Deep Neural Tangent Kernel and Laplace Kernel Have the Same RKHS[1]

Lin Chen[1] Sheng Xu[2]

[1]Simons Institute for the Theory of Computing
University of California, Berkeley

[2]Department of Statistics and Data Science,
Yale University

BLISS Seminar, Jan 29, 2021

[1]Accepted to ICLR'21.

Outline

Outline

Kernel and RKHS

Definition (Kernel)

$K : E \times E \to \mathbb{R}$ is a kernel on E if $K(x,y) = K(y,x)$ and

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0\,,$$

holds for any $x_1, \ldots, x_n \in E$ and $c_1, \ldots, c_n \in \mathbb{R}$.

Definition (Reproducing Kernel Hilbert Space (RKHS))

An RKHS on E is a Hilbert space $\mathcal{H} \subseteq \{f : E \to \mathbb{R}\}$ with a function $K : E \times E \to \mathbb{R}$, which is called the *reproducing kernel*, enjoying the *reproducing property*

$$K_x = K(\cdot, x) \in \mathcal{H} \quad \forall x \in E\,,$$
$$f(x) = \langle f, K_x \rangle_{\mathcal{H}} \quad \forall x \in E, f \in \mathcal{H}\,.$$

Kernel Determines an RKHS

Theorem

For any kernel $K : E \times E \to \mathbb{R}$, there exists a uniquely determined RKHS $\mathcal{H}_K(E)$ admitting the reproducing kernel K on E.

- Consider the pre-Hilbert space

$$\mathcal{H}_{0,K}(E) = \mathrm{span}\left\{K_x \mid x \in E\right\} \triangleq \left\{f = \sum_{i=1}^{n} a_i K(x_i, \cdot)\right\}.$$

- Define the inner product $\langle K_x, K_y\rangle = K(x, y)$ and extend it linearly

$$\left\langle \sum_{i=1}^{n} a_i K_{x_i}, \sum_{j=1}^{m} b_j K_{y_j} \right\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j K(x_i, y_j).$$

- Complete $\mathcal{H}_{0,K}(E)$ and get $\mathcal{H}_K(E)$.

Outline

Inner Product Kernel

- $K(x, y)$ is an *inner product kernel* if $\exists \tilde{K} : [-1, 1] \to \mathbb{R}$ such that
$$K(x, y) = \tilde{K}(x^\top y) .$$
- We only consider inner product kernels on $\mathbb{S}^{d-1}$, and therefore $x^\top y \in [-1, 1]$ for $x, y \in \mathbb{S}^{d-1} \triangleq \{x \in \mathbb{R}^d \mid \|x\| = 1\}$.
- We abuse the notation and use $K(z)$ to denote $\tilde{K}(z)$.
- Power series of $K(z)$ around 0 have all non-negative coefficients.

Example (Laplace Kernel on $\mathbb{S}^{d-1}$)

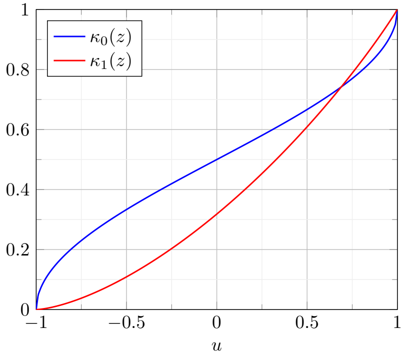
$$\begin{aligned} K_{\mathrm{Lap}}(x, y) &= e^{-c_1 \|x-y\|} = e^{-c_1 \sqrt{\|x-y\|^2}} \\ &= e^{-c_1 \sqrt{\|x\|^2 + \|y\|^2 - 2x^\top y}} = e^{-c_1 \sqrt{2(1-x^\top y)}} \\ &= e^{-c_2 \sqrt{1-z}} . \end{aligned}$$

Arc-cosine Kernels of Degree 0 and 1

- Arc-cosine kernels of degree 0 and 1 [CS09] are given by

$$\kappa_0(z) = \frac{1}{\pi}(\pi - \arccos(z)),$$
$$\kappa_1(z) = \frac{1}{\pi}\left(z \cdot (\pi - \arccos(z)) + \sqrt{1 - z^2}\right).$$

They are an example of inner product kernel.



Outline

Neural Tangent Kernel (NTK) with ReLU Activation

- Gradient flow on normally initialized, fully connected, infinitely wide NN = kernel regression with respect to the NTK.

