MIT 9.520/6.7910
Statistical Learning Theory and Applications

Class 08: Infinite Width Neural Networks & RKBS

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.

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.

Outline

Infinite width neural networks in L^2

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

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

[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 in L^2

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

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

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

[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 in L^2

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

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

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

- Linear parametrization

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

[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 with L^2 norm

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

- Linear function space:

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

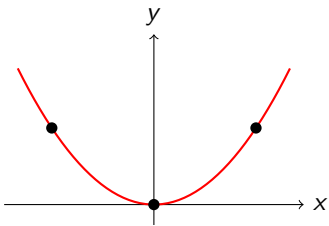
- The L^2 norm

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

induces a norm on the function f_v.

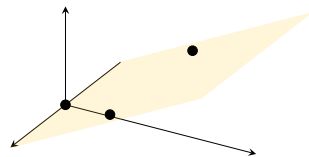
Neural networks and feature maps

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



input space $\mathcal{X}$

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



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

RKHSs of neural networks

$$f_v(x) = \langle v, \Phi(x) \rangle_{L^2(\mathbb{R}^d \times R)}, \qquad \|v\|_{L^2(R^d \to \mathbb{R})} \mapsto \|f_v\|$$

There is a corresponding RKHS with reproducing kernel

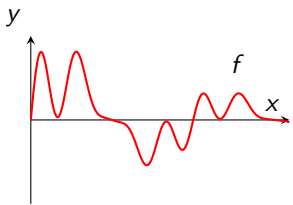
$$k(x, x') = \int_{\mathbb{R}^d \times \mathbb{R}} \sigma(w^\top x + b)\sigma(w^\top x' + b)\, d\tau(w, b).$$

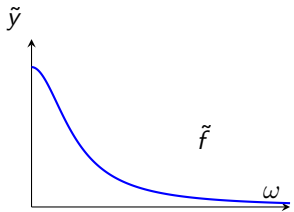
e.g. ReLU (with Gaussian weights) gives the arc-cos kernel.

Inductive bias and universality

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

RKHS norm characterization using Fourier transform [3].



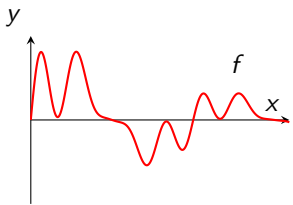


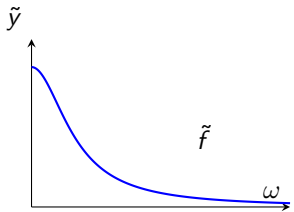
[3] For translation inviariant kernels. Analogous results for homogenous rotation invariant kernels.

Inductive bias and universality

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

RKHS norm characterization using Fourier transform [3].





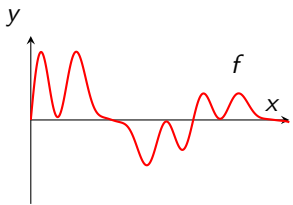
- Polynomial decay (Sobolev spaces), exponential decay (analytic functions).

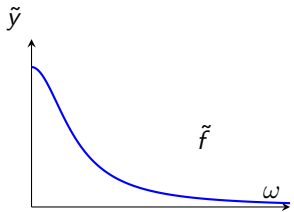
[3] For translation inviariant kernels. Analogous results for homogenous rotation invariant kernels.

Inductive bias and universality

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

RKHS norm characterization using Fourier transform [3].





- Polynomial decay (Sobolev spaces), exponential decay (analytic functions).
- Universality is ensured.

[3] For translation inviariant kernels. Analogous results for homogenous rotation invariant kernels.

Are neural networks[4] kernel methods?

- Some extensions (NTK, deep, convolutional). Inductive bias not always clear.

[4] One hidden layer!

Are neural networks[4] kernel methods?

- Some extensions (NTK, deep, convolutional). Inductive bias not always clear.
- These results suggest (shallow) neural networks should perform no better than kernel methods.

