A Brief Introduction to the Neural Tangent Kernel*

Benjamin Bowman

May 2023

1 A Short History of the Neural Tangent Kernel

The typical operating regime in deep learning is to optimize an overparameterized network via gradient-based optimization. This has been phenomenally successful in practice but leads to a number of theoretical challenges. The first is that neural networks have a highly nonlinear parameterization which leads to optimization objectives that are nonconvex [1, 2]. The nonconvexity of the optimization makes proving theoretical guarantees for gradient optimization a tall task. Furthermore overparameterized networks are able to interpolate arbitrary labels [3], and the VC-dimension of typical networks grows at least linearly with the number of parameters [4, 5]. As a consequence, classical complexity based measures from statistical learning theory such as Rademacher complexity or VC-dimension lead to vacuous generalization bounds [6]. Thus understanding modern deep learning will require innovations beyond the classical theories of both optimization and generalization.

The aforementioned challenges at first make the prospect of establishing a theoretical understanding of deep learning seem dismal. However, there was evidence as far back as the 1990s that overparameterized networks may be amenable to theoretical analysis. [7, 8] demonstrated that the network outputs converge to a Gaussian process as the number of hidden units approaches infinity. This led to a line of research studying the connection between Gaussian processes, kernel methods, neural network representations, and deep learning [9, 10, 11]. In a similar vein [12] exhibited decreasing generalization error while increasing the network width, suggesting that overparameterized networks may have a more subtle form of capacity control.

While progress was made towards understanding neural network representations via the infinite-width limit, an understanding of the optimization dynamics was still lacking. A breakthrough emerged in 2018 when [13] demonstrated that the optimization dynamics are governed via a time-dependent kernel coined the

*An excerpt from the Ph.D. thesis “On the Spectral Bias of Neural Networks in the Neural Tangent Kernel Regime” by Benjamin Bowman

"Neural Tangent Kernel (NTK)", which in the infinite-width limit becomes constant throughout training. In this limiting setting the network parameterization becomes approximately linear [14], and bounding the smallest eigenvalue of the NTK throughout training is sufficient to prove global convergence of gradient descent. In fact, almost concurrently with [13] the authors in [15] had used this technique to prove the first global convergence guarantee for gradient descent applied to a network trained on general data. The NTK had been studied earlier by the work [16] which demonstrated that the squared loss satisfies a Polyak-Lojasiewicz (PL) inequality in any region where the smallest eigenvalue of the NTK is bounded below. This analysis ties back to a well known technique in nonconvex optimization that establishing a PL-inequality is sufficient for proving convergence of gradient descent provided that the gradient is Lipschitz [17]. The innovation in [15] was to prove that the gradient descent trajectory remains in a region where a PL-inequality holds, as well as an innovative technique of bounding the number of activation patterns that change for a ReLU network as a substitute for the Lipschitz property.

2 Global Convergence Guarantees via the NTK

2.1 The Neural Tangent Kernel and PL-inequalities

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

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

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

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

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

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

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

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

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

which is the continuous-time analog of gradient descent. Speaking loosely, one can view gradient flow as gradient descent in the limit of vanishing step sizes. A key insight of [13, 15] was to analyze the gradient descent dynamics in function space (i.e. the evolution of the neural network predictions) as opposed to parameter space. In this vein we will let $u_\theta, y \in \mathbb{R}^n$ be defined by

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

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

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

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

We will let

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

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

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

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

We define

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

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

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

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

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

Therefore

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

We now note that

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

Then by Grönwall's inequality [18]

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

Now assume that

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

Then we have that

$$L(t) \leq L(0)\exp(-ct). \tag{1}$$

Thus we have just shown that lower bounding $\lambda_{\min}\left(H_t\right)$ uniformly in time is sufficient for establishing convergence of gradient flow to a global minimum when optimizing the squared loss. The quantity c provides an estimate for the convergence rate. The bound (1) is analagous to linear convergence in discrete time.

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

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

Suppose

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

Then similar to before

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

Assuming L is bounded below it follows that

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

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

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

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

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

Thus by the Cauchy-Schwarz inequality we have

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

Thus if

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

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

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

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

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

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

$$\alpha |L(u) - L(u^*)|^\beta \leq \|\nabla_u L(u)\|_2 \tag{2}$$

for some $\alpha, \beta > 0$. Let $\sigma_{min}(J_t)$ denote the smallest singular value of J_t. Then (2) implies

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

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

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

Since

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

assuming L is bounded below we have that

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

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

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

where θ^* is a parameter corresponding to a global minimum, then gradient descent with constant step size converges to a global minimum [17].

