests lie in analyzing the limit rather than the conditions for convergence to said limit, for simplicity we consider only the sequential width limit. As per [LBN18, Eq. 4], the covariance between the preactivations of the αth and βth neurons at layer $l \geq 2$ for any input pair $\mathbf{x}, \mathbf{y} \in \mathbb{R}$ are described by the following kernel,

$$\begin{aligned}\Sigma^{(l)}_{\alpha\beta}(\mathbf{x}, \mathbf{y}) &:= \mathbb{E}[g^{(l)}_{\alpha}(\mathbf{x}) g^{(l)}_{\beta}(\mathbf{y})] \\ &= \delta_{\alpha=\beta} \left(\sigma_w^2 \mathbb{E}_{g^{(l-1)} \sim \mathcal{GP}(0, \Sigma^{l-1})}[\phi(g^{(l-1)}_{\alpha}(\mathbf{x})) \phi(g^{(l-1)}_{\beta}(\mathbf{y}))] + \sigma_b^2 \right).\end{aligned}$$

We refer to this kernel as the Gaussian kernel. As each neuron is identically distributed and the covariance between pairs of neurons is 0 unless $\alpha = \beta$, moving forward we drop the subscript and discuss only the covariance between the preactivations of an arbitrary neuron given two inputs. As per the discussion by [LBN18, Section 2.3], the expectations involved in the computation of these Gaussian kernels can be computed with respect to a bivariate Gaussian distribution, whose covariance matrix has three distinct entries: the variance of a preactivation of $\mathbf{x}$ at the previous layer, $\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x})$, the variance of a preactivation of $\mathbf{y}$ at the previous layer, $\Sigma^{(l)}(\mathbf{y}, \mathbf{y})$, and the covariance between preactivations of $\mathbf{x}$ and $\mathbf{y}$, $\Sigma^{(l-1)}(\mathbf{x}, \mathbf{y})$. Therefore the Gaussian kernel, or covariance function, and its derivative, which we will require later for our analysis of the NTK, can be computed via the the following recurrence relations, see for instance [LBN18, JGH18, ADH19b, NMM21],

$$\begin{aligned}\Sigma^{(1)}(\mathbf{x}, \mathbf{y}) &= \gamma_w^2 \mathbf{x}^T \mathbf{x} + \gamma_b^2, \\ \mathbf{A}^{(l)}(\mathbf{x}, \mathbf{y}) &= \begin{bmatrix} \Sigma^{(l-1)}(\mathbf{x}, \mathbf{x}) & \Sigma^{(l-1)}(\mathbf{x}, \mathbf{y}) \\ \Sigma^{(l-1)}(\mathbf{y}, \mathbf{x}) & \Sigma^{(l-1)}(\mathbf{x}, \mathbf{x}) \end{bmatrix}, \\ \Sigma^{(l)}(\mathbf{x}, \mathbf{y}) &= \sigma_w^2 \mathbb{E}_{(B_1, B_2) \sim \mathcal{N}(0, \mathbf{A}^{(l)}(\mathbf{x}, \mathbf{y}))}[\phi(B_1) \phi(B_2)] + \sigma_b^2, \\ \dot{\Sigma}^{(l)}(\mathbf{x}, \mathbf{y}) &= \sigma_w^2 \mathbb{E}_{(B_1, B_2) \sim \mathcal{N}(0, \mathbf{A}^{(l)}(\mathbf{x}, \mathbf{y}))} \left[\phi'(B_1) \phi'(B_2)\right].\end{aligned} \tag{4.11}$$

4.6.1.2 Neural Tangent Kernel (NTK)

As discussed in the Section 4.1, under Assumption 4.2.1 $\tilde{\Theta}^{(l)}$ converges in probability to a deterministic limit, which we denote $\Theta^{(l)}$. This deterministic limit kernel can be expressed in terms of the Gaussian kernels and their derivatives from Section 4.6.1.1 via the following recurrence relationships [JGH18, Theorem 1],

$$\begin{aligned}\Theta^{(1)}(\mathbf{x},\mathbf{y}) &= \Sigma^{(1)}(\mathbf{x},\mathbf{y}),\\ \Theta^{(l)}(\mathbf{x},\mathbf{y}) &= \Theta^{(l-1)}(\mathbf{x},\mathbf{y})\dot{\Sigma}^{(l)}(\mathbf{x},\mathbf{y}) + \Sigma^{(l)}(\mathbf{x},\mathbf{y})\\ &= \Sigma^{(l)}(\mathbf{x},\mathbf{y}) + \sum_{h=1}^{l-1}\Sigma^{(h)}(\mathbf{x},\mathbf{y})\left(\prod_{h'=h+1}^{l}\dot{\Sigma}^{(h')}(\mathbf{x},\mathbf{y})\right)\ \forall l\in[2,L+1].\end{aligned} \tag{4.12}$$

A useful expression for the NTK matrix, which is a straightforward extension and generalization of [NMM21, Lemma 3.1], is provided in Lemma 4.6.1 below.

Lemma 4.6.1. *(Based on [NMM21, Lemma 3.1]) Under Assumption 4.2.1, a sequence of positive semidefinite matrices $(\mathbf{G}_l)_{l=1}^{L+1}$ in $\mathbb{R}^{n\times n}$, and the related sequence $(\dot{\mathbf{G}}_l)_{l=2}^{L+1}$ also in $\mathbb{R}^{n\times n}$, can be constructed via the following recurrence relationships,*

$$\begin{aligned}\mathbf{G}_1 &= \gamma_w^2\mathbf{X}\mathbf{X}^T + \gamma_b^2\boldsymbol{1}_{n\times n},\\ \mathbf{G}_2 &= \sigma_w^2\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)}[\phi(\mathbf{X}\mathbf{w})\phi(\mathbf{X}\mathbf{w})^T] + \sigma_b^2\boldsymbol{1}_{n\times n},\\ \dot{\mathbf{G}}_2 &= \sigma_w^2\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_n)}[\phi'(\mathbf{X}\mathbf{w})\phi'(\mathbf{X}\mathbf{w})^T],\\ \mathbf{G}_l &= \sigma_w^2\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_n)}[\phi(\sqrt{\mathbf{G}_{l-1}}\mathbf{w})\phi(\sqrt{\mathbf{G}_{l-1}}\mathbf{w})^T] + \sigma_b^2\boldsymbol{1}_{n\times n},\ l\in[3,L+1],\\ \dot{\mathbf{G}}_l &= \sigma_w^2\mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_n)}[\phi'(\sqrt{\mathbf{G}_{l-1}}\mathbf{w})\phi'(\sqrt{\mathbf{G}_{l-1}}\mathbf{w})^T],\ l\in[3,L+1].\end{aligned} \tag{4.13}$$

The sequence of NTK matrices $(\mathbf{K}_l)_{l=1}^{L+1}$ can in turn be written using the following recurrence relationship,

$$\begin{aligned}n\mathbf{K}_1 &= \mathbf{G}_1,\\ n\mathbf{K}_l &= \mathbf{G}_l + n\mathbf{K}_{l-1}\odot\dot{\mathbf{G}}_l\\ &= \mathbf{G}_l + \sum_{i=1}^{l-1}\left(\mathbf{G}_i\odot\left(\odot_{j=i+1}^{l}\dot{\mathbf{G}}_j\right)\right).\end{aligned} \tag{4.14}$$

Proof. For the sequence $(\mathbf{G}_l)_{l=1}^{L+1}$ it suffices to prove for any $i, j \in [n]$ and $l \in [L+1]$ that

$$[\mathbf{G}_l]_{i,j} = \Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_j)$$

and $\mathbf{G}_l$ is positive semi-definite. We proceed by induction, considering the base case $l = 1$ and comparing (4.13) with (4.11) then it is evident that

$$[\mathbf{G}_1]_{i,j} = \Sigma^{(1)}(\mathbf{x}_i, \mathbf{x}_j).$$

In addition, $\mathbf{G}_1$ is also clearly positive semi-definite as for any $\mathbf{u} \in \mathbb{R}^n$

$$\mathbf{u}^T \mathbf{G}_1 \mathbf{u} = \gamma_w^2 \left\| \mathbf{X}^T \mathbf{u} \right\|^2 + \gamma_b^2 \left\| \mathbf{1}_n^T \mathbf{u} \right\|^2 \geq 0.$$

We now assume the induction hypothesis is true for $\mathbf{G}_{l-1}$. We will need to distinguish slightly between two cases, $l = 2$ and $l \in [3, L+1]$. The proof of the induction step in either case is identical. To this end, and for notational ease, let $\mathbf{V} = \mathbf{X}$, $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ when $l = 2$, and $\mathbf{V} = \sqrt{\mathbf{G}_{l-1}}$, $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_n)$ for $l \in [3, L+1]$. In either case we let $\mathbf{v}_i$ denote the ith row of $\mathbf{V}$. For any $i, j \in [n]$

$$[\mathbf{G}_l]_{ij} = \sigma_w^2 \mathbb{E}_{\mathbf{w}}[\phi(\mathbf{v}_i^T \mathbf{w})\phi(\mathbf{v}_j^T \mathbf{w})] + \sigma_b^2.$$

Now let $B_1 = \mathbf{v}_i^T \mathbf{w}$, $B_2 = \mathbf{v}_j^T \mathbf{w}$ and observe for any $\alpha_1, \alpha_2 \in \mathbb{R}$ that $\alpha_1 B_1 + \alpha_2 B_2 = \sum_k^n (\alpha_1 v_{ik} + \alpha_2 v_{jk}) w_k \sim \mathcal{N}(0, \|\alpha_1 \mathbf{v}_i + \alpha_2 \mathbf{v}_j\|^2)$. Therefore the joint distribution of (B_1, B_2) is a mean 0 bivariate normal distribution. Denoting the covariance matrix of this distribution as $\tilde{\mathbf{A}} \in \mathbb{R}^{2\times 2}$, then $[\mathbf{G}_l]_{ij}$ can be expressed as

$$[\mathbf{G}_l]_{ij} = \sigma_w^2 \mathbb{E}_{(B_1, B_2) \sim \tilde{\mathbf{A}}}[\phi(B_1)\phi(B_2)] + \sigma_b^2.$$

To prove $[\mathbf{G}_l]_{i,j} = \Sigma^{(l)}$ it therefore suffices to show that $\tilde{\mathbf{A}} = \mathbf{A}^{(l)}$ as per (4.11). This follows by the induction hypothesis as

$$\mathbb{E}[B_1^2] = \mathbf{v}_i^T \mathbf{v}_i = [\mathbf{G}_{l-1}]_{ii} = \Sigma^{(l-1)}(\mathbf{x}_i, \mathbf{x}_i),$$
$$\mathbb{E}[B_2^2] = \mathbf{v}_j^T \mathbf{v}_j = [\mathbf{G}_{l-1}]_{jj} = \Sigma^{(l-1)}(\mathbf{x}_j, \mathbf{x}_j),$$
$$\mathbb{E}[B_1 B_2] = \mathbf{v}_i^T \mathbf{v}_j = [\mathbf{G}_{l-1}]_{ij} = \Sigma^{(l-1)}(\mathbf{x}_i, \mathbf{x}_j).$$

Finally, $\mathbf{G}_l$ is positive semi-definite as long as $\mathbb{E}_{\mathbf{w}}[\phi(\mathbf{V}\mathbf{w})\phi(\mathbf{V}\mathbf{w})^T]$ is positive semi-definite. Let $M(\mathbf{w}) = \phi(\mathbf{V}\mathbf{w}) \in \mathbb{R}^{n\times n}$ and observe for any $\mathbf{w}$ that $M(\mathbf{w})M(\mathbf{w})^T$ is positive semi-definite. Therefore $\mathbb{E}_{\mathbf{w}}[M(\mathbf{w})M(\mathbf{w})^T]$ must also be positive semi-definite. Thus the inductive step is complete and we may conclude for $l \in [L+1]$ that

$$[\mathbf{G}_l]_{i,j} = \Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_j). \tag{4.15}$$

For the proof of the expression for the sequence $(\dot{\mathbf{G}}_l)_{l=2}^{L+1}$ it suffices to prove for any $i, j \in [n]$ and $l \in [L+1]$ that

$$[\dot{\mathbf{G}}_l]_{i,j} = \dot{\Sigma}^{(l)}(\mathbf{x}_i, \mathbf{x}_j).$$

By comparing (4.13) with (4.11) this follows immediately from (4.15). Therefore with (4.13) proven (4.14) follows from (4.12). □

4.6.1.3 Unit Variance Initialization

The initialization scheme for a neural network, particularly a deep neural network, needs to be designed with some care in order to avoid either vanishing or exploding gradients during training [GB10, HZR15, MM16, LBO12]. Some of the most popular initialization strategies used in practice today, in particular [LBO12] and [GB10] initialization, first model the preactivations of the network as Gaussian random variables and then select the network hyperparameters in order that the variance of these idealized preactivations is fixed at one. Under Assumption 4.2.1 this idealized model on the preactivations is actually realized and if we additionally assume the conditions of Assumption 4.3.1 hold then likewise the variance of the preactivations at every layer will be fixed at one. To this end, and as in [PLR16, MAT22], consider the function $V\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined as

$$V(q) = \sigma_w^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\phi\left(\sqrt{q}Z\right)^2\right] + \sigma_b^2. \tag{4.16}$$

Noting that V is another expression for $\Sigma^{(l)}(\mathbf{x}, \mathbf{x})$, derived via a change of variables as per [PLR16], the sequence of variances $(\Sigma^{(l)}(\mathbf{x}, \mathbf{x}))_{l=2}^{L}$ can therefore be generated as follows,

$$\Sigma^{(l)}(\mathbf{x}, \mathbf{x}) = V(\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x})). \tag{4.17}$$

The linear correlation $\rho^{(l)} : \mathbb{R}^d \times \mathbb{R}^d \to [-1, 1]$ between the preactivations of two inputs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we define as

$$\rho^{(l)}(\mathbf{x}, \mathbf{y}) = \frac{\Sigma^{(l)}(\mathbf{x}, \mathbf{y})}{\sqrt{\Sigma^{(l)}(\mathbf{x}, \mathbf{x})\Sigma^{(l)}(\mathbf{y}, \mathbf{y})}}. \tag{4.18}$$

Assuming $\Sigma^{(l)}(\mathbf{x}, \mathbf{x}) = \Sigma^{(l)}(\mathbf{y}, \mathbf{y}) = 1$ for all $l \in [L + 1]$, then $\rho^{(l)}(\mathbf{x}, \mathbf{y}) = \Sigma^{(l)}(\mathbf{x}, \mathbf{y})$. Again as in [MAT22] and analogous to (4.16), with $Z_1, Z_2 \sim \mathcal{N}(0, 1)$ independent, $U_1 := Z_1$, $U_2(\rho) := (\rho Z_1 + \sqrt{1 - \rho^2} Z_2)$ [3] we define the correlation function $R : [-1, 1] \to [-1, 1]$ as

$$R(\rho) = \sigma_w^2 \mathbb{E}[\phi(U_1)\phi(U_2(\rho))] + \sigma_b^2. \tag{4.19}$$

Noting under these assumptions that R is equivalent to $\Sigma^{(l)}(\mathbf{x}, \mathbf{y})$, the sequence of correlations $(\rho^{(l)}(\mathbf{x}, \mathbf{y}))_{l=2}^{L}$ can thus be generated as

$$\rho^{(l)}(\mathbf{x}, \mathbf{y}) = R(\rho^{(l-1)}(\mathbf{x}, \mathbf{y})).$$

As observed in [PLR16, SGG17], $R(1) = V(1) = 1$, hence $\rho = 1$ is a fixed point of R. We remark that as all preactivations are distributed as $\mathcal{N}(0, 1)$, then a correlation of one between preactivations implies they are equal. The stability of the fixed point $\rho = 1$ is of particular significance in the context of initializing deep neural networks successfully. Under mild conditions on the activation function one can compute the derivative of R, see e.g., [PLR16, SGG17, MAT22], as follows,

$$R'(\rho) = \sigma_w^2 \mathbb{E}[\phi'(U_1)\phi'(U_2(\rho))]. \tag{4.20}$$

Observe that the expression for $\dot{\Sigma}^{(l)}$ and R' are equivalent via a change of variables [PLR16], and therefore the sequence of correlation derivatives may be computed as

$$\dot{\Sigma}^{(l)}(\mathbf{x}, \mathbf{y}) = R'(\rho^{(l)}(\mathbf{x}, \mathbf{y})).$$

With the relevant background material now in place we are in a position to prove Lemma 4.6.2.

[3] We remark that U_1, U_2 are dependent and identically distributed as $U_1, U_2 \sim \mathcal{N}(0, 1)$.

Lemma 4.6.2. *Under Assumptions 4.2.1 and 4.3.1 and defining* $\chi = \sigma_w^2 \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi'(Z)^2] \in \mathbb{R}_{>0}$*, then for all* $i, j \in [n]$*,* $l \in [L+1]$

- $[\mathbf{G}_{n,l}]_{ij} \in [-1, 1]$ *and* $[\mathbf{G}_{n,l}]_{ii} = 1$*,*
- $[\dot{\mathbf{G}}_{n,l}]_{ij} \in [-\chi, \chi]$ *and* $[\dot{\mathbf{G}}_{n,l}]_{ii} = \chi$*.*

Furthermore, the NTK is a dot product kernel, meaning $\Theta(\mathbf{x}_i, \mathbf{x}_j)$ *can be written as a function of the inner product between the two inputs,* $\Theta(\mathbf{x}_i^T \mathbf{x}_j)$*.*

Proof. Recall from Lemma 4.6.1 and its proof that for any $l \in [L+1]$, $i, j \in [n]$ $[\mathbf{G}_{n,l}]_{ij} = \Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_j)$ and $[\dot{\mathbf{G}}_{n,l}]_{ij} = \dot{\Sigma}^{(l)}(\mathbf{x}_i, \mathbf{x}_j)$. We first prove by induction $\Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_i) = 1$ for all $l \in [L+1]$. The base case $l = 1$ follows as

$$\Sigma^{(1)}(\mathbf{x}, \mathbf{x}) = \gamma_w^2 \mathbf{x}^T \mathbf{x} + \gamma_b^2 = \gamma_w^2 + \gamma_b^2 = 1.$$

Assume the induction hypothesis is true for layer $l-1$. With $Z \sim \mathcal{N}(0, 1)$, then from (4.16) and (4.17)

$$\begin{aligned}
\Sigma^{(l)}(\mathbf{x}, \mathbf{x}) &= V(\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x})) \\
&= \sigma_w^2 \mathbb{E}\left[\phi^2\left(\sqrt{\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x})} Z\right)\right] + \sigma_b^2 \\
&= \sigma_w^2 \mathbb{E}\left[\phi^2(Z)\right] + \sigma_b^2 \\
&= 1,
\end{aligned}$$

thus the inductive step is complete. As an immediate consequence it follows that $[\mathbf{G}_l]_{ii} = 1$. Also, for any $i, j \in [n]$ and $l \in [L+1]$,

$$\Sigma^{(l)}(\mathbf{x}_i, \mathbf{x}_j) = \rho^{(l)}(\mathbf{x}_i, \mathbf{x}_j) = R(\rho^{(l-1)}(\mathbf{x}_i, \mathbf{x}_j)) = R(...R(R(\mathbf{x}_i^T \mathbf{x}_j))).$$

Thus we can consider $\Sigma^{(l)}$ as a univariate function of the input correlation $\Sigma : [-1, 1] \to [-1, 1]$ and also conclude that $[\mathbf{G}_l]_{ij} \in [-1, 1]$. Furthermore,

$$\dot{\Sigma}^{(l)}(\mathbf{x}_i, \mathbf{x}_j) = R'(\rho^{(l)}(\mathbf{x}_i, \mathbf{x}_j)) = R'(R(...R(R(\mathbf{x}_i^T \mathbf{x}_j)))),$$

which likewise implies $\dot{\Sigma}$ is a dot product kernel. Recall now the random variables introduced to define R: $Z_1, Z_2 \sim \mathcal{N}(0,1)$ are independent and $U_1 = Z_1$, $U_2 = (\rho Z_1 + \sqrt{1-\rho^2} Z_2)$. Observe U_1, U_2 are dependent but identically distributed as $U_1, U_2 \sim \mathcal{N}(0,1)$. For any $\rho \in [-1,1]$ then applying the Cauchy-Schwarz inequality gives

$$|R'(\rho)|^2 = \sigma_w^4 \left|\mathbb{E}[\phi'(U_1)\phi'(U_2)]\right|^2 \leq \sigma_w^4 \mathbb{E}[\phi'(U_1)^2]\mathbb{E}[\phi'(U_2)^2] = \sigma_w^4 \mathbb{E}[\phi'(U_1)^2]^2 = |R'(1)|^2.$$

As a result, under the assumptions of the lemma $\dot{\Sigma}^{(l)} : [-1,1] \to [-\chi, \chi]$ and $\dot{\Sigma}^{(l)}(\mathbf{x}_i, \mathbf{x}_i) = \chi$. From this it immediately follows that $[\dot{\mathbf{G}}_l]_{ij} \in [-\chi, \chi]$ and $[\dot{\mathbf{G}}_l]_{ii} = \chi$ as claimed. Finally, as $\Sigma : [-1,1] \to [-1,1]$ and $\dot{\Sigma} : [-1,1] \to [-\chi, \chi]$ are dot product kernels, then from (4.12) the NTK must also be a dot product kernel and furthermore a univariate function of the pairwise correlation of its input arguments. □

The following corollary, which follows immediately from Lemma 4.6.2 and (4.14), characterizes the trace of the NTK matrix in terms of the trace of the input gram.

Corollary 4.6.3. *Under the same conditions as Lemma 4.6.2, suppose ϕ and σ_w^2 are chosen such that $\chi = 1$. Then*

$$Tr(\mathbf{K}_{n,l}) = l. \tag{4.21}$$

4.6.1.4 Hermite Expansions

We say that a function $f : \mathbb{R} \to \mathbb{R}$ is square integrable w.r.t. the standard Gaussian measure $\gamma = e^{-x^2/2}/\sqrt{2\pi}$ if $\mathbb{E}_{x \sim \mathcal{N}(0,1)}[f(x)^2] < \infty$. We denote by $L^2(\mathbb{R}, \gamma)$ the space of all such functions. The probabilist's Hermite polynomials are given by

$$H_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2}, \quad k = 0, 1, \ldots.$$

The first three Hermite polynomials are $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = (x^2 - 1)$. Let $h_k(x) = \frac{H_k(x)}{\sqrt{k!}}$ denote the normalized probabilist's Hermite polynomials. The normalized Hermite polynomials form a complete orthonormal basis in $L^2(\mathbb{R}, \gamma)$ [OD14, §11]: in all

that follows, whenever we reference the Hermite polynomials, we will be referring to the normalized Hermite polynomials. The Hermite expansion of a function $\phi \in L^2(\mathbb{R}, \gamma)$ is given by

$$\phi(x) = \sum_{k=0}^{\infty} \mu_k(\phi) h_k(x), \tag{4.22}$$

where

$$\mu_k(\phi) = \mathbb{E}_{X \sim \mathcal{N}(0,1)}[\phi(X) h_k(X)] \tag{4.23}$$

is the kth normalized probabilist's Hermite coefficient of ϕ. In what follows we shall make use of the following identities.

$$\forall k \geq 1,\ h_k'(x) = \sqrt{k} h_{k-1}(x), \tag{4.24}$$

$$\forall k \geq 1,\ x h_k(x) = \sqrt{k+1} h_{k+1}(x) + \sqrt{k} h_{k-1}(x). \tag{4.25}$$

$$\begin{aligned} h_k(0) &= \begin{cases} 0, & \text{if } k \text{ is odd} \\ \frac{1}{\sqrt{k!}} (-1)^{\frac{k}{2}} (k-1)!! & \text{if } k \text{ is even} \end{cases}, \\ \text{where } k!! &= \begin{cases} 1, & k \leq 0 \\ k \cdot (k-2) \cdots 5 \cdot 3 \cdot 1, & k > 0 \text{ odd} \\ k \cdot (k-2) \cdots 6 \cdot 4 \cdot 2, & k > 0 \text{ even} \end{cases}. \end{aligned} \tag{4.26}$$

We also remark that the more commonly encountered physicist's Hermite polynomials, which we denote $\tilde{H}_k$, are related to the normalized probablist's polynomials as follows,

$$h_k(z) = \frac{2^{-k/2} \tilde{H}_k(z/\sqrt{2})}{\sqrt{k!}}.$$

The Hermite expansion of the activation function deployed will play a key role in determining the coefficients of the NTK power series. In particular, the Hermite coefficients of ReLU are as follows.

Lemma 4.6.4. *[DFS16] For $\phi(z) = \max\{0, z\}$ the Hermite coefficients are given by*

$$\mu_k(\phi) = \begin{cases} 1/\sqrt{2\pi}, & k = 0, \\ 1/2, & k = 1, \\ (k-3)!!/\sqrt{2\pi k!}, & k \text{ even and } k \geq 2, \\ 0, & k \text{ odd and } k > 3. \end{cases} \tag{4.27}$$

4.6.2 Expressing the NTK as a Power Series

4.6.2.1 Deriving a Power Series for the NTK

We will require the following minor adaptation of [NM20, Lemma D.2]. We remark this result was first stated for ReLU and Softplus activations in the work of [OS20, Lemma H.2].

