Lecture 8: Non-parametric regression

Announcements

- HW 1 solutions are up, grades released next week
- Project proposals due end of today
- Lecture slides for this week and Friday will be updated by end of this week - apologies

Plan for today

- Non-parametric prediction error bound
    - Intuition for critical radius
    - Examples: sparse linear regression, Lipschitz
- Example non-parametric function space: Reproducing kernel Hilbert spaces (RKHS)
- Recap of kernels and examples for RKHS
- Friday: prediction error bound for RKHS

Recap: (Non-)parametric regression setting

- Square loss and constrained regression
- Fixed design, i.e. only care about prediction on training inputs $x_1, \dots, x_n$
- Gaussian observation noise, i.e. $W = Y - f^\star(X) \in \mathcal{N}(0, \sigma^2)$
- Today, analyze minimizer of the square loss
  $\widehat{f} = \arg\min_{f \in \mathcal{F}} R_n(f) := \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2$
  (and later also with penalty
  $\widehat{f} = \arg\min_{f \in \mathcal{F}} R_n(f) := \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}$)
- Evaluation: Prediction error of some f on fixed design points

$$\|f - f^\star\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f(x_i) - f^\star(x_i))^2 = \mathbb{E}_Y R_n(f) - \sigma^2 = R(f) - R(f^\star)$$

Recap: Motivation for localized Gaussian complexity

- Define $\hat{\Delta} = \widehat{f} - f^\star$ for simplicity, and the space $\mathcal{F}^\star = \{f - f^\star : f \in \mathcal{F}\}$
- Furthermore we assume that $\mathcal{F}^\star$ is star-shaped, i.e. for any $f \in \mathcal{F}^\star$, we have $\alpha f \in \mathcal{F}^\star$ for all $\alpha \in [0,1]$

1. Space to control is smaller than all of $\mathcal{F}^\star$ since either
   (i) $\|\hat{\Delta}\|_n \leq \delta_n$ or
   (ii) if $\|\hat{\Delta}\|_n \geq \delta_n$ then still $\|\hat{\Delta}\|_n^2 \leq \frac{2\sigma}{n}\sum_{i=1}^n w_i \hat{\Delta}(x_i)$ by basic inequality

2. Further for case (ii), if can show w.h.p.

$$\frac{2\sigma}{n}\sum_{i=1}^n w_i \hat{\Delta}(x_i) \leq 4\|\hat{\Delta}\|_n \delta_n \tag{1}$$

for all $\|\hat{\Delta}\|_n \geq \delta_n$ then we can plug that into RHS of (ii) to obtain $\|\hat{\Delta}\|_n \leq 4\delta_n$ w.h.p.

For which δ_n 2. is true

a. By star-shaped assumption on $\mathcal{F}^\star$ step (i) holds in the following:

$$\iff \sup_{\|\hat{\Delta}\|_n \geq \delta_n, \hat{\Delta} \in \mathcal{F}^\star} \frac{\sigma}{n} \sum_{i=1}^n w_i \frac{\hat{\Delta}(x_i)}{\|\hat{\Delta}\|_n} = \sup_{\|\hat{\Delta}\|_n \geq \delta_n, \hat{\Delta} \in \mathcal{F}^\star} \frac{\sigma}{n} \sum_{i=1}^n w_i \underbrace{\frac{\hat{\Delta}(x_i)\delta_n}{\|\hat{\Delta}\|_n}}_{=:\tilde{\Delta}} \frac{1}{\delta_n}$$

$$\stackrel{(i)}{=} \sup_{\|\tilde{\Delta}\|_n = \delta_n, \tilde{\Delta} \in \mathcal{F}^\star} \frac{\sigma}{n} \sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\delta_n} \leq \sup_{\|\tilde{\Delta}\|_n \leq \delta_n, \tilde{\Delta} \in \mathcal{F}^\star} \frac{\sigma}{n} \sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\delta_n}$$

b. eq. 1 follows from h.p. bound of this (localized) quantity

$$\sup_{\substack{\|\hat{\Delta}\|_n \leq \delta_n \\ \hat{\Delta} \in \mathcal{F}^\star}} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \leq \mathbb{E} \sup_{\substack{\|\hat{\Delta}\|_n \leq \delta_n \\ \hat{\Delta} \in \mathcal{F}^\star}} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) + \delta_n^2$$

and if the expectation is bounded, i.e.