2.2 Bounding the Smallest Eigenvalue of the NTK Gram Matrix

In the previous section we demonstrated that

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

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

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

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

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

Consequently, wherever $\sigma_{min}(J_\theta) > 0$ we have that each critical point of the loss is a global minimum. However neural networks are known to have spurious critical points, with saddle points being particularly prevalent [19, 20, 21, 22]. Thus the challenge for proving convergence is to demonstrate that the gradient descent trajectory remains in a region where the smallest singular value of the Jacobian, or equivalently the smallest eigenvalue of the NTK Gram matrix, is bounded below. This was the key difficulty that was overcome in the proof in [15].

It was shown in [13, 15] that under a suitable parameterization, in the infinite-width limit the matrix H_t converges to a fixed positive-definite matrix H^∞ uniformly in time. The parameterization introduced in these works has since been called the "NTK parameterization", which we introduce below. For a fully-connected network with D hidden layers, we parameterize the network as follows. Let $\theta = vec(\{W^{(l)}, b^{(l)}\}_{l=1}^{D+1})$ where $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ and $b^{(l)} \in \mathbb{R}^{n_l}$. We can then define the network output $f(x; \theta)$ via the following relations:

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

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

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

where the parameters are initialized $W_{i,j}^{(l)} \sim N(0, 1/n_l)$ and $b_i^{(l)} \sim N(0, \beta^2)$ independently. The two parameterizations can realize the same functions and are identical in distribution at initialization, however the gradients are different. For gradient descent, the standard parameterization and NTK parameterization are equivalent up to a parameter-dependent rescaling of the step-size [14]. We also note that other parameterizations have been studied, such as the "mean-field" parameterization [23]. Under the NTK parameterization under fairly general assumptions

$$H_t \to H^\infty$$

in probability uniformly on $[0, T]$ where H^∞ is a fixed positive-semidefinite matrix [13]. Given weak assumptions on the training data inputs $x_1, \ldots, x_n$ (e.g. no two inputs are parallel [15] or they are "δ-separable" [24]), we have that

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

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

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

Then we have

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

which implies that $\lambda_{min}(H_t) \geq \frac{\lambda_{min}(H^\infty)}{2}$. For simplicity, assume all layers have the same width m. At initialization it was shown in [25] that whenever the activation function is suitably smooth

$$\left\|H_0 - H^\infty\right\|_{op} = \tilde{\mathcal{O}}(n/\sqrt{m}) \tag{3}$$

with high probability. Furthermore by the results in [26, 27] it was shown that for any $R > 0$ with high probability over the initialization that

$$\left\|H_t - H_0\right\|_{op} = \tilde{\mathcal{O}}(nR^{3D}/\sqrt{m}) \tag{4}$$

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

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

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

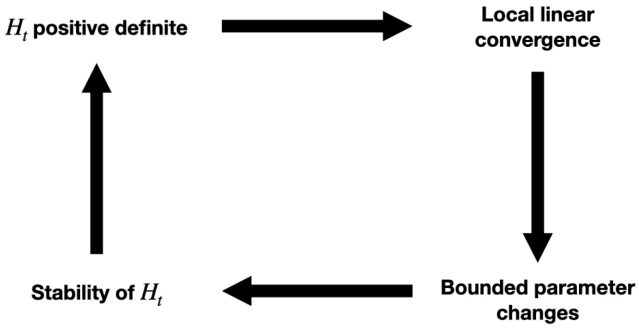

Figure 1: **A Seemingly Circular Argument** A sketch of the argument for proving convergence of gradient flow to a global minimum.

2.3 Proving Global Convergence for Gradient Flow

We will sketch the following argument for convergence of gradient flow to a global minimum, which has appeared in many different variations (e.g. [15, 25, 27]). The proof revolves around a seemingly circular argument depicted in Figure 1, which can be resolved via a continuous induction argument. By (3) if $m \gtrsim n^2 \lambda_{\min}(H^\infty)^{-2}$ we can assume

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

Fix some value $K > 0$ and let

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

We will see later that by setting K sufficiently large we can ensure that the set the supremum is taken over above is nonempty with high probability. If we can demonstrate that $T = \infty$, then we have that the smallest eigenvalue $\lambda_{\min}(H_t)$ is bounded below uniformly in time and thus we will have shown that gradient flow converges to a global minimum. Thus for the sake of contradiction assume $T < \infty$. Recall that by the results in Section 2.1 the bound $\lambda_{\min}(H_t) \geq \lambda_{\min}(H^\infty)/2$ implies that for $t \leq T$

