Reproducing Kernel Hilbert Spaces and Neural Tangent Kernels

Semih Cayci

Mathematical Foundations of Deep Learning

RWTH Aachen

Basics of Reproducing Kernel Hilbert Spaces

Neural Tangent Kernels

Hilbert Spaces

Generalize $(\mathbb{R}^d, \langle \cdot, \cdot \rangle)$ with $\langle x, x' \rangle = x^\top x'$ to infinite dimension.

Definition (Inner product)

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product if:

1. (Symmetricity) $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$,
2. (Linearity) $\langle \alpha f + \beta g, h \rangle_{\mathcal{H}} = \alpha \langle f, h \rangle + \beta \langle g, h \rangle$, $\forall \alpha, \beta \in \mathbb{R}, f, g, h \in \mathcal{H}$,
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ with equality iff $f = 0$.

Norm induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$: $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$.

Hilbert Space

A sequence $f_n \in \mathcal{H}$ converges to $f \in \mathcal{H}$ if for any $\epsilon > 0$, there exists $N = N(\epsilon)$ such that

$$\|f_n - f\|_{\mathcal{H}} < \epsilon, \ \forall n \geq N.$$

A sequence $f_n \in \mathcal{H}$ is a Cauchy sequence if for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$\|f_n - f_m\|_{\mathcal{H}} < \epsilon, \ \forall n, m \geq N.$$

Definition (Complete vector space)

A normed space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is complete if every Cauchy sequence of its elements is convergent.

Definition (Hilbert space)

A complete space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ for which $\|\cdot\|_{\mathcal{H}}$ is induced by an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is called a Hilbert space.

Mercer Kernels

Generalizing the concept of inner product.

Definition (Mercer kernel)

Let $\mathcal{X} \subset \mathbb{R}^d$ be closed. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a Mercer kernel if:

- (Symmetricity) $K(x, x') = K(x', x), \ \forall x, x' \in \mathcal{X}$.
- (Positive definiteness) $\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \geq 0$ for any $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$.

Examples:

1. $\mathcal{X} = \mathbb{R}^d$, $K(x, x') = x^\top x'$ is a Mercer kernel.
2. $\mathcal{X} = B_2(0, \rho)$ for $\rho > 0$ and $\{a_j\}$ such that $\sum_{j=0}^{\infty} a_j R^{2j} < \infty$. Then, $K(x, x') = \sum_{j=0}^{\infty} a_j (x^\top x')^j$ is a Mercer kernel.

Linear Span of Kernels

Given a (Mercer) kernel K, let $\mathcal{L}_K$ be the linear span of the set $\{K(x',\cdot) : x' \in \mathcal{X}\}$: the set of all functions $f : \mathcal{X} \to \mathbb{R}$ such that

$$f(x) = \sum_{j=1}^{n} c_j K(x, x_j),$$

for all choices of $n \in \mathbb{N}, c_1, \ldots, c_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$.

Reproducing Kernel Hilbert Spaces

For any kernel K, $\mathcal{L}_K$ can be completed into a Hilbert space.

Theorem (RKHS; Cucker and Zhou, 2007)

Let $\mathcal{X} \subset \mathbb{R}^d$ be closed, and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel. Then, there exists a unique Hilbert space $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$ s.t.:

1. Let $K_x(\cdot) = K(x, \cdot)$. Then $K_x \in \mathcal{H}_K$ and $\langle K_x, K_{x'} \rangle_K = K(x, x')$ for any $x, x' \in \mathcal{X}$.
2. The linear space $\mathcal{L}_K$ is dense in $\mathcal{H}_K$: for any $f \in \mathcal{H}_K$, there exists some $n \in \mathbb{N}$, $x_1, \cdot, x_n \in \mathcal{X}$ and $c_1, \cdot, c_n \in \mathbb{R}$ such that
$$\|f - \sum_{i=1}^{n} c_i K_{x_i}\|_K \leq \epsilon.$$
3. For any $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$, $f(x) = \langle f, K_x \rangle_K$.

$(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$ is called the RKHS associated with K.

Mercer's Representation Theorem

Kernels can be represented by series expansions in generality.

Theorem (Mercer; Hajek and Raginsky, 2021)

Suppose $\mathcal{X} \subset \mathbb{R}^d$ closed, and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel. Then, there is a sequence of continuous functions $\Phi_i : \mathcal{X} \to \mathbb{R}$ such that

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

and

$$c \in \ell^2, \;\; \sum_i c_i \Phi_i(x) = 0 \Rightarrow c = 0,$$

hold, and $(\Phi_1, \Phi_2 \ldots)$ forms an orthonormal basis for $\mathcal{H}_K$.

Transportation Mappings and Parametric RKHS

Consider a probability distribution p on Θ, and $\Phi : \mathcal{X} \times \Theta \to \mathbb{R}$.

$$\text{Let } K(x, x') = \int_\Theta \Phi(x; \theta)\Phi(x'; \theta)p(\theta)d\theta.$$

K is a Mercer kernel. Let $\mathcal{H}_K$ be the RKHS associated with K.

Proposition (Rahimi and Recht, 2008)

Let $\mathcal{G}$ be the completion of the set of all functions of the form

$$f(x) = \int_\Theta v(\theta)\Phi(x; \theta)p(\theta)d\theta,$$

such that $\int_\Theta |v(\theta)|^2 p(\theta)d\theta < \infty$, with the inner product

$$\langle f, g\rangle = \int_\Theta v(\theta)u(\theta)p(\theta)d\theta, \text{ for } g(x) = \int_\Theta u(\theta)\Phi(x; \theta)p(\theta)d\theta.$$

Then, $\mathcal{H}_K = \mathcal{G}$.