$$\mathbb{E} \sup_{\substack{\|\hat{\Delta}\|_n \leq \delta_n \\ \hat{\Delta} \in \mathcal{F}^\star}} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \leq \delta_n^2$$

Localized Gaussian complexity

Definition (Localized (empirical) Gaussian complexity)

The localized Gaussian complexity around $f^\star$ of scale δ is

$$\sigma \widetilde{\mathcal{G}}_n(\mathcal{F}^\star; \delta_n) := \sigma \widetilde{\mathcal{G}}_n(\mathcal{F}^\star(x_1^n) \cap \mathbb{B}_n(\delta_n)) = \mathbb{E} \sup_{\|\hat{\Delta}\|_n \leq \delta_n, \hat{\Delta} \in \mathcal{F}^\star} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i)$$

- Hence: Given concentration b., eq. 1, i.e. $\|\hat{\Delta}\|_n \leq 4\delta_n$ holds for all δ_n that satisfy the implicit inequality $\sigma \widetilde{\mathcal{G}}_n(\mathcal{F}^\star; \delta_n) \leq \delta_n^2$
- You can rewrite and say: $\|\hat{\Delta}\|_n \leq 4\sqrt{t}\delta_n$ holds for any $t \geq 1$ w.h.p. if δ_n is the **smallest** $\delta > 0$ such that $\sigma \widetilde{\mathcal{G}}_n(\mathcal{F}^\star; \delta) \leq \delta^2$
- All that's left to do: see that δ_n exists and show b.

Lemma (Critical radius (MW 13.6.))

For any star-shaped $\mathcal{F}$, it holds that $\frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\delta)}{\delta}$ is non-increasing and the critical inequality

$$\frac{\widetilde{\mathcal{G}}_n(\mathcal{F}; \delta)}{\delta} \leq \frac{\delta}{\sigma}$$

has a smallest solution $\delta_n > 0$ that we call the critical quantity/radius.

Illustration of localized Gaussian complexity

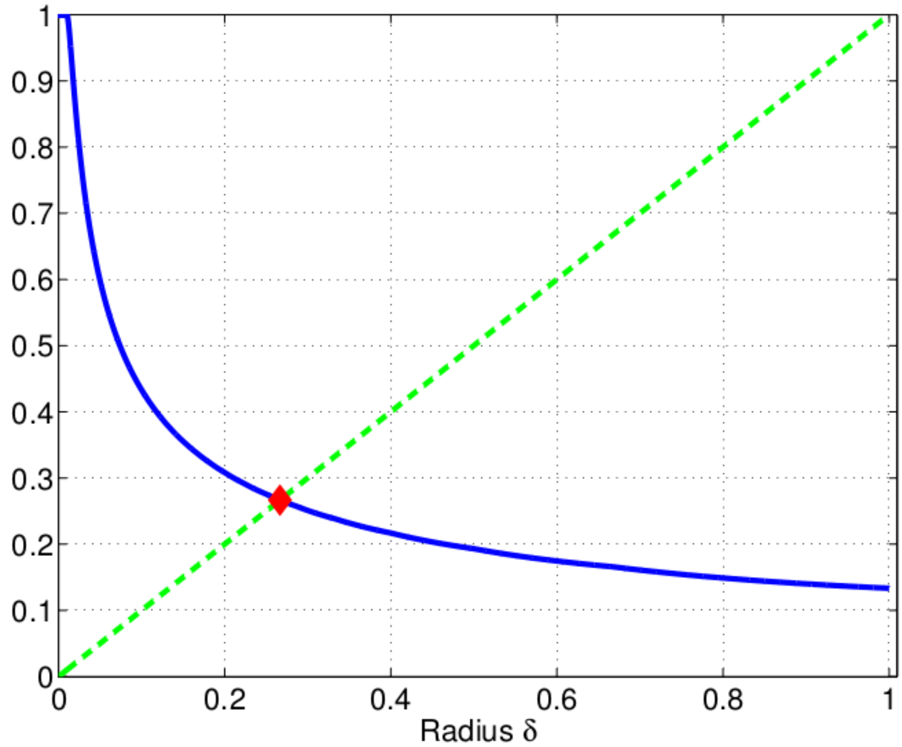

Figure 1: Blue solid: $f(\delta) = \frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\delta)}{\delta}$, Green dashed: $f(\delta) = \delta$

Prediction error bound for constrained ^{2}-loss minimizer

Theorem (Prediction error bound, MW Thm 13.5.)