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

It follows that for $t \leq T$,

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

Well then

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

It is not hard to show that the network outputs are bounded with high probability at initialization, thus assuming $\|y\| = O(\sqrt{n})$ we have that $\|\hat{r}_0\| = O(\sqrt{n})$. It follows then that there exists a quantity $R_{max} = O\left(\frac{K\sqrt{n}}{\lambda_{\min}(H^\infty)}\right)$ such that $R' \leq R_{max}$ with high probability. Well by Eq. (4) we can say with high probability for $\theta_t \in B(\theta_0, R_{max})$

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

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

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

However then

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

so that

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

Well then

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

Recall the definition of T,

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

By continuity and the maximality of T we must have that either $\lambda_{\min}(H_T) = \frac{1}{2}\lambda_{\min}(H^\infty)$ or $\|J_T\| = K$, however by (5) $\lambda_{\min}(H_T) > \frac{1}{2}\lambda_{\min}(H^\infty)$, thus it follows that $\|J_t\| = K$. However as can be seen in the Appendix (see Lemma 4.3) for any $R \geq 1$ if $\sqrt{m} \geq R$ then with high probability

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

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

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

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

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

As we saw before this implies that

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

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

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

and

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

where

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

It turns out that for general inputs $x_1, \ldots, x_n$ we have that $\lambda_{\min}\left(H^{\infty}\right) = \Omega(1)$ [28]. Thus we conclude that $m \gtrsim n^{O(D)}$ suffices to prove global convergence of gradient flow.

3 Spectral Bias

In the previous section we demonstrated that bounding the smallest eigenvalue of the NTK Gram matrix is sufficient for establishing convergence, and that the bound for the eigenvalue provides an estimate for the convergence rate of gradient descent. However, in general this is a pessimistic estimate and the convergence rate along different components will vary. From the results in Section 2.1 that we have for the squared loss

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

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

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

which has the explicit solution

$$\exp(-H^{\infty} t)\hat{r}_0. \tag{6}$$

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

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

We thus see that the convergence rate along the direction u_i is given by the eigenvalue λ_i, and consequently the directions corresponding to large eigenvalues will be learned much more quickly. It has been observed in many works (see e.g. [29, 30, 31]) that the NTK Gram matrix tends to have a small number of outlier eigenvalues and a long tail of small eigenvalues. In fact, in [32] it was proven that there are $O(1)$ eigenvalues on the same order of magnitude as the largest eigenvalue λ_1 independent of the parameter n. Consequently, there are a small number of directions that are learned much more quickly than others.

The phenomenon that eigenvectors of the NTK corresponding to large eigenvalues are learned quicker can be described as a type of "spectral bias" [33]. Classically, "spectral bias" was the title given to the phenomenon that neural networks tend to learn the low Fourier frequencies quicker during training[1] [35, 34, 36]. However, in special cases these two notions coincide. Specifically, if we let m denote the width of the network we can define

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

where the convergence is in probability over the parameter initialization [13]. K^{∞} is called the analytical Neural Tangent Kernel (NTK), and the matrix H^{∞} introduced in Section 2.1 is the Gram matrix induced by this kernel and the training data, i.e.

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

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

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

By Mercer's theorem [37] we have the decomposition

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

where $\{\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of $L^2_\rho(X)$ and each ϕ_i is an eigenfunction of $T_{K^{\infty}}$ with eigenvalue $\sigma_i \geq 0$. Whenever ρ is the uniform distribution on the sphere $X = S^{d-1}$, the eigenfunctions ϕ_i can be taken to be the spherical harmonics, which in $d = 2$ corresponds to the Fourier basis. In the work [38] it was demonstrated that in the $d = 2$ case for shallow ReLU networks the large eigenvalues of $T_{K^{\infty}}$ correspond to the low Fourier frequencies. We note that we can consider the eigenvectors u_i of H^{∞} to be empirical estimates of the eigenfunctions of $T_{K^{\infty}}$. In this case "spectral bias" in the sense of learning the low Fourier frequencies faster coincides with "spectral bias" in the sense of learning the dominant eigenvectors of the NTK Gram matrix faster.

