Deep Learning - Parameters and Functions

NTK via a Power Series Expansion

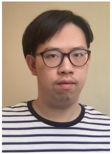
Guido Montúfar
montufar@math.ucla.edu

48th Winter Conference in Statistics, March 2024, Hemavan

UCLA
Max Planck Institute for Mathematics in the Sciences
erc European Research Council
NSF
DFG Deutsche Forschungsgemeinschaft German Research Foundation

Michael Murray

Hui Jin



Benjamin Bowman

- "Characterizing the spectrum of the NTK via a power series expansion"

Overview

- Power series expansion for the Neural Tangent Kernel of arbitrarily deep feedforward networks in the infinite width limit.
- Express coefficients of the power series depending on Hermite coefficients of activation function and depth of the network.

Overview

- Power series expansion for the Neural Tangent Kernel of arbitrarily deep feedforward networks in the infinite width limit.
- Express coefficients of the power series depending on Hermite coefficients of activation function and depth of the network.

Faster decay of Hermite coefficients	$\Rightarrow$	Faster decay of NTK coefficients
Effective rank of **NTK**	$\Leftrightarrow$	Effective rank of **Data Gram**

- Power series expansion for the Neural Tangent Kernel of arbitrarily deep feedforward networks in the infinite width limit.
- Express coefficients of the power series depending on Hermite coefficients of activation function and depth of the network.

Faster decay of Hermite coefficients	$\Rightarrow$	Faster decay of NTK coefficients
Effective rank of **NTK**	$\Leftrightarrow$	Effective rank of **Data Gram**

- Data uniform on sphere: NTK eigenvalues; impact of activation.
- Generic data: asymptotic upper bound on the NTK spectrum.

- Loss landscape of neural networks is high-dimensional, non-convex, non-smooth, ...

[1] Neyshabur, Tomioka, and Srebro 2015.

[2] Lee et al. 2018; Neal 1996.

[3] Jacot, Gabriel, and Hongler 2018.

- Loss landscape of neural networks is high-dimensional, non-convex, non-smooth, ...
- Overparametrized networks work well empirically both in the sense of parameter optimization and statistical generalization.

[1] Neyshabur, Tomioka, and Srebro 2015.

[2] Lee et al. 2018; Neal 1996.

[3] Jacot, Gabriel, and Hongler 2018.

- Loss landscape of neural networks is high-dimensional, non-convex, non-smooth, . . .
- Overparametrized networks work well empirically both in the sense of parameter optimization and statistical generalization.
- Some form of capacity control other than nr of parameters[1].

[1] Neyshabur, Tomioka, and Srebro 2015.

[2] Lee et al. 2018; Neal 1996.

[3] Jacot, Gabriel, and Hongler 2018.

- Loss landscape of neural networks is high-dimensional, non-convex, non-smooth, ...
- Overparametrized networks work well empirically both in the sense of parameter optimization and statistical generalization.
- Some form of capacity control other than nr of parameters[1].
- Infinitely wide networks correspond to Gaussian processes[2].

---

[1] Neyshabur, Tomioka, and Srebro 2015.

[2] Lee et al. 2018; Neal 1996.

[3] Jacot, Gabriel, and Hongler 2018.

- Loss landscape of neural networks is high-dimensional, non-convex, non-smooth, ...
- Overparametrized networks work well empirically both in the sense of parameter optimization and statistical generalization.
- Some form of capacity control other than nr of parameters[1].
- Infinitely wide networks correspond to Gaussian processes[2].
- Training behavior can also be described by a kernel, the NTK[3]. In the infinite-width limit, the NTK becomes deterministic at initialization and stays constant during training.

[1] Neyshabur, Tomioka, and Srebro 2015.

[2] Lee et al. 2018; Neal 1996.

[3] Jacot, Gabriel, and Hongler 2018.

- Loss landscape of neural networks is high-dimensional, non-convex, non-smooth, ...
- Overparametrized networks work well empirically both in the sense of parameter optimization and statistical generalization.
- Some form of capacity control other than nr of parameters[1].
- Infinitely wide networks correspond to Gaussian processes[2].
- Training behavior can also be described by a kernel, the NTK[3]. In the infinite-width limit, the NTK becomes deterministic at initialization and stays constant during training.
- The NTK is a tool that allows one to abstract away complexities of the parameter space.

[1] Neyshabur, Tomioka, and Srebro 2015.

[2] Lee et al. 2018; Neal 1996.

[3] Jacot, Gabriel, and Hongler 2018.

- Data, model, loss:

$$\{(\mathbf{x}_i, y_i)\}, \quad \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

$$f(\mathbf{x};\theta) = \theta^\top \mathbf{x}$$

$$L(\theta) = \frac{1}{2}\sum_{i=1}^{n}(f(\mathbf{x}_i;\theta) - y_i)^2$$

- Data, model, loss:

$$\{(\mathbf{x}_i, y_i)\}, \quad \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

$$f(\mathbf{x};\theta) = \theta^\top \mathbf{x}$$

$$L(\theta) = \frac{1}{2}\sum_{i=1}^{n}(f(\mathbf{x}_i;\theta) - y_i)^2$$

- Gradient descent:

$$\begin{aligned}\theta_{t+1} =& \theta_t - \alpha_t \nabla L(\theta_t)\\ =& \theta_t - \alpha_t \sum (f(\mathbf{x}_i;\theta) - y_i) \underbrace{\nabla f(\mathbf{x}_i;\theta_t)}_{\substack{\mathbf{x}_i \\ \text{indep. of } \theta_t}}\end{aligned}$$

For sufficiently small α_t, convergence to global optimum.