[4] One hidden layer!

Are neural networks[4] kernel methods?

- Some extensions (NTK, deep, convolutional). Inductive bias not always clear.
- These results suggest (shallow) neural networks should perform no better than kernel methods.
- Actually often true! But is it always the case?

[4] One hidden layer!

Are neural networks[4] kernel methods?

- Some extensions (NTK, deep, convolutional). Inductive bias not always clear.
- These results suggest (shallow) neural networks should perform no better than kernel methods.
- Actually often true! But is it always the case?

Are there problems for which there's a gap between shallow networks learn and kernel methods?

[4] One hidden layer!

Outline

Before we start

A Banach space $\mathcal{B}$ is:

- A linear space

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

- A normed space

$$\|f\|_{\mathcal{B}}$$

- A *complete* space

Dual spaces

The dual $\mathcal{B}^*$ is the space of linear, continuous functionals on $\mathcal{B}$.

Dual spaces

The dual $\mathcal{B}^*$ is the space of linear, continuous functionals on $\mathcal{B}$.

- Duality pairing

$$\langle \ell, f \rangle \;=\; \ell(f), \quad \ell \in \mathcal{B}^*,\; f \in \mathcal{B}.$$

Dual spaces

The dual $\mathcal{B}^*$ is the space of linear, continuous functionals on $\mathcal{B}$.

- Duality pairing
$$\langle \ell, f \rangle \;=\; \ell(f), \quad \ell \in \mathcal{B}^*,\; f \in \mathcal{B}.$$

- A Banach space is *reflexive* if
$$\mathcal{B} \;=\; \mathcal{B}^{**}.$$

Examples

- Each Hilbert space $\mathcal{H}$ is a reflexive Banach space and

$$\mathcal{H}^* = \mathcal{H}$$

(Riesz representation theorem).

Examples

- Each Hilbert space $\mathcal{H}$ is a reflexive Banach space and

$$\mathcal{H}^* = \mathcal{H}$$

(Riesz representation theorem).

- ℓ^p and $L^p(\mathbb{R}^d)$ are Banach spaces for $1 \leq p \leq \infty$.

Examples

- Each Hilbert space $\mathcal{H}$ is a reflexive Banach space and

$$\mathcal{H}^* = \mathcal{H}$$

  (Riesz representation theorem).

- ℓ^p and $L^p(\mathbb{R}^d)$ are Banach spaces for $1 \leq p \leq \infty$.

- Their dual spaces are ℓ^q and $L^q(\mathbb{R}^d)$ with

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Examples

- Each Hilbert space $\mathcal{H}$ is a reflexive Banach space and

$$\mathcal{H}^* = \mathcal{H}$$

  (Riesz representation theorem).

- ℓ^p and $L^p(\mathbb{R}^d)$ are Banach spaces for $1 \leq p \leq \infty$.

- Their dual spaces are ℓ^q and $L^q(\mathbb{R}^d)$ with

$$\frac{1}{p} + \frac{1}{q} = 1.$$

- ℓ^1/ℓ^∞ and $L^1(\mathbb{R}^d)/L^\infty(\mathbb{R}^d)$ are not reflexive.

Radon measures

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

Radon measures

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

- A Radon measure positive and finite on $\mathbb{R}^d$ (*"unnormalized probability measure"*).

Radon measures

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

- A Radon measure positive and finite on $\mathbb{R}^d$ (*"unnormalized probability measure"*).
- A signed Radon measure μ can be written as

$$\mu = \mu^+ - \mu^-,$$

with μ^+, μ^- positive Radon measures (Jordan decomposition).

