Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime

Anonymous Author(s)
Affiliation
Address
email

Abstract

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. An implication of the bounds is that the network is biased to learn the top eigenfunctions of the Neural Tangent Kernel not just on the training set but over the entire input space. This bias depends on the model architecture and input distribution alone and thus does not depend on the target function which does not need to be in the RKHS of the kernel. The result is valid for deep architectures with fully connected, convolutional, and residual layers. Furthermore the width does not need to grow polynomially with the number of samples in order to obtain high probability bounds up to a stopping time. The proof exploits the low-effective-rank property of the Fisher Information Matrix at initialization, which implies a low effective dimension of the model (far smaller than the number of parameters). We conclude that local capacity control from the low effective rank of the Fisher Information Matrix is still underexplored theoretically.

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 (Zhang et al., 2017), 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 (Neyshabur et al., 2015, 2017; Gunasekar et al., 2017). 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 (Soudry et al., 2018; Ji & Telgarsky, 2019; Nacson et al., 2019; Gunasekar et al., 2018), minimum nuclear norm bias for matrix factorization (Gunasekar et al., 2017; Li et al., 2018; Gunasekar et al., 2018), and minimum RKHS norm bias in the kernel regime (Zhang et al., 2020).

Empirically it is known that neural networks tend to learn low Fourier frequencies first and add higher frequencies only later in training (Rahaman et al., 2019; Xu et al., 2019; Yang et al., 2022), 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

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

univariate ReLU networks Basri et al. (2019, 2020) demonstrate that the dominant eigenfunctions of the Neural Tangent Kernel (NTK) (Jacot et al., 2018) correspond to the low Fourier frequencies for the uniform distribution and more generally to smoother components for nonuniform distributions. This echos the results by Williams et al. (2019) and Jin & Montúfar (2021) 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 Arora et al. (2019a), Basri et al. (2020), and Cao et al. (2021). A recent work by Bowman & Montúfar (2022) was able to demonstrate that spectral bias holds off the training set for shallow feedforward networks when the network is underparameterized. In the present work 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 Liu et al. (2020b) 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.

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 3.5 and Corollary 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 3.7 and Observation 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 3.9).

1.2 Related Work

NTK Convergence Results The NTK was introduced by Jacot et al. (2018) while almost concurrently Du et al. (2019b) 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 (Du et al., 2019b,a; Oymak & Soltanolkotabi, 2020; Allen-Zhu et al., 2019; Nguyen & Mondelli, 2020; Nguyen, 2021; Zou et al., 2020; Zou & Gu, 2019). 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 (Ji & Telgarsky, 2020) or linear width for the squared loss (E et al., 2020; Su & Yang, 2019; Bowman & Montúfar, 2022).

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. Arora et al. 2019a; Oymak et al. 2020; Li et al. 2020). Papyan (2020) 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. Karakida et al. 2021; Pennington & Worah 2018; Pennington & Bahri 2017; Fan & Wang 2020; Yang & Salman 2019). Recently the spectrum of the NTK integral operator for ReLU networks has been shown to asymptotically follow a power law (Velikanov & Yarotsky, 2021). Arora et al. (2019a) provided a generalization bound that is effective when the labels align with the top eigenvectors of the NTK. Oymak et al. (2020) were able to use the low effective rank of the NTK to obtain generalization bounds, and Li et al. (2020) 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 Yang et al. (2021). 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 Luo et al. (2020) explicitly tracked the dynamics of the infinite width shallow model in the Fourier domain. Arora et al. (2019a) 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. Cao et al. (2021) proved a similar statement for training both layers, and Basri et al. (2020) 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 (Bowman & Montúfar, 2022) 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.

2 Preliminaries

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.

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 Jacot et al. (2018). We recall the definition of the analytical NTK

$$K^\infty(x, x') := \mathbb{E}_{\theta_0 \sim \mu} \left[\langle \nabla_\theta f(x; \theta_0), \nabla_\theta f(x'; \theta_0) \rangle \right],$$

where the expectation is taken over the parameter initialization $\theta_0 \sim \mu$. The kernel K^∞ induces an integral operator $T_{K^\infty} : L^2(X, \rho) \to L^2(X, \rho)$

$$T_{K^\infty} g(x) := \int_X K^\infty(x, s) g(s) d\rho(s), \tag{1}$$

where X is our input space and ρ is the input distribution. We assume our training inputs $x_1, \ldots, x_n$ are i.i.d. samples from ρ. More generally, for a continuous kernel $K(x, x')$ we define

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

$$T_K g(x) := \int_X K(x, s) g(s) d\rho(s). \tag{2}$$

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 $x_1, \ldots, x_n$ introduces a discretization of the operator T_{K^∞}

$$T_n g(x) := \frac{1}{n} \sum_{i=1}^{n} K^\infty(x, x_i) g(x_i) = \int_X K^\infty(x, s) g(s) d\widehat{\rho}(s), \tag{3}$$

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

$$T_n^t g(x) := \frac{1}{n} \sum_{i=1}^{n} K_t(x, x_i) g(x_i) = \int_X K_t(x, s) g(s) d\widehat{\rho}(s). \tag{4}$$

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 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. Arora et al. 2019a; Oymak et al. 2020; Li et al. 2020). For ReLU networks the spectrum is known to asymptotically follow a power law $\sigma_i \sim \Lambda i^{-\nu}$ (Velikanov & Yarotsky, 2021). The goal of this work 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.