If $\mathcal{F}^\star$ is star-shaped, we have for the square loss minimizer $\widehat{f}$ for any $t \geq 1$

$$\mathbb{P}(\|\widehat{f} - f^\star\|_n^2 \geq 16t\delta_n^2) \leq e^{-\frac{nt\delta_n^2}{2\sigma^2}}$$

- Plugging in $t = O(\log \frac{1}{\delta})$ and by $\delta_n^2 \geq O(\frac{1}{n})$ (check yourself) yields that probability at least $1 - \delta$ we have $\|\widehat{f} - f^\star\|_n^2 \leq O(\log(\frac{1}{\delta})\delta_n^2)$
- As $f^\star$ is unknown, can replace $\widetilde{\mathcal{G}}_n(\mathcal{F}^\star;\delta)$ by $\widetilde{\mathcal{G}}_n(\mathcal{F} - \mathcal{F};\delta)$ (or its star hull MW Eq (13.21.)) to define critical radius δ_n
- Note: the notation for t is different from MW Thm 13.5.
- Proof follows by proof of (modified) b. and noting that $g_n(w) = \sup_{\|\hat{\Delta}\|_n \leq \sqrt{t}\delta_n} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i)$ is a Lipschitz function of Gaussians and using MW Thm 2.26 (next slide, skipped in class)

Proof of error bound: tail bounding $g_n(w)$ (skipped)

We now establish the tail bound for $g_n(w)$

1. $g_n(w)$ as a function of $w_i \sim \mathcal{N}(0,1)$ is $\frac{\sigma\sqrt{t}\delta_n}{\sqrt{n}}$-Lipschitz so that

   $\mathbb{P}(g_n(w) \geq \mathbb{E}g_n(w) + s) \leq e^{-\frac{ns^2}{2\sigma^2 t\delta_n^2}}$ (see Lecture 2 / MW Thm 2.26)

2. Furthermore $\mathbb{E}g_n(w) = \widetilde{\mathcal{G}}_n(\mathcal{F}; \sqrt{t}\delta_n)$

3. The map $\delta \to \frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\delta)}{\delta}$ is non-increasing by MW Lemma 13.6.

4. By 2. and definition of δ_n we have $\sigma\frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\sqrt{t}\delta_n)}{\sqrt{t}\delta_n} \leq \sigma\frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\delta_n)}{\delta_n} \leq \delta_n$ and setting $s = t\delta_n^2$, we obtain

$$\mathbb{P}\Big(\sup_{\|\hat{\Delta}\|_n \leq \sqrt{t}\delta_n} \frac{\sigma}{n}\sum_{i=1}^n w_i\hat{\Delta}(x_i) \geq 2t\delta_n^2\Big)$$

$$\leq \mathbb{P}\Big(\sup_{\|\hat{\Delta}\|_n \leq \sqrt{t}\delta_n} \frac{\sigma}{n}\sum_{i=1}^n w_i\hat{\Delta}(x_i) \geq \sigma\widetilde{\mathcal{G}}_n(\mathcal{F}; \sqrt{t}\delta_n) + t\delta_n^2\Big) \leq e^{-\frac{nt\delta_n^2}{2\sigma^2}} \square$$

Application 1: ℓ_0-constrained sparse linear regression

Let's say we're trying to find the best sparse linear fit

$$\widehat{f} = \arg\min_{f\in\mathcal{F}_{lin,s}} \|y - X\theta\|_n^2$$

with $\mathcal{F}_{lin,s} = \{f(\cdot) = \langle\theta, x\rangle : \|\theta\|_0 \leq s\}$

- In HW 2 we prove $\widetilde{\mathcal{G}}_n(\mathcal{F}_{lin,s};\delta) \leq O(\delta\sqrt{\frac{s\log(ed/s)}{n}})$ when $\lambda_{\max}(\frac{X_S^\top X_S}{n})$ bounded for all subsets S of size s
- Hence the critical radius has to satisfy $\frac{\widetilde{\mathcal{G}}_n(\mathcal{F}_{lin,s};\delta)}{\delta} = \sqrt{\frac{s\log(ed/s)}{n}} \leq \frac{\delta_n}{\sigma}$
- Thus using the theorem, plugging in δ_n^2 at equality, we can obtain with probability at least $1-\delta$

$$\|\widehat{f} - f^\star\|_n^2 \leq O\left(\frac{s\log(ed/s)\log 1/\delta}{n}\right)$$

Also see MW Example 13.16.