Radon measures

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

- A Radon measure positive and finite on $\mathbb{R}^d$ (*"unnormalized probability measure"*).
- A signed Radon measure μ can be written as

$$\mu = \mu^+ - \mu^-,$$

  with μ^+, μ^- positive Radon measures (Jordan decomposition).
- The total variation norm is

$$\|\mu\|_{\mathsf{TV}} = |\mu|(\mathbb{R}^d), \qquad |\mu| = \mu^+ + \mu^-.$$

Integral forms

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

Integral forms

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

- $\mu \in M(\mathbb{R}^d)$, *signed Radon measures* with norm $\|\mu\|_{\mathrm{TV}}$.

Integral forms

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

- $\mu \in M(\mathbb{R}^d)$, *signed Radon measures* with norm $\|\mu\|_{\mathrm{TV}}$.
- $f \in C_0(\mathbb{R}^d)$, continuous functions vanishing at infinity with norm $\|f\|_\infty$.

Integral forms (cont.)

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

Integral forms (cont.)

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

- $C_0(\mathbb{R}^d)$ is a Banach space and $M(\mathbb{R}^d)$ its dual (Riesz–Markov theorem).

Integral forms (cont.)

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

- $C_0(\mathbb{R}^d)$ is a Banach space and $M(\mathbb{R}^d)$ its dual (Riesz–Markov theorem).
- $C_0(\mathbb{R}^d)$ is contained in the dual of $M(\mathbb{R}^d)$.

Integral forms (cont.)

$$\langle f, \mu \rangle = \int_{\mathbb{R}^d} f(\theta)\, d\mu(\theta)$$

- $C_0(\mathbb{R}^d)$ is a Banach space and $M(\mathbb{R}^d)$ its dual (Riesz–Markov theorem).
- $C_0(\mathbb{R}^d)$ is contained in the dual of $M(\mathbb{R}^d)$.
- Not reflexive.

Outline

Infinite networks with L^2 norm

Infinite width limit

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

[5] 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 networks with L^2 norm

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 [5] $L^2(\mathbb{R}^d \times \mathbb{R})$ (not necessarily a density).

[5] 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 networks with L^2 norm

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 [5] $L^2(\mathbb{R}^d \times \mathbb{R})$ (not necessarily a density).
- σ defines a function $\Phi(x) \in L^2(\mathbb{R}^d \times \mathbb{R})$ for each $x \in \mathbb{R}^d$, that is

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

[5] 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 networks with L^2 norm

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 [5] $L^2(\mathbb{R}^d \times \mathbb{R})$ (not necessarily a density).
- σ defines a function $\Phi(x) \in L^2(\mathbb{R}^d \times \mathbb{R})$ for each $x \in \mathbb{R}^d$, that is

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

- Linear parametrization

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

[5] 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 networks with L^1 norm

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

[6] More precisely, $L^1(\mathbb{R}^d \times \mathbb{R}, \tau)$ with norm $\int_{\mathbb{R}^d \times \mathbb{R}} |v(w,b)|\, d\tau(w,b) < \infty$.

[7] Using the duality pairing between $L^1(\mathbb{R}^d \times \mathbb{R})$ and $L^\infty(\mathbb{R}^d \times \mathbb{R})$.

Infinite networks with L^1 norm

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

- v is a function in [6] $L^1(\mathbb{R}^d \times \mathbb{R})$ (not necessarily a density).

[6] More precisely, $L^1(\mathbb{R}^d \times \mathbb{R}, \tau)$ with norm $\int_{\mathbb{R}^d \times \mathbb{R}} |v(w,b)|\, d\tau(w,b) < \infty$.
[7] Using the duality pairing between $L^1(\mathbb{R}^d \times \mathbb{R})$ and $L^\infty(\mathbb{R}^d \times \mathbb{R})$.

Infinite networks with L^1 norm

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

- v is a function in [6] $L^1(\mathbb{R}^d \times \mathbb{R})$ (not necessarily a density).
- σ defines a function $\Phi(x) \in L^\infty(\mathbb{R}^d \times \mathbb{R})$ for each $x \in \mathbb{R}^d$, that is

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

[6] More precisely, $L^1(\mathbb{R}^d \times \mathbb{R}, \tau)$ with norm $\int_{\mathbb{R}^d \times \mathbb{R}} |v(w,b)|\, d\tau(w,b) < \infty$.
[7] Using the duality pairing between $L^1(\mathbb{R}^d \times \mathbb{R})$ and $L^\infty(\mathbb{R}^d \times \mathbb{R})$.

