# Paper Review: Neural Tangent Kernel

Beomsu Kim*

July 19, 2021

**Paper Information.**

- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel. In Neural Information Processing Systems, 2018.

## 1 Introduction

- In this paper, we investigate fully connected networks in the infinite-width limit, and describe the dynamics of the network function $f_\theta$ during training.

## 2 Neural Networks

- We consider fully-connected ANNs with layers numbered from 0 (input) to $L$ (output).
- $n_l$ : number of neurons in layer $l$.
- $\sigma : \mathbb{R} \to \mathbb{R}$ : Lipschitz, twice differentiable nonlinearity function, with bounded second derivative.
- $\theta$ : weights $W^{(l)} \in \mathbb{R}^{n_l \times n_{l+1}}$ and bias vectors $b^{(l)} \in \mathbb{R}^{n_l+1}$. Initialized as i.i.d. Gaussians $\mathcal{N}(0,1)$.
- $P = \sum_{l=0}^{L-1}(n_l+1)n_{l+1}$ : number of parameters.
- $\mathcal{F} = \{f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}\}$ : space of functions.
- $F^{(L)} : \mathbb{R}^P \to \mathcal{F}$ : ANN realization function, mapping parameters to the functions $f_\theta \in \mathcal{F}$.
- $p^{in} = \sum_{i=1}^N \delta_{x_i}$ : input distribution.
- $\langle f, g \rangle_{p^{in}} = \mathbb{E}_{x\sim p^{in}}[f(x)^\top g(x)]$ : bilinear form defined on $p^{in}$.
- $\|f\|_{p^{in}} = \langle f, f\rangle_{p^{in}}^{1/2}$ : seminorm defined on $p^{in}$.
- Define the functions

$$\begin{aligned}
\alpha^{(0)}(x;\theta) &= x, \\
\tilde{\alpha}^{(l+1)}(x;\theta) &= \frac{1}{\sqrt{n_l}} W^{(l)}\alpha^{(l)}(x;\theta) + \beta b^{(l)}, \\
\alpha^{(l)}(x;\theta) &= \sigma(\tilde{\alpha}^{(l)}(x;\theta)), \\
f_\theta(x) &= \tilde{\alpha}^{(L)}(x;\theta).
\end{aligned}$$

*Department of Mathematical Sciences, KAIST. Email beomsu.kim@kaist.ac.kr

## 3 Kernel Gradient

- If you are reading NTK for the first time, it is helpful to set $n_L = 1$. The concept of a multi-dimensional kernel can be confusing for $n_L > 1$.
- $C : \mathcal{F} \to \mathbb{R}$ : functional cost.
- $K : \mathbb{R}^{n_0 \times n_0} \to \mathbb{R}^{n_L \times n_L}$ : multi-dimensional kernel which satisfies $K(x, x') = K(x', x)^\top$.
- $\langle f, g \rangle_K = \mathbb{E}_{x,x' \sim p^{in}}[f(x)^\top K(x, x') g(x')]$ : inner product w.r.t. kernel $K$.
- The kernel $K$ is positive definite w.r.t. $\| \cdot \|_{p^{in}}$ if $\|f\|_{p^{in}} > 0 \implies \|f\|_K > 0$.
- $\mathcal{F}^* = \{\mu = \langle d, \cdot \rangle_{p^{in}} : d \in \mathcal{F}\}$ : the dual space of $\mathcal{F}$.
- $\Phi_K : \mathcal{F}^* \to \mathcal{F}$ is defined such that

$$\Phi_K : \langle d, \cdot \rangle_{p^{in}} \mapsto \frac{1}{N} \sum_{i=1}^{N} K(\cdot, x_i) d(x_i).$$

  $\Phi_K$ can be interpreted as a map which interpolates $d$ using the kernel $K$.
- $\partial_f^{in} C|_{f_0} = \langle d|_{f_0}, \cdot \rangle_{p^{in}}$ : functional derivative of $C$ at a point $f_0 \in \mathcal{F}$.
- $\nabla_K C|_{f_0} = \Phi_K(\partial_f^{in} C|_{f_0})$ : kernel gradient.
- In contrast to $\partial_f^{in} C$ which is only defined on the dataset, the kernel gradient generalizes to values $x$ outside the dataset thanks to the kernel $K$.
- A time-dependent function $f(t)$ follows the kernel gradient descent w.r.t. $K$ if it satisfies

$$\partial_t f(t) = -\nabla_K C|_{f(t)} = -\Phi_K(\partial_f^{in} C|_{f(t)}) = -\frac{1}{N} \sum_{i=1}^{N} K(\cdot, x_i) d|_{f(t)}(x_i).$$

- During kernel gradient descent, the cost $C(f(t))$ evolves as

$$\begin{aligned}
\partial_t C|_{f(t)} = \partial_t C(f(t)) = \partial_f^{in} C|_{f(t)}(\partial_t f(t)) &= \left\langle d|_{f(t)}, \partial_t f(t) \right\rangle_{p^{in}} \\
&= \left\langle d|_{f(t)}, -\frac{1}{N} \sum_{i=1}^{N} K(\cdot, x_i) d|_{f(t)}(x_i) \right\rangle_{p^{in}} \\
&= \frac{1}{N} \sum_{j=1}^{N} d|_{f(t)}(x_j)^\top \left( -\frac{1}{N} \sum_{i=1}^{N} K(x_j, x_i) d|_{f(t)}(x_i) \right) \\
&= -\frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} d|_{f(t)}(x_j)^\top K(x_j, x_i) d|_{f(t)}(x_i) \\
&= -\mathbb{E}_{x,x' \sim p^{in}}[d|_{f(t)}(x)^\top K(x, x') d|_{f(t)}(x')] \\
&= -\|d|_{f(t)}\|_K^2.
\end{aligned}$$

  Convergence to a critical point of $C$ is hence guaranteed if the kernel $K$ is positive definite with respect to $\| \cdot \|_{p^{in}}$ : the cost is then strictly decreasing except at points such that $\|d|_{f(t)}\|_{p^{in}} = 0$. If the cost is convex and bounded from below, the function $f(t)$ therefore converges to a global minimum as $t \to \infty$.
- For our setup, which is that of a finite dataset $x_1, \ldots, x_n \in \mathbb{R}^{n_0}$, the cost function $C$ only depends on the values of $f \in \mathcal{F}$ at the data points. Hence, the global minimum may not be unique. Specifically, any function which minimizes the cost functional on the data points is a global minimum. For instance, see Figure 2 in the paper.

