MIT 9.520/6.7910
Statistical Learning Theory and Applications

Class 07: Random Features, NTK & RKHS

Lorenzo Rosasco

The beauty of linear models

$$f_w(x) = w^\top x.$$

- Linear function space:
$$\alpha w + \beta w' \quad \mapsto \quad \alpha f_w + \beta f_{w'}.$$

- The parameter norm
$$\|f_w\| = \|w\|$$
defines a norm on the function f_w.

Neural networks: functions and parameters

$$f_\theta(x) = v^\top \sigma(Wx + b), \quad \theta = (v, W, b).$$

- Not a linear function space:

$$\alpha\theta + \beta\theta' \quad \not\mapsto \quad \alpha f_\theta + \beta f_{\theta'}.$$

- The parameter norm

$$\|\theta\|$$

  does not define a norm on the function f_θ.

Headaches

Lack of linear structure makes the study of neural networks hard.

Lack of functional norm makes it hard to understand neural networks' bias/regularization.

Are there cases in which the above problems can be tackled?

Before we start

A Hilbert space $\mathcal{H}$ is:

- A linear space

$$f, f' \in \mathcal{H} \;\; \Rightarrow \;\; \alpha f + \beta f' \in \mathcal{H}$$

- An inner product space

$$\langle f, f' \rangle_{\mathcal{H}}$$

- A *complete* space

Outline

Infinite width neural networks

$$f_\theta(x) = \sum_{j=1}^{m} v_j \, \sigma(w_j^\top x + b_j)$$

Infinite width limit[1]

$$f_v(x) = \int_{\mathbb{R}^d \times \mathbb{R}} v(w,b) \, \sigma(w^\top x + b) \, d\tau(w,b)$$

(related to *mean field* limits).

[1] τ a fixed distribution over weights, ensuring the integral is finite.

Infinite width neural networks (cont.)

Infinite width limit

$$f_v(x) = \int_{\mathbb{R}^d \times \mathbb{R}} v(w,b)\, \sigma(w^\top x + b)\, d\tau(w,b).$$

- v is a function in [2] $L^2(\mathbb{R}^d \times \mathbb{R})$ (not necessarily a density).
- σ defines a function

  $$\sigma_x(w,b) = \sigma(w^\top x + b)$$

  in $L^2(\mathbb{R}^d \times \mathbb{R})$ for each $x \in \mathbb{R}^d$.
- Linear parametrization

  $$f_v(x) = \langle v, \sigma_x \rangle_{L^2(\mathbb{R}^d \times \mathbb{R})}.$$

[2] More precisely, $L^2(\mathbb{R}^d \times \mathbb{R}, \tau)$ with inner product $\int_{\mathbb{R}^d \times \mathbb{R}} g(w,b) g'(w,b) d\tau(w,b)$.

Infinite width neural networks (cont.)

$$f_v(x) = \langle v, \sigma_x \rangle_{L^2(\mathbb{R}^d \times \mathbb{R})} .$$

- Linear function space:

$$\alpha v + \beta v' \quad \mapsto \quad \alpha f_v + \beta f_{v'}.$$

- The L^2 norm

$$\|v\|_{L^2(\mathbb{R}^d \times \mathbb{R})} .$$

induces a norm on the function f_v.

Neural tangent kernel

Neural network

$$f_\theta(x) = \sum_{j=1}^{m} v_j \, \sigma(w_j^\top x + b_j).$$

Linearized network around θ_0

$$f_\theta(x) \;\approx\; f_{\theta_0}(x) + \langle \nabla_\theta f_{\theta_0}(x), \theta - \theta_0 \rangle \, .$$

Neural tangent kernel (cont.)

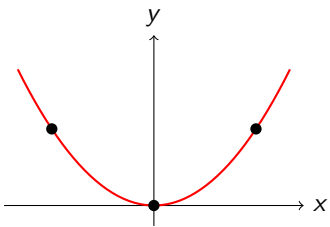
Linearized network is already linear...

$$f_\theta(x) \;\approx\; f_{\theta_0}(x) + \langle \nabla_\theta f_{\theta_0}(x), \theta - \theta_0 \rangle \,.$$

The infinite width limit can be considered... but the expression is involved!