Infinite networks with L^1 norm

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

- v is a function in [6] $L^1(\mathbb{R}^d \times \mathbb{R})$ (not necessarily a density).
- σ defines a function $\Phi(x) \in L^\infty(\mathbb{R}^d \times \mathbb{R})$ for each $x \in \mathbb{R}^d$, that is

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

- Linear parametrization[7]

$$f_v(x) = \langle v, \Phi(x) \rangle .$$

[6] More precisely, $L^1(\mathbb{R}^d \times \mathbb{R}, \tau)$ with norm $\int_{\mathbb{R}^d \times \mathbb{R}} |v(w,b)|\, d\tau(w,b) < \infty$.
[7] Using the duality pairing between $L^1(\mathbb{R}^d \times \mathbb{R})$ and $L^\infty(\mathbb{R}^d \times \mathbb{R})$.

Infinite networks with TV norm

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

[8] Using the duality pairing between $M(\mathbb{R}^d)$ and $C_0(\mathbb{R}^d \times \mathbb{R})$.

Infinite networks with TV norm

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

- μ is a measure in $M(\mathbb{R}^d)$ with norm $\|\mu\|_{\mathrm{TV}}$.

[8] Using the duality pairing between $M(\mathbb{R}^d)$ and $C_0(\mathbb{R}^d \times \mathbb{R})$.

Infinite networks with TV norm

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

- μ is a measure in $M(\mathbb{R}^d)$ with norm $\|\mu\|_{\text{TV}}$.

  σ defines a function $\Phi(x) \in C_0(\mathbb{R}^d \times \mathbb{R})$ for each $x \in \mathbb{R}^d$, that is

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

[8] Using the duality pairing between $M(\mathbb{R}^d)$ and $C_0(\mathbb{R}^d \times \mathbb{R})$.

Infinite networks with TV norm

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

- μ is a measure in $M(\mathbb{R}^d)$ with norm $\|\mu\|_{\mathrm{TV}}$.

  σ defines a function $\Phi(x) \in C_0(\mathbb{R}^d \times \mathbb{R})$ for each $x \in \mathbb{R}^d$, that is

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

- Linear parametrization[8]

  $$f_\mu(x) = \langle \mu, \Phi(x) \rangle .$$

[8] Using the duality pairing between $M(\mathbb{R}^d)$ and $C_0(\mathbb{R}^d \times \mathbb{R})$.

Remarks

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

Remarks

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

- No *"density"* needed.

Remarks

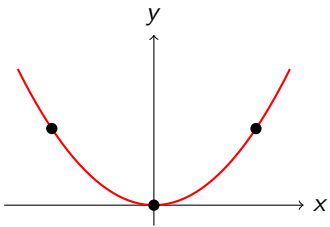
$$f_v(x) = \int_{\mathbb{R}^d \times \mathbb{R}} v(w,b)\, \sigma(w^\top x + b)\, d\tau(w,b) \qquad \text{v.s.} \qquad f_\mu(x) = \int_{\mathbb{R}^d \times \mathbb{R}} \sigma(w^\top x + b)\, d\mu(w,b).$$

- No *"density"* needed.
- Atomic measures and finite networks,

$$\mu = \sum_{j=1}^{m} v_j \delta_{(w_j^T, b_j)} \qquad \mapsto \qquad f_\mu(x) = \sum_{j=1}^{m} v_j \sigma(w_j^T x + b_j).$$

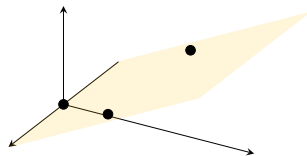
Neural networks and feature maps

$$f_v(x) = \langle \mu, \Phi(x) \rangle .$$



input space $\mathcal{X}$

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



feature space L^∞ or $\mathcal{C}_0$

Outline

Banach feature maps

Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ be a Banach space and

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

$\mathcal{F}^*$ is called a Banach feature space and Φ a Banach feature map [9].