2.3 Applicable Architectures

We now specify an architecture for our model $f(x; \theta)$. We consider deep networks of the form

$$\begin{aligned} \alpha^{(0)} &:= x, \\ \alpha^{(l)} &:= \phi_l(\theta^{(l)}, \alpha^{(l-1)}), \quad l \in [L], \\ f(x; \theta) &:= \frac{1}{\sqrt{m_L}} v^T \alpha^{(L)}, \end{aligned}$$

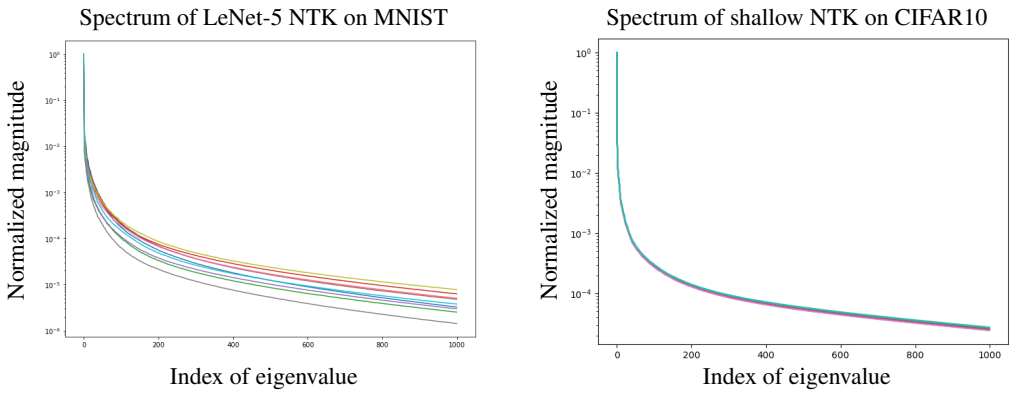

Figure 1: We plot the NTK spectrum on MNIST and CIFAR10 for two networks using 10 random parameter initializations and data batches. In both plots the x-axis represents the eigenvalue index k (linear scale) and the y-axis the normalized eigenvalue λ_k/λ_1 magnitude (log scale). To avoid numerical issues, 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.

where each $\phi_l(\theta^{(l)}, \bullet) : \mathbb{R}^{m_{l-1}} \to \mathbb{R}^{m_l}$ 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 ϕ_l has one of the following forms:

$$\text{Fully Connected} : \phi_l(\theta^{(l)}, \alpha^{(l-1)}) = \sigma\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right)$$

$$\text{Convolutional} : \phi_l(\theta^{(l)}, \alpha^{(l-1)}) = \sigma\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} * \alpha^{(l-1)}\right)$$

$$\text{Residual} : \alpha^{(l)} = \sigma\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right) + \alpha^{(l-1)}$$

Here $\theta^{(l)} = vec(W^{(l)})$ and σ is a twice continuously differentiable function such that σ and σ' 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 the appendix. 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 Hanin & Nica (2020) 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 Zhang et al. (2020) 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 (Zhang et al., 2020). 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 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 the appendix.

Assumption 3.1. *The activation σ is twice continuously differentiable and σ and σ' are Lipschitz.*

Assumption 3.2. *The input domain X is compact with strictly positive Borel measure ρ.*

Assumption 3.3. *The target function f^* satisfies $\|f^*\|_{L^\infty(X,\rho)} = O(1)$.*

Assumption 3.4. *We use the antisymmetric initilization trick so that $f(\bullet;\theta_0) \equiv 0$.*

Most activation functions except for ReLU satisfy Assumption 3.1, 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 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 simply means the target function is bounded. We believe the antisymmetric initialization specified in Assumption 3.4 is not strictly necessary but it greatly simplifies the proofs and associated bounds. To sidestep 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 2.3. We are now ready to introduce the main result.

Theorem 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* (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

$$\|r_t - \exp(-T_K t)r_0\|^2_{L^2(X,\rho)} \leq t^2 \cdot \left[4\|f^*\|^2_\infty \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right].$$

3.1 Interpretation and Consequences

Theorem 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 (Yang, 2020). 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 (Du et al., 2019b,a; Huang & Yau, 2020). Bounds that hold uniformly over x, x' of the same rate were provided by Bowman & Montúfar (2022) and Buchanan et al. (2021). A more pessimistic estimate of $1/m^{1/4}$ is provided by Arora et al. (2019b). 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 Appendix 10).

Assumption 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 3.5 we get the following corollary.

Corollary 3.7. *Let $\delta(m)$ be defined as in Assumption 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 (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 3.7 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 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 3.8. *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 3.9. *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 3.10. *Let $\delta(m)$ be defined as in Assumption 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* (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 3.10 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 3.10 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 3.5 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 3.7 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 Velikanov & Yarotsky (2021) 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 $m, n = \tilde{\Omega}(poly(\epsilon^{-1}))$.

3.2 Technical Comparison to Prior Work

Lee et al. (2019); Arora et al. (2019b) 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. Arora et al. (2019a); Cao et al. (2021); Basri et al. (2020) proved statements similar to Theorem 3.5 and Corollary 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 $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Arora et al. (2019a); Basri et al. (2020) operate in the regime where $m = \Omega(poly(n))$ and as a benefit do not need to employ a stopping time. Cao et al. (2021) 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 Bowman & Montúfar (2022), which demonstrated a version of Corollary 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.

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 (1) and (4).

Bounding Operator Deviations Bowman & Montúfar (2022) 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 3.5 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 $H(x;\theta) := \nabla^2_\theta f(x;\theta)$ 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 Liu et al. (2020b) demonstrated that for *fixed* x and $R > 0$ we have with high probability over the initialization θ_0

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|H(x,\theta)\|_{op} = \tilde{O}\left(\frac{R}{\sqrt{m}} poly(R/\sqrt{m})\right). \tag{5}$$

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 (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 Appendix 6 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

$$F := \int_X \nabla_\theta f(x;\theta_0) \nabla_\theta f(x;\theta_0)^T d\rho(x). \tag{6}$$

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 Appendix 6 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 Bowman & Montúfar (2022) 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 (Yang et al., 2021). 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.