General functions via Dudley's integral

Corollary (Dudley's integral & critical quantity - MW Cor. 13.7.)

If $\mathcal{F}$ is star-shaped, any $\delta \in [0, \sigma]$ such that

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}^\star(x_1^n) \cap \mathbb{B}_n(\delta), \|\cdot\|_n)} dt \leq \frac{\delta^2}{4\sigma}$$

satisfies the critical inequality.

Proof via chaining for localized Gaussian complexity for a $\frac{\delta^2}{4\sigma}$ cover

$$\widetilde{\mathcal{G}}_n(\mathcal{F}^\star; \delta) \leq \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}^\star(x_1^n) \cap \mathbb{B}_n(\delta), \|\cdot\|_n)} \mathrm{d}t + \frac{\delta^2}{4\sigma}$$

(skipped in class)

Application 2: General functions via Dudley's integral

1. $\mathcal{F}_L$: Lipschitz functions on $[0,1]$ and $f(0) = 0$ has $\log \mathcal{N}(\epsilon) \leq O(\frac{L}{\epsilon})$

$$\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}_L(x_1^n), \|\cdot\|_n)} \mathrm{d}t \leq \frac{1}{\sqrt{n}} \int_0^{\delta} \left(\frac{L}{t}\right)^{\frac{1}{2}} \mathrm{d}t \quad \leq \sqrt{\frac{L\delta}{n}} \overset{(!)}{\leq} \frac{\delta^2}{4\sigma^2}$$

$\rightarrow$ Rearranging terms yields $\|\widehat{f} - f^\star\|_n^2 \leq \delta_n(\mathcal{F}_L)^2 = O(\frac{L\sigma^2}{n})^{\frac{2}{3}}$

Recall how for Lipschitz functions, the "unlocalized" Dudley bound from last lec. yields $\|\widehat{f} - f^\star\|_n^2 \leq O(\frac{1}{n^{1/2}}) \rightarrow$ slower!

2. $\mathcal{F}_{1,c}$: $f \in \mathcal{F}_1$ **and** convex, has $\log \mathcal{N}(\epsilon) \leq O((\frac{1}{\epsilon})^{\frac{1}{2}})$

$$\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}_{1,c}(x_1^n), \|\cdot\|_n)} \mathrm{d}t \leq \frac{1}{\sqrt{n}} \int_0^{\delta} \left(\frac{1}{t}\right)^{\frac{1}{4}} \mathrm{d}t \quad \leq \frac{\delta^{3/4}}{\sqrt{n}} \overset{(!)}{\leq} \frac{\delta^2}{4\sigma^2}$$

$\rightarrow$ Rearranging terms yields $\delta_n(\mathcal{F}_{1,c})^2 = O((\frac{\sigma^2}{n})^{\frac{4}{5}})$

Dudley's integral in localized vs. "global" form

Comparison of how $\delta_n(\mathcal{F})$ vs. $\mathcal{R}_n(\mathcal{F})$ reflect function size differently, though in both cases we use Dudley:

- $\delta_n(\mathcal{F})$: Critical quantity reflects difference in metric entropy (size)
- $\mathcal{R}_n(\mathcal{F})$ via Dudley: If integrals $\int_0^D \sqrt{\log \mathcal{N}(t; \mathcal{F}(x_1^n), \|\cdot\|_n)} dt$ are bounded, then best is to use that and R.C. gets $\frac{1}{\sqrt{n}}$ rate. (check)
  $\rightarrow$ For both integrals are bounded, Rademacher complexity has $\frac{1}{\sqrt{n}}$
  $\rightarrow$ does not reflect size difference compared to $\delta_n(\mathcal{F})$!
- Reason: localized complexity by definition is smaller than global complexity because of extra restriction on $\|\hat{\Delta}\|_n$ norm:

$$\widetilde{\mathcal{G}}_n(\mathcal{F}^\star; \delta_n) = \mathbb{E} \sup_{\|\hat{\Delta}\|_n \leq \delta_n, \hat{\Delta} \in \mathcal{F}^\star} \frac{1}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i)$$

  where $\mathcal{F}^\star$ is "morally as large as $\mathcal{F}$"