[9] Parameter space and feature space do not coincide!

Banach feature maps and hyperplanes

Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ be a Banach space and

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

Consider functions of the form

$$f(x) = \langle v, \Phi(x) \rangle, \qquad v \in \mathcal{F},$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $\mathcal{F}$ and $\mathcal{F}^*$.

Feature maps, hyperplanes and norms

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

Define the norm

$$\|f\|_{\mathcal{B}} = \inf \left\{ \|v\|_{\mathcal{F}} \;\middle|\; f(x) = \langle v, \Phi(x) \rangle, \ \forall x \in \mathcal{X} \right\}.$$

Banach space of hyperplanes

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

$$f(x) = \langle v, \Phi(x) \rangle, \qquad v \in \mathcal{F}^*$$

with norm

$$\|f\|_{\mathcal{B}} = \inf \left\{ \|v\|_{\mathcal{F}^*} \;\middle|\; f(x) = \langle v, \Phi(x) \rangle, \ \forall x \in \mathcal{X} \right\}$$

is a Banach space.

RKBS and evaluation functionals

The space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ is such that for all $f \in \mathcal{B}$ and $x \in \mathcal{X}$

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

RKBS and evaluation functionals

The space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ is such that for all $f \in \mathcal{B}$ and $x \in \mathcal{X}$

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

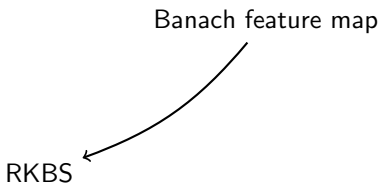
A Banach space of functions with the above property is called a Reproducing Kernel Banach Space (RKBS).

Banach feature maps and RKBS

Banach feature map

RKBS

Banach feature maps and RKBS



RKBS and evaluation functionals

The space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ is such that for all $f \in \mathcal{B}$ and $x \in \mathcal{X}$

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

RKBS and evaluation functionals

The space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ is such that for all $f \in \mathcal{B}$ and $x \in \mathcal{X}$

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

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

RKBS and evaluation functionals

The space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ is such that for all $f \in \mathcal{B}$ and $x \in \mathcal{X}$

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

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 $\delta_x \in \mathcal{B}^*$ such that

$$f(x) = \langle \delta_x, f \rangle,$$

(by the definition of the duality pairing).

RKBS and Banach feature maps

Each RKBS $\mathcal{B}$ can have multiple associated feature maps.

RKBS and Banach feature maps

Each RKBS $\mathcal{B}$ can have multiple associated feature maps.

- Canonical feature map:

$$\mathcal{F}^* = \mathcal{B}^*, \qquad \Phi(x) = \delta_x, \quad x \in \mathcal{X},$$

where $\delta_x \in \mathcal{B}^*$ is the evaluation functional.

RKBS and Banach feature maps

Each RKBS $\mathcal{B}$ can have multiple associated feature maps.

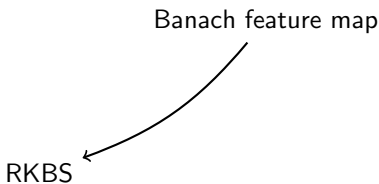
- Canonical feature map:

$$\mathcal{F}^* = \mathcal{B}^*, \qquad \Phi(x) = \delta_x, \quad x \in \mathcal{X},$$

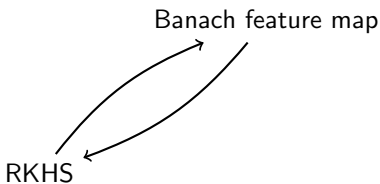
  where $\delta_x \in \mathcal{B}^*$ is the evaluation functional.