Linear functions of $\mathbf{x}$ are too restrictive, consider instead

$$\mathbf{x} \in \mathbb{R}^d \quad \longrightarrow \quad \phi(\mathbf{x}) \in \mathbb{R}^D, \quad d \ll D$$

Example:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \longrightarrow \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_1 x_2 \\ x_1 x_3 \\ \vdots \end{bmatrix}$$

Model, loss:

$$f(\mathbf{x}; \theta) = \theta^T \phi(\mathbf{x}) \qquad \text{is still linear in } \theta$$

$$\frac{1}{2} \sum_{i=1}^{n} (f(\mathbf{x}_i; \theta) - y_i)^2 \qquad \text{is still convex in } \theta$$

Kernel trick:

- In many cases we only need inner products $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$
- These may be expressible in terms of a kernel function

$$K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

  that can be computed without explicit computation of $\phi(\mathbf{x})$
- For example, polynomial kernel

$$K(\mathbf{x}, \mathbf{x}') = (c + \mathbf{x}^T \mathbf{x}')^k = \phi(\mathbf{x})^T \phi(\mathbf{x}'),$$

  feature vector $\phi(\mathbf{x})$ consists of monomials of degree $\leq k$, but product is computed in d rather than $D = \binom{d+k}{k}$

- Simple example:

$$f(\mathbf{x};\theta) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \sigma(\langle w_j, \mathbf{x} \rangle) \qquad \theta = (w_j, v_j)_{j=1}^{m}$$

$$L(\theta) = \frac{1}{2} \sum_{i=1}^{n} (f(\mathbf{x}_i;\theta) - y_i)^2$$

- Gradient descent:

$$\theta_{t+1} = \theta_t - \alpha_t \sum_{i=1}^{n} (f(\mathbf{x}_i;\theta_t) - y_i) \underbrace{\nabla f(\mathbf{x}_i;\theta_t)}_{\text{not indep. of } \theta_t}$$

- In some cases we observe "lazy training", whereby parameters remain nearly constant in t.

- So, consider 1st order Taylor expansion around θ_0:

$$f(\mathbf{x};\theta) \approx f(\mathbf{x};\theta_0) + \nabla f(\mathbf{x};\theta_0)^T(\theta - \theta_0)$$

  This is not linear in $\mathbf{x}$ but is linear in θ.
- Similar to a kernel method with feature map $\phi(\mathbf{x}) = \nabla f(\mathbf{x};\theta_0)$.
- The corresponding kernel takes the form

$$K(\mathbf{x},\mathbf{x}') = \langle \nabla f(\mathbf{x};\theta_0), \nabla f(\mathbf{x}';\theta_0)\rangle$$

- For our example $f(\mathbf{x};\theta) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \sigma(\langle w_j, \mathbf{x} \rangle)$, the feature map takes the form:

$$\nabla_{w_j} f(\mathbf{x};\theta) = \frac{1}{\sqrt{m}} v_j \sigma'(\langle w_j, \mathbf{x} \rangle)\mathbf{x}$$

$$\nabla_{v_j} f(\mathbf{x};\theta) = \frac{1}{\sqrt{m}} \sigma(\langle w_j, \mathbf{x} \rangle)$$

- The kernel takes the form:

$$K(\mathbf{x}, \mathbf{x}') = K_v(\mathbf{x}, \mathbf{x}') + K_w(\mathbf{x}, \mathbf{x}')$$

$$K_w(\mathbf{x}, \mathbf{x}') = \frac{1}{m} \sum_{j=1}^{m} v_j^2 \sigma'(\langle w_j, \mathbf{x} \rangle)\sigma'(\langle w_j, \mathbf{x} \rangle)\langle \mathbf{x}, \mathbf{x}' \rangle$$

$$K_v(\mathbf{x}, \mathbf{x}') = \frac{1}{m} \sum_{j=1}^{m} \sigma(\langle w_j, \mathbf{x} \rangle)\sigma(\langle w_j, \mathbf{x}' \rangle)$$

This may be regarded as a sample mean!