### 3.1 Random Functions Approximation

- A kernel $K$ can be approximated by a choice of $P$ random functions $f^{(p)}$ sampled independently from any distribution on $\mathcal{F}$ whose (non-centered) covariance is given by the kernel $K$:

$$\mathbb{E}[f^{(p)}(x)f^{(p)}(x')^\top] = K(x,x')$$

or equivalently,

$$\mathbb{E}[f_k^{(p)}(x)f_{k'}^{(p)}(x')] = K_{kk'}(x,x').$$

- These functions define a random linear parametrization

$$F^{lin} : \mathbb{R}^P \to \mathcal{F} : \theta \mapsto f_\theta^{lin} = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \theta_p f^{(p)}.$$

- The partial derivatives of the parametrization are given by ($e_p$ is the $p$-th standard basis vector)

$$\partial_{\theta_p} F^{lin}(\theta) = \lim_{h\to 0} \frac{F^{lin}(\theta + he_p) - F^{lin}(\theta)}{h} = \frac{1}{\sqrt{P}} f^{(p)}.$$

- Optimizing the cost $C \circ F^{lin}$ through gradient descent, the parameters follow the ODE

$$\begin{aligned}
\partial_t \theta_p(t) = -\partial_{\theta_p}(C \circ F^{lin})(\theta(t)) &= -\partial_{\theta_p} C(f_{\theta(t)}^{lin}) \\
&= -\partial_f^{in} C|_{f_{\theta(t)}^{lin}}(\partial_{\theta_p} f_{\theta(t)}^{lin}) \\
&= -\frac{1}{\sqrt{P}} \partial_f^{in} C|_{f_{\theta(t)}^{lin}}(f^{(p)}) = -\frac{1}{\sqrt{P}} \left\langle d|_{f_{\theta(t)}^{lin}}, f^{(p)} \right\rangle_{p^{in}}.
\end{aligned}$$

The first equality holds since we are performing gradient descent, i.e., the instantaneous change of $\theta_p$ at time $t$ must equal the gradient of $\theta_p$ w.r.t. the cost at time $t$.

- As a result, the function $f_{\theta(t)}^{lin}$ evolves according to

$$\begin{aligned}
\partial_t f_{\theta(t)}^{lin} = \partial_t \left( \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \theta_p(t) f^{(p)} \right) &= \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \partial_t \theta_p(t) f^{(p)} \\
&= -\frac{1}{P} \sum_{p=1}^{P} \left\langle d|_{f_{\theta(t)}^{lin}}, f^{(p)} \right\rangle_{p^{in}} f^{(p)} \\
&= -\frac{1}{P} \sum_{p=1}^{P} \frac{1}{N} \sum_{i=1}^{N} d|_{f_{\theta(t)}^{lin}}(x_i)^\top f^{(p)}(x_i) f^{(p)}(\cdot) \\
&= -\frac{1}{P} \sum_{p=1}^{P} \frac{1}{N} \sum_{i=1}^{N} (f^{(p)} \otimes f^{(p)})(\cdot, x_i) d|_{f_{\theta(t)}^{lin}}(x_i) \\
&= -\frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{P} \sum_{p=1}^{P} f^{(p)} \otimes f^{(p)} \right) (\cdot, x_i) d|_{f_{\theta(t)}^{lin}}(x_i) \\
&= -\Phi_{\tilde{K}}(\partial_f^{in} C|_{f_{\theta(t)}^{lin}}) \\
&= -\nabla_{\tilde{K}} C|_{f_{\theta(t)}^{lin}}
\end{aligned}$$

where

$$\tilde{K} = \sum_{p=1}^{P} \partial_{\theta_p} F^{lin}(\theta) \otimes \partial_{\theta_p} F^{lin}(\theta) = \frac{1}{P} \sum_{p=1}^{P} f^{(p)} \otimes f^{(p)}.$$

- This is a random $n_L$-dimensional kernel with values

$$\tilde{K}_{ii'}(x, x') = \frac{1}{P} \sum_{p=1}^{P} f_i^{(p)}(x) f_{i'}^{(p)}(x').$$

- Performing gradient descent on the cost $C \circ F^{lin}$ is therefore equivalent to performing kernel gradient descent with the tangent kernel $\tilde{K}$ in the function space.
- With $P \to \infty$, by the law of large numbers, the random kernel $\tilde{K}$ tends to the fixed kernel $K$.

$$\lim_{P\to\infty} \tilde{K}_{ii'}(x, x') = \lim_{P\to\infty} \frac{1}{P} \sum_{p=1}^{P} f_i^{(p)}(x) f_{i'}^{(p)}(x') = \mathbb{E}[f_i^{(p)}(x) f_{i'}^{(p)}(x')] = K_{ii'}(x, x').$$

Hence, this method approximates kernel gradient descent with respect to the limiting kernel $K$.

# 4 Neural Tangent Kernel

- During training, the network function $f_\theta$ evolves along the negative kernel gradient

$$\partial_t f_{\theta(t)} = -\nabla_{\Theta^{(L)}} C|_{f_{\theta(t)}}$$

with respect to the neural tangent kernel (NTK)

$$\Theta^{(L)}(\theta) = \sum_{p=1}^{P} \partial_{\theta_p} F^{(L)}(\theta) \otimes \partial_{\theta_p} F^{(L)}(\theta)$$

such that (denoting $\Theta^{(L)}(\theta)$ by $\Theta^{(L)}$ for brevity),

$$\Theta^{(L)}(x, x') = \sum_{p=1}^{P} \partial_{\theta_p} F^{(L)}(\theta)(x) \otimes \partial_{\theta_p} F^{(L)}(\theta)(x') = \sum_{p=1}^{P} \partial_{\theta_p} f_\theta(x) \otimes \partial_{\theta_p} f_\theta(x').$$

$\Theta^{(L)}(\theta)$ can be derived by following the steps in Section 3.1 with $F^{(L)}$ in place of $F^{lin}$.