- Alternative feature maps can be constructed in specific examples.

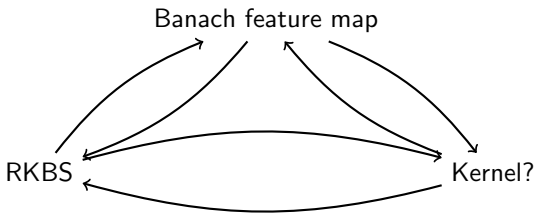
RKBS and Banach feature maps



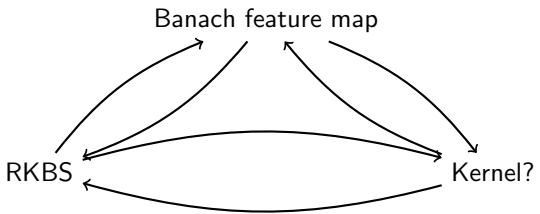
RKBS and Banach feature maps



What about kernels?

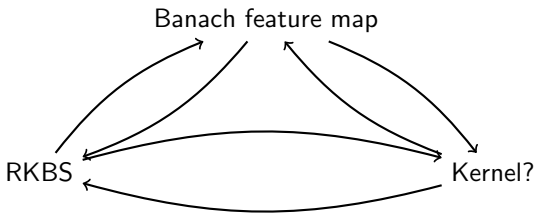


What about kernels?



- In RKHS, $\mathcal{H} = \mathcal{H}^*$ and $k(x, x') = \langle \delta_x, \delta_{x'} \rangle_{\mathcal{H}}$ is symmetric and positive definite.

What about kernels?



- In RKHS, $\mathcal{H} = \mathcal{H}^*$ and $k(x, x') = \langle \delta_x, \delta_{x'} \rangle_{\mathcal{H}}$ is symmetric and positive definite.

- In RKBS, $\mathcal{B} \neq \mathcal{B}^*$ and $k(x, x') = \langle \delta_x, \delta^*_{x'} \rangle$ is in general neither symmetric nor positive definite.

RKBS

There's no kernel in RKBS.

RKBS

There's no kernel in RKBS.

Hence no "dual" view.

RKBS

There's no kernel in RKBS.

Hence no "dual" view.

RKBS provide mathematical structure, one byproduct is a representer theorem.

RKBS

There's no kernel in RKBS.

Hence no "dual" view.

RKBS provide mathematical structure, one byproduct is a representer theorem.

Back to neural networks first.

Outline

L^2 RKHS

Consider an RKHS $\mathcal{H}$ of functions of the form

$$f(x) = \langle \mu, \Phi(x) \rangle = \int_{\mathbb{R}^d \times \mathbb{R}} v(w,b)\ \sigma(w^\top x + b)\, d\tau(w,b)$$

with norm

$$\|f\|_{\mathcal{H}} = \inf \left\{ \|v\|_{L^2(\mathbb{R}^d \times \mathbb{R})} \;\middle|\; f(x) = \langle \mu, \Phi(x) \rangle,\ \forall x \in \mathcal{X} \right\}.$$

L^1 RKBS

Consider an RKBS $\mathcal{B}_1$ of functions of the form

$$f(x) = \langle \mu, \Phi(x) \rangle = \int_{\mathbb{R}^d \times \mathbb{R}} v(w,b)\ \sigma(w^\top x + b)\, d\tau(w,b)$$

with norm

$$\|f\|_{\mathcal{B}_1} = \inf \left\{ \|v\|_{L^1(\mathbb{R}^d \times \mathbb{R})} \;\middle|\; f(x) = \langle \mu, \Phi(x) \rangle,\ \forall x \in \mathcal{X} \right\}.$$