Lemma 4.6.5. *For arbitrary $n, d \in \mathbb{N}$, let $\mathbf{A} \in \mathbb{R}^{n \times d}$. For $i \in [n]$, we denote the ith row of $\mathbf{A}$ as $\mathbf{a}_i$, and further assume that $\|\mathbf{a}_i\| = 1$. Let $\phi : \mathbb{R} \to \mathbb{R}$ satisfy $\phi \in L^2(\mathbb{R}, \gamma)$ and define*

$$\mathbf{M} = \mathbb{E}_{\mathbf{w} \sim \mathcal{N}(0, \boldsymbol{I}_n)}[\phi(\mathbf{A}\mathbf{w})\phi(\mathbf{A}\mathbf{w})^T] \in \mathbb{R}^{n \times n}.$$

Then the matrix series

$$\mathbf{S}_K = \sum_{k=0}^{K} \mu_k^2(\phi) \left(\mathbf{A}\mathbf{A}^T\right)^{\odot k}$$

converges uniformly to $\mathbf{M}$ as $K \to \infty$.

The proof of Lemma 4.6.5 follows exactly as in [NM20, Lemma D.2], and is in fact slightly simpler due to the fact we assume the rows of $\mathbf{A}$ are unit length and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ instead of $\sqrt{d}$ and $\mathbf{w} \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ respectively. For the ease of the reader, we now recall the following definitions, which are also stated in Section 4.3. Letting $\bar{\alpha}_l := (\alpha_{p,l})_{p=0}^{\infty}$ denote a sequence of real coefficients, then

$$F(p, k, \bar{\alpha}_l) := \begin{cases} 1 & k = 0 \text{ and } p = 0, \\ 0 & k = 0 \text{ and } p \geq 1, \\ \sum_{(j_i) \in \mathcal{J}(p,k)} \prod_{i=1}^{k} \alpha_{j_i, l} & k \geq 1 \text{ and } p \geq 0, \end{cases} \tag{4.28}$$

where

$$\mathcal{J}(p,k) := \{(j_i)_{i\in[k]} \ : \ j_i \geq 0 \ \forall i \in [k], \ \sum_{i=1}^{k} j_i = p\}$$

for all $p \in \mathbb{Z}_{\geq 0}$, $k \in \mathbb{Z}_{\geq 1}$.

We are now ready to derive power series for elements of $(\mathbf{G}_l))_{l=1}^{L+1}$ and $(\dot{\mathbf{G}}_l))_{l=2}^{L+1}$.

Lemma 4.6.6. *Under Assumptions 4.2.1 and 4.3.1, for all $l \in [2, L+1]$*

$$\mathbf{G}_l = \sum_{k=0}^{\infty} \alpha_{k,l} (\mathbf{X}\mathbf{X}^T)^{\odot k}, \tag{4.29}$$

where the series for each element $[\mathbf{G}_l]_{ij}$ converges absolutely and the coefficients $\alpha_{p,l}$ are nonnegative. The coefficients of the series (4.29) for all $p \in \mathbb{Z}_{\geq 0}$ can be expressed via the following recurrence relationship,

$$\alpha_{p,l} = \begin{cases} \sigma_w^2 \mu_p^2(\phi) + \delta_{p=0}\sigma_b^2, & l = 2, \\ \sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3. \end{cases} \tag{4.30}$$

Furthermore,

$$\dot{\mathbf{G}}_l = \sum_{k=0}^{\infty} \upsilon_{k,l} (\mathbf{X}\mathbf{X}^T)^{\odot k}, \tag{4.31}$$

where likewise the series for each entry $[\dot{\mathbf{G}}_l]_{ij}$ converges absolutely and the coefficients $\upsilon_{p,l}$ for all $p \in \mathbb{Z}_{\geq 0}$ are nonnegative and can be expressed via the following recurrence relationship,

$$\upsilon_{p,l} = \begin{cases} \sigma_w^2 \mu_p^2(\phi'), & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3. \end{cases} \tag{4.32}$$

Proof. We start by proving (4.29) and (4.30). Proceeding by induction, consider the base case $l = 2$. From Lemma 4.6.1

$$\mathbf{G}_2 = \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\phi(\mathbf{X}\mathbf{w})\phi(\mathbf{X}\mathbf{w})^T] + \sigma_b^2 \mathbf{1}_{n\times n}.$$

By the assumptions of the lemma, the conditions of Lemma 4.6.5 are satisfied and therefore

$$\begin{aligned} \mathbf{G}_2 &= \sigma_w^2 \sum_{k=0}^{\infty} \mu_k^2(\phi) \left(\mathbf{X}\mathbf{X}^T\right)^{\odot k} + \sigma_b^2 \mathbf{1}_{n\times n} \\ &= \alpha_{0,2}\mathbf{1}_{n\times n} + \sum_{k=1}^{\infty} \alpha_{k,2} \left(\mathbf{X}\mathbf{X}^T\right)^{\odot k}. \end{aligned}$$

Observe the coefficients $(\alpha_{k,2})_{k\in\mathbb{Z}_{\geq 0}}$ are nonnegative. Therefore, for any $i,j \in [n]$ using Lemma 4.6.2 the series for $[\mathbf{G}_l]_{ij}$ satisfies

$$\sum_{k=0}^{\infty} |\alpha_{k,2}| \left|\langle \mathbf{x}_i, \mathbf{x}_j\rangle^k\right| \leq \sum_{k=0}^{\infty} \alpha_{k,2}\langle \mathbf{x}_i, \mathbf{x}_i\rangle^k \quad = [\mathbf{G}_l]_{ii} = 1 \tag{4.33}$$

and so must be absolutely convergent. With the base case proved we proceed to assume the inductive hypothesis holds for arbitrary $\mathbf{G}_l$ with $l \in [2, L]$. Observe

$$\mathbf{G}_{l+1} = \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_n)}[\phi(\mathbf{A}\mathbf{w})\phi(\mathbf{A}\mathbf{w})^T] + \sigma_b^2 \mathbf{1}_{n\times n},$$

where $\mathbf{A}$ is a matrix square root of $\mathbf{G}_l$, meaning $\mathbf{G}_l = \mathbf{A}\mathbf{A}$. Recall from Lemma 4.6.1 that $\mathbf{G}_l$ is also symmetric and positive semi-definite, therefore we may additionally assume, without loss of generality, that $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric, which conveniently implies $\mathbf{G}_{n,l} = \mathbf{A}\mathbf{A}^T$. Under the assumptions of the lemma the conditions for Lemma 4.6.2 are satisfied and as a result $[\mathbf{G}_{n,l}]_{ii} = \|\mathbf{a}_i\| = 1$ for all $i \in [n]$, where we recall $\mathbf{a}_i$ denotes the ith row of $\mathbf{A}$. Therefore we may again apply Lemma 4.6.1,

$$\begin{aligned}
\mathbf{G}_{l+1} &= \sigma_w^2 \sum_{k=0}^{\infty} \mu_k^2(\phi) \left(\mathbf{A}\mathbf{A}^T\right)^{\odot k} + \sigma_b^2 \mathbf{1}_{n\times n} \\
&= (\sigma_w^2\mu_0^2(\phi) + \sigma_b^2)\mathbf{1}_{n\times n} + \sigma_w^2 \sum_{k=1}^{\infty} \mu_k^2(\phi) \left(\mathbf{G}_{n,l}\right)^{\odot k} \\
&= (\sigma_w^2\mu_0^2(\phi) + \sigma_b^2)\mathbf{1}_{n\times n} + \sigma_w^2 \sum_{k=1}^{\infty} \mu_k^2(\phi) \left(\sum_{m=0}^{\infty} \alpha_{m,l} (\mathbf{X}\mathbf{X}^T)^{\odot m}\right)^{\odot k},
\end{aligned}$$

where the final equality follows from the inductive hypothesis. For any pair of indices $i,j \in [n]$

$$[\mathbf{G}_{l+1}]_{ij} = (\sigma_w^2\mu_0^2(\phi) + \sigma_b^2) + \sigma_w^2 \sum_{k=1}^{\infty} \mu_k^2(\phi) \left(\sum_{m=0}^{\infty} \alpha_{m,l} \langle \mathbf{x}_i, \mathbf{x}_j\rangle^m\right)^k.$$

By the induction hypothesis, for any $i,j \in [n]$ the series $\sum_{m=0}^{\infty} \alpha_{m,l}\langle \mathbf{x}_i, \mathbf{x}_j\rangle^m$ is absolutely convergent. Therefore, from the Cauchy product of power series and for any $k \in \mathbb{Z}_{\geq 0}$ we have

$$\left(\sum_{m=0}^{\infty} \alpha_{m,l} \langle \mathbf{x}_i, \mathbf{x}_j\rangle^m\right)^k = \sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l)\langle \mathbf{x}_i, \mathbf{x}_j\rangle^p, \tag{4.34}$$

where $F(p,k,\bar{\alpha}_l)$ is defined in (4.4). By definition, $F(p,k,\bar{\alpha}_l)$ is a sum of products of positive coefficients, and therefore $|F(p,k,\bar{\alpha}_l)| = F(p,k,\bar{\alpha}_l)$. In addition, recall again by Assumption 4.3.1 and Lemma 4.6.2 that $[\mathbf{G}_l]_{ii} = 1$. As a result, for any $k \in \mathbb{Z}_{\geq 0}$, as $|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| \leq 1$

$$\sum_{p=0}^{\infty} |F(p,k,\bar{\alpha}_l)\langle \mathbf{x}_i, \mathbf{x}_j \rangle^p| \leq \left(\sum_{m=0}^{\infty} \alpha_{m,l} \right)^k = [\mathbf{G}_{n,l}]_{ii} = 1 \tag{4.35}$$

and therefore the series $\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l)\langle \mathbf{x}_i, \mathbf{x}_j \rangle^p$ converges absolutely. Recalling from the proof of the base case that the series $\sum_{p=1}^{\infty} \alpha_{p,2}$ is absolutely convergent and has only non-negative elements, we may therefore interchange the order of summation in the following,

$$\begin{aligned}
[\mathbf{G}_{l+1}]_{ij} &= (\sigma_w^2 \mu_0^2(\phi) + \sigma_b^2) + \sigma_w^2 \sum_{k=1}^{\infty} \mu_k^2(\phi) \left(\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l)\langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \right) \\
&= \alpha_{0,2} + \sum_{k=1}^{\infty} \alpha_{k,2} \left(\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l)\langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \right) \\
&= \alpha_{0,2} + \sum_{p=0}^{\infty} \left(\sum_{k=1}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p.
\end{aligned}$$

Recalling the definition of $F(p,k,l)$ in (4.4), in particular $F(0,0,\bar{\alpha}_l) = 1$ and $F(p,0,\bar{\alpha}_l) = 0$ for $p \in \mathbb{Z}_{\geq 1}$, then

$$\begin{aligned}
[\mathbf{G}_{l+1}]_{ij} &= \left(\alpha_{0,2} + \sum_{k=1}^{\infty} \alpha_{k,2} F(0,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^0 + \sum_{p=1}^{\infty} \left(\sum_{k=1}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\
&= \left(\sum_{k=0}^{\infty} \alpha_{k,2} F(0,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^0 + \sum_{p=1}^{\infty} \left(\sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\
&= \sum_{p=0}^{\infty} \left(\sum_{k=0}^{\infty} \alpha_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\
&= \sum_{p=0}^{\infty} \alpha_{p,l+1} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p.
\end{aligned}$$

As the indices $i, j \in [n]$ were arbitrary we conclude that

$$\mathbf{G}_{l+1} = \sum_{p=0}^{\infty} \alpha_{p,l+1} \left(\mathbf{X}\mathbf{X}^T \right)^{\odot p}$$

as claimed. In addition, by inspection and using the induction hypothesis it is clear that the coefficients $(\alpha_{p,l+1})_{p=0}^{\infty}$ are nonnegative. Therefore, by an argument identical to (4.33), the series for each entry of $[\mathbf{G}_{l+1}]_{ij}$ is absolutely convergent. This concludes the proof of (4.29) and (4.30).

We now turn our attention to proving the (4.31) and (4.32). Under the assumptions of the lemma the conditions for Lemmas 4.6.1 and 4.6.5 are satisfied and therefore for the base case $l = 2$

$$\begin{aligned}\dot{\mathbf{G}}_2 &= \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_n)}[\phi'(\mathbf{X}\mathbf{w})\phi'(\mathbf{X}\mathbf{w})^T] \\ &= \sigma_w^2 \sum_{k=0}^{\infty} \mu_k^2(\phi') \left(\mathbf{X}\mathbf{X}^T\right)^{\odot k} \\ &= \sum_{k=0}^{\infty} \upsilon_{k,2} \left(\mathbf{X}\mathbf{X}^T\right)^{\odot k}.\end{aligned}$$

By inspection the coefficients $(\upsilon_{p,2})_{p=0}^{\infty}$ are nonnegative and as a result by an argument again identical to (4.33) the series for each entry of $[\dot{\mathbf{G}}_2]_{ij}$ is absolutely convergent. For $l \in [2, L]$, from (4.29) and its proof there is a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ such that $\mathbf{G}_l = \mathbf{A}\mathbf{A}^T$. Again applying Lemma 4.6.5

$$\begin{aligned}\dot{\mathbf{G}}_{n,l+1} &= \sigma_w^2 \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_n)}[\phi'(\mathbf{A}\mathbf{w})\phi'(\mathbf{A}\mathbf{w})^T] \\ &= \sigma_w^2 \sum_{k=0}^{\infty} \mu_k^2(\phi') \left(\mathbf{A}\mathbf{A}^T\right)^{\odot k} \\ &= \sum_{k=0}^{\infty} \upsilon_{k,2} \left(\mathbf{G}_{n,l}\right)^{\odot k} \\ &= \sum_{k=0}^{\infty} \upsilon_{k,2} \left(\sum_{p=0}^{\infty} \alpha_{p,l} \left(\mathbf{X}\mathbf{X}^T\right)^{\odot p}\right)^{\odot k}.\end{aligned}$$

Analyzing now an arbitrary entry $[\dot{\mathbf{G}}_{l+1}]_{ij}$, by substituting in the power series expression for

$\mathbf{G}_l$ from (4.29) and using (4.34) we have

$$\begin{aligned}
[\dot{\mathbf{G}}_{l+1}]_{ij} &= \sum_{k=0}^{\infty} \upsilon_{k,2} \left(\sum_{p=0}^{\infty} \alpha_{p,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \right)^k \\
&= \sum_{k=0}^{\infty} \upsilon_{k,2} \left(\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \right) \\
&= \sum_{p=0}^{\infty} \left(\sum_{k=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_l) \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\
&= \sum_{p=0}^{\infty} \upsilon_{p,l+1} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p.
\end{aligned}$$

Note that exchanging the order of summation in the third equality above is justified as for any $k \in \mathbb{Z}_{\geq 0}$ by (4.35) we have $\sum_{p=0}^{\infty} F(p,k,\bar{\alpha}_l) |\langle \mathbf{x}_i, \mathbf{x}_j \rangle|^p \leq 1$ and therefore

$$\sum_{k=0}^{\infty} \sum_{p=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_l) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p$$

converges absolutely. As the indices $i, j \in [n]$ were arbitrary we conclude that

$$\dot{\mathbf{G}}_{l+1} = \sum_{p=0}^{\infty} \upsilon_{p,l+1} \left(\mathbf{X}\mathbf{X}^T \right)^{\odot p}$$

as claimed. Finally, by inspection the coefficients $(\upsilon_{p,l+1})_{p=0}^{\infty}$ are nonnegative, therefore, and again by an argument identical to (4.33), the series for each entry of $[\dot{\mathbf{G}}_{n,l+1}]_{ij}$ is absolutely convergent. This concludes the proof. □

We are now prove the key result of Section 4.3.

Theorem 4.3.2. *Under Assumptions 4.2.1 and 4.3.1, for all $l \in [L+1]$*

$$n\mathbf{K}_l = \sum_{p=0}^{\infty} \kappa_{p,l} \left(\mathbf{X}\mathbf{X}^T \right)^{\odot p}. \tag{4.5}$$

The series for each entry $n[\mathbf{K}_l]_{ij}$ converges absolutely and the coefficients $\kappa_{p,l}$ are nonnegative and can be evaluated using the recurrence relationships

$$\kappa_{p,l} = \begin{cases} \delta_{p=0}\gamma_b^2 + \delta_{p=1}\gamma_w^2, & l = 1, \\ \alpha_{p,l} + \sum_{q=0}^{p} \kappa_{q,l-1} \upsilon_{p-q,l}, & l \in [2, L+1], \end{cases} \tag{4.6}$$

where

$$\alpha_{p,l} = \begin{cases} \sigma_w^2 \mu_p^2(\phi) + \delta_{p=0}\sigma_b^2, & l = 2, \\ \sum_{k=0}^{\infty} \alpha_{k,2} F(p, k, \bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{4.7}$$

and

$$\upsilon_{p,l} = \begin{cases} \sigma_w^2 \mu_p^2(\phi'), & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p, k, \bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{4.8}$$

are likewise nonnegative for all $p \in \mathbb{Z}_{\geq 0}$ *and* $l \in [2, L+1]$.

Proof. We proceed by induction. The base case $l = 1$ follows trivially from Lemma 4.6.1. We therefore assume the induction hypothesis holds for an arbitrary $l - 1 \in [1, L]$. From (4.14) and Lemma 4.6.6

$$\begin{aligned} n\mathbf{K}_l &= \mathbf{G}_l + n\mathbf{K}_{l-1} \odot \dot{\mathbf{G}}_l \\ &= \left(\sum_{p=0}^{\infty} \alpha_{p,l} \left(\mathbf{X}\mathbf{X}^T \right)^{\odot p} \right) + \left(n \sum_{q=0}^{\infty} \kappa_{q,l-1} \left(\mathbf{X}\mathbf{X}^T \right)^{\odot q} \right) \odot \left(\sum_{w=0}^{\infty} \upsilon_{w,l} \left(\mathbf{X}\mathbf{X}^T \right)^{\odot w} \right). \end{aligned}$$

Therefore, for arbitrary $i, j \in [n]$

$$[n\mathbf{K}_l]_{ij} = \sum_{p=0}^{\infty} \alpha_{p,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p + \left(n \sum_{q=0}^{\infty} \kappa_{q,l-1} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^q \right) \left(\sum_{w=0}^{\infty} \upsilon_{w,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^w \right).$$

Observe $n \sum_{q=0}^{\infty} \kappa_{q,l-1} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^q = \Theta^{(l-1)}(\mathbf{x}_i, \mathbf{x}_j)$ and therefore the series must converge due to the convergence of the NTK. Furthermore, $\sum_{w=0}^{\infty} \upsilon_{w,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^w = [\dot{\mathbf{G}}_{n,l}]_{ij}$ and therefore is absolutely convergent by Lemma 4.6.6. As a result, by Merten's Theorem the product of these two series is equal to their Cauchy product. Therefore

$$\begin{aligned} [n\mathbf{K}_l]_{ij} &= \sum_{p=0}^{\infty} \alpha_{p,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p + \sum_{p=0}^{\infty} \left(\sum_{q=0}^{p} \kappa_{q,l-1} \upsilon_{p-q,l} \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\ &= \sum_{p=0}^{\infty} \left(\alpha_{p,l} + \sum_{q=0}^{p} \kappa_{q,l-1} \upsilon_{p-q,l} \right) \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p \\ &= \sum_{p=0}^{\infty} \kappa_{p,l} \langle \mathbf{x}_i, \mathbf{x}_j \rangle^p, \end{aligned}$$

from which the (4.5) immediately follows. □

4.6.2.2 Analyzing the Coefficients of the NTK Power Series

In this section we study the coefficients of the NTK power series stated in Theorem 4.3.2. Our first observation is that, under additional assumptions on the activation function ϕ, the recurrence relationship (4.6) can be simplified in order to depend only on the Hermite expansion of ϕ.

Lemma 4.6.7. *Under Assumption 4.3.3 the Hermite coefficients of ϕ' satisfy*

$$\mu_k(\phi') = \sqrt{k+1}\mu_{k+1}(\phi)$$

for all $k \in \mathbb{Z}_{\geq 0}$.

Proof. Note for each $n \in \mathbb{N}$ as ϕ is absolutely continuous on $[-n, n]$ it is differentiable a.e. on $[-n, n]$. It follows by the countable additivity of the Lebesgue measure that ϕ is differentiable a.e. on $\mathbb{R}$. Furthermore, as ϕ is polynomially bounded we have $\phi \in L^2(\mathbb{R}, e^{-x^2/2}/\sqrt{2\pi})$. Fix $a > 0$. Since ϕ is absolutely continuous on $[-a, a]$ it is of bounded variation on $[-a, a]$. Also note that $h_k(x)e^{-x^2/2}$ is of bounded variation on $[-a, a]$ due to having a bounded derivative. Thus we have by Lebesgue-Stieltjes integration-by-parts (see e.g. [Fol99, Chapter 3])

$$\begin{aligned}
&\int_{-a}^{a} \phi'(x)h_k(x)e^{-x^2/2}dx \\
&= \phi(a)h_k(a)e^{-a^2/2} - \phi(-a)h_k(-a)e^{-a^2/2} + \int_{-a}^{a} \phi(x)[xh_k(x) - h_k'(x)]e^{-x^2/2}dx \\
&= \phi(a)h_k(a)e^{-a^2/2} - \phi(-a)h_k(-a)e^{-a^2/2} + \int_{-a}^{a} \phi(x)\sqrt{k+1}h_{k+1}(x)e^{-x^2/2}dx,
\end{aligned}$$

where in the last line above we have used the fact that (4.24) and (4.25) imply that $xh_k(x) - h_k'(x) = \sqrt{k+1}h_{k+1}(x)$. Thus we have shown

$$\begin{aligned}
&\int_{-a}^{a} \phi'(x)h_k(x)e^{-x^2/2}dx \\
&= \phi(a)h_k(a)e^{-a^2/2} - \phi(-a)h_k(-a)e^{-a^2/2} + \int_{-a}^{a} \phi(x)\sqrt{k+1}h_{k+1}(x)e^{-x^2/2}dx.
\end{aligned}$$

We note that since $|\phi(x)h_k(x)| = \mathcal{O}(|x|^{\beta+k})$ we have that as $a \to \infty$ the first two terms above vanish. Thus by sending $a \to \infty$ we have

$$\int_{-\infty}^{\infty} \phi'(x)h_k(x)e^{-x^2/2}dx = \int_{-\infty}^{\infty} \sqrt{k+1}\phi(x)h_{k+1}(x)e^{-x^2/2}dx.$$

After dividing by $\sqrt{2\pi}$ we get the desired result. □

In particular, under Assumption 4.3.3, and as highlighted by Corollary 4.6.8, which follows directly from Lemmas 4.6.6 and 4.6.7, the NTK coefficients can be computed only using the Hermite coefficients of ϕ.

Corollary 4.6.8. *Under Assumptions 4.2.1, 4.3.1 and 4.3.3, for all $p \in \mathbb{Z}_{\geq 0}$*

$$\upsilon_{p,l} = \begin{cases} (p+1)\alpha_{p+1,2}, & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p,k,\bar{\alpha}_{l-1}), & l \geq 3. \end{cases} \tag{4.36}$$

With these results in place we proceed to analyze the decay of the coefficients of the NTK for depth two networks. As stated in the main text, the decay of the NTK coefficients depends on the decay of the Hermite coefficients of the activation function deployed. This in turn is strongly influenced by the behavior of the tails of the activation function. To this end we roughly group activation functions into three categories: growing tails, flat or constant tails and finally decaying tails. Analyzing each of these groups in full generality is beyond the scope of this paper, we therefore instead study the behavior of ReLU, Tanh and Gaussian activation functions, being prototypical and practically used examples of each of these three groups respectively. We remark that these three activation functions satisfy Assumption 4.3.3. For typographical ease we let $\omega_\sigma(z) := (1/\sqrt{2\pi\sigma^2}) \exp(-z^2/(2\sigma^2))$ denote the Gaussian activation function with variance σ^2.

Lemma 4.3.4. *Under Assumptions 4.2.1 and 4.3.1,*

1. *if $\phi(z) = ReLU(z)$, then $\kappa_{p,2} = \delta_{(\gamma_b>0)\cup(p \text{ even})}\Theta(p^{-3/2})$,*

2. *if* $\phi(z) = Tanh(z)$, *then* $\kappa_{p,2} = \mathcal{O}\left(\exp\left(-\frac{\pi\sqrt{p-1}}{2}\right)\right)$,

3. *if* $\phi(z) = \omega_\sigma(z)$, *then* $\kappa_{p,2} = \delta_{(\gamma_b>0)\cup(p\ even)}\Theta(p^{1/2}(\sigma^2+1)^{-p})$.