- However, in contrast to $F^{lin}$, the realization function $F^{(L)}$ of ANNs is not linear.
- As a consequence, the derivatives $\partial_{\theta_p} F^{(L)}(\theta)$ and the NTK depend on the parameters $\theta$.

## 4.1 Initialization

- The output functions $f_{\theta,i}$ for $i = 1, \ldots, n_L$ tend to i.i.d. Gaussian processes in the infinite limit.
- The first key result is that in the limit $n_1, \ldots, n_{L-1} \to \infty$, the NTK converges in probability to a deterministic limiting kernel.

## 4.2 Training

- The second key result is that the NTK stays asymptotically constant during training.
- In general, the parameters can be updated according to a training direction $d_t \in \mathcal{F}$.

$$\partial_t \theta_p(t) = \left\langle \partial_{\theta_p} F^{(L)}(\theta(t)), d_t \right\rangle_{p^{in}}$$

- In the case of gradient descent,

$$\begin{aligned}\partial_t \theta_p(t) = -\partial_{\theta_p}(C \circ F^{(L)})(\theta(t)) &= -\partial_{\theta_p} C(f_{\theta(t)}) \\ &= -\partial_f^{in} C|_{f_{\theta(t)}}(\partial_{\theta_p} f_{\theta(t)}) \\ &= \langle \partial_{\theta_p} f_{\theta(t)}, -d|_{f_{\theta(t)}} \rangle_{p^{in}} \\ &= \left\langle \partial_{\theta_p} F^{(L)}(\theta(t)), -d|_{f_{\theta(t)}} \right\rangle_{p^{in}}\end{aligned}$$

  and so

$$d_t = -d|_{f_{\theta(t)}}.$$

- The limiting NTK is positive definite if the span of the derivatives $\partial_{\theta_p} F^{(L)}$ becomes dense in $\mathcal{F}$ w.r.t. the $p^{in}$ norm as the width grows to infinity. See the Proposition below.

**Proposition.** *Let $\mathcal{F}$ be an infinite-dimensional vector space equipped with an inner product. Let $\{f_n\}_{n=1}^{\infty}$ be a set of vectors in $\mathcal{F}$ such that its span is dense in $\mathcal{F}$. That is, finite linear combinations of elements in $\{f_n\}_{n=1}^{\infty}$ is are dense in $\mathcal{F}$. Define (assuming it exists)*

$$\Theta = \sum_{n=1}^{\infty} f_n \otimes f_n.$$

*Then $\|g\| > 0$ implies that $\|g\|_\Theta > 0$, i.e., $\Theta$ is positive definite.*

*Proof.* Without loss of generality, we assume $\{f_n\}_{n=1}^{\infty}$ is linearly independent. By the Gram-Schmidt process, we may also assume $\{f_n\}_{n=1}^{\infty}$ is orthonormal. Suppose there is $g \in \mathcal{F}$ such that $\|g\|_\Theta = 0$, i.e.,

$$0 = \|g\|_\Theta^2 = \sum_{n=1}^{\infty} g^\top (f_n \otimes f_n) g = \sum_{n=1}^{\infty} \langle f_n, g \rangle^2 .$$

(We need to check whether we can interchange the inner product with $g$ and the infinite sum, but I'm being lazy here.) This implies that

$$\langle f_n, g \rangle = 0 \quad \forall n.$$

Since $\{f_n\}_{n=1}^{\infty}$ is an orthonormal basis of $\mathcal{F}$, $g = 0$ and so $\|g\| = 0$. (See, for instance, Theorem 4.2.3 of *Real Analysis* by Stein and Shakarchi.) □

# 5 Least-Squares Regression

- Given a goal function $f^*$ and input distribution $p^{in}$, the least-squares regression cost is

$$C(f) = \frac{1}{2}\|f - f^*\|_{p^{in}}^2 = \frac{1}{2}\mathbb{E}_{x \sim p^{in}}[\|f(x) - f^*(x)\|^2].$$

- We are interested in the behavior of a function $f_t$ during kernel gradient descent with a kernel $K$.

$$\partial_t f_t = \Phi_K(\langle f^* - f, \cdot \rangle_{p^{in}})$$

If we define the map $\Pi : f \mapsto \Phi_K(\langle f, \cdot \rangle_{p^{in}})$, this differential equation is equivalent to

$$\partial_t f_t = \Pi(f - f^*).$$

Since $\Pi$ is a linear operator on $\mathcal{F}$, the solution of this differential equation can be expressed as

$$f_t = f^* + e^{-t\Pi}(f_0 - f^*)$$

where

$$e^{-t\Pi} = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} \Pi^k.$$

For instance, see `https://en.wikipedia.org/wiki/Matrix_differential_equation`.

- If $\Pi$ can be diagonalized by eigenfunctions $f^{(i)}$ with eigenvalues $\lambda_i$, the exponential $e^{-t\Pi}$ has the same eigenfunctions with eigenvalues $e^{-t\lambda_i}$. Specifically,

$$e^{-t\Pi}(f^{(i)}) = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} \Pi^k(f^{(i)}) = \sum_{k=0}^{\infty} \frac{(-t\lambda_i)^k}{k!} f^{(i)} = e^{-t\lambda_i} f^{(i)}.$$

- For a finite dataset $x_1, \ldots, x_N$ of size $N$, the map $\Pi$ takes the form

$$\Pi(f)_k(x) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k'=1}^{n_L} f_{k'}(x_i) K_{kk'}(x_i, x).$$

This is from the definition of the maps $\Pi$ and $\Phi_K$.