NTK with ReLU activation $N_k(x, y)$ [JGH18; Gei+20; BM19]:

$$\Sigma_k(x,y) = \sqrt{\Sigma_{k-1}(x,x)\Sigma_{k-1}(y,y)}\kappa_1\left(\frac{\Sigma_{k-1}(x,y)}{\sqrt{\Sigma_{k-1}(x,x)\Sigma_{k-1}(y,y)}}\right)$$

$$N_k(x,y) = \Sigma_k(x,y) + N_{k-1}(x,y)\kappa_0\left(\frac{\Sigma_{k-1}(x,y)}{\sqrt{\Sigma_{k-1}(x,x)\Sigma_{k-1}(y,y)}}\right) + \beta^2\,,$$

- Initial conditions:

$$N_0(x,y) = x^\top y + \beta^2, \quad \Sigma_0(x,y) = x^\top y\,.$$

Lemma
$\Sigma_k(x,x) = 1$ *for any* $x \in \mathbb{S}^{d-1}$ *and* $k \geq 0$.

ReLU NTK on Sphere

If $x, y \in \mathbb{S}^{d-1}$, using $\Sigma_k(x, x) = 1$, we have ReLU NTK N_k on $\mathbb{S}^{d-1}$ is an inner product kernel:

$$N_k(z) = \kappa_1^{(k)}(z) + N_{k-1}(z)\kappa_0(\kappa_1^{(k-1)}(z)) + \beta^2 \,,$$

where

$$\kappa_1^{(k)}(z) \triangleq \underbrace{\kappa_1(\kappa_1(\cdots \kappa_1(\kappa_1}_{k}(z))\cdots))$$

is the k-th iterate of $\kappa_1(z)$.

Outline

RKHS of ReLU NTK = RKHS of Laplace Kernel

Theorem

Let $\mathcal{H}_{\mathrm{Lap}}(\mathbb{S}^{d-1})$ and $\mathcal{H}_{N_k}(\mathbb{S}^{d-1})$ be the RKHS associated with the Laplace kernel $K_{\mathrm{Lap}}(x, y) = e^{-c\|x-y\|}$ $(c > 0)$ and NTK N_k on $\mathbb{S}^{d-1}$. Then the two spaces include the same set of functions:

$$\mathcal{H}_{\mathrm{Lap}}(\mathbb{S}^{d-1}) = \mathcal{H}_{N_k}(\mathbb{S}^{d-1}), \quad \forall k \geq 1 .$$

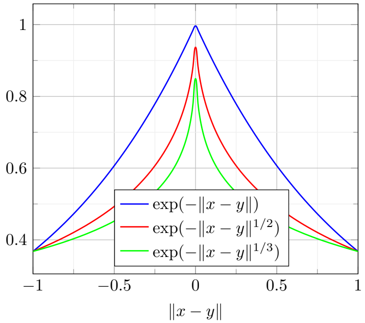
- RKHS characterizes the expressive power of a kernel to some extent. This is a negative result for the kernel regime of neural nets.

More Non-Smoothness Results in Larger RKHS

Theorem

Let $\mathcal{H}_{K_{\exp}^{\gamma,\sigma}}(\mathbb{S}^{d-1})$ and $\mathcal{H}_{K_{\exp}^{\gamma,\sigma}}(\mathbb{R}^d)$ be the RKHS associated with the exponential power kernel $K_{\exp}^{\gamma,\sigma}(x,y) = \exp\left(-\frac{\|x-y\|^\gamma}{\sigma}\right)$ $(\gamma, \sigma > 0)$ on $\mathbb{S}^{d-1}$ and $\mathbb{R}^d$, respectively.

- *If $0 < \gamma_1 < \gamma_2 < 2$, $\mathcal{H}_{K_{\exp}^{\gamma_2,\sigma_2}}(\mathbb{S}^{d-1}) \subseteq \mathcal{H}_{K_{\exp}^{\gamma_1,\sigma_1}}(\mathbb{S}^{d-1})$.*
- *If $0 < \gamma_1 < \gamma_2 < 2$ are rational, $\mathcal{H}_{K_{\exp}^{\gamma_2,\sigma_2}}(\mathbb{R}^d) \subseteq \mathcal{H}_{K_{\exp}^{\gamma_1,\sigma_1}}(\mathbb{R}^d)$.*



Outline

Comparing Power Series Coefficients

Lemma ([Aro50])

Let K_1, K_2 be two kernels. Then $\mathcal{H}_{K_1} \subseteq \mathcal{H}_{K_2}$ iff $\exists \gamma > 0$ s.t.

$$K_1 \preccurlyeq \gamma^2 K_2 \, .$$

- $K_1 \preccurlyeq \gamma^2 K_2$ means $\gamma^2 K_2 - K_1$ is a kernel.

