Precise Requirements for the Validity of the Neural Tangent Kernel Approximation

Abstract

This research investigates the conditions under which the neural tangent kernel (NTK) approximation remains valid when employing the square loss function for model training. Within the framework of lazy training, as introduced by Chizat et al., we demonstrate that a model, rescaled by a factor of $\alpha = O(T)$, maintains the validity of the NTK approximation up to a training time of T. This finding refines the earlier result from Chizat et al., which necessitated a larger rescaling factor of $\alpha = O(T^2)$, and establishes the preciseness of our established bound.

1 Introduction

In contemporary machine learning practice, the weights w of expansive neural network models $f_w : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ are trained using gradient-based optimizers. However, a comprehensive theoretical understanding remains elusive due to the non-linear nature of the training dynamics, which complicates analysis. To bridge this gap, an approximation to these dynamics, termed the NTK approximation, was introduced, and its validity for infinitely wide networks trained via gradient descent was demonstrated. The NTK approximation has proven highly influential, offering theoretical insights into various phenomena, including deep learning's capacity to memorize training data, the manifestation of spectral bias in neural networks, and the differential generalization capabilities of diverse architectures. Nevertheless, empirical evidence suggests that the training dynamics of neural networks frequently deviate from the NTK approximation's predictions. Consequently, it becomes crucial to delineate the precise conditions under which the NTK approximation remains applicable. This paper seeks to address the following inquiry:

Is it possible to establish precise conditions that guarantee the validity of the NTK approximation?

1.1 The Lazy Training Framework

The work demonstrated that the NTK approximation is applicable to the training of any differentiable model, provided the model's outputs are rescaled appropriately. This rescaling ensures that significant changes in the model's outputs can occur even with minor adjustments to the weights. The validity of the NTK approximation for models of infinite width stems from this observation, as the model is inherently rescaled as its width approaches infinity.

Consider a smoothly parameterized model $h : \mathbb{R}^p \to \mathcal{F}$, where $\mathcal{F}$ is a separable Hilbert space. Let $\alpha > 0$ be a parameter governing the model's rescaling, which should be considered large. We train the rescaled model αh using gradient flow to minimize a smooth loss function $R : \mathcal{F} \to \mathbb{R}^+$. The weights $w(t) \in \mathbb{R}^p$ are initialized at $w(0) = w_0$ and evolve according to the gradient flow:

$$\frac{dw}{dt} = -\frac{1}{\alpha^2}\nabla_w R(\alpha h(w(t))). \tag{1}$$

Define the linear approximation of the model around the initial weights w_0 as:

$$\bar{h}(w) = h(w_0) + Dh(w_0)(w - w_0), \tag{2}$$

where Dh is the first derivative of h with respect to w. Let $\bar{w}(t)$ be weights initialized at $\bar{w}(0) = w_0$ that evolve according to the gradient flow from training the rescaled linearized model $\alpha\bar{h}$:

$$\frac{d\bar{w}}{dt} = -\frac{1}{\alpha^2}\nabla_{\bar{w}} R(\alpha\bar{h}(\bar{w}(t))). \tag{3}$$

The NTK approximation asserts that:

$$\alpha h(w(t)) \approx \alpha\bar{h}(\bar{w}(t)). \tag{4}$$

In essence, this implies that the linearization of the model h remains valid throughout the training process. This greatly simplifies the analysis of training dynamics, as the model $\bar{h}$ is linear in its parameters, allowing the evolution of $\bar{h}(\bar{w})$ to be understood through a kernel gradient flow in function space.