Neural Tangent Kernels – ReLU

Recall: $\nabla_W F(x; W, c) = \left(\frac{c_i x \mathbb{1}\{W_i^\top x \geq 0\}}{\sqrt{m}}\right)_{1 \leq i \leq m}$, and

$$F(x; W, c) = \nabla_W^\top F(x; W(0), c)\big(W - W(0)\big) + \Delta(W, m),$$

where $|\Delta(W, m)| = \mathcal{O}(1/\sqrt{m})$ with high probability.

Proposition

Under symmetric Xavier initialization, for any $x, x' \in \mathbb{R}^d$,

$$\lim_{m \to \infty} \Big\langle \nabla_W F(x; W(0), c), \nabla_W F(x'; W(0), c) \Big\rangle$$
$$= \int_{\mathbb{R}^d} \langle \Phi(x; w_0), \Phi(x'; w_0) \rangle p(w_0) dw_0, \text{ a.s.},$$

where $p(w) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp(-\frac{\|w\|_2^2}{2})$ and $\Phi(x; w) = x \mathbb{1}\{w^\top x \geq 0\}$.

Transportation Mapping Characterization of NTK

Corollary

Let $v : \mathbb{R}^d \to \mathbb{R}^d$ with $\mathbb{E}_{w_0 \sim \mathcal{N}(0, I_d)} \|v(w_0)\|_2^2 < \infty$. Then, the completion of the functions of type:

$$g(x) = \mathbb{E}_{w_0 \sim \mathcal{N}(0, I_d)}[\langle v(w_0), \Phi(x; w_0) \rangle],$$

is equal to the unique RKHS associated with the neural tangent kernel K.

Remark: It is easy to see that

$$\|g\|_K^2 = \mathbb{E}_{w_0 \sim \mathcal{N}(0, I_d)} \|v(w_0)\|_2^2,$$

and if

$$\sup_{w \in \mathbb{R}^d} \|v(w)\|_2 \le \alpha < \infty,$$

then $\|g\|_K \le \alpha$. Also, $|g(x)| \le \alpha \|x\|_2$ from Cauchy-Schwarz and triangle inequality.

Approximation of NTK RKHS with Neural Networks

Let $g(x) = \mathbb{E}[v^\top(w_0)\Phi(x; w_0)]$ with $\sup_w \|v(w)\|_2 \leq \alpha$. Define $W = W(0) + U$ with

$$U_i = \frac{1}{\sqrt{m}} c_i v(W_i(0)), \;\; i = 1, 2, \ldots, m.$$

Then, by Cauchy-Schwarz, $\max_{1 \leq i \leq m} \|W_i - W_i(0)\|_2 \leq \alpha/\sqrt{m}$, therefore, from previous lecture:

$$F(x; W, c) = \nabla_W^\top F(x; W(0), c)\big[W - W(0)\big] + \mathcal{O}(1/\sqrt{m}).$$

Important: Now, notice that

$$\nabla_W^\top F(x; W(0), c)[W - W(0)] = \frac{1}{m}\sum_{i=1}^{m} \langle v(W_i(0)), \Phi(x; W_i(0))\rangle.$$

Approximation of NTK RKHS with Neural Networks

Theorem

Let $g \in \mathcal{H}_K$ with a transportation mapping $v : \mathbb{R}^d \to \mathbb{R}^d$ such that $\sup_w \|v(w)\|_2 \leq \alpha$. Then, there exists a neural network of width m with symmetric Xavier initialization such that for any $\delta \in (0, 1)$,

$$|F(x; W, c) - g(x)| \leq \alpha\sqrt{\frac{\log(3/\delta)}{m}} + \frac{1+\alpha}{\sqrt{m}}\left(\alpha\|x\|_2 + \sqrt{\log(3/\delta)}\right),$$

with probability at least $1 - \delta$.

Proof idea: Let $W_i = W_i(0) + c_i v(W_i(0))/\sqrt{m}$ for $i = 1, 2, \ldots, m$.

$$\begin{aligned}|F(x; W, c) - g(x)| = &\underbrace{|F(x; W, c) - \nabla_W^\top F(x; W(0), c)(W - W(0))|}_{\text{from previous lecture}} \\ &+ \underbrace{|\nabla_W^\top F(x; W(0), c)(W - W(0)) - g(x)|}_{\text{Hoeffding's inequality}}\end{aligned}$$

Summary

The infinite-width limit of randomly initialized neural networks is functions of type

$$g(x) = \int_{\mathbb{R}^d} \langle v(w_0), \Phi(x; w_0) \rangle p(w_0) dw_0.$$

With infinitely-wide neural networks of width m, there exists (W, c) such that

$$|F(x; W, c) - g(x)| = \mathcal{O}\left(\frac{1}{\sqrt{m}}\right).$$

- Optimization: Let $g \in \mathcal{H}_K$. Can we find $(\tilde{W}, \tilde{c})$ such that

  $$\frac{1}{2n} \sum_{j=1}^{n} \Big(F(x_i, \tilde{W}, \tilde{c}) - g(x_i) \Big)^2,$$

  is minimized?
- Approximation: How rich is the class of functions $x \mapsto \int_{\mathbb{R}^d} \langle v(w_0), \Phi(x; w_0) \rangle p(w_0) dw_0$?