Proof. Recall (4.9),

$$\kappa_{p,2} = \sigma_w^2(1+\gamma_w^2 p)\mu_p^2(\phi) + \sigma_w^2\gamma_b^2(1+p)\mu_{p+1}^2(\phi) + \delta_{p=0}\sigma_b^2.$$

In order to bound $\kappa_{p,2}$ we proceed by using Lemma 4.6.4 to bound the square of the Hermite coefficients. We start with ReLU. Note Lemma 4.6.4 actually provides precise expressions for the Hermite coefficients of ReLU, however, these are not immediately easy to interpret. Observe from Lemma 4.6.4 that above index $p = 2$ all odd indexed Hermite coefficients are 0. It therefore suffices to bound the even indexed terms, given by

$$\mu_p(ReLU) = \frac{1}{\sqrt{2\pi}}\frac{(p-3)!!}{\sqrt{p!}}.$$

Observe from (4.26) that for p even

$$h_p(0) = (-1)^{p/2}\frac{(p-1)!!}{\sqrt{p!}},$$

therefore

$$\mu_p(ReLU) = \frac{1}{\sqrt{2\pi}}\frac{(p-3)!!}{\sqrt{p!}} = \frac{1}{\sqrt{2\pi}}\frac{|h_p(0)|}{p-1}.$$

Analyzing now $|h_p(0)|$,

$$\frac{(p-1)!!}{\sqrt{p!}} = \frac{\prod_{i=1}^{p/2}(2i-1)}{\sqrt{\prod_{i=1}^{p/2}(2i-1)2i}} = \sqrt{\frac{\prod_{i=1}^{p/2}(2i-1)}{\prod_{i=1}^{p/2}2i}} = \sqrt{\frac{(p-1)!!}{p!!}}.$$

Here, the expression inside the square root is referred to in the literature as the Wallis ratio, for which the following lower and upper bounds are available [Kaz56],

$$\sqrt{\frac{1}{\pi(p+0.5)}} < \frac{(p-1)!!}{p!!} < \sqrt{\frac{1}{\pi(p+0.25)}}. \tag{4.37}$$

As a result

$$|h_p(0)| = \Theta(p^{-1/4})$$

and therefore

$$\mu_p(ReLU) = \begin{cases} \Theta(p^{-5/4}), & p \text{ even}, \\ 0, & p \text{ odd}. \end{cases}$$

As $(p+1)^{-3/2} = \Theta(p^{-3/2})$, then from (4.9)

$$\begin{aligned} \kappa_{p,2} &= \Theta((p\mu_p^2(ReLU) + \delta_{\gamma_b>0}(p+1)\mu_{p+1}^2(ReLU))) \\ &= \Theta((\delta_{p \text{ even}} p^{-3/2} + \delta_{(p \text{ odd})\cap(\gamma_b>0)}(p+1)^{-3/2})) \\ &= \Theta\left(\delta_{(p \text{ even})\cup((p \text{ odd})\cap(\gamma_b>0))} p^{-3/2}\right) \\ &= \delta_{(p \text{ even})\cup(\gamma_b>0)} \Theta\left(p^{-3/2}\right) \end{aligned}$$

as claimed in item *1.*

We now proceed to analyze $\phi(z) = Tanh(z)$. From [PSG20, Corollary F.7.1]

$$\mu_p(Tanh') = \mathcal{O}\left(\exp\left(-\frac{\pi\sqrt{p}}{4}\right)\right).$$

As Tanh satisfies the conditions of Lemma 4.6.7

$$\mu_p(Tanh) = p^{-1/2}\mu_{p-1}(Tanh') = \mathcal{O}\left(p^{-1/2}\exp\left(-\frac{\pi\sqrt{p-1}}{4}\right)\right).$$

Therefore the result claimed in item *2.* follows as

$$\begin{aligned} \kappa_{p,2} &= \mathcal{O}((p\mu_p^2(Tanh) + (p+1)\mu_{p+1}^2(Tanh))) \\ &= \mathcal{O}\left(\exp\left(-\frac{\pi\sqrt{p-1}}{2}\right) + \exp\left(-\frac{\pi\sqrt{p}}{2}\right)\right) \\ &= \mathcal{O}\left(\exp\left(-\frac{\pi\sqrt{p-1}}{2}\right)\right). \end{aligned}$$

Finally, we now consider $\phi(z) = \omega_\sigma(z)$ where $\omega_\sigma(z)$ is the density function of $\mathcal{N}(0, \sigma^2)$. Similar to ReLU, analytic expressions for the Hermite coefficients of $\omega_\sigma(z)$ are known see e.g., Theorem 2.9 in [Dav21],

$$\mu_p^2(\omega_\sigma) = \begin{cases} \frac{p!}{((p/2)!)^2 2^p 2\pi(\sigma^2+1)^{p+1}}, & \text{p even}, \\ 0, & \text{p odd}. \end{cases}$$

For p even

$$(p/2)! = p!!2^{-p/2}.$$

Therefore

$$\frac{p!}{(p/2)!(p/2)!} = 2^p \frac{p!}{p!!p!!} = 2^p \frac{(p-1)!!}{p!!}.$$

As a result, for p even and using (4.37), it follows that

$$\mu_p^2(\omega_\sigma) = \frac{(\sigma^2+1)^{-(p+1)}}{2\pi} \frac{(p-1)!!}{p!!} \quad = \Theta(p^{-1/2}(\sigma^2+1)^{-p}).$$

Finally, since $(p+1)^{1/2}(\sigma^2+1)^{-p-1} = \Theta(p^{1/2}(\sigma^2+1)^{-p})$, then from (4.9)

$$\begin{aligned}
\kappa_{p,2} &= \Theta((p\mu_p^2(\omega_\sigma) + \delta_{\gamma_b>0}(p+1)\mu_{p+1}^2(\omega_\sigma))) \\
&= \Theta\left(\delta_{(p \text{ even})\cup((p \text{ odd})\cap(\gamma_b>0))} p^{1/2}(\sigma^2+1)^{-p}\right) \\
&= \delta_{(p \text{ even})\cup(\gamma_b>0)} \Theta\left(p^{1/2}(\sigma^2+1)^{-p}\right)
\end{aligned}$$

as claimed in item *3.* □

4.6.2.3 Numerical Approximation via a Truncated NTK Power Series and Interpretation of 4.2

Currently, computing the infinite-width NTK requires either a) explicit evaluation of the Gaussian integrals highlighted in (4.13), b) numerical approximation of these same integrals such as in [LBN18], or c) approximation via a sufficiently wide yet still finite-width network, see for instance [EWS22, NSS22]. These Gaussian integrals (4.13) can be solved solved analytically only for a minority of activation functions, notably ReLU as discussed for example by [ADH19b], while the numerical integration and finite-width approximation approaches are relatively computationally expensive. The truncated NTK power series we define as analogous to (4.5) but with the series involved being computed only up to the Tth element. Once the top T coefficients are computed, then for any input correlation the NTK can be approximated by evaluating the corresponding finite degree T polynomial.

For an arbitrary pair $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ let $\rho = \mathbf{x}^T\mathbf{y}$ denote their linear correlation. Under Assumptions 4.2.1, 4.3.1 and 4.3.3, for all $l \in [2, L+1]$ the T-truncated NTK power series $\hat{\Theta}_T^{(l)} : [-1, 1] \to \mathbb{R}$ is defined as

$$\Theta_T^{(l)}(\rho) = \sum_{p=0}^{T} \hat{\kappa}_{p,l}\rho^p, \tag{4.38}$$

and whose coefficients are defined via the following recurrence relation,

$$\hat{\kappa}_{p,l} = \begin{cases} \delta_{p=0}\gamma_b^2 + \delta_{p=1}\gamma_w^2, & l = 1, \\ \hat{\alpha}_{p,l} + \sum_{q=0}^{p} \hat{\kappa}_{q,l-1}\hat{\upsilon}_{p-q,l}, & l \in [2, L+1]. \end{cases} \tag{4.39}$$

Here, with $\bar{\hat{\alpha}}_{l-1} = (\hat{\alpha}_{p,l-1})_{p=0}^T$,

$$\hat{\alpha}_{p,l} := \begin{cases} \sigma_w^2\mu_p^2(\phi) + \delta_{p=0}\sigma_b^2, & l = 2, \\ \sum_{k=0}^{T} \hat{\alpha}_{k,2}F(p, k, \bar{\hat{\alpha}}_{l-1}), & l \geq 3 \end{cases} \tag{4.40}$$

and

$$\hat{\upsilon}_{p,l} := \begin{cases} \sqrt{p+1}\hat{\alpha}_{p+1,2}, & l = 2, \\ \sum_{k=0}^{T} \sqrt{k+1}\hat{\alpha}_{p+1,2}F(p, k, \bar{\hat{\alpha}}_l), & l \geq 3. \end{cases} \tag{4.41}$$

In order to analyze the performance and potential of the truncated NTK for numerical approximation, we compute it for ReLU and compare it with its analytical expression [ADH19b]. To recall this result, let

$$\begin{aligned} R(\rho) &:= \frac{\sqrt{1-\rho^2} + \rho \cdot \arcsin(\rho)}{\pi} + \frac{\rho}{2}, \\ R'(\rho) &:= \frac{\arcsin(\rho)}{\pi} + \frac{1}{2}. \end{aligned}$$

Under Assumptions 4.2.1 and 4.3.1, with $\phi(z) = ReLU(z)$, $\gamma_w^2 = 1$, $\sigma_w^2 = 2$, $\sigma_b^2 = \gamma_b^2 = 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{S}^d$ and $\rho_1 := \mathbf{x}^T\mathbf{y}$, then $\Theta_1(\mathbf{x}, \mathbf{y}) = \rho$ and for all $l \in [2, L+1]$

$$\begin{aligned} \rho_l &= R(\rho_{l-1}), \\ \Theta_l(\mathbf{x}, \mathbf{y}) &= \rho_l + \rho_{l-1}R'(\rho_{l-1}). \end{aligned} \tag{4.42}$$

Turning our attention to Figure 4.2, we observe particularly for input correlations $|\rho| \approx 0.5$ and below then the truncated ReLU NTK power series achieves machine level precision. For $|\rho| \approx 1$ higher order coefficients play a more significant role. As the truncated ReLU NTK power series approximates these coefficients less well the overall approximation of the ReLU NTK is worse. We remark also that negative correlations have a smaller absolute error as odd indexed terms cancel with even index terms: we emphasize again that in Figure 4.2 we plot the absolute not relative error. In addition, for $L = 1$ there is symmetry in the absolute error for positive and negative correlations as $\alpha_{p,2} = 0$ for all odd p. One also observes that approximation accuracy goes down with depth, which is due to the error in the coefficients at the previous layer contributing to the error in the coefficients at the next, thereby resulting in an accumulation of error with depth. Also, and certainly as one might expect, a larger truncation point T results in overall better approximation. Finally, as the decay in the Hermite coefficients for ReLU is relatively slow, see e.g., Table 4.1 and Lemma 4.3.4, we expect the truncated ReLU NTK power series to perform worse relative to the truncated NTK's for other activation functions.

4.6.3 Effective Rank of Power Series Kernels

Recall that for a positive semidefinite matrix $\mathbf{A}$ we define the *effective rank* [HHV22] via the following ratio

$$\text{eff}(\mathbf{A}) := \frac{Tr(\mathbf{A})}{\lambda_1(\mathbf{A})}.$$

We consider a kernel Gram matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ that has the following power series representation in terms of an input gram matrix $\mathbf{X}\mathbf{X}^T$

$$n\mathbf{K} = \sum_{i=0}^{\infty} c_i (\mathbf{X}\mathbf{X}^T)^{\odot i}.$$

Whenever $c_0 \neq 0$ the effective rank of $\mathbf{K}$ is $O(1)$, as displayed in the following theorem.

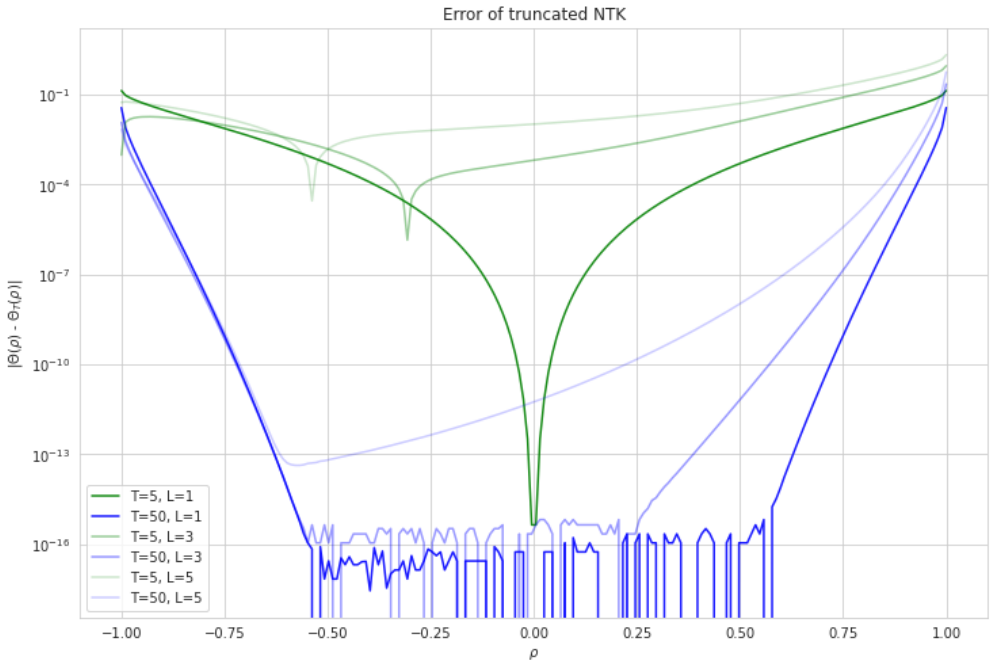

Figure 4.2: **NTK Approximation via Truncation** Absolute error between the analytical ReLU NTK and the truncated ReLU NTK power series as a function of the input correlation ρ for two different values of the truncation point T and three different values for the depth L of the network. Although the truncated NTK achieves a uniform approximation error of only 10^{-1} on $[-1, 1]$, for $|\rho| \leq 0.5$, which we remark is more typical for real world data, $T = 50$ suffices for the truncated NTK to achieve machine level precision.

Theorem 4.4.1. *Assume that we have a kernel Gram matrix* $\mathbf{K}$ *of the form*

$$n\mathbf{K} = \sum_{p=0}^{\infty} c_p (\mathbf{X}\mathbf{X}^T)^{\odot p}$$

where $c_0 \neq 0$. *Furthermore, assume the input data* $\mathbf{x}_i$ *are normalized so that* $\|\mathbf{x}_i\| = 1$ *for all* $i \in [n]$. *Then*

$$\text{eff}(\mathbf{K}) \leq \frac{\sum_{p=0}^{\infty} c_p}{c_0}.$$

Proof. By linearity of trace we have that

$$Tr(n\mathbf{K}) = \sum_{i=0}^{\infty} c_i Tr((\mathbf{X}\mathbf{X}^T)^{\odot i}) = n \sum_{i=0}^{\infty} c_i$$

where we have used the fact that $Tr((\mathbf{X}\mathbf{X}^T)^{\odot i}) = n$ for all $i \in \mathbb{N}$. On the other hand

$$\lambda_1(n\mathbf{K}) \geq \lambda_1(c_0(\mathbf{X}\mathbf{X}^T)^0) = \lambda_1(c_0\mathbf{1}_{n\times n}) = nc_0.$$

Thus we have that

$$\text{eff}(\mathbf{K}) = \frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})} = \frac{Tr(n\mathbf{K})}{\lambda_1(n\mathbf{K})} \leq \frac{\sum_{i=0}^{\infty} c_i}{c_0}.$$

□

The above theorem demonstrates that the constant term $c_0\mathbf{1}_{n\times n}$ in the kernel leads to a significant outlier in the spectrum of $\mathbf{K}$. However this fails to capture how the structure of the input data $\mathbf{X}$ manifests in the spectrum of $\mathbf{K}$. For this we will examine the centered kernel matrix $\widetilde{\mathbf{K}} := \mathbf{K} - \frac{c_0}{n}\mathbf{1}\mathbf{1}^T$. Using a very similar argument as before we can demonstrate that the effective rank of $\widetilde{\mathbf{K}}$ is controlled by the effective rank of the input data gram $\mathbf{X}\mathbf{X}^T$. This is formalized in the following theorem.

Theorem 4.4.3. *Assume that we have a kernel Gram matrix* $\mathbf{K}$ *of the form*

$$n\mathbf{K} = \sum_{p=0}^{\infty} c_p(\mathbf{X}\mathbf{X}^T)^{\odot p}$$

where $c_1 \neq 0$. *Furthermore, assume the input data* $\mathbf{x}_i$ *are normalized so that* $\|\mathbf{x}_i\| = 1$ *for all* $i \in [n]$. *Then the centered kernel* $\widetilde{\mathbf{K}} := \mathbf{K} - \frac{c_0}{n}\mathbf{1}_{n\times n}$ *satisfies*

$$\text{eff}(\widetilde{\mathbf{K}}) \leq \text{eff}(\mathbf{X}\mathbf{X}^T)\frac{\sum_{p=1}^{\infty} c_p}{c_1}.$$

Proof. By the linearity of the trace we have that

$$Tr(n\widetilde{\mathbf{K}}) = \sum_{i=1}^{\infty} c_i Tr((\mathbf{X}\mathbf{X}^T)^{\odot i}) = Tr(\mathbf{X}\mathbf{X}^T)\sum_{i=1}^{\infty} c_i$$

where we have used the fact that $Tr((\mathbf{X}\mathbf{X}^T)^{\odot i}) = Tr(\mathbf{X}\mathbf{X}^T) = n$ for all $i \in [n]$. On the other hand we have that

$$\lambda_1(n\widetilde{\mathbf{K}}) \geq \lambda_1(c_1\mathbf{X}\mathbf{X}^T) = c_1\lambda_1(\mathbf{X}\mathbf{X}^T).$$

Thus we conclude

$$\text{eff}(\widetilde{\mathbf{K}}) = \frac{Tr(\widetilde{\mathbf{K}})}{\lambda_1(\widetilde{\mathbf{K}})} = \frac{Tr(n\widetilde{\mathbf{K}})}{\lambda_1(n\widetilde{\mathbf{K}})} \leq \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \frac{\sum_{i=1}^{\infty} c_i}{c_1}.$$

□

4.6.4 Effective Rank of the NTK for Finite-width Networks

4.6.4.1 Notation and Definitions

We will let $[k] := \{1, 2, \ldots, k\}$. We consider a neural network

$$\sum_{\ell=1}^{m} a_\ell \phi(\langle \mathbf{w}_\ell, \mathbf{x} \rangle)$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{w}_\ell \in \mathbb{R}^d$, $a_\ell \in \mathbb{R}$ for all $\ell \in [m]$ and ϕ is a scalar valued activation function. The network we present here does not have any bias values in the inner-layer, however the results we will prove later apply to the nonzero bias case by replacing $\mathbf{x}$ with $[\mathbf{x}^T, 1]^T$. We let $\mathbf{W} \in \mathbb{R}^{m \times d}$ be the matrix whose ℓ-th row is equal to $\mathbf{w}_\ell$ and $\mathbf{a} \in \mathbb{R}^m$ be the vector whose ℓ-th entry is equal to a_ℓ. We can then write the neural network in vector form

$$f(\mathbf{x}; \mathbf{W}, \mathbf{a}) = \mathbf{a}^T \phi(\mathbf{W}\mathbf{x})$$

where ϕ is understood to be applied entry-wise.

Suppose we have n training data inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$. We will let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the matrix whose i-th row is equal to $\mathbf{x}_i$. Let $\theta_{inner} = vec(\mathbf{W})$ denote the row-wise vectorization of the inner-layer weights. We consider the Jacobian of the neural networks predictions on the training data with respect to the inner layer weights:

$$\mathbf{J}_{inner}^T = \left[\frac{\partial f(\mathbf{x}_1)}{\partial \theta_{inner}}, \frac{\partial f(\mathbf{x}_2)}{\partial \theta_{inner}}, \ldots, \frac{\partial f(\mathbf{x}_n)}{\partial \theta_{inner}} \right]$$

Similarly we can look at the analagous quantity for the outer layer weights

$$\mathbf{J}_{outer}^T = \left[\frac{\partial f(\mathbf{x}_1)}{\partial \mathbf{a}}, \frac{\partial f(\mathbf{x}_2)}{\partial \mathbf{a}}, \ldots, \frac{\partial f(\mathbf{x}_n)}{\partial \mathbf{a}} \right] = \phi\left(\mathbf{W}\mathbf{X}^T\right).$$

Our first observation is that the per-example gradients for the inner layer weights have a nice Kronecker product representation

$$\frac{\partial f(\mathbf{x})}{\partial \theta_{inner}} = \begin{bmatrix} a_1\phi'(\langle \mathbf{w}_1, \mathbf{x}\rangle) \\ a_2\phi'(\langle \mathbf{w}_2, \mathbf{x}\rangle) \\ \cdots \\ a_m\phi'(\langle \mathbf{w}_m, \mathbf{x}\rangle) \end{bmatrix} \otimes \mathbf{x}.$$

For convenience we will let

$$\mathbf{Y}_i := \begin{bmatrix} a_1\phi'(\langle \mathbf{w}_1, \mathbf{x}_i\rangle) \\ a_2\phi'(\langle \mathbf{w}_2, \mathbf{x}_i\rangle) \\ \cdots \\ a_m\phi'(\langle \mathbf{w}_m, \mathbf{x}_i\rangle) \end{bmatrix}.$$

where the dependence of $\mathbf{Y}_i$ on the parameters $\mathbf{W}$ and $\mathbf{a}$ is suppressed (formally $\mathbf{Y}_i = \mathbf{Y}_i(\mathbf{W}, \mathbf{a})$). This way we may write

$$\frac{\partial f(\mathbf{x}_i)}{\partial \theta_{inner}} = \mathbf{Y}_i \otimes \mathbf{x}_i.$$

We will study the NTK with respect to the inner-layer weights

$$\mathbf{K}_{inner} = \mathbf{J}_{inner}\mathbf{J}_{inner}^T$$

and the same quantity for the outer-layer weights

$$\mathbf{K}_{outer} = \mathbf{J}_{outer}\mathbf{J}_{outer}^T.$$

For a hermitian matrix $\mathbf{A}$ we will let $\lambda_i(\mathbf{A})$ denote the ith largest eigenvalue of $\mathbf{A}$ so that $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$. Similarly for an arbitrary matrix $\mathbf{A}$ we will let $\sigma_i(\mathbf{A})$ to the ith largest singular value of $\mathbf{A}$. For a matrix $\mathbf{A} \in \mathbb{R}^{r\times k}$ we will let $\sigma_{min}(\mathbf{A}) = \sigma_{\min(r,k)}$.

4.6.4.2 Effective Rank

For a positive semidefinite matrix $\mathbf{A}$ we define the *effective rank* [HHV22] of $\mathbf{A}$ to be the quantity

$$\text{eff}(\mathbf{A}) := \frac{Tr(\mathbf{A})}{\lambda_1(\mathbf{A})}.$$

The effective rank quantifies how many eigenvalues are on the order of the largest eigenvalue. We have the Markov-like inequality

$$|\{i : \lambda_i(\mathbf{A}) \geq c\lambda_1(\mathbf{A})\}| \leq c^{-1}\frac{Tr(\mathbf{A})}{\lambda_1(\mathbf{A})}$$

and the eigenvalue bound

$$\frac{\lambda_i(\mathbf{A})}{\lambda_1(\mathbf{A})} \leq \frac{1}{i}\frac{Tr(\mathbf{A})}{\lambda_1(\mathbf{A})}.$$

Let $\mathbf{A}$ and $\mathbf{B}$ be positive semidefinite matrices. Then we have

$$\frac{Tr(\mathbf{A}+\mathbf{B})}{\lambda_1(\mathbf{A}+\mathbf{B})} \leq \frac{Tr(\mathbf{A})+Tr(\mathbf{B})}{\max\left(\lambda_1(\mathbf{A}),\lambda_1(\mathbf{B})\right)} \leq \frac{Tr(\mathbf{A})}{\lambda_1(\mathbf{A})} + \frac{Tr(\mathbf{B})}{\lambda_1(\mathbf{B})}.$$

Thus the effective rank is subadditive for positive semidefinite matrices.

We will be interested in bounding the effective rank of the NTK. Let $\mathbf{K} = \mathbf{J}\mathbf{J}^T = \mathbf{J}_{outer}\mathbf{J}_{outer}^T + \mathbf{J}_{inner}\mathbf{J}_{inner}^T = \mathbf{K}_{outer} + \mathbf{K}_{inner}$ be the NTK matrix with respect to all the network parameters. Note that by subadditivity

$$\frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})} \leq \frac{Tr(\mathbf{K}_{outer})}{\lambda_1(\mathbf{K}_{outer})} + \frac{Tr(\mathbf{K}_{inner})}{\lambda_1(\mathbf{K}_{inner})}.$$

In this vein we will control the effective rank of $\mathbf{K}_{inner}$ and $\mathbf{K}_{outer}$ separately.