Lemma ([Sch42; Bin73])

Suppose that $K(x, y) = f(x^\top y)$ where $x, y \in \mathbb{S}^{d-1}$ and $f \in C([-1, 1])$. Then K is a kernel on $\mathbb{S}^{d-1}$ for every d iff $f(u) = \sum_{k=0}^{\infty} a_k u^k$, in which $a_k \geq 0$ and $\sum_{k=0}^{\infty} a_k < \infty$.

- To show $\mathcal{H}_{N_k}(\mathbb{S}^{d-1}) \subseteq \mathcal{H}_{\text{Lap}}(\mathbb{S}^{d-1})$, it suffices to show $\exists \gamma > 0$, $\gamma^2 K_{\text{Lap}} - N_k$ is a kernel, or the power series coefficients of $\gamma^2 K_{\text{Lap}} - N_k$ is ≥ 0.

Decay Rate of Power Series Coefficients

- If $K(z) = \sum_{n \geq 0} a_n z^n$, write
$$[z^n]K(z) = a_n \,.$$
- We will show both $[z^n]K_{\mathrm{Lap}}(z)$ and $[z^n]N_k(z)$ are of order $n^{-3/2}$. Then $\exists \gamma > 0$ s.t. $[z^n]\left(\gamma^2 K_{\mathrm{Lap}}(z) - N_k(z)\right) \geq 0$ and the other way round.
- We use *singularity analysis* to get decay rate of power series coefficients.
- Key idea**:** Decay rate of power series coefficients of $K(z)$ is determined by the asymptotic around the dominant singularities (singularities closest to 0) of $K(z)$ (as a complex function).
- We will extend the inner product kernel to $\mathbb{C}$.

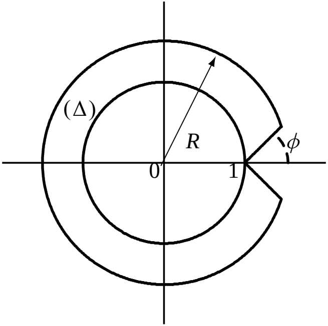
Singularity Analysis

Definition ([FS09])

For $R > 1$ and $\phi \in (0, \pi/2)$, the Δ-*domain* $\Delta(\phi, R)$ is defined by

$$\Delta(\phi, R) \triangleq \{z \in \mathbb{C} \mid |z| < R, z \neq 1, |\arg(z-1)| > \phi\} .$$

For a complex number $\zeta \neq 0$, a Δ-domain at ζ is the image by the mapping $z \mapsto \zeta z$ of $\Delta(\phi, R)$ for some $R > 1$ and $\phi \in (0, \pi/2)$. A function is Δ-*analytic* at ζ if it is analytic on a Δ-domain at ζ.



Transfer (One Dominant Singularity)

Lemma (Corollary VI.1 [FS09])

If f is Δ-analytic at its dominant singularity 1 and

$$f(z) \sim (1-z)^{-\alpha}, \quad \text{as } z \to 1, z \in \Delta$$

with $\alpha \notin \{0, -1, -2, \dots\}$, we have

$$[z^n]f(z) \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}\,.$$

Example (Laplace Kernel)

$K_{\mathrm{Lap}}(x,y) = e^{-c\sqrt{1-z}}$ is analytic on $\mathbb{C} \setminus [1,\infty)$. As $z \to 1$, we have

$$\frac{K_{\mathrm{Lap}}(z) - 1}{-c} \sim \sqrt{1-z}\,,$$

$$[z^n]K_{\mathrm{Lap}}(z) \sim \frac{c}{2\sqrt{\pi}} n^{-3/2}\,.$$

A Dictionary of Singularity Analysis vs. Decay Rate

- The following dictionary is only a small portion of Figure VI.5 [FS09]. There is a systematic approach to transferring singularity analysis to decay rate.