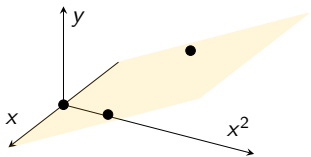
Linear models on features

$$f(x) = \langle v, \Phi(x) \rangle_{\mathcal{F}} .$$



input space $\mathcal{X}$

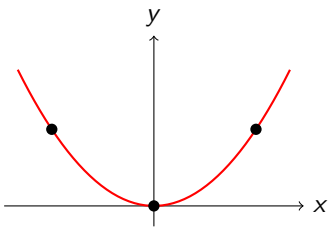
$$\Phi(x) = (x, x^2)$$



feature space $\mathcal{F}$

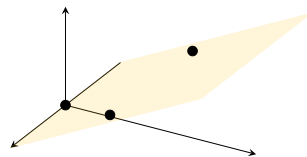
Linear models for neural networks

$$f(x) = \langle v, \Phi(x) \rangle_{L^2(\mathbb{R}^d \times R)} .$$



input space $\mathcal{X}$

$$\Phi(x) = \sigma\left((\cdot)^T x + (\cdot)\right)$$



feature space $L^2(\mathbb{R}^d \times R)$

Outline

Feature maps

Let $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ be a Hilbert space, and

$$\Phi : \mathcal{X} \to \mathcal{F}.$$

$\mathcal{F}$ is called a feature space and Φ a feature map.

Feature maps and hyperplanes

Let $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ be a Hilbert space, and

$$\Phi : \mathcal{X} \to \mathcal{F}.$$

Consider functions of the form

$$f(x) = \langle v, \Phi(x) \rangle_{\mathcal{F}} .$$

with $v \in \mathcal{F}$.

Feature maps, hyperplanes and norms

$$f(x) = \langle v, \Phi(x) \rangle_{\mathcal{F}} .$$

Define the norm

$$\|f\|_{\mathcal{H}} = \inf\{\|v\|_{\mathcal{F}} \mid f(x) = \langle v, \Phi(x) \rangle_{\mathcal{F}}, \quad \forall x \in \mathcal{X}\}.$$

Hilbert space of hyperplanes

It can be proved that the space $\mathcal{H}$ of functions

$$f(x) = \langle v, \Phi(x) \rangle_{\mathcal{F}}$$

with norm

$$\|f\|_{\mathcal{H}} = \inf\{\|v\|_{\mathcal{F}} \mid f(x) = \langle v, \Phi(x) \rangle_{\mathcal{F}}, \quad \forall x \in \mathcal{X}\}$$

is a Hilbert space (with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$).

Feature maps and RKHS

The space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is such that for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}} .$$

The above property is remarkable and **not** satisfied by common spaces like $L^2(\mathbb{R}^d)$ or $\mathcal{C}(\mathbb{R}^d)$.

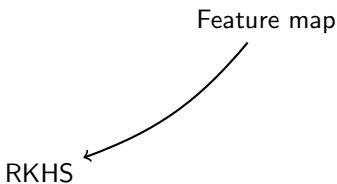
A Hilbert space of functions with the above property is called a Reproducing Kernel Hilbert Space (RKHS).

Feature maps and RKHS

Feature map

RKHS

Feature maps and RKHS



RKHS and evaluation functionals

The space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is such that for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}} .$$

For each $x \in \mathcal{X}$, evaluation functionals $f \mapsto f(x)$ are linear and continuous.

Then, for each $x \in \mathcal{X}$, there exists $\delta_x \in \mathcal{H}$ such that

$$f(x) = \langle f, \delta_x \rangle_{\mathcal{H}} ,$$

(Riesz Representation Theorem).

RKHS and feature maps

Each RKHS $\mathcal{H}$ has an infinite number of associated feature maps.

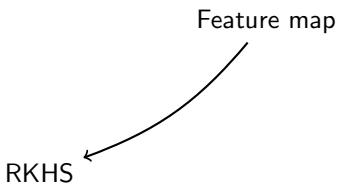
- Let $\mathcal{F} = \mathcal{H}$ and for all $x \in \mathcal{X}$

$$\Phi(x) = \delta_x.$$

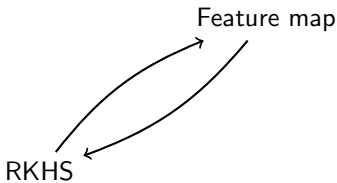
- Let $(e_j)_j$ be an orthonormal basis in $\mathcal{H}$. Let $\mathcal{F} = \ell^2$ and for all $x \in \mathcal{X}$

$$\Phi(x) = (e_j(x))_j$$

RKHS and feature maps



RKHS and feature maps



Outline

RKHS and reproducing kernels

The space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is such that for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$

$$|f(x)| \leq C_x \|f\|_{\mathcal{H}} .$$

Equivalently[3], $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a Hilbert space with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

- For all $x \in \mathcal{X}$, $k_x = k(x, \cdot) \in \mathcal{H}$.
- For all $f \in \mathcal{H}_k$, $x \in \mathcal{X}$,

  $$f(x) = \langle f, k_x \rangle_{\mathcal{H}_k}$$

  called the reproducing property.