The validity of the NTK approximation is contingent on the magnitude of the rescaling parameter α. Intuitively, a larger α implies that the weights need not deviate significantly from their initialization to induce substantial changes in the model's output, thereby prolonging the validity of the linearization. This regime of training, where weights remain close to their initialization, is referred to as "lazy training." The following bound was established, where $R_0 = R(\alpha h(w_0)))$ is the loss at initialization, and $\kappa = T\alpha^{-1}\text{Lip}(Dh)\sqrt{R_0}$ is a quantity that will also feature in our main results:

Proposition 1.1. Let $R(y) = \frac{1}{2}\|y - y^*\|_2^2$ be the square loss, where $y^* \in \mathcal{F}$ are the target labels. Assume that h is Lip(h)-Lipschitz and that Dh is Lip(Dh)-Lipschitz in a ball of radius ρ around w_0. Then, for any time $0 \le T \le \alpha\rho/(\text{Lip}(h)\sqrt{R_0})$,

$$\|\alpha h(w(T)) - \alpha\bar{h}(\bar{w}(T))\| \le T\text{Lip}(h)^2\kappa R_0. \tag{5}$$

As α approaches infinity, κ tends to 0, rendering the right-hand side of the inequality small and validating the NTK approximation.

1.2 Our Contributions

Our primary contribution is the refinement of the bound for extended time scales. We establish the following theorem:

Theorem 1.2 (NTK Approximation Error Bound). Let $R(y) = \frac{1}{2}\|y - y^*\|_2^2$ be the square loss. Assume that Dh is Lip(Dh)-Lipschitz in a ball of radius ρ around w_0. Then, at any time $0 \le T \le \alpha^2\rho^2/R_0$,

$$\|\alpha h(w(T)) - \alpha\bar{h}(\bar{w}(T))\| \le \min(6\kappa\sqrt{R_0}, 8R_0). \tag{6}$$

Furthermore, we demonstrate that this bound is tight up to a constant factor.

Theorem 1.3 (Converse to Theorem 1.2). For any α, T, Lip(Dh), and R_0, there exists a model $h : \mathbb{R} \to \mathbb{R}$, a target $y^* \in \mathbb{R}$, and an initialization $w_0 \in \mathbb{R}$ such that, for the risk $R(y) = \frac{1}{2}(y - y^*)^2$, the initial risk is $R(\alpha h(w_0)) = R_0$, the derivative map Dh is Lip(Dh)-Lipschitz, and

$$\|\alpha h(w(T)) - \alpha\bar{h}(\bar{w}(T))\| \ge \min\left(\frac{1}{5}\kappa\sqrt{R_0}, \frac{1}{5}R_0\right). \tag{7}$$

In contrast to prior work, our bound does not depend on the Lipschitz constant of h, and it exhibits a more favorable dependence on T. Specifically, if Lip(Dh), Lip(h), and R_0 are bounded by constants, our result indicates that the NTK approximation, up to an error of $O(\epsilon)$, holds for times $T = O(\alpha\epsilon)$, whereas the previously known bound was valid for $T = O(\sqrt{\alpha\epsilon})$. Given the practical interest in long training times $T \gg 1$, our result demonstrates that the NTK approximation is valid for significantly longer time horizons than previously recognized.

2 Application to Neural Networks

The bound established in Theorem 1.2 is applicable to the lazy training of any differentiable model. As a specific example, we detail its application to neural networks. We parameterize the networks in the mean-field regime, where the NTK approximation does not hold even as the width approaches infinity. Consequently, the NTK approximation is valid only when training is conducted in the lazy regime.

Let $f_w : \mathbb{R}^d \to \mathbb{R}$ be a 2-layer network of width m in the mean-field parametrization, with activation function $\sigma : \mathbb{R} \to \mathbb{R}$,

$$f_w(x) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m} a_i\sigma(\sqrt{m}\langle x, u_i\rangle). \tag{8}$$

The weights are $w = (a, U)$ for $a = [a_1, \ldots, a_m]$ and $U = [u_1, \ldots, u_m]$. These are initialized at w_0 with i.i.d. Unif$[-1/\sqrt{m}, 1/\sqrt{m}]$ entries. Given training data $(x_1, y_1), \ldots, (x_n, y_n)$, we train the weights of the network with the mean-squared loss

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\ell(f_w(x_i), y_i), \quad \ell(a, b) = \frac{1}{2}(a - b)^2. \tag{9}$$

In the Hilbert space notation, we let $\mathcal{H} = \mathbb{R}^n$, so that the gradient flow training dynamics with loss (6) correspond to the gradient flow dynamics (1) with the following model and loss function

$$h(w) = \frac{1}{\sqrt{n}}[f_w(x_1), \ldots, f_w(x_n)] \in \mathbb{R}^n, \quad R(v) = \frac{1}{2}\left\|v - \frac{y}{\sqrt{n}}\right\|_2^2. \tag{10}$$

Under certain regularity assumptions on the activation function (satisfied, for instance, by the sigmoid function) and a bound on the weights, it can be shown that Lip(Dh) is bounded.