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.

Limitations 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 (Baratin et al., 2021; Atanasov et al., 2022), 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.

Broader impacts We do not foresee any negative societal impacts of characterizing the spectral bias of neural networks. To the contrary we believe that cataloging the properties that networks are biased towards in a variety of regimes will be essential to developing fair and interpretable artificial intelligence over the long-term.

[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 $\hat{\rho}$.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

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 (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/arora19a.html.

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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper/2019/file/dbc4d84bfcfe2284ba11beffb853a8c4-Paper.pdf.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1NvflqAdoom.

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 (eds.), *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. URL https://proceedings.mlr.press/v130/baratin21a.html.

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. ISSN 1532-4435.

Ronen Basri, 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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/5ac8bb8a7d745102a978c5f8ccdb61b8-Paper.pdf.

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 (eds.), *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. URL https://proceedings.mlr.press/v119/basri20a.html.

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

Sam Buchanan, Dar Gilboa, and John Wright. Deep networks and the multiple manifold problem. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=O-6Pm_d_Q-.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2205–2211. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/304. URL https://doi.org/10.24963/ijcai.2021/304. Main Track.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings*

of Machine Learning Research, pp. 1675–1685. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/du19c.html.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=S1eK3i09YQ.

Ilya Dumer. Covering an ellipsoid with equal balls. *J. Comb. Theory Ser. A*, 113(8):1667–1676, nov 2006. ISSN 0097-3165. doi: 10.1016/j.jcta.2006.03.021. URL https://doi.org/10.1016/j.jcta.2006.03.021.

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.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/58191d2a914c6dae66371c9dcdc91b41-Paper.pdf.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/0e98aeeb54acf612b9eb4e48a269814c-Paper.pdf.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgndT4KwB.

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. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In Hal Daumé III and Aarti Singh (eds.), *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. URL https://proceedings.mlr.press/v119/huang20l.html.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In Alina Beygelzimer and Daniel Hsu (eds.), *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. URL https://proceedings.mlr.press/v99/ji19a.html.

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. URL https://openreview.net/forum?id=HygegyrYwH.

Hui Jin and Guido Montúfar. Implicit bias of gradient descent for mean squared error regression with wide neural networks, 2021. arXiv:2006.07356.

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. ISSN 0899-7667. doi: 10.1162/neco_a_01411. URL https://doi.org/10.1162/neco_a_01411. _eprint: https://direct.mit.edu/neco/article-pdf/33/8/2274/1930880/neco_a_01411.pdf.

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 (eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87 – 90. IOS Press, 2016.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

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. doi: 10.1109/5.726791.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4313–4324. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/li20j.html.

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 (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2–47. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/li18a.html.

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, 2020a. URL https://arxiv.org/abs/2010.01092.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15954–15964. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/file/b7ae8fecf15b8b6c3c69eceae636d203-Paper.pdf.

Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. On the exact computation of linear frequency principle dynamics and its generalization, 2020. URL https://arxiv.org/abs/2010.08153.

Mor Shpigel Nacson, J. Lee, Suriya Gunasekar, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *AISTATS*, 2019.

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 (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6614.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning, 2017. arXiv:1705.03071.

Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. In Marina Meila and Tong Zhang (eds.), *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. URL https://proceedings.mlr.press/v139/nguyen21a.html.

Quynh 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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11961–11972. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/8abfe8ac9ec214d68541fcb888c0b4c3-Paper.pdf.

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):84–105, 2020.

Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the Jacobian, 2020. URL https://openreview.net/forum?id=ryl5CJSFPS.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020. URL http://jmlr.org/papers/v21/20-933.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach 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*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In Doina Precup and Yee Whye Teh (eds.), *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. URL https://proceedings.mlr.press/v70/pennington17a.html.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/18bb68e2b38e4a8ce7cf4f6b2625768c-Paper.pdf.

Fernando Pérez and Brian E. Granger. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.53. URL https://ipython.org.

Gilles Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge Tracts in Mathematics. Cambridge University Press, 1989. doi: 10.1017/CBO9780511662454.

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 (eds.), *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. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(30):905–934, 2010. URL http://jmlr.org/papers/v11/rosasco10a.html.

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. URL http://jmlr.org/papers/v19/18-188.html.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/253f7b5d921338af34da817c00f42753-Paper.pdf.

Matus Telgarsky. Deep learning theory lecture notes. https://mjt.cs.illinois.edu/dlt/, 2021. Version: 2021-10-27 v0.0-e7150f2d (alpha).

Maksim Velikanov and Dmitry Yarotsky. Explicit loss asymptotics in the gradient descent training of neural networks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 2570–2582, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/14faf969228fc18fcd4fcf59437b0c97-Abstract.html.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, chapter 5. Cambridge University Press, 2012.

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. doi: 10.1038/s41592-019-0686-2.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/1f6419b1cbe79c71410cb320fc094775-Paper.pdf.

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 (eds.), *Neural Information Processing*, pp. 264–274, Cham, 2019. Springer International Publishing.

Ge Yang, Anurag Ajay, and Pulkit Agrawal. Overcoming the spectral bias of neural value approximation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vIC-xLFuM6.

Greg Yang. Tensor programs II: Neural tangent kernel for any architecture, 2020. URL https://arxiv.org/abs/2006.14548.

Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks, 2019. URL https://arxiv.org/abs/1907.10599.

Rubing Yang, Jialin Mao, and Pratik Chaudhari. Does the data induce capacity control in deep learning? *ArXiv*, abs/2110.14163, 2021.

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. URL https://openreview.net/forum?id=Sy8gdB9xx.