- Then for an infinitely wide network, by law of large numbers, convergence to the expectation:

$$K_w(\mathbf{x},\mathbf{x}') \xrightarrow{m\to\infty} \mathbb{E}[v^2\sigma'(\langle w,\mathbf{x}\rangle)\sigma'(\langle w,\mathbf{x}\rangle)\langle\mathbf{x},\mathbf{x}'\rangle]$$
$$K_v(\mathbf{x},\mathbf{x}') \xrightarrow{m\to\infty} \mathbb{E}[\sigma(\langle w,\mathbf{x}\rangle)\sigma(\langle w,\mathbf{x}'\rangle)]$$

- For example if σ ReLU and $w_j \sim$ rotation invariant distribution:

$$K_w(\mathbf{x},\mathbf{x}') = \frac{1}{2\pi}\langle\mathbf{x},\mathbf{x}'\rangle\mathbb{E}[v^2](\pi-\vartheta)$$
$$K_v(\mathbf{x},\mathbf{x}') = \frac{\|\mathbf{x}\|\|\mathbf{x}'\|\mathbb{E}[\|w\|^2]}{2\pi d}((\pi-\vartheta)\cos(\vartheta)+\sin(\vartheta))$$

Here ϑ is the angle between $\mathbf{x}$ and $\mathbf{x}'$

Gradient dynamics

- Consider the gradient flow:

$$\tfrac{d}{dt}\theta_t = -\nabla L(\theta_t)$$

- Squared error loss:

$$\begin{aligned} L(\theta) =& \tfrac{1}{2}\|\hat{y} - y\|^2, \quad \hat{y}, y \in \mathbb{R}^n \\ \nabla L(\theta) =& \nabla \hat{y}\,(\hat{y} - y) \end{aligned}$$

- Dynamics of the parameters:

$$\tfrac{d}{dt}\theta_t = -\nabla \hat{y}\,(\hat{y} - y)$$

- Dynamics of the predictions $\hat{y}$:

$$\begin{aligned} \frac{d\hat{y}}{dt} = \frac{d\hat{y}}{d\theta}\frac{d\theta}{dt} =& \ \ \nabla \hat{y}^T \tfrac{d}{dt}\theta \\ =& -\underbrace{\nabla \hat{y}^T \nabla \hat{y}}_{K}(\hat{y} - y) \end{aligned}$$

- If K is approximately constant, then for the residual $r = \hat{y} - y$:

$$\begin{aligned} \tfrac{d}{dt} r \approx & - K_{\theta_0} \cdot r \\ r_t = & r_0 \, e^{-K_{\theta_0} t} \end{aligned}$$

- If K is positive definite, $K_{\theta_0} > 0$, then linear convergence to 0 with rate determined by the least eigenvalue.
- Moreover, spectrum and eigenfunctions,

$$K_{\theta_0} = \sum_{i=1}^{n} \lambda_i \xi_i \xi_i^\top, \quad \lambda_n \geq \cdots \geq \lambda_1 > 0,$$

give convergence of r along different components.

- If K is approximately constant, then for the residual $r = \hat{y} - y$:

$$\begin{aligned} \tfrac{d}{dt} r \approx& - K_{\theta_0} \cdot r \\ r_t =& r_0 \, e^{-K_{\theta_0} t} \end{aligned}$$

- If K is positive definite, $K_{\theta_0} > 0$, then linear convergence to 0 with rate determined by the least eigenvalue.
- Moreover, spectrum and eigenfunctions,

$$K_{\theta_0} = \sum_{i=1}^{n} \lambda_i \xi_i \xi_i^\top, \quad \lambda_n \geq \cdots \geq \lambda_1 > 0,$$

give convergence of r along different components.

- Thus, interest in stability, least eigenvalue and spectrum!

Parameters vs functions

Gradient descent for $\theta \mapsto f_\theta \mapsto \ell(f_\theta) = L(\theta)$

Parameters vs functions

Gradient descent for $\theta \mapsto f_\theta \mapsto \ell(f_\theta) = L(\theta)$

- Parameter $\frac{d}{dt}\theta = -\nabla_\theta L = -J^T \nabla_f \ell(f)$ (Jacobian $J = \nabla_\theta f^T$)
- Prediction $\frac{d}{dt} f = -J \frac{d}{dt}\theta = -JJ^T \nabla_f \ell(f)$
- Loss $\frac{d}{dt} L = -\nabla_f \ell(f)^T JJ^T \nabla_f \ell(f)$

Parameters vs functions

Gradient descent for $\theta \mapsto f_\theta \mapsto \ell(f_\theta) = L(\theta)$

- Parameter $\frac{d}{dt}\theta = -\nabla_\theta L = -J^T \nabla_f \ell(f)$ (Jacobian $J = \nabla_\theta f^T$)
- Prediction $\frac{d}{dt} f = -J \frac{d}{dt}\theta = -JJ^T \nabla_f \ell(f)$
- Loss $\frac{d}{dt} L = -\nabla_f \ell(f)^T JJ^T \nabla_f \ell(f)$

In turn, for $\nabla_f \ell(f) = f - y = r$ and $JJ^T = \sum_i \lambda_i \xi_i \xi_i^T$

- If $\lambda_i \geq \epsilon$, eventually $r = 0$
- The ith component of r drops at rate λ_i

- Fully-connected network with L hidden layers and linear output.
- For a given input $\mathbf{x} \in \mathbb{R}^d$,

preactivation | activation

$$
\begin{aligned}
g^{(1)}(\mathbf{x}) &= \gamma_w \mathbf{W}^{(1)}\mathbf{x} + \gamma_b \mathbf{b}^{(1)}, & f^{(1)}(\mathbf{x}) &= \phi\left(g^{(1)}(\mathbf{x})\right), \\
g^{(l)}(\mathbf{x}) &= \tfrac{\sigma_w}{\sqrt{m_{l-1}}}\mathbf{W}^{(l)} f^{(l-1)}(\mathbf{x}) + \sigma_b \mathbf{b}^{(l)}, & f^{(l)}(\mathbf{x}) &= \phi\left(g^{(l)}(\mathbf{x})\right), \\
g^{(L+1)}(\mathbf{x}) &= \tfrac{\sigma_w}{\sqrt{m_L}}\mathbf{W}^{(L+1)} f^{(L)}(\mathbf{x}), & f^{(L+1)}(\mathbf{x}) &= g^{(L+1)}(\mathbf{x}).
\end{aligned}
\tag{1}
$$