Lemma 2.1 (Bound on Lip(Dh) for mean-field 2-layer network). Suppose there exists a constant K such that (i) the activation function σ is bounded and has bounded derivatives $\|\sigma\|_\infty, \|\sigma'\|_\infty, \|\sigma''\|_\infty, \|\sigma'''\|_\infty \le K$, (ii) the weights have bounded norm $\|U\|_a \le K$, and (iii) the data points have bounded norm $\|x\| \le K$. Then there exists a constant K' depending only on K such that

$$\text{Lip}(Dh) \le K'. \tag{11}$$

Since the assumptions of Theorem 1.2 are met, we obtain the following corollary for the lazy training dynamics of the 2-layer mean-field network.

Corollary 2.2 (Lazy training of 2-layer mean-field network). Suppose the conditions of Lemma 2.1 hold, and also that the labels are bounded in norm $\|y\| \leq c$. Then there exist constants $C, c > 0$ depending only on K such that for any time $0 \leq T \leq c\alpha^2$,

$$\|\alpha h(w(T)) - \alpha \bar{h}(\bar{w}(T))\| \leq C \min(T/\alpha, 1). \tag{12}$$

Training in the NTK parametrization corresponds to training the model $\sqrt{m} f_w$, where f_w is the network in the mean-field parametrization. This is equivalent to setting the lazy training parameter $\alpha = \sqrt{m}$ in the mean-field setting. Therefore, under the NTK parametrization with width m, the bound in Corollary 2.2 indicates that the NTK approximation is valid until training time $O(m)$ and the error bound is $O(T/\sqrt{m})$.

3 Proof Ideas

3.1 Proof Ideas for Theorem 1.2

To provide intuition for our proof, we first outline the approach used in the original proof. Define residuals $r(t), \bar{r}(t) \in \mathcal{F}$ under training the original rescaled model $\alpha h(w(t))$ and the linearized rescaled model $\alpha \bar{h}(\bar{w}(t))$ as $r(t) = y^* - \alpha h(w(t))$ and $\bar{r}(t) = y^* - \alpha \bar{h}(\bar{w}(t))$. These evolve according to

$$\frac{dr}{dt} = -K_t r \quad \text{and} \quad \frac{d\bar{r}}{dt} = -K_0 \bar{r}, \tag{13}$$

where $K_t := Dh(w(t))Dh(w(t))^*$ is the time-dependent kernel. To compare these trajectories, it was observed that, since K_0 is positive semidefinite,

$$\frac{d}{dt}\|r - \bar{r}\|_2^2 = -\langle r - \bar{r}, K_t r - K_0 \bar{r}\rangle \leq -\langle r - \bar{r}, (K_t - K_0) r\rangle \tag{14}$$

which, dividing both sides by $\|r - \bar{r}\|$ and using $\|r\| \leq \sqrt{R_0}$, implies

$$\frac{d}{dt}\|r - \bar{r}\| \leq \|K_t - K_0\|\|r\| \leq 2\text{Lip}(h)\text{Lip}(Dh)\|w - w_0\|\sqrt{R_0}. \tag{15}$$

Using the Lipschitzness of the model, it was further shown that the weight change is bounded by $\|w(t) - w_0\| \leq t\sqrt{R_0}\text{Lip}(h)/\alpha$. Plugging this into (7) yields the bound in Proposition 1.1,

$$\|\alpha h(w(T)) - \alpha \bar{h}(\bar{w}(T))\| = \|r(T) - \bar{r}(T)\| \leq 2\text{Lip}(h)^2\text{Lip}(Dh)R_0\alpha^{-1}\int_0^T t dt = T^2\text{Lip}(h)^2\text{Lip}(Dh)R_0/\alpha. \tag{16}$$

First attempt: strengthening of the bound for long time horizons We demonstrate how to strengthen this bound to hold for longer time horizons by employing an improved bound on the movement of the weights. Consider the following bound on the weight change.