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 (eds.), *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. URL https://proceedings.mlr.press/v107/zhang20a.html.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/6a61d423d02a1c56250dc23ae7ff12f3-Paper.pdf.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine learning*, 109:467–492, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2 and Section 3
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs are included in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix 11
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In Figure 1 we plot the curve for each of the 10 runs of the experiment.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix 11
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix 11
   (b) Did you mention the license of the assets? [Yes] See Appendix 11
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Appendix 11
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix 11
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix 11
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

The appendix is organized as follows.

- In Appendix 6 we bound the $L^2(X, \rho)$ metric entropy of the linearized model. This is necessary to bound the operator deviation $T_K - T_n$.
- In Appendix 7 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 Appendix 8 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 3.5.
- In Appendix 9 we put the aforementioned results together to prove Theorem 3.5.
- In Appendix 10 we explain the merit of Assumption 3.6.
- In Appendix 11 we describe the details of our experiments with a link to the relevant code.

6 Covering Number for the Linearized Model

Our approach to generalization will be based on metric entropy (see, e.g., Wainwright, 2019), a fundamental tool in learning theory. We recall some basic definitions.

Definition 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 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 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 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

$$\|f_{lin}(\bullet;\theta_1) - f_{lin}(\bullet;\theta_2)\|^2_{L^2(X,\nu)} = (\theta_1 - \theta_2)^T \left[\int_X g(x)g(x)^T d\nu(x)\right](\theta_1 - \theta_2)$$
$$= (\theta_1 - \theta_2)^T M (\theta_1 - \theta_2) = \|\theta_1 - \theta_2\|^2_M.$$

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 $\|\theta - \theta'\|_M \le \epsilon$. Well then for any $h \in \mathcal{C}_A$ we can consider $\hat{\theta}(h)$ and choose $\theta' \in N$ such that $\epsilon \ge \left\|\hat{\theta}(h) - \theta'\right\|_M = \left\|f_{lin}(\bullet;\hat{\theta}(h)) - f_{lin}(\bullet;\theta')\right\|_{L^2(X,\nu)}$. 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 $\|f_{lin}(\bullet;\theta) - h(\bullet)\|_{L^2(X,\nu)} \le \epsilon$. However since $h(\bullet) = f_{lin}(\bullet;\hat{\theta}(h))$ we have that $\epsilon \ge \left\|f_{lin}(\bullet;\theta) - f_{lin}(\bullet;\hat{\theta}(h))\right\|_2 = \left\|\theta - \hat{\theta}(h)\right\|_M$. 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 $\sigma_1 \ge \cdots \ge \sigma_p \ge 0$ be the eigenvalues of $M^{1/2}$. We note that

$$\{v \in \mathbb{R}^p : \|v\|_M \le 1\} = \left\{v \in \mathbb{R}^p : \sum_{i=1}^p \sigma_i^2 v_i^2 \le 1\right\}.$$

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

$$E_a := \left\{v \in \mathbb{R}^p : \sum_{i=1}^p \frac{v_i^2}{a_i^2} \le 1\right\}$$

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 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 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 Dumer (2006). 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 Pisier (1989) which also appeared in Gordon et al. (1987, 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 6.6 (Dumer 2006; Pisier 1989; Gordon et al. 1987)**.** *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 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 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 $P(E_a) + B^{\mu_\gamma}_{\gamma/2} \subset \left(1 + \frac{\gamma}{2(1-\gamma)}\right) P(E_a)$. Therefore

$$n \leq \frac{vol\left(P(E_a) + B^{\mu_\gamma}_{\gamma/2}\right)}{vol(B^{\mu_\gamma}_{\gamma/2})} \leq \frac{vol\left[\left(1 + \frac{\gamma}{2(1-\gamma)}\right) P(E_a)\right]}{vol(B^{\mu_\gamma}_{\gamma/2})} = \left(\frac{2}{\gamma} + \frac{1}{1-\gamma}\right)^{\mu_\gamma} \prod_{i \in J_\gamma} a_i.$$

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 6.5 and 6.6 we see that the covering number $\mathcal{N}(B^p_R, \|\bullet\|_M, \epsilon)$ 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 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 6.4 and 6.5 that $\mathcal{N}(\mathcal{C}, \|\bullet\|_{L^2(X,\nu)}, \epsilon) = \mathcal{N}(B^p_R, \|\bullet\|_M, \epsilon) = \mathcal{N}(E_{\frac{R}{\epsilon}\sigma}, B^p_1)$ 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 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 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

$$\log\left|\frac{R}{\epsilon}M^{1/2}\right| = \log\left(\prod_{\sigma_i > \epsilon/R}\frac{R}{\epsilon}\sigma_i\right)$$
$$\leq \tilde{p}(M^{1/2}, \epsilon/R)\log\left(\frac{R\sigma_1}{\epsilon}\right)$$
$$\leq \tilde{p}\left(M^{1/2}, \frac{3\epsilon}{4R}\right)\log\left(\frac{R\sigma_1}{\epsilon}\right).$$

□

7 Bounding the Network Hessian and other Technical Items

7.1 Main Hessian Bound

For a fixed input x we will let $H(x,\theta) := \nabla^2_\theta f(x;\theta)$ 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 Liu et al. (2020a, Theorem 3.3), which we state here explicitly for reference.

Theorem 7.1 (Reformulation of Liu et al. 2020a, Theorem 3.3)**.** *Let $f(x;\theta)$ be a general neural network of the form specified in Section 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 7.1 We note that our statement of Theorem 7.1 is not exactly the same as the result of Liu et al. (2020a, Theorem 3.3). Liu et al. (2020a) do not explicitly write the failure probability and the dependence of the Hessian bound on R in the statement of the theorem. In Theorem 7.1 we write the failure probability and dependence on the radius R according to the proof[2] provided by the authors Liu et al. (2020a). 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 7.2). This condition on the initial weight matrices appears in the proof by Liu et al. (2020a). The authors Liu et al. (2020a) 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 7.1 there are constants that depend exponentially on the network depth L. For this reason it 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 7.1. 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 7.2). 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

[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.