4.6.4.3 Effective Rank of Inner-layer NTK

We will show that the effective rank of inner-layer NTK is bounded by a multiple of the effective rank of the data input gram $\mathbf{X}\mathbf{X}^T$. We introduce the following meta-theorem that we will use to prove various corollaries later

Theorem 4.6.9. *Set* $\alpha := \sup_{\|\mathbf{b}\|=1} \left[\min_{j\in[n]} |\langle \mathbf{Y}_j, \mathbf{b}\rangle|\right]$. *Assume* $\alpha > 0$. *Then*

$$\frac{\min_{i\in[n]} \|\mathbf{Y}_i\|_2^2 \, Tr(\mathbf{X}\mathbf{X}^T)}{\max_{i\in[n]} \|\mathbf{Y}_i\|_2^2 \, \lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}_{inner})}{\lambda_1\left(\mathbf{K}_{inner}\right)} \leq \frac{\max_{i\in[n]} \|\mathbf{Y}_i\|_2^2}{\alpha^2} \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

Proof. We will first prove the upper bound. We first observe that

$$Tr(\mathbf{K}_{inner}) = \sum_{i=1}^{n} \left\| \frac{\partial f(\mathbf{x}_i)}{\partial \theta_{inner}} \right\|_2^2 = \sum_{i=1}^{n} \|\mathbf{Y}_i \otimes \mathbf{x}_i\|_2^2 = \sum_{i=1}^{n} \|\mathbf{Y}_i\|_2^2 \, \|\mathbf{x}_i\|_2^2$$
$$\leq \max_{j\in[n]} \|\mathbf{Y}_j\|_2^2 \sum_{i=1}^{n} \|\mathbf{x}_i\|_2^2 = \max_{j\in[n]} \|\mathbf{Y}_j\|_2^2 \, Tr(\mathbf{X}\mathbf{X}^T).$$

Recall that

$$\lambda_1\left(\mathbf{K}_{inner}\right) = \lambda_1\left(\mathbf{J}_{inner}\mathbf{J}_{inner}^T\right) = \lambda_1\left(\mathbf{J}_{inner}^T\mathbf{J}_{inner}\right).$$

Well

$$\mathbf{J}_{inner}^T\mathbf{J}_{inner} = \sum_{i=1}^{n} \frac{\partial f(\mathbf{x}_i)}{\partial \theta_{inner}} \frac{\partial f(\mathbf{x}_i)}{\partial \theta_{inner}}^T = \sum_{i=1}^{n} \left[\mathbf{Y}_i \otimes \mathbf{x}_i\right] \left[\mathbf{Y}_i \otimes \mathbf{x}_i\right]^T$$
$$= \sum_{i=1}^{n} \left[\mathbf{Y}_i\mathbf{Y}_i^T\right] \otimes \left[\mathbf{x}_i\mathbf{x}_i^T\right].$$

Well then we may use the fact that

$$\lambda_1(\mathbf{J}_{inner}^T\mathbf{J}_{inner}) = \max_{\|\mathbf{b}\|_2=1} \mathbf{b}^T\mathbf{J}_{inner}^T\mathbf{J}_{inner}\mathbf{b}.$$

Let $\mathbf{b}_1 \in \mathbb{R}^m$ and $\mathbf{b}_2 \in \mathbb{R}^d$ be vectors that we will optimize later satisfying $\|\mathbf{b}_1\|_2 \|\mathbf{b}_2\|_2 = 1$. Then we have that $\|\mathbf{b}_1 \otimes \mathbf{b}_2\| = 1$ and

$$(\mathbf{b}_1 \otimes \mathbf{b}_2)^T\mathbf{J}_{inner}^T\mathbf{J}_{inner}(\mathbf{b}_1 \otimes \mathbf{b}_2) = \sum_{i=1}^{n} (\mathbf{b}_1 \otimes \mathbf{b}_2)^T \left(\left[\mathbf{Y}_i\mathbf{Y}_i^T\right] \otimes \left[\mathbf{x}_i\mathbf{x}_i^T\right]\right) (\mathbf{b}_1 \otimes \mathbf{b}_2)$$
$$= \sum_{i=1}^{n} \left[\mathbf{b}_1^T\mathbf{Y}_i\mathbf{Y}_i^T\mathbf{b}_1\right] \left[\mathbf{b}_2^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{b}_2\right] \geq \left[\min_{j\in[n]} \mathbf{b}_1^T\mathbf{Y}_j\mathbf{Y}_j^T\mathbf{b}_1\right] \sum_{i=1}^{n} \mathbf{b}_2^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{b}_2$$
$$= \left[\min_{j\in[n]} \mathbf{b}_1^T\mathbf{Y}_j\mathbf{Y}_j^T\mathbf{b}_1\right] \mathbf{b}_2^T \left[\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\right] \mathbf{b}_2 = \left[\min_{j\in[n]} \mathbf{b}_1^T\mathbf{Y}_j\mathbf{Y}_j^T\mathbf{b}_1\right] \mathbf{b}_2\mathbf{X}^T\mathbf{X}\mathbf{b}_2.$$

Pick $\mathbf{b}_2$ so that $\|\mathbf{b}_2\| = 1$ and

$$\mathbf{b}_2\mathbf{X}^T\mathbf{X}\mathbf{b}_2 = \lambda_1(\mathbf{X}^T\mathbf{X}) = \lambda_1(\mathbf{X}\mathbf{X}^T).$$

Thus for this choice of $\mathbf{b}_2$ we have

$$\lambda_1(\mathbf{J}_{inner}^T\mathbf{J}_{inner}) \geq (\mathbf{b}_1 \otimes \mathbf{b}_2)^T\mathbf{J}_{inner}^T\mathbf{J}_{inner}(\mathbf{b}_1 \otimes \mathbf{b}_2) \geq$$
$$\left[\min_{j\in[n]} \mathbf{b}_1^T\mathbf{Y}_j\mathbf{Y}_j^T\mathbf{b}_1\right] \mathbf{b}_2\mathbf{X}^T\mathbf{X}\mathbf{b}_2 = \left[\min_{j\in[n]} \mathbf{b}_1^T\mathbf{Y}_j\mathbf{Y}_j^T\mathbf{b}_1\right] \lambda_1(\mathbf{X}\mathbf{X}^T).$$

Now note that $\alpha^2 = \sup_{\|\mathbf{b}_1\|=1} \left[\min_{j\in[n]} \mathbf{b}_1^T\mathbf{Y}_j\mathbf{Y}_j^T\mathbf{b}_1\right]$. Thus by taking the sup over $\mathbf{b}_1$ in our previous bound we have

$$\lambda_1(\mathbf{K}_{inner}) = \lambda_1(\mathbf{J}_{inner}^T\mathbf{J}_{inner}) \geq \alpha^2\lambda_1(\mathbf{X}\mathbf{X}^T).$$

Thus combined with our previous result we have

$$\frac{Tr(\mathbf{K}_{inner})}{\lambda_1(\mathbf{K}_{inner})} \leq \frac{\max_{i\in[n]} \|\mathbf{Y}_i\|_2^2}{\alpha^2} \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

We now prove the lower bound.

$$Tr(\mathbf{K}_{inner}) = \sum_{i=1}^n \left\| \frac{\partial f(\mathbf{x}_i)}{\partial \theta_{inner}} \right\|_2^2 = \sum_{i=1}^n \|\mathbf{Y}_i \otimes \mathbf{x}_i\|_2^2 = \sum_{i=1}^n \|\mathbf{Y}_i\|_2^2 \|\mathbf{x}_i\|_2^2$$
$$\geq \min_{j\in[n]} \|\mathbf{Y}_j\|_2^2 \sum_{i=1}^n \|\mathbf{x}_i\|_2^2 = \min_{j\in[n]} \|\mathbf{Y}_j\|_2^2 Tr(\mathbf{X}\mathbf{X}^T).$$

Let $\mathbf{Y} \in \mathbb{R}^{n\times m}$ be the matrix whose ith row is equal to $\mathbf{Y}_i$. Then observe that

$$\mathbf{K}_{inner} = [\mathbf{Y}\mathbf{Y}^T] \odot [\mathbf{X}\mathbf{X}^T]$$

where $\odot$ denotes the entry-wise Hadamard product of two matrices. We now recall that if $\mathbf{A}$ and $\mathbf{B}$ are two positive semidefinite matrices we have [OS20, Lemma 2]

$$\lambda_1(\mathbf{A} \odot \mathbf{B}) \leq \max_{i\in[n]} \mathbf{A}_{i,i}\lambda_1(\mathbf{B}).$$

Applying this to $\mathbf{K}_{inner}$ we get that

$$\lambda_1(\mathbf{K}_{inner}) \leq \max_{i\in[n]} \|\mathbf{Y}_i\|_2^2 \lambda_1(\mathbf{X}\mathbf{X}^T).$$

Combining this with our previous result we get

$$\frac{\min_{i\in[n]} \|\mathbf{Y}_i\|_2^2 Tr(\mathbf{X}\mathbf{X}^T)}{\max_{i\in[n]} \|\mathbf{Y}_i\|_2^2 \lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}_{inner})}{\lambda_1(\mathbf{K}_{inner})}.$$

□

We can immediately get a useful corollary that applies to the ReLU activation function

Corollary 4.6.10. *Set* $\alpha := \sup_{\|\mathbf{b}\|=1} \left[\min_{j\in[n]} |\langle \mathbf{Y}_j, \mathbf{b}\rangle|\right]$ *and* $\gamma_{max} := \sup_{x\in\mathbb{R}} |\phi'(x)|$. *Assume* $\alpha > 0$ *and* $\gamma_{max} < \infty$. *Then*

$$\frac{\alpha^2}{\gamma_{max}^2 \|\mathbf{a}\|_2^2} \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}_{inner})}{\lambda_1(\mathbf{K}_{inner})} \leq \frac{\gamma_{max}^2 \|\mathbf{a}\|_2^2}{\alpha^2} \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

Proof. Note that the hypothesis on $|\phi'|$ gives $\|\mathbf{Y}_i\|_2^2 \leq \gamma_{max}^2 \|\mathbf{a}\|_2^2$ for all $i \in [n]$. Moreover by Cauchy-Schwarz we have that $\min_{i\in[n]} \|\mathbf{Y}_i\|_2 \geq \alpha$. Thus by theorem 4.6.9 we get the desired result. □

If ϕ is a leaky ReLU type activation (say like those used in [NM20]) Theorem 4.6.9 translates into an even simpler bound

Corollary 4.6.11. *Suppose* $\phi'(x) \in [\gamma_{min}, \gamma_{max}]$ *for all* $x \in \mathbb{R}$ *where* $\gamma_{min} > 0$. *Then*

$$\frac{\gamma_{min}^2 Tr(\mathbf{X}\mathbf{X}^T)}{\gamma_{max}^2 \lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}_{inner})}{\lambda_1(\mathbf{K}_{inner})} \leq \frac{\gamma_{max}^2}{\gamma_{min}^2} \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

Proof. We will lower bound

$$\alpha := \sup_{\|\mathbf{b}\|=1} \left[\min_{j\in[n]} |\langle \mathbf{Y}_j, \mathbf{b}\rangle|\right]$$

so that we can apply Corollary 4.6.10. Set $\mathbf{b} = \mathbf{a}/\|\mathbf{a}\|_2$. Then we have that

$$\langle \mathbf{Y}_j, \mathbf{b}\rangle = \sum_{\ell=1}^m a_\ell \phi'(\langle \mathbf{w}_\ell, \mathbf{x}_j\rangle) a_\ell / \|\mathbf{a}\|_2 \geq \frac{\gamma_{min}}{\|\mathbf{a}\|_2} \sum_{\ell=1}^m a_\ell^2 = \gamma_{min} \|\mathbf{a}\|_2 .$$

Thus $\alpha \geq \gamma_{min} \|\mathbf{a}\|_2$. The result then follows from Corollary 4.6.10. □

To control α in Theorem 4.6.9 when ϕ is the ReLU activation function requires a bit more work. To this end we introduce the following lemma.

Lemma 4.6.12. *Assume* $\phi(x) = ReLU(x)$. *Let* $R_{min}, R_{max} > 0$ *and define* $\tau = \{\ell \in [m] : |a_\ell| \in [R_{min}, R_{max}]\}$. *Set* $T = \min_{i\in[n]} \sum_{\ell\in\tau} \mathbb{I}[\langle \mathbf{x}_i, \mathbf{w}_\ell\rangle \geq 0]$. *Then*

$$\alpha := \sup_{\|\mathbf{b}\|=1} \left[\min_{i\in[n]} |\langle \mathbf{Y}_i, \mathbf{b}\rangle|\right] \geq \frac{R_{min}^2}{R_{max}} \frac{T}{|\tau|^{1/2}}.$$

Proof. Let $\mathbf{a}_\tau$ be the vector such that $(\mathbf{a}_\tau)_\ell = a_\ell \mathbb{I}[\ell \in \tau]$. Then note that

$$\langle \mathbf{Y}_j, \mathbf{a}_\tau / \|\mathbf{a}_\tau\|_2 \rangle = \frac{1}{\|\mathbf{a}_\tau\|} \sum_{\ell \in \tau} a_\ell^2 \mathbb{I}[\langle \mathbf{w}_\ell, \mathbf{x}_j \rangle \geq 0] \geq$$
$$\frac{R_{min}^2}{\|\mathbf{a}_\tau\|} \sum_{\ell \in \tau} \mathbb{I}[\langle \mathbf{w}_\ell, \mathbf{x}_j \rangle \geq 0] \geq \frac{R_{min}^2}{\|\mathbf{a}_\tau\|_2} T \geq \frac{R_{min}^2}{R_{max} |\tau|^{1/2}} T.$$

□

Roughly what Lemma 4.6.12 says is that α is controlled when there is a set of inner-layer neurons that are active for each data point whose outer layer weights are similar in magnitude. Note that in [DZP19], [ADH19a], [OFL19], [LSO20], [XLS17] and [OS20] the outer layer weights all have fixed constant magnitude. Thus in that case we can set $R_{min} = R_{max}$ in Lemma 4.6.12 so that $\tau = [m]$. In this setting we have the following result.

Theorem 4.6.13. *Assume $\phi(x) = ReLU(x)$. Suppose $|a_\ell| = R > 0$ for all $\ell \in [m]$. Furthermore suppose $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are independent random vectors such that $\mathbf{w}_\ell / \|\mathbf{w}_\ell\|$ has the uniform distribution on the sphere for each $\ell \in [m]$. Also assume $m \geq \frac{4 \log(n/\epsilon)}{\delta^2}$ for some $\delta, \epsilon \in (0,1)$. Then with probability at least $1 - \epsilon$ we have that*

$$\frac{(1-\delta)^2}{4} \text{eff}(\mathbf{X}\mathbf{X}^T) \leq \text{eff}(\mathbf{K}_{inner}) \leq \frac{4}{(1-\delta)^2} \text{eff}(\mathbf{X}\mathbf{X}^T).$$

Proof. Fix $j \in [n]$. Note by the assumption on the $\mathbf{w}_\ell$'s we have that

$$\mathbb{I}[\langle \mathbf{w}_1, \mathbf{x}_j \rangle \geq 0], \ldots, \mathbb{I}[\langle \mathbf{w}_m, \mathbf{x}_j \rangle \geq 0]$$

are i.i.d. Bernouilli random variables taking the values 0 and 1 with probability 1/2. Thus by the Chernoff bound for Binomial random variables we have that

$$\mathbb{P}\left(\sum_{\ell=1}^m \mathbb{I}[\langle \mathbf{w}_\ell, \mathbf{x}_j \rangle \geq 0] \leq \frac{m}{2}(1-\delta) \right) \leq \exp\left(-\delta^2 \frac{m}{4} \right).$$

Thus taking the union bound over every $j \in [n]$ we get that if $m \geq \frac{4 \log(n/\epsilon)}{\delta^2}$ then

$$\min_{j \in [n]} \sum_{\ell=1}^m \mathbb{I}[\langle \mathbf{w}_\ell, \mathbf{x}_j \rangle \geq 0] \geq \frac{m}{2}(1-\delta)$$

holds with probability at least $1-\epsilon$. Now note that if we set $R_{min} = R_{max} = R$ we have that $\tau = [m]$ where τ is defined as it is in Lemma 4.6.12. In this case by our previous bound we have that T as defined in Lemma 4.6.12 satisfies $T \geq \frac{m}{2}(1-\delta)$ with probability at least $1-\epsilon$. In this case the conclusion of Lemma 4.6.12 gives us

$$\alpha \geq Rm^{1/2}\frac{(1-\delta)}{2} = \|\mathbf{a}\|_2 \frac{(1-\delta)}{2}.$$

Thus by Corollary 4.6.10 and the above bound for α we get the desired result. □

We will now use Lemma 4.6.12 to prove a bound in the case of Gaussian initialization.

Lemma 4.6.14. *Assume $\phi(x) = ReLU(x)$. Suppose that $a_\ell \sim N(0,\nu^2)$ for each $\ell \in [m]$ i.i.d. Furthermore suppose $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are random vectors independent of each other and $\mathbf{a}$ such that $\mathbf{w}_\ell / \|\mathbf{w}_\ell\|$ has the uniform distribution on the sphere for each $\ell \in [m]$. Set $p = \mathbb{P}_{z\sim N(0,1)}(|z| \in [1/2,1]) \approx 0.3$. Assume*

$$m \geq \frac{4\log(n/\epsilon)}{\delta^2(1-\delta)p}$$

for some $\epsilon, \delta \in (0,1)$. Then with probability at least $(1-\epsilon)^2$ we have that

$$\alpha := \sup_{\|\mathbf{b}\|=1}\left[\min_{i\in[n]}|\langle \mathbf{Y}_i, \mathbf{b}\rangle|\right] \geq \frac{\nu}{8}(1-\delta)^{3/2}p^{1/2}m^{1/2}.$$

Proof. Set $R_{min} = \nu/2$ and $R_{max} = \nu$. Now set

$$p = \mathbb{P}_{a\sim N(0,\nu^2)}(|a| \in [R_{min}, R_{max}]) = 2\mathbb{P}_{z\sim N(0,1)}\left(z \in \left[\frac{R_{min}}{\nu}, \frac{R_{max}}{\nu}\right]\right)$$
$$= 2\mathbb{P}_{z\sim N(0,1)}(z \in [1/2,1]) \approx 0.3.$$

Now define $\tau = \{\ell \in [m] : |a_\ell| \in [R_{min}, R_{max}]\}$. We have by the Chernoff bound for binomial random variables

$$\mathbb{P}(|\tau| \leq (1-\delta)mp) \leq \exp\left(-\delta^2\frac{mp}{2}\right).$$

Thus if $m \geq \log\left(\frac{1}{\epsilon}\right)\frac{2}{p\delta^2}$ (a weaker condition than the hypothesis on m) then we have that $|\tau| \geq (1-\delta)mp$ with probability at least $1-\epsilon$. From now on assume such a τ has been

observed and view it as fixed so that the only remaining randomness is over the $\mathbf{w}_\ell$'s. Now set $T = \min_{i\in[n]} \sum_{\ell\in\tau} \mathbb{I}\left[\langle \mathbf{x}_i, \mathbf{w}_\ell\rangle \geq 0\right]$. By the Chernoff bound again we get that for fixed $i \in [n]$

$$\mathbb{P}\left(\sum_{\ell\in\tau} \mathbb{I}\left[\langle \mathbf{x}_i, \mathbf{w}_\ell\rangle \geq 0\right] \leq \frac{(1-\delta)}{2}|\tau|\right) \leq \exp\left(-\delta^2\frac{|\tau|}{4}\right).$$

Thus by taking the union bound over $i \in [n]$ we get

$$\begin{aligned}\mathbb{P}\left(T \leq \frac{(1-\delta)}{2}|\tau|\right) &\leq n\exp\left(-\delta^2\frac{|\tau|}{4}\right)\\ &\leq n\exp\left(-\delta^2\frac{(1-\delta)mp}{4}\right).\end{aligned}$$

Thus if we consider τ as fixed and $m \geq \frac{4\log(n/\epsilon)}{\delta^2(1-\delta)p}$ then with probability at least $1-\epsilon$ over the sampling of the $\mathbf{w}_\ell$'s we have that

$$T \geq \frac{(1-\delta)}{2}|\tau|.$$

In this case by lemma 4.6.12 we have that

$$\begin{aligned}\alpha := \sup_{\|\mathbf{b}\|=1}\left[\min_{i\in[n]}|\langle \mathbf{Y}_i, \mathbf{b}\rangle|\right] &\geq \frac{R_{min}^2}{R_{max}}\frac{T}{|\tau|^{1/2}}\\ &\geq \frac{\nu}{8}(1-\delta)^{3/2}m^{1/2}p^{1/2}.\end{aligned}$$

Thus the above holds with probability at least $(1-\epsilon)^2$. □

This lemma now allows us to bound the effective rank of $\mathbf{K}_{inner}$ in the case of Gaussian initialization.

Theorem 4.6.15. *Assume $\phi(x) = ReLU(x)$. Suppose that $a_\ell \sim N(0,\nu^2)$ for each $\ell \in [m]$ i.i.d. Furthermore suppose $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are random vectors independent of each other and $\mathbf{a}$ such that $\mathbf{w}_\ell / \|\mathbf{w}_\ell\|$ has the uniform distribution on the sphere for each $\ell \in [m]$. Set $p = \mathbb{P}_{z\sim N(0,1)}(|z| \in [1/2, 1]) \approx 0.3$. Let $\epsilon, \delta \in (0,1)$. Then there exists absolute constants $c, K > 0$ such that if*

$$m \geq \frac{4\log(n/\epsilon)}{\delta^2(1-\delta)p}$$

then with probability at least $1 - 3\epsilon$ *we have that*

$$\frac{1}{C}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}_{inner})}{\lambda_1\left(\mathbf{K}_{inner}\right)} \leq C\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}$$

where

$$C = \frac{64}{(1-\delta)^3 p}\left[1 + \frac{\max\{c^{-1}K\log(1/\epsilon), mK\}}{m}\right].$$

Proof. By Bernstein's inequality

$$\mathbb{P}\left(\|\mathbf{a}/\nu\|_2^2 - m \geq t\right) \leq \exp\left[-c \cdot \min\left(\frac{t^2}{mK^2}, \frac{t}{K}\right)\right]$$

where c is an absolute constant. Set $t = \max\{c^{-1}K\log(1/\epsilon), mK\}$ so that the right hand side of the above inequality is bounded by ϵ. Thus by Lemma 4.6.14 and the union bound we can ensure that with probability at least

$$1 - \epsilon - [1 - (1-\epsilon)^2] = 1 - 3\epsilon + \epsilon^2 \geq 1 - 3\epsilon$$

that $\|\mathbf{a}/\nu\|_2^2 \leq m + t$ and the conclusion of Lemma 4.6.14 hold simultaneously. In that case

$$\frac{\|\mathbf{a}\|_2^2}{\alpha^2} \leq \frac{\nu^2[m+t]}{\frac{\nu^2}{64}(1-\delta)^3 mp} = \frac{64}{(1-\delta)^3 p}\left[1 + \frac{t}{m}\right] = C.$$

Thus by Corollary 4.6.10 we get the desired result. □

By fixing $\delta > 0$ in the previous theorem we get the immediate corollary

Corollary 4.6.16. *Assume* $\phi(x) = ReLU(x)$. *Suppose that* $a_\ell \sim N(0, \nu^2)$ *for each* $\ell \in [m]$ *i.i.d. Furthermore suppose* $\mathbf{w}_1, \ldots, \mathbf{w}_m$ *are random vectors independent of each other and* $\mathbf{a}$ *such that* $\mathbf{w}_\ell / \|\mathbf{w}_\ell\|$ *has the uniform distribution on the sphere for each* $\ell \in [m]$. *Then there exists an absolute constant* $C > 0$ *such that* $m = \Omega(\log(n/\epsilon))$ *ensures that with probability at least* $1 - \epsilon$

$$\frac{1}{C}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}_{inner})}{\lambda_1\left(\mathbf{K}_{inner}\right)} \leq C\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

4.6.4.4 Effective Rank of Outer-layer NTK

Throughout this section $\phi(x) = ReLU(x)$. Our goal of this section, similar to before, is to bound the effective rank of $\mathbf{K}_{outer}$ by the effective rank of the input data gram $\mathbf{X}\mathbf{X}^T$. In this section we will use often make use of the basic identities

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_F$$

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$$

$$Tr(\mathbf{A}\mathbf{A}^T) = Tr(\mathbf{A}^T\mathbf{A}) = \|\mathbf{A}\|_F^2$$

$$\|\mathbf{A}\|_2 = \left\|\mathbf{A}^T\right\|_2$$

$$\lambda_1(\mathbf{A}^T\mathbf{A}) = \lambda_1(\mathbf{A}\mathbf{A}^T) = \|\mathbf{A}\|_2^2.$$

To begin bounding the effective rank of $\mathbf{K}_{outer}$, we prove the following lemma.