Proposition 3.1 (Bound on weight change, implicit in proof of Theorem 2.2).

$$\|w(T) - w_0\| \leq \sqrt{TR_0/\alpha} \quad \text{and} \quad \|\bar{w}(T) - w_0\| \leq \sqrt{TR_0/\alpha}. \tag{17}$$

Proof of Proposition 3.1. By (a) Cauchy-Schwarz, and (b) the nonnegativity of the loss R,

$$\|w(T) - w(0)\| \leq \int_0^T \left\|\frac{dw}{dt}\right\| dt \overset{(a)}{\leq} \sqrt{T\int_0^T \left\|\frac{dw}{dt}\right\|^2 dt} = \sqrt{-\frac{T}{\alpha^2}\int_0^T \frac{d}{dt}R(\alpha h(w(t)))dt} \overset{(b)}{\leq} \sqrt{TR_0/\alpha}. \tag{18}$$

The bound for $\bar{w}$ is analogous.

This bound (8) has the advantage of $\sqrt{t}$ dependence (instead of linear t dependence) and does not depend on $\text{Lip}(h)$. Plugging it into (7), we obtain

$$\|\alpha h(w(T)) - \alpha \bar{h}(\bar{w}(T))\| \leq 2\text{Lip}(h)\text{Lip}(Dh)R_0\alpha^{-1}\int_0^T \sqrt{t} dt = \frac{4}{3}T^{3/2}\text{Lip}(h)\text{Lip}(Dh)R_0/\alpha. \tag{19}$$

This improves over Proposition 1.1 for long time horizons, as the time dependence scales as $T^{3/2}$ instead of T^2. However, it still depends on the Lipschitz constant $\text{Lip}(h)$ and falls short of the linear in T dependence of Theorem 1.2.

Second attempt: new approach to prove Theorem 1.2 To avoid dependence on $\text{Lip}(h)$ and achieve a linear dependence in T, we develop a new approach. We cannot use (7), which was central to the original proof, as it depends on $\text{Lip}(h)$. Furthermore, to achieve linear T dependence using (7), we would need $\|w - w_0\| = O(1)$ for a constant independent of the time horizon, which is not true unless the problem is well-conditioned.

In the full proof in Appendix A, we bound $\|r(T) - \bar{r}(T)\|$, which requires working with a product integral formulation of the dynamics of r to handle the time-varying kernels K_t. The main technical innovation in the proof is Theorem A.8, which is a new, general bound on the difference between product integrals.

To avoid the technical complications of the appendix, we provide some intuitions here by proving a simplified theorem that does not imply the main result. We show:

Theorem 3.2 (Simplified variant of Theorem 1.2). Consider $r'(t) \in \mathcal{F}$ initialized as $r'(0) = r(0)$ and evolving as $\frac{dr'}{dt} = -K_T r'$. Then,

$$\|r'(T) - \bar{r}(T)\| \leq \min(3\kappa\sqrt{R_0}, 8R_0). \tag{20}$$

Intuitively, if we can prove in Theorem 3.2 that $r'(T)$ and $\bar{r}(T)$ are close, then the same should hold for $r(T)$ and $\bar{r}(T)$ as in Theorem 1.2. For convenience, define the operators

$$A = Dh(w_0)^* \quad \text{and} \quad B = Dh(w(T))^* - Dh(w_0)^*. \tag{21}$$

Since the kernels do not vary in time, the closed-form solution is

$$r'(t) = e^{-(A+B)^*(A+B)t} r(0) \quad \text{and} \quad \bar{r}(t) = e^{-A^*At} r(0) \tag{22}$$

We prove that the time evolution operators for r' and $\bar{r}$ are close in operator norm.

Lemma 3.3. For any $t \geq 0$, we have $\|e^{-(A+B)^*(A+B)t} - e^{-A^*At}\| \leq 2\sqrt{t}\|B\|$.