Non-parametric regression for kernel spaces $\mathcal{F}$

- Motivation 1: Non-parametric regression specific function spaces $\mathcal{F}$ for which we can actually find global minimizer $\widehat{f}$?
- Motivation 2: Intro to ML course: *implementable* transition from linear to featurized regression via kernel trick
- Motivation 3: From research: one standard way to think about NN is that it's just doing kernel regression in an RKHS. Actually, convolutional neural tangent kernels (based on NN) can predict CIFAR10 with ~90% test accuracy

Reproducing Kernel Hilbert spaces (RKHS) are nice (in low dimensions) because we have good analysis tools to get bounds (can even use to approximate neural networks)

Caveats/limits: "fail" for high-dimensional data (ask us if interested), only hold for close to initialization for neural networks

Plan for now

- RKHS primer:
    - Definition
    - RKHS via kernels
    - Representer theorem
- From function space to RKHS (Examples)
- Next time: RKHS as an example for non-parametric prediction error bounds

Reproducing Kernel Hilbert spaces

For generic (say e.g. Lipschitz, or non-decreasing) function spaces its super complicated to search in since infinite dimensional

$\rightarrow$ RKHS have nice reproducing property that enables efficient search since one can write solution easily in closed form with matrix vectors

Recall: Hilbert space $\mathcal{F}$ with $f : \mathcal{X} \rightarrow \mathbb{R}$ is a vector space with

- a valid inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ that is symmetric, additive
- $\langle f, f \rangle_{\mathcal{F}} \geq 0$ for all f, equality iff $f = 0$

Definition (Reproducing kernel Hilbert space - MW Def 12.12.)

A Hilbert space with $f : \mathcal{X} \rightarrow \mathbb{R}$ with evaluation functional that is bounded and linear, i.e. for all $x \in \mathcal{X}$ there exists $L_x : \mathcal{F} \rightarrow \mathbb{R}$ with $L_x(f) = f(x)$ and $|L_x(f)| \leq M_x \|f\|_{\mathcal{F}}$ for all $f \in \mathcal{F}$ for some $M_x < \infty$

$\rightarrow$ can (i) design RKHS via a kernel directly, or (ii) take Hilbert space satisfying abstract definition in last slide and find kernel "in hindsight"

(i) RKHS induced by kernels (recap)

Definition (Reminder - psd kernels)

A bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a valid kernel iff $\mathcal{K}$ is symmetric and psd, i.e. for $x_1, \dots x_n$, kernel matrix $K \in \mathbb{R}^{n \times n}$ with $K_{ij} := \mathcal{K}(x_i, x_j)$ is psd

Examples for kernels:

- inner product kernels such as polynomial kernels, but also NTK
- RBF kernels such as α-exponential kernels $e^{-\frac{\|x-y\|_2^\alpha}{\tau}}$ with bandwidth parameter τ (Gaussian $\alpha = 2$, Laplacian $\alpha = 1$)

Theorem (RKHS induced by kernel - MW Thm 12.11.)

Given any psd kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there is a unique Hilbert space $\mathcal{F}_\mathcal{K}$ in which $\mathcal{K}$ is reproducing, i.e. for all $x \in \mathcal{X}$, $f(x) = \langle f, \mathcal{K}(\cdot, x)\rangle_\mathcal{F}$ for all $f \in \mathcal{F}$ and $\mathcal{K}(\cdot, x) \in \mathcal{F}$. We call it the (reproducing kernel) Hilbert space induced by (or associated with) $\mathcal{K}$.

(i) RKHS "induced" via kernel

Given $\mathcal{K}$, how may the induced RKHS $\mathcal{F}_\mathcal{K}$ look like?