Lemma 4.6.17. *Assume $\phi(x) = ReLU(x)$ and $\mathbf{W}$ is full rank with $m \geq d$. Then*

$$\frac{\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2}{\left[\|\phi(\mathbf{W}\mathbf{X}^T)\|_2 + \|\phi(-\mathbf{W}\mathbf{X}^T)\|_2\right]^2} \leq \frac{\|\mathbf{W}\|_2^2}{\sigma_{min}(\mathbf{W})^2} \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

Proof. First note that

$$\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2 \leq \left\|\mathbf{W}\mathbf{X}^T\right\|_F^2 \leq \|\mathbf{W}\|_2^2 \left\|\mathbf{X}^T\right\|_F^2 = \|\mathbf{W}\|_2^2 Tr(\mathbf{X}\mathbf{X}^T).$$

Pick $\mathbf{b} \in \mathbb{R}^d$ such that $\|\mathbf{b}\|_2 = 1$ and $\|\mathbf{X}\mathbf{b}\|_2 = \|\mathbf{X}\|_2$. Since $\mathbf{W}^T$ is full rank we may set $\mathbf{u} = (\mathbf{W}^T)^\dagger \mathbf{b}$ so that $\mathbf{W}^T\mathbf{u} = \mathbf{b}$ where $\|\mathbf{u}\|_2 \leq \sigma_{min}(\mathbf{W}^T)^{-1}$ where $\sigma_{min}(\mathbf{W}^T)$ is the smallest *nonzero* singular value of $\mathbf{W}^T$. Well then

$$\begin{aligned}\|\mathbf{X}\|_2 = \|\mathbf{X}\mathbf{b}\|_2 = \left\|\mathbf{X}\mathbf{W}^T\mathbf{u}\right\|_2 \leq \left\|\mathbf{X}\mathbf{W}^T\right\|_2 \|\mathbf{u}\|_2 \leq \left\|\mathbf{X}\mathbf{W}^T\right\|_2 \sigma_{min}(\mathbf{W}^T)^{-1} \\ = \left\|\mathbf{W}\mathbf{X}^T\right\|_2 \sigma_{min}(\mathbf{W})^{-1}.\end{aligned}$$

Now using the fact that $x = \phi(x) - \phi(-x)$ we have that

$$\left\|\mathbf{W}\mathbf{X}^T\right\|_2 = \left\|\phi(\mathbf{W}\mathbf{X}^T) - \phi(-\mathbf{W}\mathbf{X}^T)\right\|_2 \leq \left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2 + \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_2.$$

Thus combined with our previous results gives

$$\|\mathbf{X}\|_2 \leq \sigma_{min}(\mathbf{W})^{-1}\left[\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2 + \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_2\right].$$

Therefore

$$\begin{aligned}\frac{\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2}{\sigma_{min}(\mathbf{W})^{-2}\left[\|\phi(\mathbf{W}\mathbf{X}^T)\|_2 + \|\phi(-\mathbf{W}\mathbf{X}^T)\|_2\right]^2} &\leq \frac{\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2}{\|\mathbf{X}\|_2^2} \\ &\leq \frac{\|\mathbf{W}\|_2^2 Tr(\mathbf{X}\mathbf{X}^T)}{\|\mathbf{X}\|_2^2} = \|\mathbf{W}\|_2^2 \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}\end{aligned}$$

which gives us the desired result. □

Corollary 4.6.18. *Assume $\phi(x) = ReLU(x)$ and $\mathbf{W}$ is full rank with $m \geq d$. Then*

$$\frac{\max\left(\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2, \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_F^2\right)}{\max\left(\|\phi(\mathbf{W}\mathbf{X}^T)\|_2^2, \|\phi(-\mathbf{W}\mathbf{X}^T)\|_2^2\right)} \leq 4\frac{\|\mathbf{W}\|_2^2}{\sigma_{min}(\mathbf{W})^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

Proof. Using the fact that

$$\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2 + \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_2 \leq 2\max\left(\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2, \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_2\right)$$

and lemma 4.6.17 we have that

$$\frac{\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2}{4\max\left(\|\phi(\mathbf{W}\mathbf{X}^T)\|_2^2, \|\phi(-\mathbf{W}\mathbf{X}^T)\|_2^2\right)} \leq \frac{\|\mathbf{W}\|_2^2}{\sigma_{min}(\mathbf{W})^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

Note that the right hand side and the denominator of the left hand side do not change when you replace $\mathbf{W}$ with $-\mathbf{W}$. Therefore by using the above bound for both $\mathbf{W}$ and $-\mathbf{W}$ as the weight matrix separately we can conclude

$$\frac{\max\left(\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2, \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_F^2\right)}{4\max\left(\|\phi(\mathbf{W}\mathbf{X}^T)\|_2^2, \|\phi(-\mathbf{W}\mathbf{X}^T)\|_2^2\right)} \leq \frac{\|\mathbf{W}\|_2^2}{\sigma_{min}(\mathbf{W})^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

□

Corollary 4.6.19. *Assume $\phi(x) = ReLU(x)$ and $m \geq d$. Suppose $\mathbf{W}$ and $-\mathbf{W}$ have the same distribution. Then conditioned on $\mathbf{W}$ being full rank we have that with probability at least $1/2$*

$$\frac{Tr(\mathbf{K}_{outer})}{\lambda_1(\mathbf{K}_{outer})} \leq 4\frac{\|\mathbf{W}\|_2^2}{\sigma_{min}(\mathbf{W})^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

Proof. Fix $\mathbf{W}$ where $\mathbf{W}$ is full rank. We have by corollary 4.6.18 that either

$$\frac{\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2}{\|\phi(\mathbf{W}\mathbf{X}^T)\|_2^2} \leq 4\frac{\|\mathbf{W}\|_2^2}{\sigma_{min}(\mathbf{W})^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}$$

holds or

$$\frac{\left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_F^2}{\|\phi(-\mathbf{W}\mathbf{X}^T)\|_2^2} \leq 4\frac{\|\mathbf{W}\|_2^2}{\sigma_{min}(\mathbf{W})^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}$$

(the first holds in the case where $\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2^2 \geq \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_2^2$ and the second in the case $\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2^2 < \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_2^2$). Since $\mathbf{W}$ and $-\mathbf{W}$ have the same distribution, it follows that the first inequality must hold at least $1/2$ of the time. From

$$\frac{Tr(\mathbf{K}_{outer})}{\lambda_1(\mathbf{K}_{outer})} = \frac{\left\|\mathbf{J}_{outer}^T\right\|_F^2}{\|\mathbf{J}_{outer}^T\|_2^2} = \frac{\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2}{\|\phi(\mathbf{W}\mathbf{X}^T)\|_2^2}$$

we get the desired result. □

We now note that when $\mathbf{W}$ is rectangular shaped and the entries of $\mathbf{W}$ are i.i.d. Gaussians that $\mathbf{W}$ is full rank with high probability and $\sigma_{min}(\mathbf{W})^{-2}\|\mathbf{W}\|_2^2$ is well behaved. We recall the result from [Ver12]:

Theorem 4.6.20. *Let $\mathbf{A}$ be a $N \times n$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$ one has*

$$\sqrt{N} - \sqrt{n} - t \leq \sigma_{min}(\mathbf{A}) \leq \sigma_1(\mathbf{A}) \leq \sqrt{N} + \sqrt{n} + t.$$

Corollary 4.6.19 gives us a bound that works at least half the time. However, we would like to derive a bound that holds with high probability. We will have that when $m \gtrsim n$ we have sufficient concentration of the largest singular value of $\phi(\mathbf{W}\mathbf{X}^T)$ to prove such a bound. We recall the result from [Ver12] (Remark 5.40):

Theorem 4.6.21. *Assume that* $\mathbf{A}$ *is an* $N \times n$ *matrix whose rows* $\mathbf{A}_i$ *are independent subgaussian random vectors in* $\mathbb{R}^n$ *with second moment matrix* Σ*. Then for every* $t \geq 0$*, the following inequality holds with probability at least* $1 - 2\exp(-ct^2)$

$$\left\| \frac{1}{N}\mathbf{A}^*\mathbf{A} - \Sigma \right\|_2 \leq \max(\delta, \delta^2) \quad \textit{where} \quad \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}$$

where $C = C_K, c = c_K > 0$ *depend only on* $K := \max_i \|\mathbf{A}_i\|_{\psi_2}$.

We will use theorem 4.6.21 in the following lemma.

Lemma 4.6.22. *Assume* $\phi(x) = ReLU(x)$*. Let* $\mathbf{A} = \phi(\mathbf{W}\mathbf{X}^T)$ *and* $M = \max_{i\in[n]} \|\mathbf{x}_i\|_2$*. Suppose that* $\mathbf{w}_1, \ldots, \mathbf{w}_m \sim N(0, \nu^2 I_d)$ *i.i.d. Set* $K = M\nu\sqrt{n}$ *and define*

$$\Sigma := \mathbb{E}_{\mathbf{w}\sim N(0,\nu^2 I)}[\phi(\mathbf{X}\mathbf{w})\phi(\mathbf{w}^T\mathbf{X}^T)].$$

Then for every $t \geq 0$ *the following inequality holds with probability at least* $1 - 2\exp(-c_K t^2)$

$$\left\| \frac{1}{m}\mathbf{A}^T\mathbf{A} - \Sigma \right\|_2 \leq \max(\delta, \delta^2) \quad \textit{where} \quad \delta = C_K\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}},$$

where $c_K, C_K > 0$ *are absolute constants that depend only on* K*.*

Proof. We will let $\mathbf{A}_{\ell:}$ denote the ℓth row of $\mathbf{A}$ (considered as a column vector). Note that

$$\mathbf{A}_{\ell:} = \phi(\mathbf{X}\mathbf{w}_\ell).$$

We immediately get that the rows of $\mathbf{A}$ are i.i.d. We will now bound $\|\mathbf{A}_{\ell:}\|_{\psi_2}$. Let $\mathbf{b} \in \mathbb{R}^n$ such that $\|\mathbf{b}\|_2 = 1$. Then

$$\begin{aligned}
\|\langle \phi(\mathbf{X}\mathbf{w}_\ell), \mathbf{b}\rangle\|_{\psi_2} &= \left\| \sum_{i=1}^n \phi(\langle \mathbf{x}_i, \mathbf{w}_\ell\rangle) b_i \right\|_{\psi_2} \\
&\leq \sum_{i=1}^n |b_i| \, \|\phi(\langle \mathbf{x}_i, \mathbf{w}_\ell\rangle)\|_{\psi_2} \leq \sum_{i=1}^n |b_i| \, \|\langle \mathbf{x}_i, \mathbf{w}_\ell\rangle\|_{\psi_2} \\
&\leq \sum_{i=1}^n |b_i| C \|\mathbf{x}_i\|_2 \nu \leq CM\nu \|\mathbf{b}\|_1 \leq CM\nu\sqrt{n}
\end{aligned}$$

where $C > 0$ is an absolute constant. Set $K := M\nu\sqrt{n}$. Well then by theorem 4.6.21 we have the following. For every $t \geq 0$ the following inequality holds with probability at least $1 - 2\exp(-c_K t^2)$

$$\left\| \frac{1}{m}\mathbf{A}^T\mathbf{A} - \Sigma \right\|_2 \leq \max(\delta, \delta^2) \quad \text{where} \quad \delta = C_K\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}.$$

□

We are now ready to prove a high probability bound for the effective rank of $\mathbf{K}_{outer}$.

Theorem 4.6.23. *Assume $\phi(x) = ReLU(x)$ and $m \geq d$. Let $M = \max_{i\in[n]} \|\mathbf{x}_i\|_2$. Suppose that $\mathbf{w}_1, \ldots, \mathbf{w}_m \sim N(0, \nu^2 I_d)$ i.i.d. Set $K = M\nu\sqrt{n}$*

$$\Sigma := \mathbb{E}_{\mathbf{w}\sim N(0,\nu^2 I)}[\phi(\mathbf{X}\mathbf{w})\phi(\mathbf{w}^T\mathbf{X}^T)]$$

$$\delta = C_K\left[\sqrt{\frac{n}{m}} + \sqrt{\frac{\log(2/\epsilon)}{m}}\right]$$

where $\epsilon > 0$ is small. Now assume

$$\sqrt{m} > \sqrt{d} + \sqrt{2\log(2/\epsilon)}$$

and

$$\max(\delta, \delta^2) \leq \frac{1}{2}\lambda_1(\Sigma).$$

Then with probability at least $1 - 3\epsilon$

$$\frac{Tr(\mathbf{K}_{outer})}{\lambda_1(\mathbf{K}_{outer})} \leq 12\left(\frac{\sqrt{m} + \sqrt{d} + t_1}{\sqrt{m} - \sqrt{d} - t_1}\right)^2 \frac{Tr(\mathbf{X}^T\mathbf{X})}{\lambda_1(\mathbf{X}^T\mathbf{X})}.$$

Proof. By theorem 4.6.20 with $t_1 = \sqrt{2\log(2/\epsilon)}$ we have that with probability at least $1 - \epsilon$ that

$$\sqrt{m} - \sqrt{d} - t_1 \leq \sigma_{min}(\mathbf{W}/\nu) \leq \sigma_1(\mathbf{W}/\nu) \leq \sqrt{m} + \sqrt{d} + t_1. \tag{4.43}$$

The above inequalities and the hypothesis on m imply that $\mathbf{W}$ is full rank.

Let $\mathbf{A} = \phi(\mathbf{W}\mathbf{X}^T)$ and $\tilde{\mathbf{A}} = \phi(-\mathbf{W}\mathbf{X}^T)$. Set $t_2 = \sqrt{\frac{\log(2/\epsilon)}{c_K}}$ where c_K is defined as in theorem 4.6.22. Note that $\mathbf{A}$ and $\tilde{\mathbf{A}}$ are identical in distribution. Thus by theorem 4.6.22 and the union bound we get that with probability at least $1 - 2\epsilon$

$$\left\|\frac{1}{m}\mathbf{A}^T\mathbf{A} - \Sigma\right\|_2, \left\|\frac{1}{m}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} - \Sigma\right\|_2 \leq \max(\delta, \delta^2) =: \rho \tag{4.44}$$

where

$$\delta = C_K\sqrt{\frac{n}{m}} + \frac{t_2}{\sqrt{m}}.$$

By our previous results and the union bound we can ensure with probability at least $1 - 3\epsilon$ that the bounds (4.43) and (4.44) all hold simultaneously. In this case we have

$$\left\|\frac{1}{m}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right\|_2 \leq \left\|\frac{1}{m}\mathbf{A}^T\mathbf{A}\right\|_2 + 2\rho$$
$$= \left\|\frac{1}{m}\mathbf{A}^T\mathbf{A}\right\|_2\left[1 + \frac{2\rho}{\left\|\frac{1}{m}\mathbf{A}^T\mathbf{A}\right\|_2}\right] \leq \left\|\frac{1}{m}\mathbf{A}^T\mathbf{A}\right\|_2\left[1 + \frac{2\rho}{\lambda_1(\Sigma) - \rho}\right].$$

Assuming $\rho \leq \lambda_1(\Sigma)/2$ we have by the above bound

$$\left\|\frac{1}{m}\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right\|_2 \leq 3\left\|\frac{1}{m}\mathbf{A}^T\mathbf{A}\right\|_2.$$

Now note that

$$\left\|\mathbf{A}^T\mathbf{A}\right\|_2 = \left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2^2 \quad \left\|\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\right\|_2 = \left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_2^2$$

so that our previous bound implies

$$\left\|\phi(-\mathbf{W}\mathbf{X}^T)\right\|_2^2 \leq 3\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2^2.$$

Then we have by corollary 4.6.18 that

$$\frac{Tr(\mathbf{K}_{outer})}{\lambda_1(\mathbf{K}_{outer})} = \frac{\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_F^2}{\left\|\phi(\mathbf{W}\mathbf{X}^T)\right\|_2^2} \leq 12\frac{\|\mathbf{W}\|_2^2}{\sigma_{min}(\mathbf{W})^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}$$
$$\leq 12\left(\frac{\sqrt{m} + \sqrt{d} + t_1}{\sqrt{m} - \sqrt{d} - t_1}\right)^2\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

□

From the above theorem we get the following corollary.

Corollary 4.6.24. *Assume $\phi(x) = ReLU(x)$ and $n \geq d$. Suppose that $\mathbf{w}_1, \ldots, \mathbf{w}_m \sim N(0, \nu^2 I_d)$ i.i.d. Fix $\epsilon > 0$ small. Set $M = \max_{i\in[n]} \|\mathbf{x}_i\|_2$. Then*

$$m = \Omega\left(\max(\lambda_1(\Sigma)^{-2}, 1)\max(n, \log(1/\epsilon))\right)$$

and

$$\nu = O(1/M\sqrt{m})$$

suffices to ensure that with probability at least $1 - \epsilon$

$$\frac{Tr(\mathbf{K}_{outer})}{\lambda_1(\mathbf{K}_{outer})} \leq C\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}$$

where $C > 0$ is an absolute constant.

4.6.4.5 Bound for the Combined NTK

Based on the results in the previous two sections, we can now bound the effective rank of the combined NTK gram matrix $\mathbf{K} = \mathbf{K}_{inner} + \mathbf{K}_{outer}$.

Theorem 4.4.5. *Assume $\phi(x) = ReLU(x)$ and $n \geq d$. Fix $\epsilon > 0$ small. Suppose that $\mathbf{w}_1, \ldots, \mathbf{w}_m \sim N(0, \nu_1^2 I_d)$ i.i.d. and $a_1, \ldots, a_m \sim N(0, \nu_2^2)$. Set $M = \max_{i\in[n]} \|\mathbf{x}_i\|_2$, and let*

$$\Sigma := \mathbb{E}_{\mathbf{w}\sim N(0,\nu_1^2 I)}[\phi(\mathbf{X}\mathbf{w})\phi(\mathbf{w}^T\mathbf{X}^T)].$$

Then

$$m = \Omega\left(\max(\lambda_1(\Sigma)^{-2}, 1)\max(n, \log(1/\epsilon))\right), \quad \nu_1 = O(1/M\sqrt{m})$$

suffices to ensure that, with probability at least $1 - \epsilon$ over the sampling of the parameter initialization,

$$\text{eff}(\mathbf{K}) \leq C \cdot \text{eff}(\mathbf{X}\mathbf{X}^T),$$

where $C > 0$ is an absolute constant.

Proof. This follows from the union bound and Corollaries 4.6.16 and 4.6.24. □

4.6.4.6 Magnitude of the Spectrum

By our results in sections 4.6.4.3 and 4.6.4.4 we have that $m \gtrsim n$ suffices to ensure that

$$\frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})} \lesssim \frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \leq d.$$

Well note that

$$i\frac{\lambda_i(\mathbf{K})}{\lambda_1(\mathbf{K})} \leq \frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})} \lesssim d.$$

If $i \gg d$ then $\lambda_i(\mathbf{K})/\lambda_1(\mathbf{K})$ is small. Thus the NTK only has $O(d)$ large eigenvalues. The smallest eigenvalue $\lambda_n(\mathbf{K})$ of the NTK has been of interest in proving convergence guarantees [DLL19, DZP19, OS20]. By our previous inequality

$$\frac{\lambda_n(\mathbf{K})}{\lambda_1(\mathbf{K})} \lesssim \frac{d}{n}.$$

Thus in the setting where $m \gtrsim n \gg d$ we have that the smallest eigenvalue will be driven to zero relative to the largest eigenvalue. Alternatively we can view the above inequality as a lower bound on the condition number

$$\frac{\lambda_1(\mathbf{K})}{\lambda_n(\mathbf{K})} \gtrsim \frac{n}{d}.$$

We will first bound the analytical NTK in the setting when the outer layer weights have fixed constant magnitude. This is the setting considered by [XLS17], [ADH19a], [DZP19], [OFL19], [LSO20], and [OS20].

Theorem 4.6.25. *Let $\phi(x) = ReLU(x)$ and assume $\mathbf{X} \neq 0$. Let $\mathbf{K}^{\infty}_{inner} \in \mathbb{R}^{n\times n}$ be the analytical NTK, i.e.*

$$(\mathbf{K}^{\infty}_{inner})_{i,j} := \langle \mathbf{x}_i, \mathbf{x}_j\rangle \mathbb{E}_{\mathbf{w}\sim N(0,I_d)}\left[\phi'(\langle \mathbf{x}_i, \mathbf{w}\rangle)\phi'(\langle \mathbf{x}_j, \mathbf{w}\rangle)\right].$$

Then

$$\frac{1}{4}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}^{\infty}_{inner})}{\lambda_1\left(\mathbf{K}^{\infty}_{inner}\right)} \leq 4\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

Proof. We consider the setting where $|a_\ell| = 1/\sqrt{m}$ for all $\ell \in [m]$ and $\mathbf{w}_\ell \sim N(0, I_d)$ i.i.d.. As was shown by [JGH18], [DZP19] in this setting we have that if we fix the training data $\mathbf{X}$ and send $m \to \infty$ we have that

$$\|\mathbf{K}_{inner} - \mathbf{K}_{inner}^{\infty}\|_2 \to 0$$

in probability. Therefore by continuity of the effective rank we have that

$$\frac{Tr(\mathbf{K}_{inner})}{\lambda_1(\mathbf{K}_{inner})} \to \frac{Tr(\mathbf{K}_{inner}^{\infty})}{\lambda_1(\mathbf{K}_{inner}^{\infty})}$$

in probability. Let $\eta > 0$. Then there exists an $M \in \mathbb{N}$ such that $m \geq M$ implies that

$$\left| \frac{Tr(\mathbf{K}_{inner})}{\lambda_1(\mathbf{K}_{inner})} - \frac{Tr(\mathbf{K}_{inner}^{\infty})}{\lambda_1(\mathbf{K}_{inner}^{\infty})} \right| \leq \eta \tag{4.45}$$

with probability greater than $1/2$. Now fix $\delta \in (0,1)$. On the other hand by Theorem 4.6.13 with $\epsilon = 1/4$ we have that if $m \geq \frac{4}{\delta^2}\log(4n)$ then with probability at least $3/4$ that

$$\frac{(1-\delta)^2}{4}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}_{inner})}{\lambda_1(\mathbf{K}_{inner})} \leq \frac{4}{(1-\delta)^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}. \tag{4.46}$$

Thus if we set $m = \max(\frac{4}{\delta^2}\log(4n), M)$ we have with probability at least $3/4 - 1/2 = 1/4$ that (4.45) and (4.46) hold simultaneously. In this case we have that

$$\frac{(1-\delta)^2}{4}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} - \eta \leq \frac{Tr(\mathbf{K}_{inner}^{\infty})}{\lambda_1(\mathbf{K}_{inner}^{\infty})} \leq \frac{4}{(1-\delta)^2}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} + \eta$$

Note that the above argument runs through for any $\eta > 0$ and $\delta \in (0,1)$. Thus we may send $\eta \to 0^+$ and $\delta \to 0^+$ in the above inequality to get

$$\frac{1}{4}\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)} \leq \frac{Tr(\mathbf{K}_{inner}^{\infty})}{\lambda_1(\mathbf{K}_{inner}^{\infty})} \leq 4\frac{Tr(\mathbf{X}\mathbf{X}^T)}{\lambda_1(\mathbf{X}\mathbf{X}^T)}.$$

□

We thus have the following corollary about the conditioning of the analytical NTK.