[3] Again Riesz Representation Theorem.

RKHS and PD functions

Each reproducing kernel is symmetric and positive definite (PD)[4].

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite if

- For all points $x_1, \ldots, x_n \in \mathcal{X}$, the matrix $\widehat{K}_{i,j} = k(x_i, x_j)$ is positive semidefinite, i.e.

$$a^\top \widehat{K} a \geq 0, \quad \forall a \in \mathbb{R}^n.$$

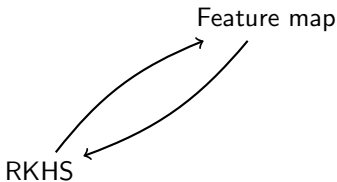
- Equivalently

$$\sum_{i,j=1}^{n} k(x_i, x_j) a_i a_j \geq 0,$$

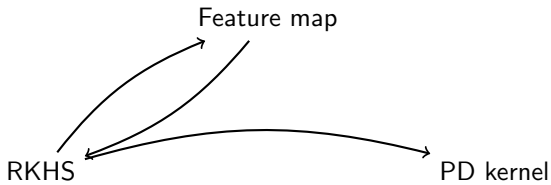
for all $a_1, \ldots, a_n \in \mathbb{R}$, $x_1, \ldots, x_n \in \mathcal{X}$.

[4]Note that $k(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$.

RKHS and PD kernels



RKHS and PD kernels



PD functions and RKHS

Each symmetric and PD kernel defines a unique RKHS (Moore-Aronszajn theorem).

Main steps:

- Consider the space $\mathcal{H}$ of functions of the form

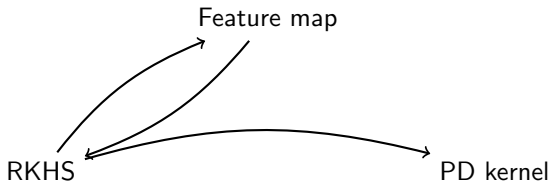
$$f(x) = \sum_{j=1}^{N} c_j k(x, x_j).$$

- Define

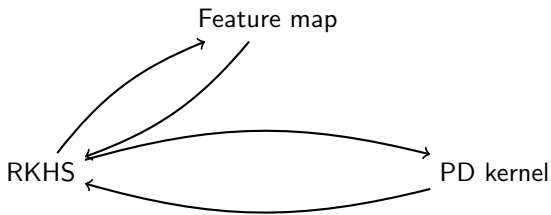
$$\langle f, g \rangle_{\mathcal{H}} = \sum_{j=1}^{N} \sum_{i=1}^{M} c_j b_i k(x_i, x_j).$$

- *Complete* into a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$.

PD kernels and RKHS



PD kernels and RKHS



Outline

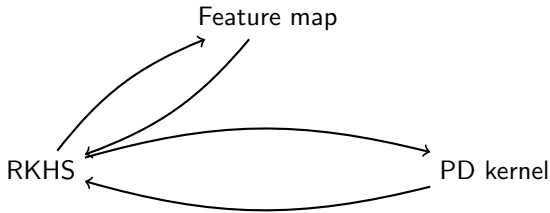
PD functions and feature maps

Each PD function k defines an RKHS $\mathcal{H}$ of which it is reproducing kernel.

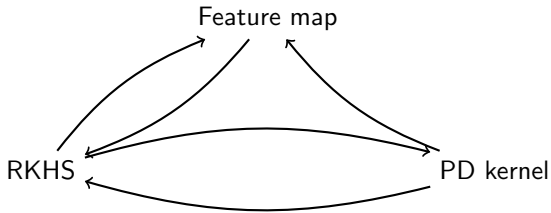
Let $\mathcal{F} = \mathcal{H}$ and for all $x \in \mathcal{X}$

$$\Phi(x) = k_x.$$

PD kernels and feature maps



PD kernels and feature maps



Feature maps and PD functions

Each feature map $\Phi : \mathcal{X} \to \mathcal{F}$ defines a PD kernel

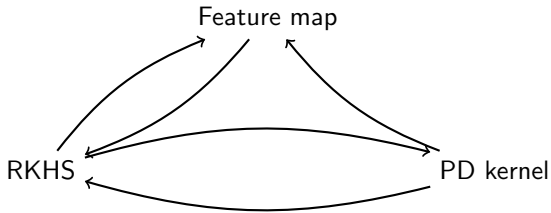
$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}} .$$