- The map $\Pi$ has at most $Nn_L$ positive eigenfunctions since the dimension of $(K_{kk'}(x_i, x_j))_{ik,jk'}$ is $Nn_L \times Nn_L$. The eigenfunctions are the kernel principal components $f^{(1)}, \ldots, f^{(Nn_L)}$ of the data with respect to the kernel $K$. To understand what this means, consider the case $n_L = 1$ so

$$\Pi(f)(x) = \frac{1}{N} \sum_{i=1}^{N} f(x_i) K(x_i, x).$$

Each eigenfunction $f^{(n)}$ must satisfy

$$\lambda_n f^{(n)}(x_j) = \Pi(f^{(n)})(x_j) = \frac{1}{N} \sum_{i=1}^{N} f^{(n)}(x_i) K(x_i, x_j)$$

or in matrix form,

$$N\lambda_n (f^{(n)}(x_j))_j = (K(x_i, x_j))_{i,j} (f^{(n)}(x_j))_j.$$

This coincides with Equation (8) in `https://people.eecs.berkeley.edu/~wainwrig/stat241b/scholkopf_kernel.pdf`. The corresponding eigenvalues $\lambda_i$ is the variance captured by the component.

- Decomposing the difference

$$f^* - f_0 = \Delta_f^0 + \Delta_f^1 + \cdots + \Delta_f^{Nn_L}$$

along the eigenspaces of $\Pi$, the trajectory of the function $f_t$ reads

$$f_t = f^* + \Delta_f^0 + \sum_{i=1}^{Nn_L} e^{-t\lambda_i}\Delta_f^i,$$

where $\Delta_f^0$ is in the kernel (null-space) of $\Pi$ and $\Delta_f^i \propto f^{(i)}$.

- Note that by the linearity of the map $e^{-t\Pi}$, if $f_0$ is initialized with a Gaussian distribution (as in the case for ANNs in the infinite-width limit), then $f_t$ is Gaussian for all times $t$.
- Assuming that the kernel is positive definite on the data (implying that the $Nn_L \times Nn_L$ Gram matrix $\tilde{K} = (K_{kk'}(x_i, x_j))_{ik,jk'}$ is invertible), as $t \to \infty$ limit, we get that

$$f_\infty = f^* + \Delta_f^0 = f_0 - \sum_{i=1}^{Nn_L} \Delta_f^i$$

takes the form

$$f_{\infty,k}(x) = \kappa_{x,k}^\top \tilde{K}^{-1} y^* + (f_0(x) - \kappa_{x,k}^\top \tilde{K}^{-1} y_0)$$

with the $Nn_L$-vectors $\kappa_{x,k}$, $y^*$, and $y_0$ given by

$$\begin{aligned} \kappa_{x,k} &= (K_{kk'}(x, x_i))_{i,k'} \\ y^* &= (f_k^*(x_i))_{i,k} \\ y_0 &= (f_{0,k}(x_i))_{i,k}. \end{aligned}$$

- The expression for $f_\infty(x)$ might be confusing. Let us analyze the expression. We observe that

$$\sum_{i=1}^{Nn_L} \Delta_f^i$$

corresponds to the projection of $f^* - f_0$ onto the range of $\Pi$. This is because, by definition,

$$\sum_{i=1}^{Nn_L} \Delta_f^i = (f^* - f_0) - \Delta_f^0,$$

and $\Delta_f^0$ is the component of $f^* - f_0$ in the null space of $\Pi$. To verify the relation

$$\sum_{i=1}^{Nn_L} \Delta_f^i = \kappa_{x,k}^\top \tilde{K}^{-1}(y^* - y_0),$$

we thus need to check two facts.

  - $\kappa_{x,k}^\top \tilde{K}^{-1}(y^* - y_0)$ is in the range of $\Pi$.[1]
  - $\Pi(\kappa_{x,k}^\top \tilde{K}^{-1}(y^* - y_0)) = \Pi(f)$.

Recall that

$$\Pi(f)_k(x) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k'=1}^{n_L} f_{k'}(x_i) K_{kk'}(x_i, x).$$

[1]Here, I am assuming the null space and the range of $\Pi$ is orthogonal. This holds for linear operators on finite-dimensional vector spaces.

The range of $\Pi$ is the collection of functions that can be expressed as a linear combination of $\{K_{kk'}(x_i, x)\}_{i,k'}$. We indeed observe that $\kappa_{x,k}^\top \tilde{K}^{-1}(y^* - y_0)$ is a linear combination of $\{K_{kk'}(x_i, x)\}_{i,k'}$. Moreover, it can easily be calculated that

$$\kappa_{x_i,k}^\top \tilde{K}^{-1}(y^* - y_0) = (f^* - f_0)(x_i)$$

and so the second point is true as well.

- The first term, $\kappa_{x,k}^\top \tilde{K}^{-1} y^*$, has an important statistical interpretation. It is the maximum-a-posteriori (MAP) estimate given a Gaussian prior on functions $f_k \sim \mathcal{N}(0, \Theta_\infty^{(L)})$ and the conditions $f_k(x_i) = f_k^*(x_i)$. For instance, see Equation (2.19) in `http://www.gaussianprocess.org/gpml/chapters/RW.pdf`. Equivalently, it is equal to the kernel ridge regression as the regularization goes to zero.
- The second term $(f_0(x) - \kappa_{x,k}^\top \tilde{K}^{-1} y_0)$ is a centered Gaussian whose variance vanishes on the points of the dataset. Indeed, since $f_{0,k} \sim \mathcal{N}(0, \Theta_\infty^{(L)})$,

$$\mathbb{E}[f_0(x) - \kappa_{x,k}^\top \tilde{K}^{-1} y_0] = \mathbb{E}[f_0(x)] - \kappa_{x,k}^\top \tilde{K}^{-1} \mathbb{E}[y_0] = 0.$$

Also,

$$f_0(x_i) - \kappa_{x_i,k}^\top \tilde{K}^{-1} y_0 = f_0(x_i) - f_0(x_i) = 0$$

so the variance vanishes on the points of the dataset.
- This shows that in the infinite-width limit, we are essentially fitting a Gaussian process to the dataset! Indeed, see Figure 2.