- The idea: First define the following set of functions

$$\mathcal{F}_{\text{pre}} = \{\sum_{i=1}^N \alpha_i \mathcal{K}(\cdot, x_i) : N \in \mathbb{N}, \alpha \in \mathbb{R}^N, x_1, \dots, x_N \in \mathcal{X}\} \text{ and}$$

defining inner product for $f = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j \mathcal{K}(\cdot, \tilde{x}_i)$

$$\langle f, g\rangle_{\mathcal{F}_{\text{pre}}} = \sum_{i=1}^{\ell}\sum_{j=1}^{m} \alpha_i \beta_j \mathcal{K}(x_i, \tilde{x}_j)$$

- We call $\mathcal{F}_\mathcal{K}$ its completion, that is the space including limit objects of all Cauchy sequences in $\mathcal{F}_{\text{pre}}$ (sometimes omitting the subscript)
- $\mathcal{K}$ satisfies the following *reproducing property* in $\mathcal{F}_\mathcal{K}$ since $\langle \mathcal{K}(x_i, \cdot), \mathcal{K}(x_j, \cdot)\rangle_{\mathcal{F}_\mathcal{K}} = \mathcal{K}(x_i, x_j) \to$ for any $f = \sum_{l=1}^m \beta_l \mathcal{K}(x_l, \cdot)$

$$f(x) = \sum_{l=1}^m \beta_l \langle \mathcal{K}(x_l, \cdot), \mathcal{K}(x, \cdot)\rangle_{\mathcal{F}_\mathcal{K}} = \langle \sum_{l=1}^m \beta_l \mathcal{K}(x_l, \cdot), \mathcal{K}(x, \cdot)\rangle_{\mathcal{F}_\mathcal{K}} = \langle f, \mathcal{K}(x, \cdot)\rangle_{\mathcal{F}_\mathcal{K}}$$

Rewriting the (penalized) empirical risk for RKHS

Given the corresponding kernel of an RKHS, we can easily find (the or a, dependent on $\lambda \geq 0$) minimizer $\widehat{f}$ for kernel (ridge) regression by searching only in a subset $\mathcal{F}_S$.

Proposition (Representer Theorem - MW Prop. 12.33.)

A global empirical risk minimizer in $\mathcal{F}_\mathcal{K}$ for any loss is in $\mathcal{F}_S := \text{span}\{\mathcal{K}(x_1, \cdot), \ldots, \mathcal{K}(x_n, \cdot)\}$. Further the minimizer of empirical risk (with any loss) with an additive RKHS norm penalty lies in $\mathcal{F}_S$.

Hence, we rewrite $f(x) = \sum_{i=1}^n \alpha_i \mathcal{K}(x_i, x)$ for some $\alpha \in \mathbb{R}^n$ and search over $\mathbb{R}^n$ instead!

$$\min_{f \in \mathcal{F}_\mathcal{K}} \frac{1}{2n} \|y - f(x_1^n)\|_2^2 + \lambda \|f\|_{\mathcal{F}_\mathcal{K}}^2 = \min_{f \in \mathcal{F}_S} \frac{1}{2n} \|y - f(x_1^n)\|_2^2 + \lambda \|f\|_{\mathcal{F}_\mathcal{K}}^2$$

$$= \min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|y - K\alpha\|_2^2 + \lambda \alpha^\top K \alpha$$

Neighbor-Q: How about when $\lambda = 0$, does the minimizer still lie in $\mathcal{F}_S$? Isn't this a parametric problem again with parameters α?

Proof of Representer Theorem for RKHS (skipped)

- We can write $f \in \mathcal{F}_\mathcal{K}$ using the orthogonal decomposition of $\mathcal{F}_\mathcal{K} = \mathcal{F}_S \bigoplus \mathcal{F}_{S^\perp}$, i.e. $f = f_S + f_{S^\perp}$ with $f_S \in \mathcal{F}_S$ etc.
- By the reproducing property and orthogonality between $\mathcal{F}_S, \mathcal{F}_{S^\perp}$, we have $f(x_i) = \langle f_S + f_{S^\perp}, \mathcal{K}(x_i, \cdot)\rangle_{\mathcal{F}_\mathcal{K}} = \langle f_S, \mathcal{K}(x_i, \cdot)\rangle_{\mathcal{F}_\mathcal{K}}$ so that