The space $\mathcal{H}$ of functions $f(x) = \langle v, \Phi(x) \rangle_{\mathcal{F}}$ with norm

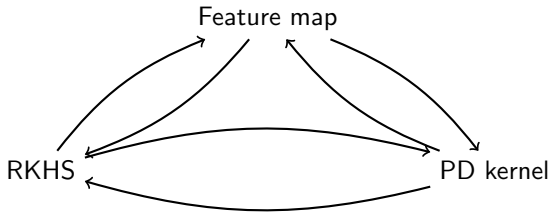
$$\|f\|_{\mathcal{H}} = \inf\{\|v\|_{\mathcal{F}} \mid f(x) = \langle v, \Phi(x) \rangle_{\mathcal{F}}, \quad \forall x \in \mathcal{X}\}$$

is the RKHS with reproducing kernel k.

Feature maps and PD kernels



Feature maps and PD kernels



Outline

Representer theorem

Let $\mathcal{H}$ be an RKHS with kernel k. Consider the regularized ERM problem,

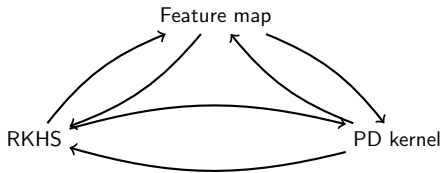
$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 .$$

The minimizer $\hat{f}_\lambda$ of the above functional has the form

$$\hat{f}_\lambda(x) = \sum_{i=1}^{n} c_i k(x, x_i).$$

The simple proof relies on an orthogonal decomposition.

Few remarks



- Feature maps describe a general form of linear models.
- RKHS theory highlights the rich mathematical structure of these spaces.
- Kernels provide a "dual" perspective, with a direct practical value (representer theorem).

Few more remarks

Many more connections:

- Mercer's theorem (Karhunen–Loève expansion).
- Gaussian processes.
- Cameron–Martin spaces.

Outline

Infinite width neural networks and kernels

$$f_v(x) = \int_{\mathbb{R}^d \times \mathbb{R}} v(w,b)\, \sigma(w^\top x + b)\, d\tau(w,b), \qquad v \in L^2(\mathbb{R}^d \times R)$$

The corresponding kernel is

$$k(x,x') = \int_{\mathbb{R}^d \times \mathbb{R}} \sigma(w^\top x + b)\sigma(w^\top x' + b)\, d\tau(w,b).$$

Random features kernels

$$k(x, x') = \int_{\mathbb{R}^d \times \mathbb{R}} \sigma(w^\top x + b)\sigma(w^\top x' + b)\, d\tau(w, b).$$

If v is a probability density,

$$k(x, x') \approx k_m(x, x') = \frac{1}{m} \sum_{j=1}^{m} \sigma(w_j^\top x + b_j)\, \sigma(w_j^\top x' + b_j),$$

where $(w_j, b_j) \sim v$, $j = 1, \ldots m$.

Random features

$$k_m(x, x') = \frac{1}{m} \sum_{j=1}^{m} \sigma(w_j^\top x + b_j)\, \sigma(w_j^\top x' + b_j).$$

The corresponding feature map is

$$\Phi_m(x) = \frac{1}{\sqrt{m}} \big(\sigma(w_1^\top x + b_1), \ldots, \sigma(w_m^\top x + b_m)\big).$$

Random features and neural networks

$$\Phi_m(x) = \frac{1}{\sqrt{m}}\big(\sigma(w_1^\top x + b_1), \ldots, \sigma(w_m^\top x + b_m)\big).$$

The corresponding hyperplanes are neural nets

$$f_m(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} v_j \, \sigma(w_j^\top x + b_j),$$

with random hidden weights $(w_j, b_j) \sim \nu$, $j = 1, \ldots, m$, and trainable output weights v_j.

ReLU and Arc-cos kernel

Kernels associated to neural networks can be computed for many activations σ and distributions τ.

If $\sigma(t) = \max(0, t)$ and τ is standard Gaussian, then

$$k(x, x') = \frac{\|x\| \, \|x'\|}{2\pi} \Big(\sin \varphi + (\pi - \varphi) \cos \varphi \Big),$$

where $\varphi = \arccos\Big(\frac{\langle x, x' \rangle}{\|x\| \|x'\|} \Big)$.

This is the *arc-cos kernel*.

Fourier features and Gaussian kernel

Kernels associated to neural networks can be computed for many activations σ and distributions τ.