# A Appendix

- We study the limit of the NTK as $n_1, \ldots, n_{L-1} \to \infty$ sequentially. That is, we first take $n_1 \to \infty$, then $n_2 \to \infty$, and so on. This leads to much simpler proofs, but our results could in principle by strengthened to the more general setting when $\min(n_1, \ldots, n_{L-1}) \to \infty$.
- A natural choice of convergence to study the NTK is w.r.t. the operator norm on kernels

$$\|K\|_{op} = \max_{\|f\|_{p^{in}} \le 1} \|f\|_K = \max_{\|f\|_{p^{in}} \le 1} \sqrt{\mathbb{E}_{x,x' \sim p^{in}}[f(x)^\top K(x,x') f(x')]}.$$

- Define

$$\mathbf{f} := (f(x_1), \ldots, f(x_N)) \in \mathbb{R}^{Nn_L}, \qquad \mathbf{K} := (K_{kk'}(x_i, x_j))_{k,k' < n_L, i,j < N} \in \mathbb{R}^{Nn_L \times Nn_L}.$$

We then have

$$\|f\|_{p^{in}}^2 = \langle f, f \rangle_{p^{in}} = \mathbb{E}_{x \sim p^{in}}[\|f(x)\|_2^2] = \frac{1}{N} \sum_{i=1}^N \|f(x_i)\|_2^2 = \frac{1}{N} \|(f(x_1), \ldots, f(x_N))\|^2 = \frac{1}{N} \|\mathbf{f}\|_2^2.$$

Also, note that

$$\mathbb{E}_{x,x'}[f(x)^\top K(x,x') f(x')] = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N f(x_i)^\top K(x_i, x_j) f(x_j) = \frac{1}{N^2} \mathbf{f}^\top \mathbf{K} \mathbf{f}$$

and so

$$\|K\|_{op} = \frac{1}{N} \max_{\|\mathbf{f}\|_2 \le \sqrt{N}} \sqrt{\mathbf{f}^\top \mathbf{K} \mathbf{f}} = \frac{1}{\sqrt{N}} \max_{\|\mathbf{f}\|_2 \le 1} \sqrt{\mathbf{f}^\top \mathbf{K} \mathbf{f}}$$

Hence, $\|K\|_{op}$ is equal to the (scaled) leading eigenvalue of the $Nn_L \times Nn_L$ Gram matrix $\mathbf{K}$.

## A.1 Asymptotics at Initialization

**Proposition 1.** *For a network of depth $L$ at initialization, with a Lipschitz nonlinearity $\sigma$, and in the limit as $n_1, \ldots, n_{L-1} \to \infty$ sequentially, the output functions $f_{\theta,k}$, for $k = 1, \ldots, n_L$, tend (in law) to i.i.d. centered Gaussian processes of covariance $\Sigma^{(L)}$, where $\Sigma^{(L)}$ is defined recursively by*

$$\Sigma^{(1)}(x, x') = \frac{1}{n_0} x^\top x' + \beta^2,$$
$$\Sigma^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}(0, \Sigma^{(L)})}[\sigma(f(x))\sigma(f(x'))] + \beta^2$$

*where $\mathcal{N}(0, \Sigma^{(L)})$ denotes the centered Gaussian process with covariance function $\Sigma^{(L)}$.*

*Proof.* We prove the result by induction. When $L = 1$, there are no hidden layers and $f_\theta$ is a random affine function of the form

$$f_\theta(x) = \frac{1}{\sqrt{n_0}} W^{(0)} x + \beta b^{(0)}.$$

For any index $i$, we thus have

$$\mathbb{E}[f_{\theta,i}(x)] = \mathbb{E}\left[\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} W_{i,j}^{(0)} x_j + \beta b_i^{(0)}\right] = \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \mathbb{E}[W_{i,j}^{(0)}] x_j + \beta \mathbb{E}[b_i^{(0)}] = 0$$

because the parameters are initialized from $\mathcal{N}(0, 1)$. It follows that

$$\begin{aligned}
\Sigma^{(1)}(x, x') &= \mathbb{E}[f_{\theta,i}(x) f_{\theta,i}(x')] - \mathbb{E}[f_{\theta,i}(x)]\mathbb{E}[f_{\theta,i}(x')] \\
&= \mathbb{E}[f_{\theta,i}(x) f_{\theta,i}(x')] \\
&= \frac{1}{n_0} \sum_{j=1}^{n_0} \sum_{j'=1}^{n_0} \mathbb{E}[W_{i,j}^{(0)} W_{i,j'}^{(0)}] x_j x'_{j'} + \frac{\beta}{\sqrt{n_0}} \sum_{j=1}^{n_0} \mathbb{E}[W_{i,j}^{(0)} \beta_i^{(0)}](x_j + x'_j) + \beta^2 \mathbb{E}[(b_i^{(0)})^2] \\
&= \frac{1}{n_0} x^\top x' + \beta^2
\end{aligned}$$

which proves $f_{\theta,i} \sim \mathcal{N}(0, \Sigma^{(1)})$. This concludes the base case.

The key to the induction step is to consider an $(L+1)$-network as the following composition. An $L$-network $\mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ mapping the input to the pre-activations $\tilde{\alpha}_i^{(L)}$, followed by an elementwise application of the nonlinearity $\sigma$ and then a random affine map $\mathbb{R}^{n_L} \to \mathbb{R}^{n_{L+1}}$. The induction hypothesis gives in the limit as sequentially $n_1, \ldots, n_{L-1} \to \infty$, the preactivations $\tilde{\alpha}_i^{(L)}$ tend to i.i.d. Gaussian processes with covariance $\Sigma^{(L)}$. Then, conditioned on $\tilde{\alpha}_i^{(L)} \sim \mathcal{N}(0, \Sigma^{(L)})$, the outputs

$$f_{\theta,i}(x) = \frac{1}{\sqrt{n_L}} \sum_{j=1}^{n_L} W_{i,j}^{(L)} \sigma(\tilde{\alpha}_j^{(L)}(x; \theta)) + \beta b_i^{(L)}$$

are i.i.d. centered Gaussians with covariance

$$\tilde{\Sigma}^{(L+1)}(x, x') = \mathbb{E}[f_{\theta,i}(x) f_{\theta,i}(x')] = \frac{1}{n_L} \sigma(\tilde{\alpha}^{(L)}(x; \theta))^\top \sigma(\tilde{\alpha}^{(L)}(x; \theta)) + \beta^2.$$