[29] had quantified the extent in which finite-width networks approximate the idealized infinite-width dynamics that are given by the evolution described in (6). However, this equation only describes the network on the training set $x_1, \ldots, x_n$. Let f^* be our target function so that $y_i = f^*(x_i)$. We are interested in describing the behavior of the residual $r_t(x) := f(x; \theta_t) - f^*(x)$ for an arbitrary input x. Informally speaking, in the limit of infinite data the matrix H^{∞} converges to the integral operator $T_{K^{\infty}}$ and the empirical residual $\hat{r}_t$ converges

[1] This has also been called the "Frequency Principle" [34].

to the full residual r_t. In this idealized setting the evolution described in (6) becomes

$$r_t = \exp(-T_{K^\infty} t) r_0. \tag{7}$$

Assuming (7) holds we have

$$\langle r_t, \phi_i \rangle_{L^2_\rho} = \langle \exp(-T_{K^\infty} t) r_0, \phi_i \rangle_{L^2_\rho} = \exp(-\sigma_i t) \langle r_0, \phi_i \rangle_{L^2_\rho}. \tag{8}$$

Thus under the evolution described in (7) we have that the eigenfunctions ϕ_i are learned at rates corresponding to their eigenvalues σ_i. In contrast to (6), (7) and (8) describe the dynamics of the residual over the entire input domain and not just the training set. Thus in this limiting setting the network exhibits a stronger form of spectral bias that determines the behavior of the network over the entire input domain. The works [39] and [40] quantified to what extent the finite-width network trained on finitely many samples exhibits the behavior of the idealized limit of infinite width and infinite data described in (7).

4 Limitations and Challenges of NTK Analysis

The paper that introduced the Neural Tangent Kernel [13] has become one of the most highly cited works in deep learning theory, with the NTK having attracted both fanaticism and criticism [41]. While the NTK has greatly enhanced the understanding of the optimization dynamics of wide networks [15, 42, 24, 43, 44, 45, 46, 47, 14], this analysis breaks down whenever the depth of the network scales in tandem with the width [48], which is known to achieve better performance in practice [49, 50]. Furthermore NTK analysis is only applicable when training with small learning rates, with more moderate learning rates leading to distinct behavior [51]. It is also known that in practice the NTK deviates to adapt to the target function [52, 53, 54], which stands in contrast to the infinite-width behavior where the NTK is constant. Establishing a theoretical framework that can handle more realistic scalings for the depth and learning rate which also makes allowances for feature learning remains an active challenge. Nevertheless, infinite-width networks achieve compelling performance and serve well as a first approximation of the average behavior of finite-width models [55], suggesting that the Neural Tangent Kernel will remain a fundamental tool in deep learning theory.

References

[1] E. D. Sontag and H. J. Sussmann, “Backpropagation can give rise to spurious local minima even for networks without hidden layers,” *Complex Systems*, vol. 3, pp. 91–106, 1989.

[2] E. D. Sontag and H. J. Sussmann, “Back propagation separates where perceptrons do,” *Neural Networks*, vol. 4, no. 2, pp. 243–249, 1991.

[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. 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.

[4] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks," *Journal of Machine Learning Research*, vol. 20, no. 63, pp. 1–17, 2019.

[5] P. Koiran and E. Sontag, "Neural networks with quadratic vc dimension," in *Advances in Neural Information Processing Systems* (D. Touretzky, M. Mozer, and M. Hasselmo, eds.), vol. 8, MIT Press, 1995.

[6] M. Anthony and P. L. Bartlett, *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 2002.

[7] R. M. Neal, *Bayesian Learning for Neural Networks*. Berlin, Heidelberg: Springer-Verlag, 1996.

[8] C. Williams, "Computing with infinite networks," in *Advances in Neural Information Processing Systems* (M. Mozer, M. Jordan, and T. Petsche, eds.), vol. 9, MIT Press, 1996.

[9] Y. Cho and L. Saul, "Kernel methods for deep learning," in *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, eds.), vol. 22, Curran Associates, Inc., 2009.

[10] A. Daniely, R. Frostig, and Y. Singer, "Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity," in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016.

[11] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, "Deep neural networks as Gaussian processes," in *International Conference on Learning Representations*, 2018.

[12] B. Neyshabur, R. Tomioka, and N. Srebro, "In search of the real inductive bias: On the role of implicit regularization in deep learning," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.