- Weight matrices $\mathbf{W}^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$, bias vectors $\mathbf{b}^{(l)} \in \mathbb{R}^{m_l}$, parameters up to lth layer $\theta_l = (\mathbf{W}^{(h)}, \mathbf{b}^{(h)})_{h=1}^{l} \in \mathbb{R}^p$.
- Activation function $\phi \colon \mathbb{R} \to \mathbb{R}$ applied elementwise.
- Hyperparameters $\gamma_w, \sigma_w \in \mathbb{R}_{>0}$, $\gamma_b, \sigma_b \in \mathbb{R}_{\geq 0}$.

Assumption 1

1. At initialization all network parameters are iid $\mathcal{N}(0, 1)$.
2. Activation fct $\phi \in L^2(\mathbb{R}, \gamma)$ differentiable a.e., $\phi' \in L^2(\mathbb{R}, \gamma)$.
3. Widths sent to infinity in sequence, $m_1 \to \infty, \ldots, m_L \to \infty$.

- We denote by $L^2(\mathbb{R}, \gamma)$ the space of functions $\phi \colon \mathbb{R} \to \mathbb{R}$ with

$$\mathbb{E}_{X \sim \mathcal{N}(0,1)}[\phi(X)^2] < \infty.$$

- Item 2 is satisfied for ReLU, Tanh, Softplus,...

- The **NTK** of $f^{(l)}$ at layer $l \in [L+1]$ is $\tilde{\Theta}^{(l)}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$\tilde{\Theta}^{(l)}(\mathbf{x}, \mathbf{y}) := \langle \nabla_{\theta_l} f^{(l)}(\mathbf{x}), \nabla_{\theta_l} f^{(l)}(\mathbf{y}) \rangle \tag{2}$$

[4] Jacot, Gabriel, and Hongler 2018.

[5] Arora et al. 2019; Lee et al. 2019; Woodworth et al. 2020.

- The **NTK** of $f^{(l)}$ at layer $l \in [L+1]$ is $\tilde{\Theta}^{(l)} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$\tilde{\Theta}^{(l)}(\mathbf{x}, \mathbf{y}) := \langle \nabla_{\theta_l} f^{(l)}(\mathbf{x}), \nabla_{\theta_l} f^{(l)}(\mathbf{y}) \rangle \tag{2}$$

- Under Assumption 1, for any $l \in [L+1]$,
  - $\tilde{\Theta}^{(l)}$ converges in probability to a deterministic $\Theta^{(l)}$[4].
  - Network behaves like kernelized linear predictor during training[5].

[4] Jacot, Gabriel, and Hongler 2018.

[5] Arora et al. 2019; Lee et al. 2019; Woodworth et al. 2020.

- The **NTK** of $f^{(l)}$ at layer $l \in [L+1]$ is $\tilde{\Theta}^{(l)} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$\tilde{\Theta}^{(l)}(\mathbf{x}, \mathbf{y}) := \langle \nabla_{\theta_l} f^{(l)}(\mathbf{x}), \nabla_{\theta_l} f^{(l)}(\mathbf{y}) \rangle \tag{2}$$

- Under Assumption 1, for any $l \in [L+1]$,
  - $\tilde{\Theta}^{(l)}$ converges in probability to a deterministic $\Theta^{(l)}$[4].
  - Network behaves like kernelized linear predictor during training[5].

- The (infinite width limit) **NTK matrix** at layer $l \in [L+1]$ for data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$ is

$$[\mathbf{K}_l]_{ij} = \frac{1}{n} \Theta^{(l)}(\mathbf{x}_i, \mathbf{x}_j), \quad \forall (i,j) \in [n] \times [n]. \tag{3}$$

[4] Jacot, Gabriel, and Hongler 2018.

[5] Arora et al. 2019; Lee et al. 2019; Woodworth et al. 2020.

Assumption 2

1. The hyperparameters of the network satisfy

$$\gamma_w^2+\gamma_b^2=1,\ \sigma_w^2\mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi(Z)^2]\leq 1,\ \sigma_b^2=1-\sigma_w^2\mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi(Z)^2].$$

2. The data is normalized so that $\|\mathbf{x}_i\|=1$ for all $i\in[n]$.

- This ensures the preactivation of each neuron has unit variance, reminiscent of init to avoid vanishing/exploding gradients.
- Under Assumption 2 we write the NTK as an analytic power series on $[-1,1]$ and derive expressions for the coefficients.