In other words, conditioned on $\tilde{\alpha}_i^{(L)} \sim \mathcal{N}(0, \Sigma^{(L)})$, $f_{\theta,i} \sim \mathcal{N}(0, \tilde{\Sigma}^{(L+1)})$. The proof for this is identical to the proof for the base case. Specifically, repeat the proof for the base case with $\sigma(\tilde{\alpha}^{(L)}(x; \theta))$ in place of $x$. By the law of large numbers, as $n_L \to \infty$, this covariance tends in probability to the expectation

$$\begin{aligned}\Sigma^{(L+1)}(x,x') &:= \lim_{n_L\to\infty} \tilde{\Sigma}^{L+1}(x,x') \\ &= \lim_{n_L\to\infty} \frac{1}{n_L}\sigma(\tilde{\alpha}^{(L)}(x;\theta))^\top \sigma(\tilde{\alpha}^{(L)}(x;\theta)) + \beta^2 \\ &= \lim_{n_L\to\infty} \frac{1}{n_L}\sum_{i=1}^{n_L} \sigma(\tilde{\alpha}_i^{(L)}(x;\theta))\sigma(\tilde{\alpha}_i^{(L)}(x';\theta)) + \beta^2 \\ &= \mathbb{E}_{f\sim\mathcal{N}(0,\Sigma^{(L)})}[\sigma(f(x))\sigma(f(x'))] + \beta^2\end{aligned}$$

because $\tilde{\alpha}_i^{(L)}$ are i.i.d. samples from $\mathcal{N}(0,\Sigma^{(L)})$. In particular, the covariance is deterministic and hence independent of $\alpha^{(L)}$. As a consequence, the conditioned and unconditioned distributions of $f_{\theta,i}$ are equal in the limit. They are i.i.d. centered Gaussian processes of covariance $\Sigma^{(L+1)}$. This concludes the induction step and so the proposition is proved. □

**Theorem 1.** *For a network of depth $L$ at initialization, with a Lipschitz nonlinearity $\sigma$, and in the limit as the layers width $n_1,\dots,n_{L-1}\to\infty$ sequentially, the NTK $\Theta^{(L)}$ converges in probability to a deterministic limiting kernel*

$$\Theta^{(L)} \to \Theta_\infty^{(L)} \otimes Id_{n_L}.$$

*The scalar kernel $\Theta_\infty^{(L)} : \mathbb{R}^{n_0}\times\mathbb{R}^{n_0}\to\mathbb{R}$ is defined recursively by*

$$\begin{aligned}\Theta_\infty^{(1)}(x,x') &= \Sigma^{(1)}(x,x') \\ \Theta_\infty^{(L+1)}(x,x') &= \Theta^{(L)}(x,x')\dot{\Sigma}^{(L+1)}(x,x') + \Sigma^{(L+1)}(x,x'),\end{aligned}$$

*where*

$$\dot{\Sigma}^{(L+1)}(x,x') = \mathbb{E}_{f\sim\mathcal{N}(0,\Sigma^{(L)})}[\dot{\sigma}(f(x))\dot{\sigma}(f(x'))],$$

*where $\dot{\sigma}$ denotes the derivative of $\sigma$.*

*Proof.* We first remark that

$$\Theta_\infty^{(L)} \otimes Id_{n_L}$$

is just the $n_L \times n_L$ identity matrix multiplied by the scalar kernel $\Theta_\infty^{(L)}$.

Let us begin the proof. Recall that the NTK is given by

$$\Theta^{(L)}(x,x') = \sum_{p=1}^{P} \partial_{\theta_p} f_\theta(x) \otimes \partial_{\theta_p} f_\theta(x').$$

It follows that

$$\Theta_{kk'}(x,x') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\theta,k}(x) \cdot \partial_{\theta_p} f_{\theta,k'}(x').$$

From here, we proceed by induction. When $L=1$,

$$f_\theta(x) = \frac{1}{\sqrt{n_0}} W^{(0)} x + \beta b^{(0)}.$$

The NTK is a sum over the entries of $W^{(0)}$ and those of $b^{(0)}$. Since

$$\partial_{W_{ij}^{(0)}} f_{\theta,k}(x) = \frac{1}{\sqrt{n_0}} x_i \delta_{jk}, \qquad \partial_{b_j^{(0)}} f_{\theta,k}(x) = \beta \delta_{jk},$$

we have

$$\begin{aligned}
\Theta_{kk'}(x,x') &= \sum_{i=1}^{n_0}\sum_{j=1}^{n_1} \partial_{W_{ij}^{(0)}} f_{\theta,k}(x) \cdot \partial_{W_{ij}^{(0)}} f_{\theta,k'}(x') + \sum_{j=1}^{n_1} \partial_{b_j^{(0)}} f_{\theta,k}(x) \cdot \partial_{b_j^{(0)}} f_{\theta,k'}(x') \\
&= \frac{1}{n_0} \sum_{i=1}^{n_0}\sum_{j=1}^{n_1} x_i x'_i \delta_{jk}\delta_{jk'} + \beta^2 \sum_{j=1}^{n_1} \delta_{jk}\delta_{jk'} \\
&= \frac{1}{n_0} x^\top x \delta_{kk'} + \beta^2 \delta_{kk'} \\
&= \Sigma^{(1)}(x,x')\delta_{kk'} \\
&= (\Sigma^{(1)}(x,x') \otimes I_{n_1})_{kk'}
\end{aligned}$$

and so

$$\Theta(x,x') = \Sigma^{(1)}(x,x') \otimes I_{n_1}.$$

This concludes the base case.