TV RKBS

Consider an RKBS $\mathcal{B}$ of functions of the form

$$f(x) = \langle \mu, \Phi(x) \rangle = \int_{\mathbb{R}^d \times \mathbb{R}} \sigma(w^\top x + b)\, d\mu(w, b)$$

with norm

$$\|f\|_{\mathcal{B}} = \inf \left\{ \|\mu\|_{\mathrm{TV}} \;\middle|\; f(x) = \langle \mu, \Phi(x) \rangle,\ \forall x \in \mathcal{X} \right\}.$$

Outline

TV RKBS

Consider a TV RKBS of functions of the form

$$f(x) = \langle \mu, \Phi(x) \rangle = \int_{\mathbb{R}^d \times \mathbb{R}} \sigma(w^\top x + b)\, d\mu(w, b).$$

with norm

$$\|f\|_{\mathcal{B}} = \inf \left\{ \|\mu\|_{\mathrm{TV}} \;\middle|\; f(x) = \langle \mu, \Phi(x) \rangle,\ \forall x \in \mathcal{X} \right\}.$$

Representer theorem in RKBS

Let $\mathcal{B}$ be the ReLU TV RKBS and $\ell(y, \cdot)$ convex and coercive for all $y \in \mathbb{R}$. Consider the regularized ERM problem,

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

Representer theorem in RKBS

Let $\mathcal{B}$ be the ReLU TV RKBS and $\ell(y, \cdot)$ convex and coercive for all $y \in \mathbb{R}$. Consider the regularized ERM problem,

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

A minimizer of the above functional has the form

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

What's the value of this representer theorem?

What's the value of this representer theorem?

- Unlike the one for RKHS, it does not provides computational insights.

What's the value of this representer theorem?

- Unlike the one for RKHS, it does not provides computational insights.

- It shows that finite networks of width m are optimal when learning with ERM.

What's the value of this representer theorem?

- Unlike the one for RKHS, it does not provides computational insights.
- It shows that finite networks of width m are optimal when learning with ERM.
- It suggests that TV RKBS norm is the inductive bias for finite networks.

Equivalent norms

$$\min_{f \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{B}} \quad \Leftrightarrow \quad \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \lambda R(\theta).$$

Equivalent norms

$$\min_{f \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{B}} \quad \Leftrightarrow \quad \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \lambda R(\theta).$$

For ReLU the following regularization are equivalent:

- ℓ^2 weight decay

$$R(\theta) = \sum_{j=1}^{m} (|v_j|^2 + \|w_j\|^2 + |b_j|^2)$$

Equivalent norms

$$\min_{f \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{B}} \quad \Leftrightarrow \quad \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \lambda R(\theta).$$

For ReLU the following regularization are equivalent:

- ℓ^2 weight decay

$$R(\theta) = \sum_{j=1}^{m} (|v_j|^2 + \|w_j\|^2 + |b_j|^2)$$

- Path norm

$$R(\theta) = \sum_{j=1}^{m} |v_j| \, \|w_j\| \, |b_j|.$$

Equivalent norms

$$\min_{f \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{B}} \quad \Leftrightarrow \quad \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \lambda R(\theta).$$

For ReLU the following regularization are equivalent:

- ℓ^2 weight decay

$$R(\theta) = \sum_{j=1}^{m} (|v_j|^2 + \|w_j\|^2 + |b_j|^2)$$

- Path norm

$$R(\theta) = \sum_{j=1}^{m} |v_j| \, \|w_j\| \, |b_j|.$$

- ℓ^1 with normalized hidden weights,

$$R(\theta) = \sum_{j=1}^{m} |v_j|, \qquad \|w_j\| = 1, |b_j| = 1, \quad j = 1, \ldots, m.$$