Hermite expansion

- The normalized probabilist's Hermite polynomials are defined as

$$h_k(x) = \frac{(-1)^k e^{x^2/2}}{\sqrt{k!}} \frac{d^k}{dx^k} e^{-x^2/2}, \quad k = 0, 1, \dots$$

  These form a complete orthonormal basis in $L^2(\mathbb{R}, \gamma)$[6].
- The Hermite expansion of a function $\phi \in L^2(\mathbb{R}, \gamma)$ is given by

$$\phi(x) = \sum_{k=0}^{\infty} \mu_k(\phi) h_k(x),$$

  with Hermite coefficients

$$\mu_k(\phi) = \mathbb{E}_{X \sim \mathcal{N}(0,1)}[\phi(X) h_k(X)].$$

[6] O'Donnell 2014.

Notations

- Denote the Hadamard (entrywise) product by $\mathbf{X} \odot \mathbf{Y}$ and

$$\mathbf{X}^{\odot p} = \mathbf{X} \odot \mathbf{X} \odot \cdots \odot \mathbf{X}.$$

- Given a Hermitian or symmetric matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$, we adopt the convention that $\lambda_i(\mathbf{X})$ denotes the ith largest eigenvalue,

$$\lambda_1(\mathbf{X}) \geq \lambda_2(\mathbf{X}) \geq \cdots \geq \lambda_n(\mathbf{X}).$$

- For a square matrix we let $Tr(\mathbf{X}) = \sum_{i=1}^{n} [\mathbf{X}]_{ii}$ denote the trace.
- $\mathbf{X}\mathbf{X}^T$ is the Gram matrix of the input data.

Theorem 1

Under Assumptions 1 and 2, for all $l \in [L+1]$

$$n\mathbf{K}_l = \sum_{p=0}^{\infty} \kappa_{p,l} \left(\mathbf{X}\mathbf{X}^T\right)^{\odot p}. \tag{4}$$

The series for each entry $n[\mathbf{K}_l]_{ij}$ converges absolutely.

The $\kappa_{p,l}$ are nonnegative and are expressed by following recurrence relation depending on the Hermite coefficients $\mu_p(\phi)$ and $\mu_p(\phi')$.

The coefficients of the power series (4) are given by

$$\kappa_{p,l} = \begin{cases} \delta_{p=0}\gamma_b^2 + \delta_{p=1}\gamma_w^2, & l = 1, \\ \alpha_{p,l} + \sum_{q=0}^{p} \kappa_{q,l-1}\upsilon_{p-q,l}, & l \in [2, L+1], \end{cases} \tag{5}$$

where

$$\alpha_{p,l} = \begin{cases} \sigma_w^2\mu_p^2(\phi) + \delta_{p=0}\sigma_b^2, & l = 2, \\ \sum_{k=0}^{\infty} \alpha_{k,2} F(p, k, \bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{6}$$

and

$$\upsilon_{p,l} = \begin{cases} \sigma_w^2\mu_p^2(\phi'), & l = 2, \\ \sum_{k=0}^{\infty} \upsilon_{k,2} F(p, k, \bar{\alpha}_{l-1}), & l \geq 3, \end{cases} \tag{7}$$

are likewise nonnegative for all $p \in \mathbb{Z}_{\geq 0}$ and $l \in [2, L+1]$, where for a sequence of reals $\bar{a} = (a_j)_{j=0}^{\infty}$ and any $p, k \in \mathbb{Z}_{\geq 0}$,

- set of k-tuples of nonnegative integers which sum to p

$$\mathcal{J}(p,k) = \Big\{(j_i)_{i\in[k]} \colon j_i \geq 0\ \forall i \in [k],\ \sum_{i=1}^{k} j_i = p\Big\}, \quad \forall p \in \mathbb{Z}_{\geq 0}, k \in \mathbb{N}$$

- sum of ordered products of k-tuples of $\bar{a}$ whose indices sum to p

$$F(p,k,\bar{a}) = \begin{cases} 1, & k = 0 \text{ and } p = 0, \\ 0, & k = 0 \text{ and } p \geq 1, \\ \sum_{(j_i)\in\mathcal{J}(p,k)} \prod_{i=1}^{k} a_{j_i}, & k \geq 1 \text{ and } p \geq 0. \end{cases} \tag{8}$$

Assumption 3

The activation fct $\phi\colon \mathbb{R} \to \mathbb{R}$ is absolutely continuous, differentiable a.e., poly bounded, $|\phi(x)| = \mathcal{O}(|x|^{\beta})$.

- This is satisfied by ReLU, Tanh, Sigmoid, Softplus and has minimal impact on the generality of our results.
- Under Assumption 3, $\upsilon_{p,2} = (p+1)\alpha_{p+1,2}$ and thus to compute $\kappa_{p,l}$ we do not need the Hermite coefficients of ϕ'.

Activation and NTK coefficients

To better understand the relationship between

Hermite coefficients and **NTK coefficients**

consider first the simple two-layer case, i.e., with $L = 1$,

$$\kappa_{p,2} = \sigma_w^2(1 + \gamma_w^2 p)\mu_p^2(\phi) + \sigma_w^2\gamma_b^2(1 + p)\mu_{p+1}^2(\phi) + \delta_{p=0}\sigma_b^2.$$

First few NTK coefficients

Table 1: Percentage of $\sum_{p=0}^{\infty} \kappa_{p,2}$ accounted for by the first $T+1$ NTK coefficients assuming $\gamma_w^2 = 1$, $\gamma_b^2 = 0$, $\sigma_w^2 = 1$ and $\sigma_b^2 = 1 - \mathbb{E}[\phi(Z)^2]$.