Function	Coeffcients
$(1-z)^{3/2}$	$\frac{1}{\sqrt{\pi n^5}}\left(\frac{3}{4}+\frac{45}{32n}+\frac{1155}{512n^2}+O\left(\frac{1}{n^3}\right)\right)$
$(1-z)^{1/2}$	$-\frac{1}{\sqrt{\pi n^3}}\left(\frac{1}{2}+\frac{3}{16n}+\frac{25}{256n^2}+O\left(\frac{1}{n^3}\right)\right)$
$(1-z)^{-1/2}\log\frac{1}{1-z}$	$\frac{1}{\sqrt{\pi n}}\left(\log n+\gamma+2\log 2-\frac{\log n+\gamma+2\log 2}{8n}+O\left(\frac{\log n}{n^2}\right)\right)$
$(1-z)^{-2}\log^2\frac{1}{1-z}$	$n\left(\log^2 n+2(\gamma-1)\log n+\gamma^2-2\gamma+2-\frac{\pi^2}{6}+O\left(\frac{\log n}{n}\right)\right)$

Outline

Extending Arc-Cosine Kernels to Complex Plane

- Recall

$$\kappa_0(u) = \frac{1}{\pi}(\pi - \arccos(u))\,,$$
$$\kappa_1(u) = \frac{1}{\pi}\left(u \cdot (\pi - \arccos(u)) + \sqrt{1-u^2}\right)\,.$$

  Both $\arccos(z)$ and $\sqrt{1-z^2}$ have branch points at $z = \pm 1$.

- Branch cut of $\kappa_0(z)$ and $\kappa_1(z)$ is $[1, \infty) \cup (-\infty, -1]$. They have a single-valued analytic branch on $D = \mathbb{C} \setminus [1, \infty) \setminus (-\infty, -1]$. On this branch, we have

$$\kappa_0(z) = \frac{\pi + \mathbf{i}\log(z + \mathbf{i}\sqrt{1-z^2})}{\pi}\,,$$
$$\kappa_1(z) = \frac{1}{\pi}\left[z \cdot \left(\pi + \mathbf{i}\log(z + \mathbf{i}\sqrt{1-z^2}\right) + \sqrt{1-z^2}\right]\,,$$

  where we use the principal value of the logarithm and square root.

ReLU NTK on Complex Plane

Recall ReLU NTK on $\mathbb{S}^{d-1}$ and $z = x^\top y$

$$N_k(z) = \kappa_1^{(k)}(z) + N_{k-1}(z)\kappa_0(\kappa_1^{(k-1)}(z)) + \beta^2 .$$

- Dominant singularities of $N_k(z)$ are ± 1.
- $N_k(z)$ is Δ-analytic at ± 1. **This part requires most work** because the branch cut becomes very complicated after composition.

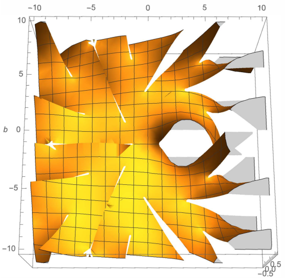


Figure: Imaginary part of $\kappa_1^{(5)}$. The slots are the branch cut.

Asymptotic of NTK Around ± 1

- As $z \to 1$

$$N_k(z) = (k+1)(z+\beta^2) - \left(\sqrt{2}(1+\beta^2)\frac{k(k+1)}{2\pi} + o(1)\right)\sqrt{1-z}\,.$$

- As $z \to -1$

$$N_k(z) = N_k(-1) + \left(\frac{\sqrt{2}(\beta^2-1)}{\pi}\prod_{j=1}^{k-1}\kappa_0(\kappa_1^j(-1)) + o(1)\right)\sqrt{1+z}\,.$$

Decay Rate of Power Series of NTK

- Multiple dominant singularities: the influence of each singularity is added up. (Theorem VI.5 [FS09]). We get
$$[z^n]N_k(z) = Cn^{-3/2}.$$
We can even determine the constant C (details in the paper).
- We conclude that $[z^n]K_{\mathrm{Lap}}(z)$ and $[z^n]N_k$ are both of order $n^{-3/2}$. Thus they have the same RKHS.

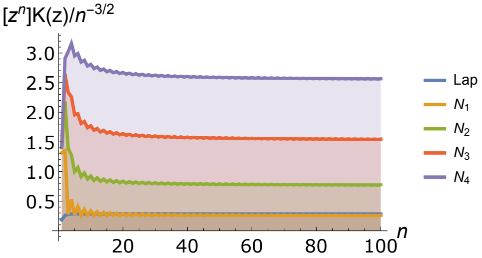


Figure: $[z^n]K(z)/n^{-3/2}$ vs. n for the Laplace kernel $K_{\mathrm{Lap}}(z) = e^{-\sqrt{2(1-z)}}$ and NTKs $N_1, \ldots, N_4$ with $\beta = 0, 1$.