Outline

Norm characterization for ReLU

Norm characterization for ReLU

For ReLU, the RKBS norm can be characterized using the Radon transform

$$\|f\|_{\mathcal{B}} \approx \|DTR\ f\|_{\mathrm{TV}}\,.$$

Norm characterization for ReLU

For ReLU, the RKBS norm can be characterized using the Radon transform

$$\|f\|_{\mathcal{B}} \approx \|DTR\ f\|_{\mathrm{TV}}\,.$$

- D is the Laplacian on b.
- T a suitable truncation operator.
- R the Radon transform,

$$(Rf)(w,b) = \int_{\{x\in\mathbb{R}^d : w^\top x = b\}} f(x)\, d\sigma(x).$$

Some intuitions

Analysis

$$\mu \quad \mapsto \quad f_\mu(x) = \int |w^\top x + b|_+ d\mu(w, b)$$

Some intuitions

Analysis

$$\mu \quad \mapsto \quad f_\mu(x) = \int |w^\top x + b|_+ d\mu(w, b)$$

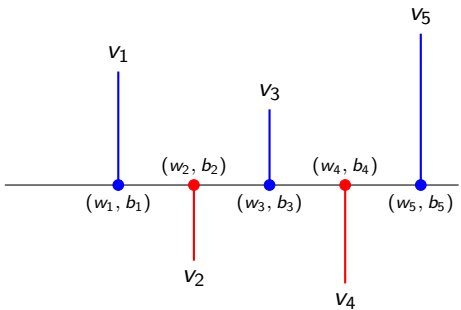
Synthesis

$$f \quad \mapsto \quad \mu = DTR\ f$$

Some intuitions (cont.)

Norm induces sparse measures

$$\|f\|_{\mathcal{B}} \approx \|DTRf\|_{\mathrm{TV}}\,.$$



Barron spaces & universality

$$\mathcal{H} \subset \mathcal{B}_1 \subset \mathcal{B}.$$

Barron spaces & universality

$$\mathcal{H} \subset \mathcal{B}_1 \subset \mathcal{B}.$$

- $\mathcal{H} \approx$ Sobolev spaces, already universal.

Barron spaces & universality

$$\mathcal{H} \subset \mathcal{B}_1 \subset \mathcal{B}.$$

- $\mathcal{H} \approx$ Sobolev spaces, already universal.
- $\mathcal{B}$ related to so called Barron spaces.

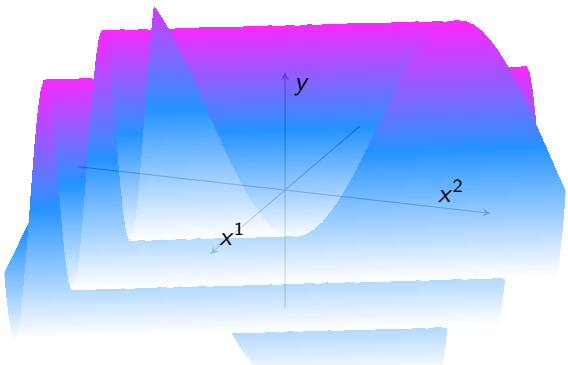
Barron spaces & universality

$$\mathcal{H} \subset \mathcal{B}_1 \subset \mathcal{B}.$$

- $\mathcal{H} \approx$ Sobolev spaces, already universal.
- $\mathcal{B}$ related to so called Barron spaces.
- $\mathcal{B}$ spaces contain more...

Multi-index models

$$f(x) = g(Bx), \quad B \in \mathbb{R}^{k \times d}, \quad k \ll d.$$



Summing up

- Infinite width neural networks can be associated with RKBS: more complex mathematical structure!
- No kernels... but linear normed spaces and an interesting representer theorem nonetheless.
- RKBS norm characterizations can give insight into inductive bias.
- Universality can be established.
- Few recent extensions for deep networks.

Summing up (cont.)

- This theory suggests a gap between kernel and neural networks.
- It suggests neural networks can adapt to learn function beyond Sobolev spaces.
- It does not address the compositional structure which is (probably) the key behind the success of convolutional networks/transformers!