Corollary 4.6.26. *Let $\phi(x) = ReLU(x)$ and assume $\mathbf{X} \neq 0$. Let $\mathbf{K}_{inner}^{\infty} \in \mathbb{R}^{n\times n}$ be the analytical NTK, i.e.*

$$(\mathbf{K}_{inner}^{\infty})_{i,j} := \langle \mathbf{x}_i, \mathbf{x}_j \rangle \mathbb{E}_{\mathbf{w}\sim N(0,I_d)}\left[\phi'(\langle \mathbf{x}_i, \mathbf{w}\rangle)\phi'(\langle \mathbf{x}_j, \mathbf{w}\rangle)\right].$$

Then

$$\frac{\lambda_n(\mathbf{K}^{\infty}_{inner})}{\lambda_1(\mathbf{K}^{\infty}_{inner})} \leq 4\frac{d}{n}.$$

4.6.5 Experimental Validation of Results on the NTK Spectrum

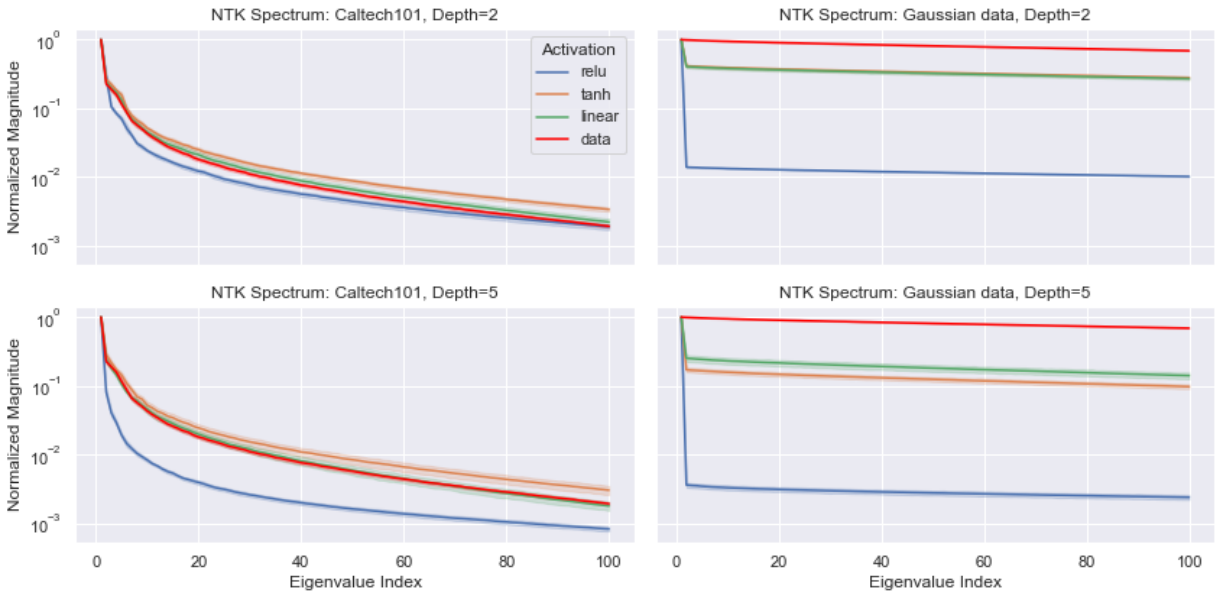

Figure 4.3: **NTK Spectrum for CNNs** We plot the normalized eigenvalues λ_p/λ_1 of the NTK Gram matrix $\mathbf{K}$ and the data Gram matrix $\mathbf{X}\mathbf{X}^T$ for Caltech101 and isotropic Gaussian datasets. To compute the NTK, we randomly initialize convolutional neural networks of depth 2 and 5 with 100 channels per layer. We use the standard parameterization and Pytorch's default Kaiming uniform initialization in order to better connect our results with what is used in practice. We consider a batch size of $n = 200$ and plot the first 100 eigenvalues. The thick part of each curve corresponds to the mean across 10 trials while the transparent part corresponds to the 95% confidence interval.

We experimentally test the theory developed in Section 4.4.1 and its implications by analyzing the spectrum of the NTK for both fully connected neural network architectures (FCNNs), the results of which are displayed in Figure 4.1, and also convolutional neural

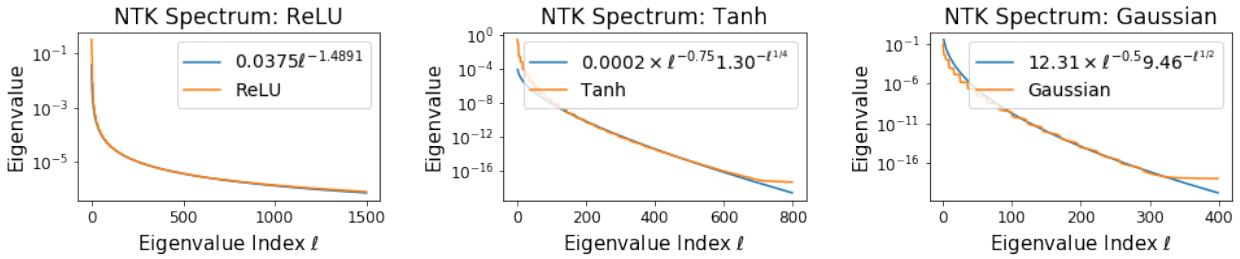

Figure 4.4: **Asymptotic NTK Spectrum** NTK spectrum of two-layer fully connected networks with ReLU, Tanh and Gaussian activations under the NTK parameterization. The orange curve is the experimental eigenvalue. The blue curves in the left shows the regression fit for the experimental eigenvalues as a function of eigenvalue index ℓ in the form of $\lambda_\ell = a\ell^{-b}$ where a and b are unknown parameters determined by regression. The blue curves in the middle shows the regression fit for the experimental eigenvalues in the form of $\lambda_\ell = a\ell^{-0.75}b^{-l^{1/4}}$. The blue curves in the right shows the regression fit for the experimental eigenvalues in the form of $\lambda_\ell = a\ell^{-0.5}b^{-l^{1/2}}$.

network architectures (CNNs), shown in Figure 4.3. For the feedforward architectures we consider networks of depth 2 and 5 with the width of all layers being set at 500. With regard to the activation function we test linear, ReLU and Tanh, and in terms of initialization we use Kaiming uniform [HZR15], which is very common in practice and is the default in PyTorch [PGM19]. For the convolutional architectures we again consider depths 2 and 5, with each layer consisting of 100 channels with the filter size set to 5x5. In terms of data, we consider 40x40 patches from both real world images, generated by applying Pytorch's `RandomResizedCrop` transform to a random batch of Caltech101 images [LAR22], as well as synthetic images corresponding to isotropic Gaussian vectors. The batch sized is fixed at 200 and we plot only the first 100 normalized eigenvalues. Each experiment was repeated 10 times. Finally, to compute the NTK we use the `functorch`[4] module in PyTorch using an algorithmic approach inspired by [NSS22].

[4] `https://pytorch.org/functorch/stable/notebooks/neural_tangent_kernels.html`

The results for convolutional neural networks show the same trends as observed in feedforward neural networks, which we discussed in Section 4.4.1. In particular, we again observe the dominant outlier eigenvalue, which increases with both depth and the size of the Gaussian mean of the activation. We also again see that the NTK spectrum inherits its structure from the data, i.e., is skewed for skewed data or relatively flat for isotropic Gaussian data. Finally, we also see that the spectrum for Tanh is closer to the spectrum for the linear activation when compared with the ReLU spectrum. In terms of differences between the CNN and FCNN experiments, we observe that the spread of the 95% confidence interval is slightly larger for convolutional nets, implying a slightly larger variance between trials. We remark that this is likely attributable to the fact that there are only 100 channels in each layer and by increasing this quantity we would expect the variance to reduce. In summary, despite the fact that our analysis is concerned with FCNNs, it appears that the broad implications and trends also hold for CNNs. We leave a thorough study of the NTK spectrum for CNNs and other network architectures to future work.

To test our theory in Section 4.4.2, we numerically plot the spectrum of NTK of two-layer feedforward networks with ReLU, Tanh, and Gaussian activations in Figure 4.4. The input data are uniformly drawn from $\mathbb{S}^2$. Notice that when $d = 2$, $k = \Theta(\ell^{1/2})$. Then Corollary 4.4.7 shows that for the ReLU activation $\lambda_\ell = \Theta(\ell^{-3/2})$, for the Tanh activation $\lambda_\ell = O\left(\ell^{-3/4}\exp(-\frac{\pi}{2}\ell^{1/4})\right)$, and for the Gaussian activation $\lambda_\ell = O(\ell^{-1/2}2^{-\ell^{1/2}})$. These theoretical decay rates for the NTK spectrum are verified by the experimental results in Figure 4.4.

4.6.6 Analysis of the Lower Spectrum: Uniform Data

Theorem 4.4.6. *[AM15] Suppose that the training data are uniformly sampled from the unit hypersphere $\mathbb{S}^d$, $d \geq 2$. If the dot-product kernel function has the expansion $K(x_1, x_2) = \sum_{p=0}^{\infty} c_p \langle x_1, x_2 \rangle^p$ where $c_p \geq 0$, then the eigenvalue of every spherical harmonic of frequency*

k is given by

$$\overline{\lambda_k} = \frac{\pi^{d/2}}{2^{k-1}} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} c_p \frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)},$$

where Γ is the gamma function.

Proof. Let $\theta(t) = \sum_{p=0}^{\infty} c_p t^p$, then $K(x_1, x_2) = \theta(\langle x_1, x_2 \rangle)$ According to Funk Hecke theorem [BJK19, Section 4.2], we have

$$\overline{\lambda_k} = \mathrm{Vol}(\mathbb{S}^{d-1}) \int_{-1}^{1} \theta(t) P_{k,d}(t) (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t, \tag{4.47}$$

where $\mathrm{Vol}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ is the volume of the hypersphere $\mathbb{S}^{d-1}$, and $P_{k,d}(t)$ is the Gegenbauer polynomial, given by

$$P_{k,d}(t) = \frac{(-1)^k}{2^k} \frac{\Gamma(d/2)}{\Gamma(k+d/2)} \frac{1}{(1-t^2)^{(d-2)/2}} \frac{d^k}{dt^k} (1-t^2)^{k+(d-2)/2},$$

and Γ is the gamma function.

From (4.47) we have

$$\begin{aligned}
\overline{\lambda_k} &= \mathrm{Vol}(\mathbb{S}^{d-1}) \int_{-1}^{1} \theta(t) P_{k,d}(t) (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t \\
&= \frac{2\pi^{d/2}}{\Gamma(d/2)} \int_{-1}^{1} \theta(t) \frac{(-1)^k}{2^k} \frac{\Gamma(d/2)}{\Gamma(k+d/2)} \frac{d^k}{dt^k} (1-t^2)^{k+(d-2)/2} \mathrm{d}t \\
&= \frac{2\pi^{d/2}}{\Gamma(d/2)} \frac{(-1)^k}{2^k} \frac{\Gamma(d/2)}{\Gamma(k+d/2)} \sum_{p=0}^{\infty} c_p \int_{-1}^{1} t^p \frac{d^k}{dt^k} (1-t^2)^{k+(d-2)/2} \mathrm{d}t.
\end{aligned} \tag{4.48}$$

Using integration by parts, we have

$$\begin{aligned}
&\int_{-1}^{1} t^p \frac{d^k}{dt^k} (1-t^2)^{k+(d-2)/2} \mathrm{d}t \\
&= t^p \frac{d^{k-1}}{dt^{k-1}} (1-t^2)^{k+(d-2)/2} \Big|_{-1}^{1} - p \int_{-1}^{1} t^{p-1} \frac{d^{k-1}}{dt^{k-1}} (1-t^2)^{k+(d-2)/2} \mathrm{d}t \\
&= -p \int_{-1}^{1} t^{p-1} \frac{d^{k-1}}{dt^{k-1}} (1-t^2)^{k+(d-2)/2} \mathrm{d}t,
\end{aligned} \tag{4.49}$$

where the last line in (4.49) holds because $\frac{d^{k-1}}{dt^{k-1}} (1-t^2)^{k+(d-2)/2} = 0$ when $t = 1$ or $t = -1$.

When $p < k$, repeat the above procedure (4.49) p times, we get

$$\begin{aligned}\int_{-1}^{1} t^p \frac{d^k}{dt^k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t &= (-1)^p p! \int_{-1}^{1} \frac{d^{k-p}}{dt^{k-p}}(1-t^2)^{k+(d-2)/2}\mathrm{d}t \\ &= (-1)^p p! \left.\frac{d^{k-p-1}}{dt^{k-p-1}}(1-t^2)^{k+(d-2)/2}\right|_{-1}^{1} \\ &= 0. \end{aligned} \tag{4.50}$$

When $p \geq k$, repeat the above procedure (4.49) k times, we get

$$\int_{-1}^{1} t^p \frac{d^k}{dt^k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t = (-1)^k p(p-1)\cdots(p-k+1)\int_{-1}^{1} t^{p-k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t. \tag{4.51}$$

When $p-k$ is odd, $t^{p-k}(1-t^2)^{k+(d-2)/2}$ is an odd function, then

$$\int_{-1}^{1} t^{p-k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t = 0. \tag{4.52}$$

When $p-k$ is even,

$$\begin{aligned}\int_{-1}^{1} t^{p-k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t &= 2\int_{0}^{1} t^{p-k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t \\ &= \int_{0}^{1} (t^2)^{(p-k-1)/2}(1-t^2)^{k+(d-2)/2}\mathrm{d}t^2 \\ &= B\left(\frac{p-k+1}{2}, k+d/2\right) \\ &= \frac{\Gamma(\frac{p-k+1}{2})\Gamma(k+d/2)}{\Gamma(\frac{p-k+1}{2}+k+d/2)}, \end{aligned} \tag{4.53}$$

where B is the beta function.

Plugging (4.53) , (4.50) and (4.52) into (4.51), we get

$$\begin{aligned}&\int_{-1}^{1} t^p \frac{d^k}{dt^k}(1-t^2)^{k+(d-2)/2}\mathrm{d}t \\ &= \begin{cases} (-1)^k p(p-1)\ldots(p-k+1)\frac{\Gamma(\frac{p-k+1}{2})\Gamma(k+d/2)}{\Gamma(\frac{p-k+1}{2}+k+d/2)}, & p-k \text{ is even and } p \geq k, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{4.54}$$

Plugging (4.54) into (4.48), we get

$$\begin{aligned}\overline{\lambda_k} &= \frac{2\pi^{d/2}}{\Gamma(d/2)}\frac{(-1)^k}{2^k}\frac{\Gamma(d/2)}{\Gamma(k+d/2)}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} c_p(-1)^k p(p-1)\dots(p-k+1)\frac{\Gamma(\frac{p-k+1}{2})\Gamma(k+d/2)}{\Gamma(\frac{p-k+1}{2}+k+d/2)}\\ &= \frac{\pi^{d/2}}{2^{k-1}}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} c_p\frac{p(p-1)\dots(p-k+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(\frac{p-k+1}{2}+k+d/2)}\\ &= \frac{\pi^{d/2}}{2^{k-1}}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} c_p\frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)}.\end{aligned}$$

□

Corollary 4.4.7. *Under the same setting as in Theorem 4.4.6,*

1. *if* $c_p = \Theta(p^{-a})$ *where* $a \geq 1$, *then* $\overline{\lambda_k} = \Theta(k^{-d-2a+2})$,
2. *if* $c_p = \delta_{(p \text{ even})}\Theta(p^{-a})$, *then* $\overline{\lambda_k} = \delta_{(k \text{ even})}\Theta(k^{-d-2a+2})$,
3. *if* $c_p = \mathcal{O}\left(\exp\left(-a\sqrt{p}\right)\right)$, *then* $\overline{\lambda_k} = \mathcal{O}\left(k^{-d+1/2}\exp\left(-a\sqrt{k}\right)\right)$,
4. *if* $c_p = \Theta(p^{1/2}a^{-p})$, *then* $\overline{\lambda_k} = \mathcal{O}\left(k^{-d+1}a^{-k}\right)$ *and* $\overline{\lambda_k} = \Omega\left(k^{-d/2+1}2^{-k}a^{-k}\right)$.

Proof of Corollary 4.6.6, part 1. We first prove $\overline{\lambda_k} = O(k^{-d-2a+2})$. Suppose that $c_p \leq Cp^{-a}$ for some constant C, then according to Theorem 4.4.6 we have

$$\overline{\lambda_k} \leq \frac{\pi^{d/2}}{2^{k-1}}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} Cp^{-a}\frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)}.$$

According to Stirling's formula, we have

$$\Gamma(z) = \sqrt{\frac{2\pi}{z}}\left(\frac{z}{e}\right)^z\left(1+O\left(\frac{1}{z}\right)\right). \tag{4.55}$$

Then for any $z \geq \frac{1}{2}$, we can find constants C_1 and C_2 such that

$$C_1\sqrt{\frac{2\pi}{z}}\left(\frac{z}{e}\right)^z \leq \Gamma(z) \leq C_2\sqrt{\frac{2\pi}{z}}\left(\frac{z}{e}\right)^z. \tag{4.56}$$

Then

$$\begin{aligned}
\overline{\lambda_k} &\le \frac{\pi^{d/2}}{2^{k-1}}\frac{C_2^2}{C_1^2}\sum_{\substack{p\ge k\\ p-k \text{ is even}}} Cp^{-a}\frac{\sqrt{\frac{2\pi}{p+1}}\left(\frac{p+1}{e}\right)^{p+1}\sqrt{\frac{2\pi}{\frac{p-k+1}{2}}}\left(\frac{\frac{p-k+1}{2}}{e}\right)^{\frac{p-k+1}{2}}}{\sqrt{\frac{2\pi}{p-k+1}}\left(\frac{p-k+1}{e}\right)^{p-k+1}\sqrt{\frac{2\pi}{\frac{p-k+1}{2}+k+d/2}}\left(\frac{\frac{p-k+1}{2}+k+d/2}{e}\right)^{\frac{p-k+1}{2}+k+d/2}} \\
&= \frac{\pi^{d/2}}{2^{k-1}}\frac{C_2^2C}{C_1^2}\sum_{\substack{p\ge k\\ p-k \text{ is even}}} p^{-a}\frac{e^{\frac{d}{2}}\sqrt{\frac{2}{p+1}}(p+1)^{p+1}\left(\frac{p-k+1}{2}\right)^{\frac{p-k+1}{2}}}{(p-k+1)^{p-k+1}\sqrt{\frac{1}{\frac{p-k+1}{2}+k+d/2}}\left(\frac{p-k+1}{2}+k+d/2\right)^{\frac{p-k+1}{2}+k+d/2}} \\
&= \frac{\pi^{d/2}}{2^{k-1}}\frac{C_2^2C}{C_1^2}\sum_{\substack{p\ge k\\ p-k \text{ is even}}} p^{-a}\frac{e^{\frac{d}{2}}2^{\frac{-p+k}{2}}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}\left(\frac{p-k+1}{2}+k+d/2\right)^{\frac{p-k}{2}+k+d/2}} \\
&= 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C}{C_1^2}\sum_{\substack{p\ge k\\ p-k \text{ is even}}} \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}}. \qquad (4.57)
\end{aligned}$$

We define

$$f_a(p) = \frac{p^{-a}(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}(p+k+1+d)^{\frac{p+k+d}{2}}}. \qquad (4.58)$$

By applying the chain rule to $e^{\log f_a(p)}$, we have that the derivative of f_a is

$$\begin{aligned}
f_a'(p) = &\frac{(p+1)^{p+\frac{1}{2}}p^{-a}}{2(p-k+1)^{\frac{p-k+1}{2}}(p+k+d+1)^{\frac{p+k+d}{2}}} \\
&\cdot\left(-\frac{2a}{p}-\frac{k+d}{(p+1)(p+k+d+1)}+\log(1+\frac{k^2-d(p-k+1)}{(p-k+1)(p+k+d+1)})\right). \qquad (4.59)
\end{aligned}$$

Let $g_a(p) = -\frac{2a}{p}-\frac{k+d}{(p+1)(p+k+d+1)}+\log(1+\frac{k^2-d(p-k+1)}{(p-k+1)(p+k+d+1)})$. Then $g_a(p)$ and $f_a'(p)$ have the same sign. Next we will show that $g_a(p)\ge 0$ for $k\le p\le \frac{k^2}{d+24a}$ when k is large enough.

First when $p\ge k$ and $\frac{k^2-d(p-k+1)}{(p-k+1)(p+k+d+1)}\ge 1$, we have

$$g_a(p) \ge -\frac{2a}{k}-\frac{k+d}{(k+1)(k+k+d+1)}+\log(2)\ge 0, \qquad (4.60)$$

when k is sufficiently large.

Second when $p\ge k$ and $0\le\frac{k^2-d(p-k+1)}{(p-k+1)(p+k+d+1)}\le 1$, since $\log(1+x)\ge\frac{x}{2}$ for $0\le x\le 1$, we

have

$$\begin{aligned} g_a(p) &\geq -\frac{2a}{p} - \frac{k+d}{(p+1)(p+k+d+1)} + \frac{k^2-d(p-k+1)}{2(p-k+1)(p+k+d+1)} \\ &\geq -\frac{2a}{p} - \frac{k+d}{(p+1)(p+k+d+1)} + \frac{k^2-dp}{2p(p+k+d+1)}. \end{aligned}$$

When $p \leq \frac{k^2}{d+24a}$, we have $k^2 - dp \geq 24ap$. Then

$$\frac{k^2-dp}{4p(p+k+d+1)} \geq \frac{24ap}{4p(p+k+d+1)} \geq \frac{6ap}{(p+1)(p+k+d+1)} \geq \frac{k+d}{(p+1)(p+k+d+1)}$$

when k is sufficiently large. Also we have

$$\frac{k^2-dp}{4r(p+k+d+1)} \geq \frac{24ap}{4r(p+k+d+1)} \geq \frac{6a}{p+k+d+1} \geq \frac{2a}{p}$$

when k is sufficiently large.

Combining all the arguments above, we conclude that $g_a(p) \geq 0$ and $f_a'(p) \geq 0$ when $k \leq p \leq \frac{k^2}{d+24a}$. Then when $k \leq p \leq \frac{k^2}{d+24a}$, we have

$$f_a(p) \leq f_a\left(\frac{k^2}{d+24a}\right). \tag{4.61}$$

When $p \geq \frac{k^2}{d+24a}$, we have

$$\begin{aligned} f_a(p) &= \frac{p^{-a}\,(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}\,(p+k+1+d)^{\frac{p+k+d}{2}}} \\ &= \frac{p^{-a}\,(p+1)^{p+\frac{1}{2}}}{((p+1)^2-k^2+d(p-k+1))^{\frac{p-k+1}{2}}\,(p+k+1+d)^{\frac{2k+d-1}{2}}} \\ &= \frac{p^{-a}\,(p+1)^{-\frac{d}{2}}}{\left(1-\frac{k^2-d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}}\left(1+\frac{k+d}{p+1}\right)^{\frac{2k+d-1}{2}}} \\ &\leq \frac{p^{-a-\frac{d}{2}}}{\left(1-\frac{k^2-d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}}}. \end{aligned}$$

If $k^2-d(p-k+1) < 0$, $\left(1-\frac{k^2-d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}} \geq 1$. If $k^2-d(p-k+1) \geq 0$, i.e., $p \leq \frac{k^2+dk-d}{d}$,

for sufficiently large k, we have

$$\begin{aligned}\left(1-\frac{k^2-d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}} &\geq \left(1-\frac{k^2-d(\frac{k^2}{d+24a}-k+1)}{(\frac{k^2}{d+24a}+1)^2}\right)^{\frac{\frac{k^2+dk-d}{d}-k+1}{2}} \\ &\geq \left(1-\frac{48a(d+24a)}{k^2}\right)^{\frac{k^2}{2d}} \\ &\geq e^{-\frac{k^2}{2d}\frac{48a(d+24a)}{k^2}} = e^{-\frac{48a(d+24a)}{2d}},\end{aligned}$$

which is a constant independent of k. Then for $p \geq \frac{k^2}{d+24a}$, we have

$$f_a(p) \leq e^{\frac{48a(d+24a)}{2d}} p^{-a-\frac{d}{2}}. \tag{4.62}$$

Finally we have

$$\begin{aligned}\overline{\lambda_k} &= 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C}{C_1^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} f_a(p) \\ &\leq O\left(\sum_{\substack{k\leq p\leq \frac{k^2}{d+24a}\\ p-k \text{ is even}}} f_a(p) + \sum_{\substack{p\geq \frac{k^2}{d+24a}\\ p-k \text{ is even}}} f_a(p)\right) \\ &\leq O\left(\left(\frac{k^2}{d+24a}-k+1\right) f_a\left(\frac{k^2}{d+24a}\right) + \sum_{\substack{p\geq \frac{k^2}{d+24a}\\ p-k \text{ is even}}} e^{\frac{48a(d+24a)}{2d}} p^{-a-\frac{d}{2}}\right) \\ &\leq O\left(\left(\frac{k^2}{d+24a}-k+1\right) e^{\frac{48a(d+24a)}{2d}} \left(\frac{k^2}{d+24a}\right)^{-a-\frac{d}{2}} + \frac{e^{\frac{48a(d+24a)}{2d}}}{a+\frac{d}{2}-1}\left(\frac{k^2}{d+24a}-1\right)^{1-a-\frac{d}{2}}\right) \\ &= O(k^{-d-2a+2}).\end{aligned}$$

Next we prove $\overline{\lambda_k} = \Omega(k^{-d-2a+2})$. Since c_p are nonnegative and $c_p = \Theta(p^{-a})$, we have that $c_p \geq C'p^{-a}$ for some constant C'. Then we have

$$\overline{\lambda_k} \geq \frac{\pi^{d/2}}{2^{k-1}} \sum_{\substack{p\geq k\\ p-k \text{ is even}}} C'p^{-a}\frac{\Gamma(p+1)\Gamma(\frac{p-k+1}{2})}{\Gamma(p-k+1)\Gamma(\frac{p-k+1}{2}+k+d/2)}. \tag{4.63}$$

According to Stirling's formula (4.55) and (4.56), using the similar argument as (4.57) we have

$$\overline{\lambda_k} \geq \frac{\pi^{d/2}}{2^{k-1}} \frac{C_1^2}{C_2^2} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} C' p^{-a} \frac{\sqrt{\frac{2\pi}{p+1}} \left(\frac{p+1}{e}\right)^{p+1} \sqrt{\frac{2\pi}{\frac{p-k+1}{2}}} \left(\frac{\frac{p-k+1}{2}}{e}\right)^{\frac{p-k+1}{2}}}{\sqrt{\frac{2\pi}{p-k+1}} \left(\frac{p-k+1}{e}\right)^{p-k+1} \sqrt{\frac{2\pi}{\frac{p-k+1}{2}+k+d/2}} \left(\frac{\frac{p-k+1}{2}+k+d/2}{e}\right)^{\frac{p-k+1}{2}+k+d/2}} \tag{4.64}$$

$$= 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_1^2 C'}{C_2^2} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} \frac{p^{-a} (p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}} (p+k+1+d)^{\frac{p+k+d}{2}}} \tag{4.65}$$

$$\geq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_1^2 C'}{C_2^2} \sum_{\substack{p \geq k^2 \\ p-k \text{ is even}}} f_a(p), \tag{4.66}$$

where $f_a(p)$ is defined in (4.58). When $p \geq k^2$, we have

$$\begin{aligned} f_a(p) &= \frac{p^{-a} (p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}} (p+k+1+d)^{\frac{p+k+d}{2}}} \\ &= \frac{p^{-a} (p+1)^{p+\frac{1}{2}}}{((p+1)^2 - k^2 + d(p-k+1))^{\frac{p-k+1}{2}} (p+k+1+d)^{\frac{2k+d-1}{2}}} \\ &\geq \frac{(p+1)^{-a-\frac{d}{2}}}{\left(1 - \frac{k^2 - d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}} \left(1 + \frac{k+d}{p+1}\right)^{\frac{2k+d-1}{2}}}. \end{aligned}$$

For sufficiently large k, $k^2 - d(p-k+1) < 0$. Then we have

$$\begin{aligned} \left(1 - \frac{k^2 - d(p-k+1)}{(p+1)^2}\right)^{\frac{p-k+1}{2}} &= \left(1 - \frac{k^2 - d(p-k+1)}{(p+1)^2}\right)^{\frac{-(p+1)^2}{k^2-d(p-k+1)} \cdot \frac{-k^2+d(p-k+1)}{(p+1)^2} \cdot \frac{p-k+1}{2}} \\ &\leq e^{\frac{-k^2+d(p-k+1)}{(p+1)^2} \cdot \frac{p-k+1}{2}} \\ &\leq e^{\frac{dp^2}{2p^2}} = e^{\frac{d}{2}} \end{aligned}$$

which is a constant independent of k. Also, for sufficiently large k, we have

$$\begin{aligned} \left(1 + \frac{k+d}{p+1}\right)^{\frac{2k+d-1}{2}} &= \left(1 + \frac{k+d}{p+1}\right)^{\frac{p+1}{k+d} \frac{k+d}{p+1} \frac{2k+d-1}{2}} \\ &\leq e^{\frac{k+d}{p+1} \frac{2k+d-1}{2}} \\ &\leq e^{\frac{3k^2}{2r}} = e^{\frac{3}{2}}. \end{aligned}$$

Then for $p \geq k^2$, we have $f_a(p) \geq e^{-\frac{d}{2}-\frac{3}{2}}(p+1)^{-a-\frac{d}{2}}$.