Here again, the key to proving the induction step is the observation that a network of depth $L+1$ is an $L$-network mapping the inputs $x$ to the preactivations on the $L$-th layer $\tilde{\alpha}^{(L)}(x)$ followed by a nonlinearity and a random affine function. For a network of depth $L+1$, let us therefore split the parameters into the parameters $\tilde{\theta}$ of the first $L$ layers and those of the last layer $(W^{(L)}, b^{(L)})$. We can then split the NTK into a sum over the parameters $\tilde{\theta}$ of the first $L$ layers and the remaining parameters $W^{(L)}$ and $b^{(L)}$.

$$\Theta_{kk'}^{(L+1)}(x,x') = \underbrace{\sum_{\tilde{\theta}_p \in \tilde{\theta}} \partial_{\tilde{\theta}_p} f_{\theta,k}(x) \cdot \partial_{\tilde{\theta}_p} f_{\theta,k'}(x')}_{①} + \underbrace{\sum_{\theta_p \in (W^{(L)}, b^{(L)})} \partial_{\theta_p} f_{\theta,k}(x) \cdot \partial_{\theta_p} f_{\theta,k'}(x')}_{②}.$$

We need to calculate ① and ② in the limit $n_1, \ldots, n_{L+1} \to \infty$.

Let us calculate ① as $n_1, \ldots, n_{L+1} \to \infty$. Since

$$f_{\theta,k}(x) = \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} W_{ik}^{(L)} \sigma(\tilde{\alpha}_i^{(L)}(x;\theta)) + \beta b_k^{(L)},$$

by the chain rule, for $\tilde{\theta}_p \in \tilde{\theta}$,

$$\partial_{\tilde{\theta}_p} f_{\theta,k}(x) = \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} \partial_{\tilde{\theta}_p} \tilde{\alpha}_i^{(L)}(x;\theta) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) W_{ik}^{(L)}.$$

It follows that the contribution of the parameters $\tilde{\theta}$ to the NTK in the limit $n_1, \ldots, n_{L-1} \to \infty$ is

$$
\begin{aligned}
① &= \frac{1}{n_L} \sum_{\tilde{\theta}_p \in \tilde{\theta}} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} \left( \partial_{\tilde{\theta}_p} \tilde{\alpha}_i^{(L)}(x;\theta) \cdot \partial_{\tilde{\theta}_p} \tilde{\alpha}_{i'}^{(L)}(x';\theta) \right) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x';\theta)) W_{ik}^{(L)} W_{i'k'}^{(L)} \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} \left[ \sum_{\tilde{\theta}_p \in \tilde{\theta}} \partial_{\tilde{\theta}_p} \tilde{\alpha}_i^{(L)}(x;\theta) \cdot \partial_{\tilde{\theta}_p} \tilde{\alpha}_{i'}^{(L)}(x';\theta) \right] \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x';\theta)) W_{ik}^{(L)} W_{i'k'}^{(L)} \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} \Theta_{ii'}^{(L)}(x,x') \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x';\theta)) W_{ik}^{(L)} W_{i'k'}^{(L)} \\
&\to \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} (\Theta_\infty^{(L)}(x,x') \otimes I_{n_L})_{ii'} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x';\theta)) W_{ik}^{(L)} W_{i'k'}^{(L)} \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{i'=1}^{n_L} \Theta_\infty^{(L)}(x,x') \delta_{ii'} \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_{i'}^{(L)}(x';\theta)) W_{ik}^{(L)} W_{i'k'}^{(L)} \\
&= \frac{1}{n_L} \sum_{i=1}^{n_L} \Theta_\infty^{(L)}(x,x') \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x';\theta)) W_{ik}^{(L)} W_{ik'}^{(L)}.
\end{aligned}
$$

At the third line, we have used the definition of the NTK of the first $L$ layers of the network. At the fourth line, we have used the induction hypothesis. Then, by the law of large numbers, as $n_L \to \infty$, this tends to its expectation

$$
\begin{aligned}
&\lim_{n_L \to \infty} \frac{1}{n_L} \sum_{i=1}^{n_L} \Theta_\infty^{(L)}(x,x') \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x';\theta)) W_{ik}^{(L)} W_{ik'}^{(L)} \\
&= \mathbb{E} \left[ \Theta_\infty^{(L)}(x,x') \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x';\theta)) W_{ik}^{(L)} W_{ik'}^{(L)} \right] \\
&= \Theta_\infty^{(L)}(x,x') \mathbb{E} \left[ \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x';\theta)) \right] \mathbb{E} \left[ W_{ik}^{(L)} W_{ik'}^{(L)} \right] \\
&= \Theta_\infty^{(L)}(x,x') \dot{\Sigma}^{(L+1)}(x,x') \delta_{kk'} \\
&= (\Theta_\infty^{(L)}(x,x') \dot{\Sigma}^{(L+1)}(x,x') \otimes I_{n_{L+1}})_{kk'}.
\end{aligned}
$$

At the third line, we have used the independence of $\dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta))\dot{\sigma}(\tilde{\alpha}_i^{(L)}(x';\theta))$ and $W_{ik}^{(L)} W_{ik'}^{(L)}$. At the fourth line, we have used Proposition 1 to evaluate the first expectation, i.e., $\tilde{\alpha}_i^{(L)} \sim \mathcal{N}(0, \Sigma^{(L)})$.

We now calculate ② in the limit $n_1, \ldots, n_{L+1} \to \infty$.