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 $\|\theta-\theta_0\|_2 \le R$ 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 7.1. Let $f(x;\theta)$ denote any model of the form specified in Section 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

$$\left\|\nabla^2_{\tilde{\theta}} f_{ASI}(x;\tilde{\theta})\right\|_{op} \le \max\left[\|H(x,\theta)\|_{op}, \|H(x,\theta')\|_{op}\right].$$

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 $\left\|\tilde{\theta} - \tilde{\theta}_0\right\|_2 \le R$ then $\|\theta-\theta_0\|_2 \le R$ and $\|\theta'-\theta_0\|_2 \le R$. Thus if θ_0 is an initialization such that the conclusion of Theorem 7.1 holds for the model $f(x;\theta)$ then the same conclusion holds for $f_{ASI}(x;\tilde{\theta})$ with initialization $\tilde{\theta}_0$.

7.2 Definition of the Convolution Operation

In this subsection we will formally define the convolution operation $*$ introduced in Section 2.3. We use the same convention for the convolution operation as Liu et al. (2020a). A convolutional layer of the network has the form

$$\alpha^{(l)} = \phi_l(\theta^{(l)}, \alpha^{(l-1)}) = \sigma\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} * \alpha^{(l-1)}\right).$$

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 ϕ_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

$$\phi(W,\alpha) = \sigma\left(\sum_{k=1}^{K}\frac{1}{\sqrt{m_{l-1}}} W^{[k]}\alpha^{[k]}\right).$$

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 $W^{[k]}$.

7.3 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 7.2. *Let $f(x;\theta)$ be a neural network of the form specified in Section 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 Vershynin (2012, 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 (Liu et al., 2020a).

Lemma 7.3. *Let $R \geq 1$ and let $f(x;\theta)$ be a neural network of the form specified in Section 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 7.4. *Let $R \geq 1$ and let $f(x;\theta)$ be a neural network of the form specified in Section 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 7.2 and Lemma 7.3. □

The following lemma bounds the kernel deviations $K^\theta - K^{\theta_0}$ in terms of the network Hessian.

Lemma 7.5. *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 7.6.

$$\|\theta_t-\theta_0\|_2\leq\frac{\sqrt{t}}{\sqrt{2}}\|\hat{r}_0\|_{\mathbb{R}^n}\leq\frac{\sqrt{t}}{\sqrt{2}}\|f^*\|_{L^\infty(X,\rho)}.$$

Proof.

$$\begin{aligned}\|\theta_t-\theta_0\|_2\leq\int_0^t\|\partial_s\theta_s\|_2\,ds=\int_0^t\|\partial_\theta L(\theta_s)\|_2\,ds\leq\sqrt{t}\left[\int_0^t\|\partial_\theta L(\theta_s)\|_2^2\,ds\right]^{1/2}\\ =\sqrt{t}\left[\int_0^t-\partial_s L(\theta_s)ds\right]^{1/2}=\sqrt{t}\left[L(\theta_0)-L(\theta_t)\right]^{1/2}\leq\sqrt{t}\left[L(\theta_0)\right]^{1/2}=\frac{\sqrt{t}}{\sqrt{2}}\|\hat{r}_0\|_{\mathbb{R}^n}\\ \leq\frac{\sqrt{t}}{\sqrt{2}}\|f^*\|_{L^\infty(X,\rho)}\end{aligned}$$

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. □

8 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 7.1 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 $S=(z_1,\dots,z_k)$ be a finite point set. We define

$$\mathcal{F}_{|S}=\{(g(z_1),\dots,g(z_k)):g\in\mathcal{F}\}$$

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. Telgarsky (2021, Theorem 13.1)

Theorem 8.1. *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 $z_1,\dots,z_n$*

$$\sup_{g\in\mathcal{F}}\left[\mathbb{E}[g(Z)]-\frac{1}{n}\sum_{i=1}^n g(z_i)\right]\le\frac{2}{n}URad(\mathcal{F}_{|S})+3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

We will also make use of the following lemma which is also classic, see e.g. Telgarsky (2021, Lemma 13.3)

Lemma 8.2. *Let $\ell:\mathbb{R}^n\to\mathbb{R}^n$ be a vector of univariate C-lipschitz functions. Then $URad((\ell\circ\mathcal{F})_{|S})\le C\cdot URad(\mathcal{F}_{|S})$.*

Using this we will now prove the following technical lemma. For the purpose of this lemma $x_1,\dots,x_n$ will be treated as fixed and the randomness will be over a ghost sample $S'=(x_1',\dots,x_n')$.

Lemma 8.3. *Let $R\ge1$ and $B=\sup_{x\in X}\sup_{\theta\in\overline{B}(\theta_0,R)}\|\nabla_\theta f(x,\theta)\|_2$. Consider $x_1,\dots,x_n\in X$ 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 $x_1',\dots,x_n'$ be sampled i.i.d. from ρ. Let $S=(x_1,\dots,x_n)$ and $S'=\{x_1',\dots,x_n'\}$ 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 $x_1',\dots,x_n'$ we have that every $g\in\mathcal{F}$ satisfies

$$\mathbb{E}_{x\sim\rho}[g(x)]\le12B^2H_{max}^2R^2+12B^4\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. We note that for $\theta\in\overline{B}(\theta_0,R)$

$$|K^\theta(x,x_i)-K^{\theta_0}(x,x_i)|^2\le[|K^\theta(x,x_i)|+|K^{\theta_0}(x,x_i)|]^2\le[2B^2]^2=4B^4.$$