Finally we have

$$\overline{\lambda_k} \geq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_1^2C'}{C_2^2} \sum_{\substack{p\geq k^2 \\ p-k \text{ is even}}} f_a(p) \tag{4.67}$$

$$\geq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_1^2C'}{C_2^2} \sum_{\substack{p\geq k^2 \\ p-k \text{ is even}}} e^{-\frac{d}{2}-\frac{3}{2}}(p+1)^{-a-\frac{d}{2}} \tag{4.68}$$

$$\geq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_1^2C'}{C_2^2}e^{-\frac{d}{2}-\frac{3}{2}}\frac{1}{2(a+\frac{d}{2}-1)}(k^2+2)^{1-a-\frac{d}{2}} \tag{4.69}$$

$$= \Omega(k^{-d-2a+2}). \tag{4.70}$$

Overall, we have $\overline{\lambda_k} = \Theta(k^{-d-2a+2})$. □

Proof of Corollary 4.6.6, part 2. It is easy to verify that $\overline{\lambda_k} = 0$ when k is even because $c_p = 0$ when $p \geq k$ and $p-k$ is even. When k is odd, the proof of Theorem 4.4.6 still applies. □

Proof of Corollary 4.6.6, part 3. Since $c_p = \mathcal{O}\left(\exp\left(-a\sqrt{p}\right)\right)$, we have that $c_p \leq Ce^{-a\sqrt{p}}$ for some constant C. Similar to (4.57), we have

$$\overline{\lambda_k} \leq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C}{C_1^2} \sum_{\substack{p\geq k \\ p-k \text{ is even}}} \frac{e^{-a\sqrt{p}}\,(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}\,(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.71}$$

Use the definition in (4.58) and let $a = 0$, we have

$$f_0(p) = \frac{(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}\,(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.72}$$

Then according to (4.61) and (4.62), for sufficiently large k, we have $f_0(p) \leq f_0\left(\frac{k^2}{d}\right)$ when $k \leq p \leq \frac{k^2}{d}$ and $f_0(p) \leq C_3p^{-\frac{d}{2}}$ for some constant C_3 when $p \geq \frac{k^2}{d}$. Then when $k \leq p \leq \frac{k^2}{d}$, we have $f_0(p) \leq f_0\left(\frac{k^2}{d}\right) \leq C_3\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}$. When $p \geq \frac{k^2}{d}$, we have $f_0(p) \leq C_3p^{-\frac{d}{2}} \leq C_3\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}$. Overall, for all $p \geq k$, we have

$$f_0(p) \leq C_3\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}. \tag{4.73}$$

Then we have

$$\overline{\lambda_k} \leq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C}{C_1^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} e^{-a\sqrt{p}}f_0(p) \tag{4.74}$$

$$\leq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C_3C}{C_1^2}\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}\sum_{\substack{p\geq k\\ p-k \text{ is even}}} e^{-a\sqrt{p}} \tag{4.75}$$

$$\leq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C_3C}{C_1^2}\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}\frac{2e^{-a\sqrt{k-1}}(a\sqrt{k-1}+1)}{a^2} \tag{4.76}$$

$$= \mathcal{O}\left(k^{-d+1/2}\exp\left(-a\sqrt{k}\right)\right). \tag{4.77}$$

□

Proof of Corollary 4.6.6, part 4. Since $c_p = \Theta(p^{1/2}a^{-p})$, we have that $c_p \leq Cp^{1/2}a^{-p}$ for some constant C. Similar to (4.57), we have

$$\overline{\lambda_k} \leq 2\pi^{d/2}\frac{2^{\frac{d}{2}}e^{\frac{d}{2}}C_2^2C}{C_1^2}\sum_{\substack{p\geq k\\ p-k \text{ is even}}}\frac{p^{1/2}a^{-p}\,(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}\,(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.78}$$

Use the definition in (4.58) and let $a = 0$, we have

$$f_0(p) = \frac{(p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}}\,(p+k+1+d)^{\frac{p+k+d}{2}}}. \tag{4.79}$$

Then according to (4.61) and (4.62), for sufficiently large k, we have $f_0(p) \leq f_0\left(\frac{k^2}{d}\right)$ when $k \leq p \leq \frac{k^2}{d}$ and $f_0(p) \leq C_3p^{-\frac{d}{2}}$ for some constant C_3 when $p \geq \frac{k^2}{d}$. Then when $k \leq p \leq \frac{k^2}{d}$, we have $p^{1/2}f_0(p) \leq p^{1/2}f_0\left(\frac{k^2}{d}\right) \leq C_3\left(\frac{k^2}{d}\right)^{1/2}\left(\frac{k^2}{d}\right)^{-\frac{d}{2}}$. When $p \geq \frac{k^2}{d}$, we have $p^{1/2}f_0(p) \leq C_3p^{1/2}p^{-\frac{d}{2}} \leq C_3\left(\frac{k^2}{d}\right)^{-\frac{d}{2}+\frac{1}{2}}$. Overall, for all $p \geq k$, we have

$$p^{1/2}f_0(p) \leq C_3\left(\frac{k^2}{d}\right)^{-\frac{d}{2}+\frac{1}{2}}. \tag{4.80}$$

Then we have

$$\overline{\lambda_k} \leq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_2^2 C}{C_1^2} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} p^{1/2} a^{-p} f_0(p) \tag{4.81}$$

$$\leq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_2^2 C_3 C}{C_1^2} \left(\frac{k^2}{d}\right)^{-\frac{d}{2}+\frac{1}{2}} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} a^{-p} \tag{4.82}$$

$$\leq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_2^2 C_3 C}{C_1^2} \left(\frac{k^2}{d}\right)^{-\frac{d}{2}+\frac{1}{2}} \frac{1}{\log a} a^{-(k-1)} \tag{4.83}$$

$$= \mathcal{O}\left(k^{-d+1} a^{-k}\right). \tag{4.84}$$

On the other hand, since $c_p = \Theta(p^{1/2} a^{-p})$, we have that $c_p \geq C' p^{1/2} a^{-p}$ for some constant C'. Similar to (4.65), we have

$$\overline{\lambda_k} \geq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_1^2 C'}{C_2^2} \sum_{\substack{p \geq k \\ p-k \text{ is even}}} \frac{p^{1/2} a^{-p} (p+1)^{p+\frac{1}{2}}}{(p-k+1)^{\frac{p-k+1}{2}} (p+k+1+d)^{\frac{p+k+d}{2}}} \tag{4.85}$$

$$\geq 2\pi^{d/2} \frac{2^{\frac{d}{2}} e^{\frac{d}{2}} C_1^2 C'}{C_2^2} \frac{k^{1/2} a^{-k} (k+1)^{k+\frac{1}{2}}}{(k-k+1)^{\frac{k-k+1}{2}} (k+k+1+d)^{\frac{k+k+d}{2}}} \tag{4.86}$$

$$= \Omega\left(\frac{k^{-d/2+1} a^{-k} (k+1)^k}{(k+k+1+d)^k}\right). \tag{4.87}$$

Since $(k+1)^k = k^k (1+1/k)^k = \Theta(k^k)$. Similarly, $(k+k+1+d)^k = \Theta((2k)^k)$. Then we have

$$\overline{\lambda_k} = \Omega\left(\frac{k^{-d/2+1} a^{-k} (k+1)^k}{(k+k+1+d)^k}\right) \tag{4.88}$$

$$= \Omega\left(\frac{k^{-d/2+1} a^{-k} k^k}{(2k)^k}\right) \tag{4.89}$$

$$= \Omega\left(k^{-d/2+1} 2^{-k} a^{-k}\right). \tag{4.90}$$

□

For the NTK of a two-layer ReLU network with $\gamma_b > 0$, then according to Lemma 4.3.4 we have $c_p = \kappa_{p,2} = \Theta(p^{-3/2})$. Therefore using Corollary 4.4.7 $\overline{\lambda_k} = \Theta(k^{-d-1})$. Notice here

that k refers to the frequency, and the number of spherical harmonics of frequency at most k is $\Theta(k^d)$. Therefore, for the ℓth largest eigenvalue λ_ℓ we have $\lambda_\ell = \Theta(\ell^{-(d+1)/d})$. This rate agrees with [BJK19] and [VY21]. For the NTK of a two-layer ReLU network with $\gamma_b = 0$, the eigenvalues corresponding to the even frequencies are 0, which also agrees with [BJK19]. Corollary 4.4.7 also shows the decay rates of eigenvalues for the NTK of two-layer networks with Tanh activation and Gaussian activation. We observe that when the coefficients of the kernel power series decay quickly then the eigenvalues of the kernel also decay quickly. As a faster decay of the eigenvalues of the kernel implies a smaller RKHS, Corollary 4.4.7 demonstrates that using ReLU results in a larger RKHS relative to using either Tanh or Gaussian activations. We numerically illustrate Corollary 4.4.7 in Figure 4.4, Section 4.6.5.

4.6.7 Analysis of the Lower Spectrum: Non-uniform Data

The purpose of this section is to prove a formal version of Theorem 4.4.8. In order to prove this result we first need the following lemma.

Lemma 4.6.27. *Let the coefficients $(c_j)_{j=0}^\infty$ with $c_j \in \mathbb{R}_{\geq 0}$ for all $j \in \mathbb{Z}_{\geq 0}$ be such that the series $\sum_{j=0}^\infty c_j \rho^j$ converges for all $\rho \in [-1, 1]$. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\|\mathbf{x}_i\| = 1$ for all $i \in [n]$, define $r := \operatorname{rank}(\mathbf{X}) \geq 2$ and the gram matrix $\mathbf{G} := \mathbf{X}\mathbf{X}^T$. Consider the kernel matrix*

$$n\mathbf{K} = \sum_{j=0}^{\infty} c_j \mathbf{G}^{\odot j}.$$

For arbitrary $m \in \mathbb{Z}_{\geq 1}$, let the eigenvalue index k satisfy $n \geq k > \operatorname{rank}(\mathbf{H}_m)$, where $\mathbf{H}_m := \sum_{j=0}^{m-1} c_j \mathbf{G}^{\odot j}$. Then

$$\lambda_k(\mathbf{K}) \leq \frac{\|\mathbf{G}^{\odot m}\|}{n} \sum_{j=m}^{\infty} c_j. \tag{4.91}$$

Proof. We start our analysis by considering $\lambda_k(n\mathbf{K})$ for some arbitrary $k \in \mathbb{N}_{\leq n}$. Let

$$\mathbf{H}_m := \sum_{j=0}^{m-1} c_j \mathbf{G}^{\odot j},$$
$$\mathbf{T}_m := \sum_{j=m}^{\infty} c_j \mathbf{G}^{\odot j}$$

be the m-head and m-tail of the Hermite expansion of $n\mathbf{K}$: clearly $n\mathbf{K} = \mathbf{H}_m + \mathbf{T}_m$ for any $m \in \mathbb{N}$. Recall that a constant matrix is symmetric and positive semi-definite, furthermore, by the Schur product theorem, the Hadamard product of two positive semi-definite matrices is positive semi-definite. As a result, $\mathbf{G}^{\odot j}$ is symmetric and positive semi-definite for all $j \in \mathbb{Z}_{\geq 0}$ and therefore $\mathbf{H}_m$ and $\mathbf{T}_m$ are also symmetric positive semi-definite matrices. From Weyl's inequality [Wey12, Satz 1] it follows that

$$n\lambda_k(\mathbf{K}) \leq \lambda_k(\mathbf{H}_m) + \lambda_1(\mathbf{T}_m). \tag{4.92}$$

In order to upper bound $\lambda_1(\mathbf{T}_m)$, observe, as $\mathbf{T}_m$ is square, symmetric and positive semi-definite, that $\lambda_1(\mathbf{T}_m) = \|\mathbf{T}_m\|$. Using the non-negativity of the coefficients $(c_j)_{j=0}^{\infty}$ and the triangle inequality we have

$$\lambda_1(\mathbf{T}_m) = \left\| \sum_{j=m}^{\infty} c_j \mathbf{G}^{\odot j} \right\| \leq \sum_{j=m}^{\infty} c_j \left\| \mathbf{G}^{\odot j} \right\|.$$

By the assumptions of the lemma $[\mathbf{G}]_{ii} = 1$ and therefore $[\mathbf{G}]_{ii}^j = 1$ for all $j \in \mathbb{Z}_{\geq 0}$. Furthermore, for any pair of positive semi-definite matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ and $k \in [n]$

$$\lambda_1(\mathbf{A} \odot \mathbf{B}) \leq \max_{i \in [n]} [\mathbf{A}]_{ii} \lambda_1(\mathbf{B}), \tag{4.93}$$

[Sch11]. Therefore, as $\max_{i \in [n]} [\mathbf{G}]_{ii} = 1$,

$$\left\| \mathbf{G}^{\odot j} \right\| = \lambda_1(\mathbf{G}^{\odot j}) = \lambda_1(\mathbf{G} \odot \mathbf{G}^{\odot (j-1)}) \leq \lambda_1(\mathbf{G}^{\odot (j-1)}) = \left\| \mathbf{G}^{\odot (j-1)} \right\|$$

for all $j \in \mathbb{N}$. As a result

$$\lambda_1(\mathbf{T}_m) \leq \left\| \mathbf{G}^{\odot m} \right\| \sum_{j=m}^{\infty} c_j.$$

Finally, we now turn our attention to the analysis of $\lambda_k(\mathbf{H}_m)$. Upper bounding a small eigenvalue is typically challenging, however, the problem simplifies when and k exceeds the rank of $\mathbf{H}_m$, as is assumed here, as this trivially implies $\lambda_k(\mathbf{H}_m) = 0$. Therefore, for $k > \text{rank}(\mathbf{H}_m)$

$$\lambda_k(\mathbf{K}) \leq \frac{\|\mathbf{G}^m\|}{n} \sum_{j=m}^{\infty} c_j$$

as claimed. □

In order to use Lemma 4.6.27 we require an upper bound on the rank of $\mathbf{H}_m$. To this end we provide Lemma 4.6.28.

Lemma 4.6.28. *Let $\mathbf{G} \in \mathbb{R}^{n \times n}$ be a symmetric, positive semi-definite matrix of rank $2 \leq r \leq d$. Define $\mathbf{H}_m \in \mathbb{R}^{n \times n}$ as*

$$\mathbf{H}_m = \sum_{j=0}^{m-1} c_j \mathbf{G}^{\odot j} \tag{4.94}$$

where $(c_j)_{j=0}^{m-1}$ is a sequence of real coefficients. Then

$$\begin{aligned} \textit{rank}\,(\mathbf{H}_m) \leq & 1 + \min\{r-1, m-1\}(2e)^{r-1} \\ & + \max\{0, m-r\} \left(\frac{2e}{r-1}\right)^{r-1} (m-1)^{r-1}. \end{aligned} \tag{4.95}$$

Proof. As $\mathbf{G}$ is a symmetric and positive semi-definite matrix, its eigenvalues are real and non-negative and its eigenvectors are orthogonal. Let $\{\mathbf{v}_i\}_{i=1}^r$ be a set of orthogonal eigenvectors for $\mathbf{G}$ and γ_i the eigenvalue associated with $\mathbf{v}_i \in \mathbb{R}^n$. Then $\mathbf{G}$ may be written as a sum of rank one matrices as follows,

$$\mathbf{G} = \sum_{i=1}^{r} \gamma_i \mathbf{v}_i \mathbf{v}_i^T.$$

As the Hadamard product is commutative, associative and distributive over addition, for

any $j \in \mathbb{Z}_{\geq 0}$ $\mathbf{G}^{\odot j}$ can also be expressed as a sum of rank 1 matrices,

$$\begin{aligned}
\mathbf{G}^{\odot j} &= \left(\sum_{i=1}^{r} \gamma_i \mathbf{v}_i \mathbf{v}_i^T \right)^{\odot j} \\
&= \left(\sum_{i_1=1}^{r} \gamma_{i_1} \mathbf{v}_{i_1} \mathbf{v}_{i_1}^T \right) \odot \left(\sum_{i_2=1}^{r} \gamma_{i_2} \mathbf{v}_{i_2} \mathbf{v}_{i_2}^T \right) \odot \cdots \odot \left(\sum_{i_j=1}^{r} \gamma_{i_j} \mathbf{v}_{i_j} \mathbf{v}_{i_j}^T \right) \\
&= \sum_{i_1, i_2 \ldots i_j = 1}^{r} \gamma_{i_1} \gamma_{i_2} \cdots \gamma_{i_r} \left(\mathbf{v}_{i_1} \mathbf{v}_{i_1}^T \right) \odot \left(\mathbf{v}_{i_2} \mathbf{v}_{i_2}^T \right) \odot \cdots \odot \left(\mathbf{v}_{i_j} \mathbf{v}_{i_j}^T \right) \\
&= \sum_{i_1, i_2, \ldots, i_j = 1}^{r} \gamma_{i_1} \gamma_{i_2} \cdots \gamma_{i_j} \left(\mathbf{v}_{i_1} \odot \mathbf{v}_{i_2} \odot \cdots \odot \mathbf{v}_{i_j} \right) \left(\mathbf{v}_{i_1} \odot \mathbf{v}_{i_2} \odot \cdots \odot \mathbf{v}_{i_j} \right)^T .
\end{aligned}$$

Note the fourth equality in the above follows from $\mathbf{v}_i \mathbf{v}_i^T = \mathbf{v}_i \otimes \mathbf{v}_i$ and an application of the mixed-product property of the Hadamard product. As matrix rank is sub-additive, the rank of $\mathbf{G}^{\odot j}$ is less than or equal to the number of distinct rank-one matrix summands. This quantity in turn is equal to the number of vectors of the form $\left(\mathbf{v}_{i_1} \odot \mathbf{v}_{i_2} \odot \cdots \odot \mathbf{v}_{i_j} \right)$, where $i_1, i_2, \ldots, i_j \in [r]$. This in turn is equivalent to computing the number of j-combinations with repetition from r objects. Via a stars and bars argument this is equal to $\binom{r+j-1}{j} = \binom{r+j-1}{r(n)-1}$. It therefore follows that

$$\begin{aligned}
\operatorname{rank}(\mathbf{G}^{\odot j}) &\leq \binom{r+j-1}{r-1} \\
&\leq \left(\frac{e(r+j-1)}{r-1} \right)^{r-1} \\
&\leq e^{r-1} \left(1 + \frac{j}{r-1} \right)^{r-1} \\
&\leq (2e)^{r-1} \left(\delta_{j \leq r-1} + \delta_{j > r-1} \left(\frac{j}{r-1} \right)^{r-1} \right).
\end{aligned}$$

The rank of $\mathbf{H}_m$ can therefore be bounded via subadditivity of the rank as

$$\begin{aligned}
\operatorname{rank}(\mathbf{H}_m) =& \operatorname{rank}\left(a_0\mathbf{1}_{n\times n} + \sum_{j=1}^{m-1} c_j \mathbf{G}^{\odot j}\right) \\
\leq& 1 + \sum_{j=1}^{m-1} \operatorname{rank}\left(\mathbf{G}^{\odot j}\right) \\
\leq& 1 + \sum_{j=1}^{m-1} (2e)^{r-1}\left(\delta_{j\leq r-1} + \delta_{j>r-1}\left(\frac{j}{r-1}\right)^{r-1}\right) \\
\leq& 1 + \min\{r-1, m-1\}(2e)^{r-1} \\
&+ \max\{0, m-r\}\left(\frac{2e}{r-1}\right)^{r-1}(m-1)^{r-1}.
\end{aligned} \tag{4.96}$$

□

As our goal here is to characterize the small eigenvalues, then as n grows we need both k and therefore m to grow as well. As a result we will therefore be operating in the regime where $m > r$. To this end we provide the following corollary.

Corollary 4.6.29. *Under the same conditions and setup as Lemma 4.6.28 with $m \geq r \geq 7$ then*

$$\operatorname{rank}(\mathbf{H}_m) < 2m^r.$$

Proof. If $r \geq 7 > 2e + 1$ then $r - 1 > 2e$. As a result from Lemma 4.6.28

$$\begin{aligned}
\operatorname{rank}(\mathbf{H}_m) &\leq 1 + (r-1)(2e)^{r-1} + (m-r)\left(\frac{2e}{r-1}\right)^{r-1}(m-1)^{r-1} \\
&< r(2e)^{r-1} + (m-1)^r \\
&< 2m^r
\end{aligned}$$

as claimed. □

Corollary 4.6.29 implies for any $k \geq 2m^r$, $k \leq n$ that we can apply Lemma 4.6.27 to upper bound the size of the kth eigenvalue. Our goal is to upper bound the decay of the smallest eigenvalue. To this end, and in order to make our bounds as tight as possible, we

therefore choose the truncation point $m(n) = \lfloor (n/2)^{1/r} \rfloor$, note this is the largest truncation which still satisfies $2m(n)^r \leq n$. In order to state the next lemma, we introduce the following pieces of notation: with $\mathcal{L} := \{\ell : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}\}$ define $U : \mathcal{L} \times \mathbb{Z}_{\geq 1} \to \mathbb{R}_{\geq 0}$ as

$$U(\ell, m) = \int_{m-1}^{\infty} \ell(x) dx.$$

Lemma 4.6.30. *Given a sequence of data points $(\mathbf{x}_i)_{i \in \mathbb{Z}_{\geq 1}}$ with $\mathbf{x}_i \in \mathbb{S}^d$ for all $i \in \mathbb{Z}_{\geq 1}$, construct a sequence of row-wise data matrices $(\mathbf{X}_n)_{n \in \mathbb{Z}_{\geq 1}}$, $\mathbf{X}_n \in \mathbb{R}^{n \times d}$, with $\mathbf{x}_i$ corresponding to the ith row of $\mathbf{X}_n$. The corresponding sequence of gram matrices we denote $\mathbf{G}_n := \mathbf{X}_n \mathbf{X}_n^T$. Let $m(n) := \lfloor (n/2)^{1/r(n)} \rfloor$ where $r(n) := \operatorname{rank}(\mathbf{X}_n)$ and suppose for all sufficiently large n that $m(n) \geq r(n) \geq 7$. Let the coefficients $(c_j)_{j=0}^{\infty}$ with $c_j \in \mathbb{R}_{\geq 0}$ for all $j \in \mathbb{Z}_{\geq 0}$ be such that 1) the series $\sum_{j=0}^{\infty} c_j \rho^j$ converges for all $\rho \in [-1, 1]$ and 2) $(c_j)_{j=0}^{\infty} = \mathcal{O}(\ell(j))$, where $\ell \in \mathcal{L}$ satisfies $U(\ell, m(n)) < \infty$ for all n and is monotonically decreasing. Consider the sequence of kernel matrices indexed by n and defined as*

$$n\mathbf{K}_n = \sum_{j=0}^{\infty} c_j \mathbf{G}_n^{\odot j}.$$

With $\nu : \mathbb{Z}_{\geq 1} \to \mathbb{Z}_{\geq 1}$ suppose $\left\| \mathbf{G}_n^{\odot m(n)} \right\| = \mathcal{O}(n^{-\nu(n)+1})$, then

$$\lambda_n(\mathbf{K}_n) = \mathcal{O}(n^{-\nu(n)} U(\ell, m(n))). \tag{4.97}$$

Proof. By the assumptions of the Lemma we may apply Lemma 4.6.27 and Corollary 4.6.29, which results in

$$\lambda_n(\mathbf{K}_n) \leq \frac{\left\| \mathbf{G}_n^{\odot m(n)} \right\|}{n} \sum_{j=m(n)}^{\infty} c_j = \mathcal{O}(n^{-\nu(n)}) \sum_{j=m(n)}^{\infty} c_j.$$

Additionally, as $(c_j)_{j=0}^{\infty} = \mathcal{O}(\ell(j))$ then

$$\begin{aligned}\lambda_n(\mathbf{K}_n) &= \mathcal{O}\left(n^{-\nu(n)} \sum_{j=m(n)}^{\infty} \ell(j)\right) \\ &= \mathcal{O}\left(n^{-\nu(n)} \int_{m(n)-1}^{\infty} \ell(x)dx\right) \\ &= \mathcal{O}\left(n^{-\nu(n)} U(\ell, m(n))\right)\end{aligned}$$

as claimed. □

Based on Lemma 4.6.27 we provide Theorem 4.6.31, which considers three specific scenarios for the decay of the power series coefficients inspired by Lemma 4.3.4.