□

### A.2 Asymptotics During Training

Given a training direction $t \mapsto d_t \in \mathcal{F}$, a neural network is trained in the following manner: the parameters $\theta_p$ are initialized as i.i.d. $\mathcal{N}(0,1)$ and follow the differential equation

$$\partial_t \theta_p(t) = \left\langle \partial_{\theta_p} F^{(L)}(\theta(t)), d_t \right\rangle_{p^{in}}.$$

**Theorem 2.** *Assume that $\sigma$ is a Lipschitz twice differentiable nonlinearity function, with bounded second derivative. For any $T$ such that the integral $\int_0^T \|d_t\|_{p^{in}} dt$ stays stochastically bounded, as $n_1, \ldots, n_{L-1} \to \infty$ sequentially, we have, uniformly for $t \in [0,T]$,*

$$\Theta^{(L)}(t) \to \Theta_\infty^{(L)} \otimes Id_{n_L}.$$

*As a consequence, in this limit, the dynamics of $f_\theta$ is described by the differential equation*

$$\partial_t f_{\theta(t)} = \Phi_{\Theta_\infty^{(L)} \otimes Id_{n_L}} \left( \langle d_t, \cdot \rangle_{p^{in}} \right).$$

*Proof.* Let $\tilde{\theta}$ be the parameters of the smaller network, and let $\theta_p \in \tilde{\theta}$. Then

$$\begin{aligned}
\partial_{\theta_p} F^{(L+1)}(\theta) &= \partial_{\theta_p} \left( \frac{1}{\sqrt{n_L}} W^{(L)} \sigma(F^{(L)}(\tilde{\theta})) + \beta b^{(L)} \right) \\
&= \frac{1}{\sqrt{n_L}} W^{(L)} \dot{\sigma}(F^{(L)}(\tilde{\theta})) \partial_{\theta_p} F^{(L)}(\tilde{\theta})
\end{aligned}$$

and so

$$\begin{aligned}
\partial_t \theta_p(t) &= \left\langle \partial_{\theta_p} F^{(L)}(\theta(t)), d_t \right\rangle_{p^{in}} \\
&= \left\langle \frac{1}{\sqrt{n_L}} W^{(L)}(t) \dot{\sigma}(F^{(L)}(\tilde{\theta}(t))) \partial_{\theta_p} F^{(L)}(\tilde{\theta}(t)), d_t \right\rangle_{p^{in}} \\
&= \left\langle \partial_{\theta_p} F^{(L)}(\tilde{\theta}(t)), \dot{\sigma}(F^{(L)}(\tilde{\theta}(t))) \left( \frac{1}{\sqrt{n_L}} W^{(L)}(t) \right)^\top d_t \right\rangle
\end{aligned}$$

which implies that the smaller network follows the training direction

$$d'_t = \dot{\sigma}(F^{(L)}(\tilde{\theta}(t))) \left( \frac{1}{\sqrt{n_L}} W^{(L)}(t) \right)^\top d_t.$$

Since $\sigma$ is a $c$-Lipschitz function, $|\dot{\sigma}| \leq c$ and so

$$\begin{aligned}
\|d'_t\|_{p^{in}} &\leq |\dot{\sigma}(F^{(L)}(\tilde{\theta}(t)))| \left\| \frac{1}{\sqrt{n_L}} W^{(L)}(t) \right\|_{op} \|d_t\|_{p^{in}} \\
&\leq c \left\| \frac{1}{\sqrt{n_L}} W^{(L)}(t) \right\|_{op} \|d_t\|_{p^{in}}.
\end{aligned}$$

From the law of large numbers,

$$\left\| \frac{1}{\sqrt{n_L}} W_i^{(L)}(0) \right\|_2^2 = \frac{1}{n_L} \sum_{j=1}^{n_L} W_{ij}^2(0) \to \mathbb{E}[W_{ij}^2(0)] = 1$$

since $W_{ij}(0)$ are i.i.d. samples from $\mathcal{N}(0,1)$. Hence, $\|\frac{1}{\sqrt{n_L}}W^{(L)}(0)\|_{op}$ is bounded.

Observe that by the triangle inequality,

$$\partial_t \|f(t)\| = \lim_{h\to 0} \frac{\|f(t+h)\| - \|f(t)\|}{h} \leq \lim_{h\to 0} \frac{\|f(t+h) - f(t)\|}{h} = \|\partial_t f(t)\|$$

and so $\partial_t\|\cdot\| \leq \|\partial_t \cdot\|$. It follows that

$$\begin{aligned} \partial_t \left\| W_i^{(L)}(t) - W_i^{(L)}(0) \right\|_2 &\leq \left\| \partial_t \left( W_i^{(L)}(t) - W_i^{(L)}(0) \right) \right\|_2 \\ &= \left\| \partial_t W_i^{(L)}(t) \right\|_2 \\ &\leq \frac{1}{\sqrt{n_L}} \|\alpha_i^{(L)}(t)\|_{p^{in}} \|d_t\|_{p^{in}}. \end{aligned}$$

$$\partial_t \left( c \left\| \tilde{\alpha}_i^{(L)}(t) - \tilde{\alpha}_i^{(L)}(0) \right\|_{p^{in}} + \left\| W_i^{(L)}(t) - W_i^{(L)}(0) \right\|_2 \right) = \partial_t(A(t) - A(0)) = \partial_t A(t) = O\left( \frac{1}{\sqrt{n_L}} \right).$$

$$\partial_{W_{ij}^{(L)}} f_{\theta(t),j'}(x) \otimes \partial_{W_{ij}^{(L)}} f_{\theta(t),j''}(x') = \frac{1}{n_L} \alpha_i^{(L)}(x;\theta(t))^2 \delta_{jj'}\delta_{jj''}$$

and so

$$\partial_t \left( \partial_{W_{ij}^{(L)}} f_{\theta(t),j'}(x) \otimes \partial_{W_{ij}^{(L)}} f_{\theta(t),j''}(x') \right) = \frac{2}{n_L} \partial_t \alpha_i^{(L)}(x;\theta(t)) \delta_{jj'}\delta_{jj''}$$

and since $|\partial_t \alpha_i^{(L)}| = O(\frac{1}{\sqrt{n_L}})$, we see that the summands vary at the rate $n_L^{-3/2}$. Since the dimension of $W^{(L)}$ is $n_L \times n_{L+1}$ (recall that $n_{L+1}$ is fixed), the sum induces a variation of the NTK of rate $\frac{1}{\sqrt{n_L}}$. $\square$

### A.3 A Priori Control During Training

### A.4 Positive-Definiteness of $\Theta_\infty^{(L)}$