$T =$	0	1	2	3	4	5
ReLU	43.944	77.277	93.192	93.192	95.403	95.403
Tanh	41.362	91.468	91.468	97.487	97.487	99.090
Sigmoid	91.557	99.729	99.729	99.977	99.977	99.997
Gaussian	95.834	95.834	98.729	98.729	99.634	99.634

- Across activation functions, the first few coefficients account for the large majority of the total NTK coefficient series.

Asymptotic rate of decay of NTK coefficients

Lemma 2

Under Assumptions 1 and 2,

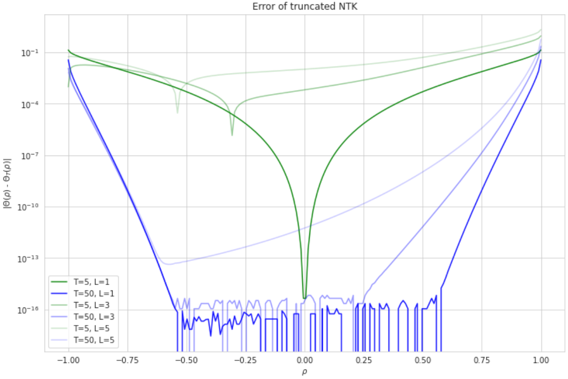
1. *if* $\phi(z) = ReLU(z)$, *then* $\kappa_{p,2} = \delta_{(\gamma_b > 0) \cup (p \text{ even})} \Theta(p^{-3/2})$,
2. *if* $\phi(z) = Tanh(z)$, *then* $\kappa_{p,2} = \mathcal{O}\left(\exp\left(-\frac{\pi\sqrt{p-1}}{2}\right)\right)$,
3. *if* $\phi(z) = \omega_\sigma(z)$, *then* $\kappa_{p,2} = \delta_{(\gamma_b > 0) \cup (p \text{ even})} \Theta(p^{1/2}(\sigma^2 + 1)^{-p})$.

Here $\omega_\sigma(z) = (1/\sqrt{2\pi\sigma^2}) \exp\left(-z^2/(2\sigma^2)\right)$

- The asymptotic rate of decay of the NTK coefficients varies significantly by activation function.

NTK approximation by truncated series

- Currently computing $\Theta^{(l)}$ requires either explicit evaluation of Gaussian integrals, or approximation, or wide networks.
- We may also use a truncated power series.



$T = 5$

$T = 50$

Figure 1: Absolute error between the analytical ReLU NTK and its truncated power series, where $\rho = \mathbf{x}^T\mathbf{y}$, truncation point T, and depth L.

Effective rank

Effective rank of the NTK

- For convenience drop subscr l and let $n\mathbf{K} = \sum_{p=0}^{\infty} c_p(\mathbf{X}\mathbf{X}^T)^{\odot p}$.
- We study the effective rank of the kernel $\mathbf{K}$, which is defined as

$$\text{eff}(\mathbf{K}) := \frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})}.$$

Effective rank of the NTK

- For convenience drop subscr l and let $n\mathbf{K} = \sum_{p=0}^{\infty} c_p (\mathbf{X}\mathbf{X}^T)^{\odot p}$.
- We study the effective rank of the kernel $\mathbf{K}$, which is defined as

$$\mathrm{eff}(\mathbf{K}) := \frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})}.$$

Theorem 3

For general activations ($\mu_0(\phi) \neq 0$) or nonzero bias networks, $c_0 \neq 0$, the effective rank is $O(1)$ as

$$\frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})} \leq \frac{\sum_{i=0}^{\infty} c_i}{c_0}.$$

For ReLU in Table 1, approx 2.3.

Effective rank of the NTK

- For convenience drop subscr l and let $n\mathbf{K} = \sum_{p=0}^{\infty} c_p(\mathbf{X}\mathbf{X}^T)^{\odot p}$.
- We study the effective rank of the kernel $\mathbf{K}$, which is defined as

$$\mathrm{eff}(\mathbf{K}) := \frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})}.$$

Theorem 3

For general activations ($\mu_0(\phi) \neq 0$) or nonzero bias networks, $c_0 \neq 0$, the effective rank is $O(1)$ as

$$\frac{Tr(\mathbf{K})}{\lambda_1(\mathbf{K})} \leq \frac{\sum_{i=0}^{\infty} c_i}{c_0}.$$

For ReLU in Table 1, approx 2.3.

Corollary 4

The largest eigenvalue $\lambda_1(\mathbf{K})$ takes up $\Omega(1)$ fraction of the trace and there are $O(1)$ eigenvalues on the order of $\lambda_1(\mathbf{K})$.

Here O and Ω is wrt n. By contrast, well-conditioned matrix has rank $\Omega(n)$.

- To understand the rest of the spectrum, we analyze the centered kernel $\tilde{\mathbf{K}} := \mathbf{K} - c_0 \mathbf{1}_{n\times n}$.