Proof of Lemma 3.3. Define $Z(\zeta) = (A + \zeta B)^*(A + \zeta B)t$. By the fundamental theorem of calculus,

$$\|e^{-(A+B)^*(A+B)t} - e^{-A^*At}\| = \|e^{Z(1)} - e^{Z(0)}\| = \left\| \int_0^1 \frac{d}{d\zeta} e^{Z(\zeta)} d\zeta \right\| \leq \sup_{\zeta \in [0,1]} \left\| \frac{d}{d\zeta} e^{Z(\zeta)} \right\|. \tag{23}$$

Using the integral representation of the exponential map,

$$\left\| \frac{d}{d\zeta} e^{Z(\zeta)} \right\| = \left\| \int_0^1 e^{(1-\tau)Z(\zeta)} \left(\frac{d}{d\zeta} Z(\zeta) \right) e^{\tau Z(\zeta)} d\tau \right\| = \left\| \int_0^1 e^{(1-\tau)Z(\zeta)} (A^*B + B^*A + 2\zeta B^*B) e^{\tau Z(\zeta)} d\tau \right\| \tag{24}$$

By symmetry under transposing and reversing time, it suffices to bound the first term. Since $\|e^{\tau Z(\zeta)}\| \leq 1$,

$$\left\| \int_0^1 e^{(1-\tau)Z(\zeta)} (A + \zeta B)^* B e^{\tau Z(\zeta)} t d\tau \right\| \leq \int_0^1 \|e^{(1-\tau)Z(\zeta)} (A + \zeta B)^*\| \|tB\| d\tau \leq 2t/e \|B\| \leq 2\sqrt{t}\|B\| \tag{25}$$

Finally, let us combine Lemma 3.3 with the weight-change bound in Proposition 3.1 to prove Theorem 3.2. Notice that the weight-change bound in Proposition 3.1 implies

$$\|B\| \leq \text{Lip}(Dh)\|w(T) - w_0\| \leq \text{Lip}(Dh)\sqrt{TR_0/\alpha}. \tag{26}$$

So Lemma 3.3 implies

$$\|r'(T) - \bar{r}(T)\| \leq 2\text{Lip}(Dh)T\sqrt{R_0}\alpha^{-1}\|r(0)\| = 2\kappa\|r(0)\|. \tag{27}$$

Combining this with $\|r'(T) - \bar{r}(T)\| \leq \|r'(T)\| + \|\bar{r}(T)\| \leq 2\sqrt{2R_0}$ implies (9). Thus, we have shown Theorem 3.2, which is the result of Theorem 1.2 if we replace r by r'. The actual proof of the theorem handles the time-varying kernel K_t and is in Appendix A.

3.2 Proof Ideas for Theorem 1.3

The converse in Theorem 1.3 is achieved in the simple case where $h(w) = aw + \frac{1}{2}bw^2$ for $a = 1/\sqrt{T}$ and $b = \text{Lip}(Dh)$, and $w_0 = 0$ and $R(y) = \frac{1}{2}(y - \sqrt{2R_0})^2$, as we show in Appendix B by direct calculation.

4 Discussion

A limitation of our result is that it applies only to gradient flow, which corresponds to SGD with infinitesimally small step size. However, larger step sizes are beneficial for generalization in practice, so it would be interesting to understand the validity of the NTK approximation in that setting. Another limitation is that our result applies only to the square loss and not to other popular losses such as the cross-entropy loss. Indeed, the known bounds in the setting of general losses require either a "well-conditioning" assumption or taking α exponential in the training time T. Can one prove bounds analogous to Theorem 1.2 for more general losses, with α depending polynomially on T, and without conditioning assumptions?

A natural question raised by our bounds in Theorems 1.2 and 1.3 is: how do the dynamics behave just outside the regime where the NTK approximation is valid? For models h where $\text{Lip}(h)$ and $\text{Lip}(Dh)$ are bounded by a constant, can we understand the dynamics in the regime where $T \approx C\alpha$ for some large constant C and $\alpha \gg C$, at the edge of the lazy training regime?