Outline

EPK On $\mathbb{S}^{d-1}$: Easy

Recall that the exponential power kernel on $\mathbb{S}^{d-1}$ is given by

$$K_{\exp}^{\gamma,\sigma}(x,y) = \exp\left(-\frac{\|x-y\|^\gamma}{\sigma}\right) = \exp\left(-\frac{(2(1-x^\top y))^{\gamma/2}}{\sigma}\right).$$

As $z \to 1$, we get

$$K_{\exp}^{\gamma,\sigma}(z) = 1 - (c + o(1))(1-z)^{\gamma/2},$$

$$[z^n]K_{\exp}^{\gamma,\sigma}(z) \sim \frac{cn^{-\gamma/2-1}}{-\Gamma(-\gamma/2)}.$$

Therefore, a smaller γ results in a larger RKHS.

Outline

Step 1: Showing Complete Monotonicity

Definition

$f(t)$ is *completely monotone* if $f \in C[0,\infty) \cap C^\infty(0,\infty)$ and satisfies $(-1)^n \frac{d^n f(t)}{dt} \geq 0$ for every $n = 0, 1, 2, \ldots$ and $t > 0$.

Lemma (Theorem 1 of Chapter 15 [CL09])

If f is completely monotone but not constant on $[0,\infty)$, then for any n distinct points $x_1, x_2, \ldots, x_n$ in any inner-product space, the matrix $A_{ij} = f(\|x_i - x_j\|^2)$ is positive definite.

- We need to show that

$$c^2 \exp(-x^{\gamma_1/2}/\sigma_1) - \exp(-x^{\gamma_2/2}/\sigma_2)$$

is completely monotone but not constant on $[0,\infty)$ for some $c > 0$.

Step 2: Inverse Laplace Transform

Lemma

$f : [0,\infty) \to [0,\infty)$ is completely monotone iff there is a nondecreasing bounded function g s.t $f(t) = \int_0^\infty e^{-st} dg(s)$.

- It suffices to check that $c^2 \exp(-x^{\gamma_1/2}/\sigma_1) - \exp(-x^{\gamma_2/2}/\sigma_2)$ is the Laplace transform of a non-negative function on $[0,\infty)$.

Lemma (Shown by term-by-term contour integration)

For $a \in (0,1)$, $f(t) \triangleq \mathscr{L}^{-1}\{\exp(-s^a)\}(t)$ exists. Moreover, $f(t)$ is continuous on $\mathbb{R}$ and satisfies $f(0) = 0$. If $t > 0$, we have

$$f(t) = \frac{1}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^{k+1}\Gamma(ak+1)\sin(\pi ak)}{k!\,t^{ak+1}}.$$

Lemma

For rational $a = \frac{p}{q} \in (0,1)$, $f(t) \sim -\frac{1}{t^{a+1}\Gamma(-a)}$ as $t \to +\infty$.

References I

- Aronszajn, Nachman (1950). "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3, pp. 337–404.
- Bietti, Alberto and Julien Mairal (2019). "On the inductive bias of neural tangent kernels". In: *Advances in Neural Information Processing Systems*, pp. 12893–12904.
- Bingham, Nicholas H (1973). "Positive definite functions on spheres". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 73. 1. Cambridge University Press, pp. 145–156.
- Cheney, Elliott Ward and William Allan Light (2009). *A course in approximation theory*. Vol. 101. American Mathematical Soc.
- Cho, Youngmin and Lawrence K Saul (2009). "Kernel methods for deep learning". In: *Advances in neural information processing systems*, pp. 342–350.

References II

- Flajolet, Philippe and Robert Sedgewick (2009). *Analytic combinatorics*. cambridge University press.
- Geifman, Amnon et al. (2020). “On the Similarity between the Laplace and Neural Tangent Kernels”. In: *arXiv preprint arXiv:2007.01580*.
- Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018). “Neural tangent kernel: Convergence and generalization in neural networks”. In: *Advances in neural information processing systems*, pp. 8571–8580.
- Schoenberg, I J (1942). “Positive definite functions on spheres”. In: *Duke Mathematical Journal* 9.1, pp. 96–108.