Theorem 5

The effective rank of the centered kernel $\tilde{\mathbf{K}}$ *is upper bounded by the effective rank of the data Gram* $\mathbf{X}\mathbf{X}^T$

$$\mathrm{eff}(\tilde{\mathbf{K}}) \leq \mathrm{eff}(\mathbf{X}\mathbf{X}^T) \frac{\sum_{p=1}^{\infty} c_p}{c_1}.$$

For ReLU in Table 1, approx 1.7.

Effective rank of the NTK

- To understand the rest of the spectrum, we analyze the centered kernel $\tilde{\mathbf{K}} := \mathbf{K} - c_0 \mathbf{1}_{n\times n}$.

Theorem 5

The effective rank of the centered kernel $\tilde{\mathbf{K}}$ is upper bounded by the effective rank of the data Gram $\mathbf{X}\mathbf{X}^T$

$$\mathrm{eff}(\tilde{\mathbf{K}}) \leq \mathrm{eff}(\mathbf{X}\mathbf{X}^T) \frac{\sum_{p=1}^{\infty} c_p}{c_1}.$$

For ReLU in Table 1, approx 1.7.

Corollary 6

Whenever the input data matrix $\mathbf{X}\mathbf{X}^T$ is approx low rank, $\tilde{\mathbf{K}}$ is also approx low rank. Since real-world data tends to be low-rank, the NTK also tends to be low-rank!

Effective rank of the NTK

- To understand the rest of the spectrum, we analyze the centered kernel $\tilde{\mathbf{K}} := \mathbf{K} - c_0 \mathbf{1}_{n \times n}$.

Theorem 5

The effective rank of the centered kernel $\tilde{\mathbf{K}}$ is upper bounded by the effective rank of the data Gram $\mathbf{X}\mathbf{X}^T$

$$\operatorname{eff}(\tilde{\mathbf{K}}) \leq \operatorname{eff}(\mathbf{X}\mathbf{X}^T) \frac{\sum_{p=1}^{\infty} c_p}{c_1}.$$

For ReLU in Table 1, approx 1.7.

Corollary 6

Whenever the input data matrix $\mathbf{X}\mathbf{X}^T$ is approx low rank, $\tilde{\mathbf{K}}$ is also approx low rank. Since real-world data tends to be low-rank, the NTK also tends to be low-rank!

Theorem 7

Also holds for finite-width shallow ReLU networks.

NTK spectrum mimics input data spectrum

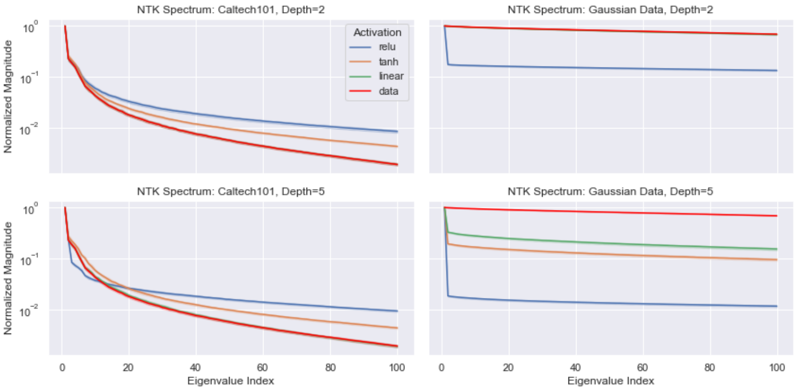


Figure 2: λ_p/λ_1 of NTK matrix $\mathbf{K}$ and data Gram $\mathbf{X}\mathbf{X}^T$.

Width 500, Kaiming uniform init, $n = 200$, mean across 10 trials and 95%

Asymptotic decay

Asymptotic decay of the spectrum

- For a dot product kernel with data uniform on a sphere, the eigenfunctions are the spherical harmonics[7].
- For a kernel function of the form $K(\mathbf{x}, \mathbf{y}) = \sum_{p=0}^{\infty} c_p \langle \mathbf{x}, \mathbf{y} \rangle^p$, Azevedo and Menegatto 2015 gave the eigenvals in terms of c_p.
- Given a specific decay rate for the coefficients c_p one may derive the decay rate of λ_k.

[7] Basri et al. 2019; Bietti and Mairal 2019.

Asymptotic decay of the spectrum

For the **uniform distribution** on the sphere $\mathbb{S}^d$, the decay of the power series coefficients determines the decay of the spectrum.

- Let $\overline{\lambda_k}$ be the eigenvalue for frequency-k spherical harmonic.
- (ReLU) if $c_p = \Theta(p^{-a})$ with $a \geq 1$, then

$$\overline{\lambda_k} = \Theta(k^{-d-2a+2}),$$

- (Tanh) if $c_p = O\left(\exp\left(-a\sqrt{p}\right)\right)$, then

$$\overline{\lambda_k} = O\left(k^{-d+1/2} \exp\left(-a\sqrt{k}\right)\right),$$

- (Gaussian) if $c_p = \Theta(p^{1/2} a^{-p})$, then

$$\overline{\lambda_k} = O\left(k^{-d+1} a^{-k}\right) \quad \text{and} \quad \overline{\lambda_k} = \Omega\left(k^{-d/2+1} 2^{-k} a^{-k}\right).$$