Therefore for all $g\in\mathcal{F}$ we have that $g(x)\in[0,4B^4]$ a.s. Then by Theorem 8.1 we have with probability at least $1-\delta$ over the sampling of $S'=\{x_1',\dots,x_n'\}$

$$\sup_{g\in\mathcal{F}}\left[\mathbb{E}_{x\sim\rho}[g(x)]-\frac{1}{n}\sum_{i=1}^n g(x_i')\right]\le\frac{2}{n}URad(\mathcal{F}_{|S'})+12B^4\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Then we note that for any $z,z'\in S\cup S'$ we by Lemma 7.5 that $\theta\in\overline{B}(\theta_0,R)$ implies

$$|K^\theta(z,z')-K^{\theta_0}(z,z')|\le2BH_{max}R.$$

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 8.4. *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

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}|K(x,x_i)-K_t(x,x_i)|^2 \\ &\leq \frac{2}{n}\sum_{i=1}^{n}|K(x,x_i)-K_0(x,x_i)|^2 + \frac{2}{n}\sum_{i=1}^{n}|K_0(x,x_i)-K_t(x,x_i)|^2.\end{aligned}$$

For conciseness let

$$\begin{aligned}h_1(x) &:= \frac{1}{n}\sum_{i=1}^{n}|K(x,x_i)-K_0(x,x_i)|^2 \\ h_2^t(x) &:= \frac{1}{n}\sum_{i=1}^{n}|K_0(x,x_i)-K_t(x,x_i)|^2.\end{aligned}$$

Then by the above we have that

$$\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 2\|\hat{r}_0\|_{\mathbb{R}^n}^2\left[\mathbb{E}_{x\sim\rho}[h_1(x)] + \mathbb{E}_{x\sim\rho}[h_2^t(x)]\right].$$

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 $Z_i := \|K(\bullet, x_i) - K_0(\bullet, x_i)\|^2_{L^2(X,\rho)}$ 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 $x_1, \ldots, x_n$

$$\frac{1}{n}\sum_{i=1}^n Z_i \leq \mathbb{E}_{x_1\sim\rho}[Z_1] + \frac{\sqrt{2}[\kappa + B^2]^2\sqrt{\log(1/\delta)}}{\sqrt{n}}. \tag{7}$$

Now note that

$$\frac{1}{n}\sum_{i=1}^n Z_i = \mathbb{E}_{x\sim\rho}[h_1(x)] \qquad \mathbb{E}_{x_1\sim\rho}[Z_1] = \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)}.$$

Thus whenever (7) holds we have that

$$\begin{aligned}\mathbb{E}_{x\sim\rho}[h_1(x)] &\leq \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \frac{\sqrt{2}[\kappa + B^2]^2\sqrt{\log(1/\delta)}}{\sqrt{n}} \\ &= \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right).\end{aligned}$$

On the other hand we have by Lemma 8.3 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 $\theta \in \overline{B}(\theta_0, R)$

$$\mathbb{E}_{x\sim\rho}\left[\frac{1}{n}\sum_{i=1}^n |K^\theta(x, x_i) - K^{\theta_0}(x, x_i)|^2\right] \leq 12B^2H^2_{max}R^2 + 12B^4\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{8}$$

Whenever the above holds we have that for any θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$ we have that

$$\mathbb{E}_{x\sim\rho}[h_2^t(x)] \leq 12B^2H^2_{max}R^2 + 12B^4\sqrt{\frac{\log(2/\delta)}{2n}} = 12B^2H^2_{max}R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right).$$

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 (7) and (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

$$\mathbb{E}_{x\sim\rho}[h_1(x)] + \mathbb{E}_{x\sim\rho}[h_2^t(x)] \leq \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + 12B^2H^2_{max}R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right).$$

Well then

$$\begin{aligned}\left\|(T_n - T_n^t)r_t\right\|^2_{L^2(X,\rho)} &\leq 2\|\hat{r}_0\|^2_{\mathbb{R}^n}\left[\mathbb{E}_{x\sim\rho}[h_1(x)] + \mathbb{E}_{x\sim\rho}[h_2^t(x)]\right] \\ &\leq 2\|\hat{r}_0\|^2_{\mathbb{R}^n}\left[\|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + 12B^2H^2_{max}R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right)\right] \\ &\leq 2\|f^*\|^2_{L^\infty(X,\rho)}\left[\|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + 12B^2H^2_{max}R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right)\right].\end{aligned}$$

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 8.4 we get the following corollary.

Corollary 8.5. *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 $\|\theta_t - \theta_0\|_2 \leq R$

$$\left\|(T_n - T_n^t) r_t\right\|_{L^2(X,\rho)}^2 \leq 2 \|f^*\|_{L^\infty(X,\rho)}^2 \|K - K_0\|_{L^2(X^2, \rho \otimes \rho)}^2 + \epsilon$$

provided that $B = \tilde{O}(1)$, $H_{max} = \tilde{O}(\epsilon^{1/2}/R)$ and $n = \tilde{\Omega}(\epsilon^{-2})$.

Proof. We have by Lemma 8.4 with probability at least $1 - \delta$ over the sampling of S, S'

$$\left\|(T_n - T_n^t) r_t\right\|_{L^2(X,\rho)}^2 \leq 2 \|f^*\|_{L^\infty(X,\rho)}^2 \left[\|K - K_0\|_{L^2(X^2, \rho \otimes \rho)}^2 + 12 B^2 H_{max}^2 R^2 + \tilde{O}\left(\frac{B^4}{\sqrt{n}}\right)\right].$$

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

$$\left\|(T_n - T_n^t) r_t\right\|_{L^2(X,\rho)}^2 \leq 2 \|f^*\|_{L^\infty(X,\rho)}^2 \|K - K_0\|_{L^2(X^2, \rho \otimes \rho)}^2 + \epsilon.$$

□

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 8.6. *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*

$$H_{max} := \max_{z \in S} \sup_{\theta \in \overline{B}(\theta_0, R)} \|H(z, \theta)\|_{op}.$$

Then

$$\sup_{g \in \mathcal{F}} \|g\|_{L^\infty(X,\rho)} \leq 2BR$$

and

$$\sup_{g \in \mathcal{F}} \max_{z \in S} |g(z)| \leq \frac{1}{2} R^2 H_{max}.$$

In particular

$$\frac{1}{n} URad((\mathcal{F} \cup -\mathcal{F})_{|S}) \leq \frac{1}{2} R^2 H_{max}.$$

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 8.7. *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 Bartlett & Mendelson (2003, 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 8.6 and 8.7.