[13] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[14] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, “Wide neural networks of any depth evolve as linear models under gradient descent,” in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[15] S. S. Du, X. Zhai, B. Poczos, and A. Singh, “Gradient descent provably optimizes over-parameterized neural networks,” in *International Conference on Learning Representations*, 2019.

[16] B. Xie, Y. Liang, and L. Song, “Diverse Neural Network Learns True Target Functions,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54 of *Proceedings of Machine Learning Research*, pp. 1216–1224, PMLR, 2017.

[17] B. Polyak, “Gradient methods for the minimisation of functionals,” *Ussr Computational Mathematics and Mathematical Physics*, vol. 3, pp. 864–878, 12 1963.

[18] T. H. Gronwall, “Note on the derivatives with respect to a parameter of the solutions of a system of differential equations,” *Annals of Mathematics*, vol. 20, no. 4, pp. 292–296, 1919.

[19] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, “Identifying and attacking the saddle point problem in high-dimensional non-convex optimization,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[20] K. Kawaguchi, “Deep learning without poor local minima,” in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[21] K. Fukumizu and S. Amari, “Local minima and plateaus in hierarchical structures of multilayer perceptrons,” *Neural Networks*, vol. 13, no. 3, pp. 317–327, 2000.

[22] B. Simsek, F. Ged, A. Jacot, F. Spadaro, C. Hongler, W. Gerstner, and J. Brea, “Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances,” in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 9722–9732, PMLR, 18–24 Jul 2021.

[23] J. Sirignano and K. Spiliopoulos, “Mean field analysis of neural networks: A law of large numbers,” *SIAM Journal on Applied Mathematics*, vol. 80, no. 2, pp. 725–752, 2020.

[24] S. Oymak and M. Soltanolkotabi, "Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, 2020.

[25] J. Huang and H.-T. Yau, "Dynamics of deep neural networks and neural tangent hierarchy," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 4542–4551, PMLR, 13–18 Jul 2020.

[26] C. Liu, L. Zhu, and M. Belkin, "On the linearity of large non-linear models: when and why the tangent kernel is constant," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 15954–15964, Curran Associates, Inc., 2020.

[27] C. Liu, L. Zhu, and M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks," *Applied and Computational Harmonic Analysis*, vol. 59, pp. 85–116, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[28] Q. Nguyen, M. Mondelli, and G. Montúfar, "Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8119–8129, PMLR, 2021.

[29] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 322–332, PMLR, 09–15 Jun 2019.

[30] S. Oymak, Z. Fabian, M. Li, and M. Soltanolkotabi, "Generalization guarantees for neural networks via harnessing the low-rank structure of the Jacobian," *CoRR*, vol. abs/1906.05392, 2019.

[31] M. Li, M. Soltanolkotabi, and S. Oymak, "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, vol. 108 of *Proceedings of Machine Learning Research*, pp. 4313–4324, PMLR, 2020.

[32] M. Murray, H. Jin, B. Bowman, and G. Montufar, "Characterizing the spectrum of the NTK via a power series expansion," in *The Eleventh International Conference on Learning Representations*, 2023.

[33] Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu, "Towards understanding the spectral bias of deep learning," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Z.-H. Zhou, ed.), pp. 2205–2211, International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[34] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao, "Training behavior of deep neural network in frequency domain," in *Neural Information Processing* (T. Gedeon, K. W. Wong, and M. Lee, eds.), (Cham), pp. 264–274, Springer International Publishing, 2019.

[35] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310, PMLR, 09–15 Jun 2019.

[36] G. Yang, A. Ajay, and P. Agrawal, "Overcoming the spectral bias of neural value approximation," in *International Conference on Learning Representations*, 2022.

[37] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 209, pp. 415–446, 1909.

[38] B. Ronen, D. Jacobs, Y. Kasten, and S. Kritchman, "The convergence rate of neural networks for learned functions of different frequencies," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[39] B. Bowman and G. Montúfar, "Implicit bias of MSE gradient optimization in underparameterized neural networks," in *International Conference on Learning Representations*, 2022.

[40] B. Bowman and G. Montufar, "Spectral bias outside the training set for deep networks in the kernel regime," in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.

[41] A. Ananthaswamy, "A new link to an old model could crack the mystery of deep learning," *Quanta Magazine*, October 2021.

[42] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685, PMLR, 2019.

[43] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 242–252, PMLR, 2019.

[44] Q. N. Nguyen and M. Mondelli, "Global convergence of deep networks with one wide layer followed by pyramidal topology," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 11961–11972, Curran Associates, Inc., 2020.