If $\sigma(t) = \cos(t)$ and $\tau = \mathcal{N}(0, 2\gamma I)$, then

$$k(x, x') = \exp\big(-\gamma \|x - x'\|^2\big),$$

which is the Gaussian kernel.

Outline

Inductive bias of RKHS norm

Infinite width networks (with random weights and L^2 norm) have associated RKHS and kernels.

The RKHS norm can be seen as the associated inductive bias.

Can we interpret it?

Norm characterization for translation invariant kernels

$$k(x, x') = \kappa(x - x')$$

Then

$$\|f\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d} \frac{|\tilde{f}(\omega)|^2}{\tilde{\kappa}(\omega)}\, d\omega,$$

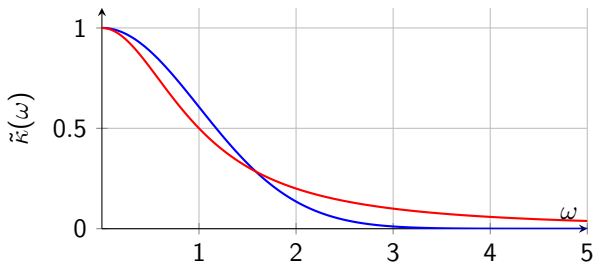
where $\tilde{f}, \tilde{\kappa}$ are Fourier transforms (Bochner's theorem).

One dimensional illustration

Let $\mathcal{X} = \mathbb{R}$, then

Gaussian kernel: $k(x, x') = \exp\big(-\frac{\gamma}{2}(x - x')^2\big)$, $\quad \tilde{\kappa}(\omega) \propto \exp\Big(-\frac{\omega^2}{2\gamma}\Big)$.

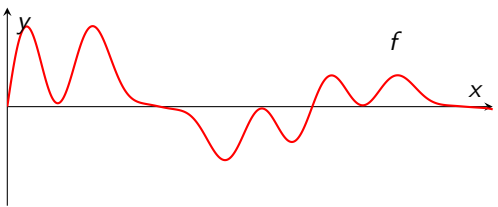
Laplace kernel: $k(x, x') = \frac{1}{2} e^{-|x - x'|}$, $\quad \tilde{\kappa}(\omega) = \dfrac{1}{1 + \omega^2}$.



One dimensional illustration (cont.)

$$\|f\|_{\mathcal{H}}^2 = \int_{\mathbb{R}} \frac{|\tilde{f}(\omega)|^2}{\frac{1}{1+\omega^2}}\, d\omega = \int_{\mathbb{R}} (1+\omega^2)\, |\tilde{f}(\omega)|^2\, d\omega = \|f\|_{L^2}^2 + \|f'\|_{L^2}^2.$$

(Plancherel formula).



Sobolev spaces and analytic functions

Sobolev spaces: Polynomial decay of the Fourier transform of the kernel corresponds to spaces of functions with controlled derivatives.

Analytic functions: Exponential decay corresponds to very smooth functions.

Paley–Wiener spaces: Kernels with compactly supported Fourier transform correspond to band-limited functions.

Back to ReLU and Arc-cos kernel

$$k(x,x') = \frac{\|x\|\,\|x'\|}{2\pi}\Big(\sin\varphi + (\pi-\varphi)\cos\varphi\Big), \quad \varphi = \arccos\Big(\frac{\langle x,x'\rangle}{\|x\|\|x'\|}\Big).$$

The translation invariant kernels analysis extends to:

- rotation–invariant kernels
$$k(x,x') = \kappa(x^\top x'),$$
(arc-cos kernel on the sphere, i.e. $\|x\| = 1$).

- homogeneous rotation–invariant kernels
$$k(x,x') = \|x\|^q\,\|x'\|^q\,\kappa\Big(\frac{x^\top x'}{\|x\|\|x'\|}\Big),$$
with $q = 1$ giving the arc-cos kernel.

Universality of translation invariant kernels

For $k(x, x') = \kappa(x - x')$,

$$\|f\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d} \frac{|\tilde{f}(\omega)|^2}{\tilde{\kappa}(\omega)} \, d\omega.$$

If $\tilde{\kappa}(\omega) > 0$ on every compact set, then $\mathcal{H}$ is dense in $C(\mathcal{X})$ for every compact $\mathcal{X} \subset \mathbb{R}^d$.

Similar results hold for homogeneous rotation–invariant kernels (Schoenberg's theorem).

Summing up

- Infinite width neural networks can be associated with RKHSs: rich mathematical structure!
- Kernels provide a "dual" representation.
- RKHS norm characterizations can give insight into inductive bias.
- Universality can be established.
- Various extensions (NTK, deep, convolutional).
- Is this the right point of view?