Lemma 8.8. *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 $S = (x_1,\ldots,x_n)$*

$$\sup_{g\in\mathcal{F}_1\cup\mathcal{F}_2}\left|\|g\|^2_{L^2(X,\rho)} - \|g\|^2_{L^2(X,\widehat{\rho})}\right| \leq 4BR^3H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right).$$

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 8.6 and 8.7 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 8.1 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 8.2 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 8.6 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 8.7 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 8.9. *Let $R \geq 1$ and let B and H_{max} be defined as in Lemma 8.8. 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

$$\Phi = 4BR^3 H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right)$$

such that with probability at least $1-\delta$ over the sampling of $x_1, \ldots, x_n$

$$\sup_{\theta \in \overline{B}(\theta_0, R)} \|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)} \leq \Gamma + \kappa\left[\sqrt{R^4 H_{max}^2 + 2\Phi} + \sqrt{4\epsilon^2 + 2\Phi}\right].$$

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 Rosasco et al. 2010, 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

$$\max_{g \in E} \|(T_n - T_K)g\|_{L^2(X,\rho)} \leq s\sqrt{\sigma_1} \leq s\sqrt{\kappa}. \tag{9}$$

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

$$\begin{aligned}\|(T_K - T_n)h\|_{L^2(X,\rho)} &\leq \|(T_K - T_n)h\|_{L^\infty(X,\rho)} \leq \|T_K h\|_{L^\infty(X,\rho)} + \|T_n h\|_{L^\infty(X,\rho)} \\ &\leq \kappa[\|h\|_{L^2(X,\rho)} + \|h\|_{L^2(X,\widehat{\rho})}].\end{aligned}$$

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 8.8. 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 8.8. Thus by Lemma 8.8 separate from the randomness before we have with probability at least $1-\delta$ over the sampling of $x_1, \ldots, x_n$

$$\sup_{g \in \mathcal{F}_1 \cup \mathcal{F}_2} \left|\|g\|^2_{L^2(X,\rho)} - \|g\|^2_{L^2(X,\widehat{\rho})}\right| \leq 4BR^3 H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right) := \Phi. \tag{10}$$

Well note that by Lemma 8.6 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 (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

$$\|h_1\|_{L^2(X,\rho)} + \|h_1\|_{L^2(X,\widehat{\rho})} \leq \sqrt{2}\sqrt{\frac{1}{2}R^4 H^2_{max} + \Phi} = \sqrt{R^4 H^2_{max} + 2\Phi}$$

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 (10) holds that

$$\|(T_K - T_n)h_1\|_{L^2(X,\rho)} \leq \kappa[\|h_1\|_{L^2(X,\rho)} + \|h_1\|_{L^2(X,\widehat{\rho})}] \leq \kappa\sqrt{R^4 H^2_{max} + 2\Phi}$$

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 (9) and (10) hold simultaneously. In which case our previous estimates combine to give us the bound

$$\begin{aligned}
&\|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)} \\
&\leq \left\|(T_K - T_n)r_{lin}(\bullet;\hat{\theta})\right\|_{L^2(X,\rho)} + \|(T_K - T_n)h_1\|_{L^2(X,\rho)} + \|(T_K - T_n)h_2\|_{L^2(X,\rho)} \\
&\leq s\sqrt{\kappa} + \kappa\left[\sqrt{R^4 H^2_{max} + 2\Phi} + \sqrt{4\epsilon^2 + 2\Phi}\right].
\end{aligned}$$

We now note that as long as (9) and (10) hold the same argument runs through for any $\theta \in \overline{B}(\theta_0, R)$. Thus with probability at least $1 - 2\delta$

$$\sup_{\theta\in\overline{B}(\theta_0,R)} \|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)} \leq s\sqrt{\kappa} + \kappa\left[\sqrt{R^4 H^2_{max} + 2\Phi} + \sqrt{4\epsilon^2 + 2\Phi}\right].$$

The desired conclusion follows by setting $\Gamma = s\sqrt{\kappa}$ and replacing δ with $\delta/2$ in the previous argument. □

From Lemma 8.9 we get the following corollary.

Corollary 8.10. *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 8.9 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

$$\Gamma = \tilde{O}\left(\frac{BR\sqrt{\log(\mathcal{N}(\mathcal{C}, L^2(X,\rho), \epsilon^{1/2}))}}{\sqrt{n}}\right),$$

$$\Phi = 4BR^3 H_{max} + \tilde{O}\left(\frac{B^2R^2}{\sqrt{n}}\right),$$

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 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

$$\begin{aligned} Tr(F) = Tr(\mathbb{E}_{x\sim\rho}[\nabla_\theta f(x;\theta_0)\nabla_\theta f(x;\theta_0)^T]) &= \mathbb{E}_{x\sim\rho} Tr(\nabla_\theta f(x;\theta_0)\nabla_\theta f(x;\theta_0)^T) \\ &= \mathbb{E}_{x\sim\rho}\left\|\nabla_\theta f(x;\theta_0)\right\|^2 \leq B^2. \end{aligned}$$