Recovers ReLU Basri et al. 2019; Bietti and Bach 2021; Geifman et al. 2020; Velikanov and Yarotsky 2021, gives rates for shallow Tanh and Gaussian

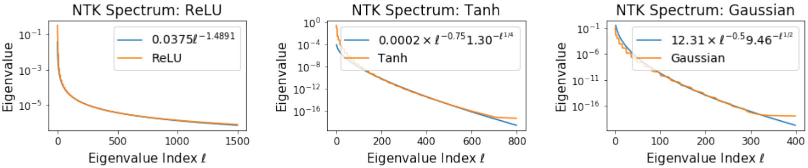


Figure 3: NTK spectrum of two-layer fully connected networks with ReLU, Tanh and Gaussian activations under the NTK parameterization.

Asymptotic decay of the spectrum

Results similar in spirit, albeit weaker, hold for **any data** on $\mathbb{S}^d$.

- Let $r(n)$ denote the rank of the data matrix.
- If $c_p = O(p^{-a})$ with $a > r(n) + 1$, then

$$\lambda_n = O\left(n^{-\frac{a-1}{r(n)}}\right),$$

- if $c_p = O(e^{-a\sqrt{p}})$, then, for any $a' < a2^{-1/2r(n)}$,

$$\lambda_n = O\left(n^{\frac{1}{2r(n)}} \exp\left(-a' n^{\frac{1}{2r(n)}}\right)\right),$$

- if $c_p = O(e^{-ap})$ then, for any $a' < a2^{-1/2r(n)}$,

$$\lambda_n = O\left(\exp\left(-a' n^{\frac{1}{r(n)}}\right)\right).$$

Conclusions

- A simple power series analysis can be used to characterize both outlier eigenvalues and asymptotic decay of the NTK spectrum.
- The NTK has a large outlier eigenvalue and $O(1)$ eigenvalues on the same order of magnitude as the largest eigenvalue.
- If the input data matrix is low rank, the NTK is also low rank.
- The asymptotic decay of the power series coefficients determines the asymptotic decay of the spectrum.
- The decay of these coefficients are in turn driven by the Hermite coefficients of the activation function and the network depth.

Further reading

- A brief intro to the NTK (Ben Bowman)
- Implicit bias of gradient descent for MSE with wide shallow ReLU nets (with Hui Jin)
- Spectral bias outside the training set for deep nets in the kernel regime (with Ben Bowman)
- Math Machine Learning seminar MPI MiS + UCLA `https://www.mis.mpg.de/events/series/math-machine-learning-seminar-mpi-mis-ucla`

Arora, Sanjeev et al. (2019). “On Exact Computation with an Infinitely Wide Neural Net”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/dbc4d84bfcfe2284ba11beffb853a8c4-Paper.pdf.

Azevedo, Douglas and Valdir A Menegatto (2015). “Eigenvalues of dot-product kernels on the sphere”. In: *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics* 3.1.

Basri, Ronen et al. (2019). “The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies”. In: *Advances in Neural Information Processing Systems 32*. Ed. by Hanna M. Wallach et al., pp. 4763–4772. URL: https://proceedings.neurips.cc/paper/2019/hash/5ac8bb8a7d745102a978c5f8ccdb61b8-Abstract.html.

- Bietti, Alberto and Francis Bach (2021). "Deep Equals Shallow for ReLU Networks in Kernel Regimes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=aDjoksTpXOP.
- Bietti, Alberto and Julien Mairal (2019). "On the Inductive Bias of Neural Tangent Kernels". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/c4ef9c39b300931b69a36fb3dbb8d60e-Paper.pdf.
- Geifman, Amnon et al. (2020). "On the Similarity between the Laplace and Neural Tangent Kernels". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1451–1461. URL: https://proceedings.neurips.cc/paper/2020/file/1006ff12c465532f8c574aeaa4461b16-Paper.pdf.

References III

Jacot, Arthur, Franck Gabriel, and Clement Hongler (2018). "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Lee, Jaehoon et al. (2018). "Deep Neural Networks as Gaussian Processes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=B1EA-M-0Z.

Lee, Jaehoon et al. (2019). "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

Murray, Michael et al. (2023). "Characterizing the spectrum of the NTK via a power series expansion". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Tvms8xrZHyR.

Neal, Radford M. (1996). *Bayesian Learning for Neural Networks*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387947248.

Neyshabur, Behnam, Ryota Tomioka, and Nathan Srebro (2015). "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*. URL: http://arxiv.org/abs/1412.6614.

O'Donnell, Ryan (2014). *Analysis of Boolean functions*. Cambridge University Press.

References V

Velikanov, Maksim and Dmitry Yarotsky (2021). "Explicit loss asymptotics in the gradient descent training of neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 2570–2582. URL: https://proceedings.neurips.cc/paper/2021/file/14faf969228fc18fcd4fcf59437b0c97-Paper.pdf.

Woodworth, Blake et al. (2020). "Kernel and Rich Regimes in Overparametrized Models". In: *Proceedings of Thirty Third Conference on Learning Theory*. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 3635–3673. URL: https://proceedings.mlr.press/v125/woodworth20a.html.