[45] Q. Nguyen, "On the proof of global convergence of gradient descent for deep relu networks with linear widths," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8056–8062, PMLR, 18–24 Jul 2021.

[46] D. Zou, Y. Cao, D. Zhou, and Q. Gu, "Gradient descent optimizes over-parameterized deep ReLU networks," *Machine learning*, vol. 109, no. 3, pp. 467–492, 2020.

[47] D. Zou and Q. Gu, "An improved analysis of training over-parameterized deep neural networks," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[48] B. Hanin and M. Nica, "Finite depth and width corrections to the neural tangent kernel," in *International Conference on Learning Representations*, 2020.

[49] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 09–15 Jun 2019.

[50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 18–24 Jul 2021.

[51] A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari, "The large learning rate phase of deep learning: the catapult mechanism," *arXiv preprint arXiv:2003.02218*, 2020.

[52] A. Baratin, T. George, C. Laurent, R. Devon Hjelm, G. Lajoie, P. Vincent, and S. Lacoste-Julien, "Implicit regularization via neural feature alignment," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (A. Banerjee and K. Fukumizu, eds.), vol. 130 of *Proceedings of Machine Learning Research*, pp. 2269–2277, PMLR, 13–15 Apr 2021.

[53] A. Atanasov, B. Bordelon, and C. Pehlevan, "Neural networks as kernel learners: The silent alignment effect," in *International Conference on Learning Representations*, 2022.

[54] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang, "High-dimensional asymptotics of feature learning: How one gradient step improves the representation," in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.

[55] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, "Finite versus infinite neural networks: an empirical study," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 15156–15172, Curran Associates, Inc., 2020.

[56] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing*, ch. 5, Cambridge University Press, 2012.

[57] C. Liu, L. Zhu, and M. Belkin, "On the linearity of large non-linear models: when and why the tangent kernel is constant," *CoRR*, vol. abs/2010.01092, 2020.

Appendix

This section will cover some technical lemmas. The following lemma bounds the operator norm of the weight matrices at initialization.

Lemma 4.1. *Let $f(x;\theta)$ be a neural network of the form specified in Section 2.2 with weight matrices $\{W^{(l)}\}_{l=1}^{D+1}$ where $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$. Furthermore let $m = \min_{l\geq 1} n_l$. Assume $m \geq d$ and $\max_l \frac{n_l}{m} \leq A$. 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 depth but is independent of the width m.

Proof. Fix a weight matrix $W \in \mathbb{R}^{n_l \times n_{l-1}}$ in the model. Following [56, Corollary 5.35] we have with probability at least $1 - 2\exp(-t^2/2)$ over the initialization

$$\|W_0\|_{op} \leq \sqrt{n_l} + \sqrt{n_{l-1}} + t$$

and thus

$$\frac{1}{\sqrt{m}} \left\|W_0^{(l)}\right\|_{op} \leq \frac{\sqrt{n_l}}{\sqrt{m}} + \frac{\sqrt{n_{l-1}}}{\sqrt{m}} + \frac{t}{\sqrt{m}} \leq 2\sqrt{A} + \frac{t}{\sqrt{m}}.$$

Thus by setting $t = \sqrt{m}$ and taking the union bound over all weight matrices in the model (which depends on the depth) we get the desired result. □

We now state for reference the following lemma which follows from the proof in [57].

Lemma 4.2. *Let $R \geq 1$ and let $f(x;\theta)$ be a neural network of the form specified in Section 2.2. Furthermore let $m = \min_{l\geq 1} n_l$. 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*

$$\sup_{x\in X} \sup_{\theta\in\overline{B}(\theta_0,R)} \|\nabla_\theta f(x;\theta)\|_2 = O\left(\max\left\{1, \frac{R}{\sqrt{m}}\right\}^{O(D)}\right).$$

In particular if $\sqrt{m} \geq R$ then

$$\sup_{x\in X} \sup_{\theta\in\overline{B}(\theta_0,R)} \|\nabla_\theta f(x;\theta)\|_2 = O\left(1\right).$$

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 4.3. *Let $R \geq 1$ and let $f(x;\theta)$ be a neural network of the form specified in Section 2.2. Furthermore let $m = \min_{l\geq 1} n_l$. Assume that $m \geq d$, $\max_l \frac{n_l}{m} = O(1)$, 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 depth but is independent of the width m

Proof. This follows immediately from Lemma 4.1 and Lemma 4.2. □