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

$$\Gamma + \kappa\left[\sqrt{R^4 H_{max}^2 + 2\Phi} + \sqrt{4\epsilon + 2\Phi}\right] = O(\epsilon^{1/2}).$$

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 8.11. *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

$$\left\|(T_K - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \leq 4\|f^*\|_{L^\infty(X,\rho)}^2 \|K - K_0\|_{L^2(X^2,\rho\otimes\rho)}^2 + \epsilon$$

provided that $B = \tilde{O}(1)$, $H_{max} = \tilde{O}(\epsilon/R^3)$ 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

$$\begin{aligned}\left\|(T_K - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 &\leq [\|(T_K - T_n)r_t\|_{L^2(X,\rho)} + \left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}]^2 \\ &\leq 2\|(T_K - T_n)r_t\|_{L^2(X,\rho)}^2 + 2\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2 \\ &\leq 2 \sup_{\theta \in \overline{B}(\theta_0,R)} \|(T_K - T_n)r(\bullet;\theta)\|_{L^2(X,\rho)}^2 + 2\left\|(T_n - T_n^t)r_t\right\|_{L^2(X,\rho)}^2.\end{aligned}$$

Well by Corollary 8.10 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 8.5 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 $\epsilon/4$. □

9 Main Result

9.1 Damped Deviations

In this subsection we will recall some definitions and results from Bowman & Montúfar (2022). The main theorems in Bowman & Montúfar (2022) 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 $\int_0^t \|g_s\|_{L^2(X,\rho)}^2 ds < \infty$. 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 Bowman & Montúfar (2022, Lemma 2.4).

Lemma 9.1. *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 Bowman & Montúfar (2022, Lemma C.8)

Lemma 9.2. *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*

$$\|P_k(r_t - \exp(-T_K t)r_0)\|_{L^2(X,\rho)} \leq \frac{1-\exp(-\sigma_k t)}{\sigma_k} \sup_{s\leq t} \|(T_K - T_n^s)r_s\|_{L^2(X,\rho)}.$$

and

$$\|r_t - \exp(-T_K t)r_0\|_{L^2(X,\rho)} \leq t \cdot \sup_{s\leq t} \|(T_K - T_n^s)r_s\|_{L^2(X,\rho)}.$$

9.2 Proof of Theorem 3.5

We are now ready to prove the main result of this paper.

Theorem 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 (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 8.11 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 7.4 and the expression $\tilde{O}(\epsilon/R^3)$ is precisely the condition on H_{max} in the conclusion of Lemma 8.11. By Lemma 8.11 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 7.1 and Lemma 7.4 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 8.11 that for any θ_t such that $\|\theta_t - \theta_0\|_2 \leq R$

$$\left\|(T_K - T_n^t)r_t\right\|^2_{L^2(X,\rho)} \leq 4\|f^*\|^2_{L^\infty(X,\rho)} \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon. \tag{11}$$

Well by Lemma 7.6 we have that $\|\theta_t - \theta_0\| \leq \frac{\sqrt{t}}{\sqrt{2}} \|f^*\|_{L^\infty(X,\rho)}$. Thus for $t \leq \frac{2R^2}{\|f^*\|^2_{L^\infty(X,\rho)}}$ we have that $\|\theta_t - \theta_0\| \leq R$. Well then by Lemma 9.2 and the inequality (11) we have that

$$\|P_k(r_t - \exp(-T_K t) r_0)\|^2_{L^2(X,\rho)}$$
$$\leq \left[\frac{1 - \exp(-\sigma_k t)}{\sigma_k}\right]^2 \cdot \left[4 \|f^*\|^2_{L^\infty(X,\rho)} \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right]$$

and

$$\|r_t - \exp(-T_K t) r_0\|^2_{L^2(X,\rho)} \leq t^2 \cdot \left[4 \|f^*\|^2_{L^\infty(X,\rho)} \|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right].$$

The desired result then follows by setting $T = \frac{2R^2}{\|f^*\|^2_{L^\infty(X,\rho)}}$. □

10 Discussion of Assumption 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

$$\|K_0 - K^\infty\|^2_{L^2(X\times X,\rho\otimes\rho)}$$
$$\leq \frac{1}{N}\sum_{i=1}^N |K_0(x_i, x_i') - K^\infty(x_i, x_i')|^2 + \tilde{O}\left(\frac{\|K_0 - K^\infty\|^2_{L^\infty(X\times X,\rho\times\rho)}}{\sqrt{N}}\right),$$

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 7.4 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 3.6 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 Du et al. (2019b) 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, Du et al. (2019a) get an analogous bound for deep fully-connected, ResNet, and convolutional networks with smooth activations. This is substantiated by the results of Huang & Yau (2020) 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. Bowman & Montúfar (2022) 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 Arora et al. (2019b) 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 Buchanan et al. (2021) 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 Buchanan et al. (2021) 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 3.6 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 3.6 by citing the aforementioned results to get statements for the cases in which they apply to.

11 Experimental Details

Architecture and Parameterization The code to produce Figure 1 is available at `https://anonymous.4open.science/r/deepspec-5564/README.md` 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 (LeCun et al., 1998) 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 $\sigma(x) = \log(1+\exp(x))$. For LeNet-5 we compute the NTK using PyTorch (Paszke et al., 2019) using the default PyTorch initialization and parameterization. For the shallow network we implement the network directly and use the Neural Tangent Kernel parameterization:

$$f(x;\theta) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m} a_i \sigma(\langle w_i, x\rangle + b_i) + b_0,$$

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 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 (LeCun et al., 1998) and the dataset for the shallow model is CIFAR-10 (Krizhevsky, 2009). 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 (Paszke et al., 2019), NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), Matplotlib (Hunter, 2007), Jupyter Notebook (Kluyver et al., 2016), IPython (Pérez & Granger, 2007), 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/`.