Theorem 4.6.31. *In the same setting, and also under the same assumptions as in Lemma 4.6.30, then*

1. *if* $c_p = \mathcal{O}(p^{-\alpha})$ *with* $\alpha > r(n) + 1$ *for all* $n \in \mathbb{Z}_{\geq 0}$ *then* $\lambda_n(\mathbf{K}_n) = \mathcal{O}\left(n^{-\frac{\alpha-1}{r(n)}}\right)$,
2. *if* $c_p = \mathcal{O}(e^{-\alpha\sqrt{p}})$, *then* $\lambda_n(\mathbf{K}_n) = \mathcal{O}\left(n^{\frac{1}{2r(n)}} \exp\left(-\alpha' n^{\frac{1}{2r(n)}}\right)\right)$ *for any* $\alpha' < \alpha 2^{-1/2r(n)}$,
3. *if* $c_p = \mathcal{O}(e^{-\alpha p})$, *then* $\lambda_n(\mathbf{K}_n) = \mathcal{O}\left(\exp\left(-\alpha' n^{\frac{1}{r(n)}}\right)\right)$ *for any* $\alpha' < \alpha 2^{-1/2r(n)}$.

Proof. First, as $[\mathbf{G}_n]_{ij} \leq 1$ then

$$\frac{\left\|\mathbf{G}^{\odot m(n)}\right\|}{n} \leq \frac{\text{Trace}(\mathbf{G}^{\odot m(n)})}{n} = 1.$$

Therefore, to recover the three results listed we now apply Lemma 4.6.30 with $\nu(n) = 0$. First, to prove 1., under the assumption $\ell(x) = x^{-\alpha}$ with $\alpha > 0$ then

$$\int_{m(n)-1}^{\infty} x^{-\alpha} dx = \frac{(m(n)-1)^{-\alpha+1}}{\alpha - 1}.$$

As a result

$$\lambda_n(\mathbf{K}_n) = \mathcal{O}\left(n^{-\frac{\alpha-1}{r(n)}}\right).$$

To prove ii), under the assumption $\ell(x) = e^{-\alpha\sqrt{x}}$ with $\alpha > 0$ then

$$\int_{m(n)-1}^{\infty} e^{-\alpha\sqrt{x}} dx = \frac{2\exp(-\alpha(\sqrt{m(n)-1})(\alpha\sqrt{m(n)-1}+1)}{\alpha^2}.$$

As a result

$$\lambda_n(\mathbf{K}_n) = \mathcal{O}\left(n^{\frac{1}{2r(n)}} \exp\left(-\alpha' n^{\frac{1}{2r(n)}}\right)\right)$$

for any $\alpha' < \alpha 2^{-1/2r(n)}$. Finally, to prove iii), under the assumption $\ell(x) = e^{-\alpha x}$ with $\alpha > 0$ then

$$\int_{m(n)-1}^{\infty} e^{-\alpha x} dx = \frac{\exp(-\alpha(m(n)-1)}{\alpha}.$$

Therefore

$$\lambda_n(\mathbf{K}_n) = \mathcal{O}\left(\exp\left(-\alpha' n^{\frac{1}{r(n)}}\right)\right)$$

again for any $\alpha' < \alpha 2^{-1/2r(n)}$. □

Unfortunately, the curse of dimensionality is clearly present in these results due to the $1/r(n)$ factor in the exponents of n. However, although perhaps somewhat loose we emphasize that these results are certainly far from trivial. In particular, while trivially we know that $\lambda_n(\mathbf{K}_n) \leq Tr(\mathbf{K}_n)/n = \mathcal{O}(n^{-1})$, in contrast, even the weakest result concerning the power law decay our result is a clear improvement as long as $\alpha > r(n) + 1$. For the other settings, i.e., those specified in 2. and 3., our results are significantly stronger.

REFERENCES

[AB02] Martin Anthony and Peter L. Bartlett. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 2002.

[ABP22] Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. "Neural Networks as Kernel Learners: The Silent Alignment Effect." In *International Conference on Learning Representations*, 2022.

[ADH19a] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. "Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks." In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 09–15 Jun 2019.

[ADH19b] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. "On Exact Computation with an Infinitely Wide Neural Net." In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[ALS19a] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "A Convergence Theory for Deep Learning via Over-Parameterization." In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 2019.

[ALS19b] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "On the Convergence Rate of Training Recurrent Neural Networks." 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.

[AM15] Douglas Azevedo and Valdir A Menegatto. "Eigenvalues of dot-product kernels on the sphere." *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, **3**(1), 2015.

[Ana21] Anil Ananthaswamy. "A New Link to an Old Model Could Crack the Mystery of Deep Learning." *Quanta Magazine*, October 2021.

[Bar94] Andrew R Barron. "Approximation and estimation bounds for artificial neural networks." *Machine learning*, **14**(1):115–133, 1994.

[BB21] Alberto Bietti and Francis Bach. "Deep Equals Shallow for ReLU Networks in Kernel Regimes." In *International Conference on Learning Representations*, 2021.

[BES22] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. "High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation." In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[BGG20] Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. "Frequency Bias in Neural Networks for Input of Non-Uniform Density." In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 685–694. PMLR, 13–18 Jul 2020.

[BGL21] Aristide Baratin, Thomas George, César Laurent, R Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. "Implicit Regularization via Neural Feature Alignment." In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2269–2277. PMLR, 13–15 Apr 2021.

[BGW21] Sam Buchanan, Dar Gilboa, and John Wright. "Deep Networks and the Multiple Manifold Problem." In *International Conference on Learning Representations*, 2021.

[BHL19] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. "Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks." *Journal of Machine Learning Research*, **20**(63):1–17, 2019.

[BJK19] Ronen Basri, David W. Jacobs, Yoni Kasten, and Shira Kritchman. "The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies." In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32*, pp. 4763–4772, 2019.

[BM03] Peter L. Bartlett and Shahar Mendelson. "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results." *J. Mach. Learn. Res.*, **3**(null):463–482, mar 2003.

[BM19] Alberto Bietti and Julien Mairal. "On the Inductive Bias of Neural Tangent Kernels." In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[BM22a] Benjamin Bowman and Guido Montúfar. "Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks." In *International Conference on Learning Representations*, 2022.

[BM22b] Benjamin Bowman and Guido Montufar. "Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime." In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[BR93] Avrim L. Blum and Ronald L. Rivest. *Training a 3-node neural network is NP-complete*, pp. 9–28. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.

[BT04] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, Boston, MA, 2004.

[BZZ18] Brian Bullins, Cyril Zhang, and Yi Zhang. "Not-So-Random Features." In *International Conference on Learning Representations*, 2018.

[CD07] Andrea Caponnetto and Ernesto De Vito. "Optimal rates for the regularized least-squares algorithm." *Foundations of Computational Mathematics*, **7**(3):331–368, 2007.

[CFW21] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. "Towards Understanding the Spectral Bias of Deep Learning." In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2205–2211. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[CLK21] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "Generalization Error Rates in Kernel Regression: The Crossover from the Noiseless to Noisy Regime." In *Advances in Neural Information Processing Systems*, 2021.

[COB19] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. "On Lazy Training in Differentiable Programming." 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.

[CS09] Youngmin Cho and Lawrence Saul. "Kernel Methods for Deep Learning." In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[CX21] Lin Chen and Sheng Xu. "Deep Neural Tangent Kernel and Laplace Kernel Have the Same RKHS." In *International Conference on Learning Representations*, 2021.

[Dav21] Tom Davis. "A General Expression for Hermite Expansions with Applications." 2021.

[DFS16] Amit Daniely, Roy Frostig, and Yoram Singer. "Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity." In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[DGM20] Yonatan Dukler, Quanquan Gu, and Guido Montúfar. "Optimization Theory for ReLU Neural Networks Trained with Normalization Layers." In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2751–2760. PMLR, 13–18 Jul 2020.

[dHR18] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. "Gaussian Process Behaviour in Wide Deep Neural Networks." In *International Conference on Learning Representations*, 2018.

[DLL19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. "Gradient Descent Finds Global Minima of Deep Neural Networks." In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 2019.

[DPG14] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization." In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[Dum06] Ilya Dumer. "Covering an Ellipsoid with Equal Balls." *J. Comb. Theory Ser. A*, **113**(8):1667–1676, nov 2006.

[DZP19] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. "Gradient Descent Provably Optimizes Over-parameterized Neural Networks." In *International Conference on Learning Representations*, 2019.

[EMW20] Weinan E, Chao Ma, and Lei Wu. "A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics." *Science China Mathematics*, **63**:1235–1258, 2020.

[EWS22] Andrew Engel, Zhichao Wang, Anand Sarwate, Sutanay Choudhury, and Tony Chiang. "TorchNTK: A Library for Calculation of Neural Tangent Kernels of PyTorch Models." 2022.

[FA00] K. Fukumizu and S. Amari. "Local minima and plateaus in hierarchical structures of multilayer perceptrons." *Neural Networks*, **13**(3):317–327, 2000.

[Fol99] G. B. Folland. *Real analysis: Modern techniques and their applications*. Wiley, New York, 1999.

[FW20] Zhou Fan and Zhichao Wang. "Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks." In *Advances in Neural Information Processing Systems*, volume 33, pp. 7710–7721. Curran Associates, Inc., 2020.

[GB10] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256. PMLR, 2010.

[GGJ22] Amnon Geifman, Meirav Galun, David Jacobs, and Ronen Basri. "On the Spectral Bias of Convolutional Neural Tangent and Gaussian Process Kernels." In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[GKS87] Y. Gordon, Hermann König, and Carsten Schütt. "Geometric and probabilistic estimates for entropy and approximation numbers of operators." *Journal of Approximation Theory*, **49**:219–239, 1987.

[GLS18] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. "Implicit Bias of Gradient Descent on Linear Convolutional Networks." In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[Gro19] T. H. Gronwall. "Note on the Derivatives with Respect to a Parameter of the Solutions of a System of Differential Equations." *Annals of Mathematics*, **20**(4):292–296, 1919.

[GWB17] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. "Implicit Regularization in Matrix Factorization." In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[GYK20] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. "On the Similarity between the Laplace and Neural Tangent Kernels." In *Advances in Neural Information Processing Systems*, volume 33, pp. 1451–1461. Curran Associates, Inc., 2020.

[HHV22] Ningyuan Teresa Huang, David W. Hogg, and Soledad Villar. "Dimensionality Reduction, Regularization, and Generalization in Overparameterized Regressions." *SIAM J. Math. Data Sci.*, **4**(1):126–152, 2022.

[HMW20] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. "Array programming with NumPy." *Nature*, **585**(7825):357–362, September 2020.

[HN20] Boris Hanin and Mihai Nica. "Finite Depth and Width Corrections to the Neural Tangent Kernel." In *International Conference on Learning Representations*, 2020.

[Hun07] J. D. Hunter. "Matplotlib: A 2D graphics environment." *Computing in Science & Engineering*, **9**(3):90–95, 2007.

[HY20] Jiaoyang Huang and Horng-Tzer Yau. "Dynamics of Deep Neural Networks and Neural Tangent Hierarchy." In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4542–4551. PMLR, 13–18 Jul 2020.

[HZL22] Insu Han, Amir Zandieh, Jaehoon Lee, Roman Novak, Lechao Xiao, and Amin Karbasi. "Fast Neural Kernel Embeddings for General Activations." In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[HZR15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

[JBM22] Hui Jin, Pradeep Kr. Banerjee, and Guido Montúfar. "Learning Curves for Gaussian Process Regression with Power-Law Priors and Targets." In *International Conference on Learning Representations*, 2022.

[JGH18] Arthur Jacot, Franck Gabriel, and Clement Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[JM21] Hui Jin and Guido Montúfar. "Implicit bias of gradient descent for mean squared error regression with wide neural networks.", 2021. arXiv:2006.07356.

[JT19] Ziwei Ji and Matus Telgarsky. "The implicit bias of gradient descent on nonseparable data." In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of*

the Thirty-Second Conference on Learning Theory, volume 99 of *Proceedings of Machine Learning Research*, pp. 1772–1798. PMLR, 25–28 Jun 2019.

[JT20] Ziwei Ji and Matus Telgarsky. "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks." In *International Conference on Learning Representations*, 2020.

[KAA20] Ryo Karakida, Shotaro Akaho, and Shun ichi Amari. "Universal statistics of Fisher information in deep neural networks: mean field approach." *Journal of Statistical Mechanics: Theory and Experiment*, **2020**(12):124005, 2020.

[KAA21] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. "Pathological Spectra of the Fisher Information Metric and Its Variants in Deep Neural Networks." *Neural Computation*, **33**(8):2274–2307, July 2021. https://direct.mit.edu/neco/article-pdf/33/8/2274/1930880/neco_a_01411.pdf.

[Kaw16] Kenji Kawaguchi. "Deep Learning without Poor Local Minima." In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[Kaz56] Donat K. Kazarinoff. "On Wallis' formula." *Edinburgh Mathematical Notes*, **40**:19–21, 1956.

[Kri09] Alex Krizhevsky. "Learning multiple layers of features from tiny images." Technical report, 2009.

[KRP16] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. "Jupyter Notebooks – a publishing format for reproducible computational workflows." In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87 – 90. IOS Press, 2016.

[KS95] Pascal Koiran and Eduardo Sontag. "Neural Networks with Quadratic VC Dimension." In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995.

[LAR22] Li, Andreeto, Ranzato, and Perona. "Caltech 101.", Apr 2022.

[LBB98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, **86**(11):2278–2324, 1998.

[LBD20] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. "The large learning rate phase of deep learning: the catapult mechanism." *arXiv preprint arXiv:2003.02218*, 2020.

[LBN18] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. "Deep Neural Networks as Gaussian Processes." In *International Conference on Learning Representations*, 2018.

[LBO12] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pp. 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet. "A Random Matrix Approach to Neural Networks." *The Annals of Applied Probability*, **28**(2):1190–1248, 2018.

[LLP93] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function." *Neural Networks*, **6**(6):861–867, 1993.

[LMX22] Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. "On the Exact Computation of Linear Frequency Principle Dynamics and Its Generalization." *SIAM Journal on Mathematics of Data Science*, **4**(4):1272–1292, 2022.

[LMZ18] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. "Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations." 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*, pp. 2–47. PMLR, 06–09 Jul 2018.

[LSO20] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. "Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks." In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4313–4324. PMLR, 2020.

[LSP20] Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. "Finite Versus Infinite Neural Networks: an Empirical Study." In *Advances in Neural Information Processing Systems*, volume 33, pp. 15156–15172. Curran Associates, Inc., 2020.

[LXS19] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent." In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[LZB20a] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. "On the linearity of large non-linear models: when and why the tangent kernel is constant." *CoRR*, **abs/2010.01092**, 2020.

[LZB20b] Chaoyue Liu, Libin Zhu, and Misha Belkin. "On the linearity of large non-linear models: when and why the tangent kernel is constant." In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pp. 15954–15964. Curran Associates, Inc., 2020.

[LZB22] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks." *Applied and Computational Harmonic Analysis*, **59**:85–116, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[MAT22] M. Murray, V. Abrol, and J. Tanner. "Activation function design for deep networks: linearity and effective initialisation." *Applied and Computational Harmonic Analysis*, **59**:117–154, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[Mer09] J. Mercer. "Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **209**:415–446, 1909.

[MM16] Dmytro Mishkin and Jiri Matas. "All you need is a good init." In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, Conference Track Proceedings*, 2016.

[Nea96] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996.

[Ngu21] Quynh Nguyen. "On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths." In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8056–8062. PMLR, 18–24 Jul 2021.

[NLG19] Mor Shpigel Nacson, J. Lee, Suriya Gunasekar, Nathan Srebro, and Daniel Soudry. "Convergence of Gradient Descent on Separable Data." In *AISTATS*, 2019.

[NM20] Quynh N Nguyen and Marco Mondelli. "Global Convergence of Deep Networks with One Wide Layer Followed by Pyramidal Topology." In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural*

Information Processing Systems, volume 33, pp. 11961–11972. Curran Associates, Inc., 2020.

[NMM21] Quynh Nguyen, Marco Mondelli, and Guido Montúfar. "Tight Bounds on the Smallest Eigenvalue of the Neural Tangent Kernel for Deep ReLU Networks." In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8119–8129. PMLR, 2021.

[NSS22] Roman Novak, Jascha Sohl-Dickstein, and Samuel S Schoenholz. "Fast Finite Width Neural Tangent Kernel." In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17018–17044. PMLR, 17–23 Jul 2022.

[NTS15] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning." In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[NTS17] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. "Geometry of Optimization and Implicit Regularization in Deep Learning.", 2017. arXiv:1705.03071.

[NXB19] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. "Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes." In *7th International Conference on Learning Representations*. OpenReview.net, 2019.

[OD14] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

[OFL19] Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. "Generalization Guarantees for Neural Networks via Harnessing the Low-rank Structure of the Jacobian." *CoRR*, **abs/1906.05392**, 2019.

[OS20] Samet Oymak and Mahdi Soltanolkotabi. "Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks." *IEEE Journal on Selected Areas in Information Theory*, **1**(1), 2020.

[Pap20] Vardan Papyan. "Traces of Class/Cross-Class Structure Pervade Deep Learning Spectra." *Journal of Machine Learning Research*, **21**(252):1–64, 2020.

[PB17] Jeffrey Pennington and Yasaman Bahri. "Geometry of Neural Network Loss Surfaces via Random Matrix Theory." In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2798–2806. PMLR, 06–11 Aug 2017.

[PG07] Fernando Pérez and Brian E. Granger. "IPython: a System for Interactive Scientific Computing." *Computing in Science and Engineering*, **9**(3):21–29, May 2007.

[PGM19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

[Pis89] Gilles Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge Tracts in Mathematics. Cambridge University Press, 1989.

[PLR16] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. "Exponential expressivity in deep neural networks through transient chaos." In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[Pol63] Boris Polyak. "Gradient methods for the minimisation of functionals." *Ussr Computational Mathematics and Mathematical Physics*, **3**:864–878, 12 1963.

[PSG20] Abhishek Panigrahi, Abhishek Shetty, and Navin Goyal. "Effect of Activation Functions on the Training of Overparametrized Neural Nets." In *International Conference on Learning Representations*, 2020.

[PW17] Jeffrey Pennington and Pratik Worah. "Nonlinear random matrix theory for deep learning." In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[PW18] Jeffrey Pennington and Pratik Worah. "The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network." In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[RBA19] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. "On the Spectral Bias of Neural Networks." In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 09–15 Jun 2019.

[RBV10] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. "On Learning with Integral Operators." *Journal of Machine Learning Research*, **11**(30):905–934, 2010.

[RJK19] Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. "The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies." 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.

[RKH21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision." In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[RR08a] Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines." In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[RR08b] Ali Rahimi and Benjamin Recht. "Uniform approximation of functions with random bases." In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 555–561, 2008.

[SAD22] James Benjamin Simon, Sajant Anand, and Mike Deweese. "Reverse Engineering the Neural Tangent Kernel." In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20215–20231. PMLR, 17–23 Jul 2022.

[Sch11] J. Schur. "Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen." *Journal für die reine und angewandte Mathematik*, **140**:1–28, 1911.

[SGG17] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. "Deep Information Propagation." In *International Conference on Learning Representations (ICLR)*, 2017.

[SGJ21] Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clement Hongler, Wulfram Gerstner, and Johanni Brea. "Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances." In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9722–9732. PMLR, 18–24 Jul 2021.

[SH21] Meyer Scetbon and Zaid Harchaoui. "A spectral analysis of dot-product kernels." In *International conference on artificial intelligence and statistics*, pp. 3394–3402. PMLR, 2021.

[SHN18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. "The Implicit Bias of Gradient Descent on Separable Data." *Journal of Machine Learning Research*, **19**(70):1–57, 2018.

[SS89] Eduardo D. Sontag and Héctor J. Sussmann. "Backpropagation Can Give Rise To Spurious Local Minima Even For Networks Without Hidden Layers." *Complex Systems*, **3**:91–106, 1989.

[SS91] Eduardo D. Sontag and Héctor J. Sussmann. "Back propagation separates where perceptrons do." *Neural Networks*, **4**(2):243–249, 1991.

[SS20] Justin Sirignano and Konstantinos Spiliopoulos. "Mean Field Analysis of Neural Networks: A Law of Large Numbers." *SIAM Journal on Applied Mathematics*, **80**(2):725–752, 2020.

[SY19] Lili Su and Pengkun Yang. "On Learning Over-parameterized Neural Networks: A Functional Approximation Perspective." 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.

[SY20] Zhao Song and Xin Yang. "Quadratic Suffices for Over-parametrization via Matrix Chernoff Bound.", 2020. arXiv:1906.03593.

[Tel21] Matus Telgarsky. "Deep learning theory lecture notes." `https://mjt.cs.illinois.edu/dlt/`, 2021. Version: 2021-10-27 v0.0-e7150f2d (alpha).

[TL19] Mingxing Tan and Quoc Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019.

[Ver12] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices." In *Compressed Sensing*, chapter 5. Cambridge University Press, 2012.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[VGO20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods*, **17**:261–272, 2020.

[VY21] Maksim Velikanov and Dmitry Yarotsky. "Explicit loss asymptotics in the gradient descent training of neural networks." In *Advances in Neural Information Processing Systems*, volume 34, pp. 2570–2582. Curran Associates, Inc., 2021.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[Wey12] Hermann Weyl. "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)." *Mathematische Annalen*, **71**(4):441–479, 1912.

[WGL20] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. "Kernel and Rich Regimes in Overparametrized Models." In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 09–12 Jul 2020.

[Wil96] Christopher Williams. "Computing with Infinite Networks." In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[WTP19] Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. "Gradient Dynamics of Shallow Univariate ReLU Networks." 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.

[WZE17] Lei Wu, Zhanxing Zhu, and Weinan E. "Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes." *CoRR*, **abs/1706.10239**, 2017.

[XLS17] Bo Xie, Yingyu Liang, and Le Song. "Diverse Neural Network Learns True Target Functions." In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1216–1224. PMLR, 2017.

[XZX19] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. "Training Behavior of Deep Neural Network in Frequency Domain." In Tom Gedeon, Kok Wai Wong, and Minho Lee, editors, *Neural Information Processing*, pp. 264–274, Cham, 2019. Springer International Publishing.

[YAA22] Ge Yang, Anurag Ajay, and Pulkit Agrawal. "Overcoming The Spectral Bias of Neural Value Approximation." In *International Conference on Learning Representations*, 2022.

[Yan20] Greg Yang. "Tensor Programs II: Neural Tangent Kernel for Any Architecture.", 2020.

[YMC22] Rubing Yang, Jialin Mao, and Pratik Chaudhari. "Does the Data Induce Capacity Control in Deep Learning?" In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25166–25197. PMLR, 17–23 Jul 2022.

[YS19] Greg Yang and Hadi Salman. "A Fine-Grained Spectral Perspective on Neural Networks.", 2019.

[ZBH17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization." In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[ZCZ20] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. "Gradient descent optimizes over-parameterized deep ReLU networks." *Machine learning*, **109**(3):467–492, 2020.

[ZG19] Difan Zou and Quanquan Gu. "An Improved Analysis of Training Over-parameterized Deep Neural Networks." 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.

[ZXL20] Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. "A type of generalization error induced by initialization in deep neural networks." In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pp. 144–164. PMLR, 20–24 Jul 2020.