$$\min_{f_S + f_{S^\perp} \in \mathcal{F}_\mathcal{K}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (f_S + f_{S^\perp})(x_i)\|_2^2 + \lambda \|f_S + f_{S^\perp}\|_{\mathcal{F}_\mathcal{K}}^2$$

$$\geq \min_{f_S \in \mathcal{F}_S} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_S(x_i)) + \lambda \|f_S\|_{\mathcal{F}_\mathcal{K}}^2$$

because $\|f_S\|_{\mathcal{F}_\mathcal{K}} < \|f_S + f_{S^\perp}\|_{\mathcal{F}_\mathcal{K}}$ and with equality only if $\lambda = 0$ $\square$

Reproducing property in RKHS: $\langle \mathcal{K}_x(\cdot), f\rangle_\mathcal{F} = f(x)$ for all $f \in \mathcal{F}$
$\to$ convergence in $\mathcal{F}$ pointwise convergence
$\to$ reduces to n-dim regression problem

ii) From function class (RKHS) to kernel

Theorem (Existence of kernel, MW Thm 12.13)

Given an RKHS $\mathcal{F}$, there is a unique psd kernel $\mathcal{K}_{\mathcal{F}}$ that satisfies the reproducing property

Proof (skipped during class):

- By the Riesz representation theorem there exists a unique R_x with $L_x(f) = \langle R_x, f \rangle_{\mathcal{F}}$
- The corresponding kernel $\mathcal{K}_{\mathcal{F}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of $\mathcal{F}$ reads $\mathcal{K}_{\mathcal{F}}(x, y) = \langle R_x, R_y \rangle = R_x(y)$ and is psd, symmetric
- $\mathcal{F}_{\mathcal{K}}$ also has bounded evaluation functionals where $M_x = \sqrt{\mathcal{K}(x, x)}$ via Cauchy Schwarz
- $\mathcal{F}_{\mathcal{K}}$ is the only Hilbert space in which $\mathcal{K}$ satisfies the reproducing property $\langle \mathcal{K}_x(\cdot), f \rangle_{\mathcal{F}} = f(x)$ for all $f \in \mathcal{F}$ (MW Thm 12.11)

ii) From function class (RKHS) to kernel: Examples

1. Is $\mathcal{F}_{lin} = \{f : f(x) = \langle w, x \rangle, w \in \mathbb{R}^d\}$ an RKHS?
   - Propose $\mathcal{K}(x, y) = \langle x, z \rangle$ as a reproducing kernel
   - Following discussion about $\mathcal{F}_{\text{pre}}$ we define for $f = \langle w_f, \cdot, \rangle$ and $g = \langle w_g, \cdot \rangle$ the inner product $\langle f, g \rangle = w_f^\top w_g$
   - By definition the $\mathcal{K}$ then satisfies the reproducing property: $\langle f(\cdot), \langle \cdot, z \rangle \rangle = w_f^\top z = f(z)$
2. Is $\mathcal{L}^2([0, 1])$ an RKHS?
   - Does not converge point-wise, necessary for all RKHS: that is if $f_n \to f$ in the Hilbert norm, then it also does for every x by boundedness of evaluation functional
3. Some restrictions on $\mathcal{L}^2([0, 1])$ can fix that: Sobolev space on $[0, 1]$ $\mathcal{W}_2^1([0, 1]) = \{f : [0, 1] \to \mathbb{R} \mid f(0) = 0, f' \in \mathcal{L}^2([0, 1])\}$ where derivative exists almost everywhere
   - IP $\langle f, g \rangle = \int_0^1 f'(x) g'(z) \mathrm{d}z$ (interpretable)
   - Sobolev kernel: $\mathcal{K}(x, y) = \min\{x, y\}$
   - Checking it's reproducing: $\langle f(\cdot), \min\{\cdot, z\} \rangle = \int_0^1 f'(x) \mathbb{1}_{x \le z} \mathrm{d}x = \int_0^z f'(x) \mathrm{d}x = f(z)$
   - can extend to higher order derivatives / smoothness (HW 3)

References

Reproducing Kernel Hilbert spaces:

- MW Chapter 12
- SC Chapter 4

Non-parametric regression:

- MW Chapter 13

Recap: kernel trick (skipped in class)

The following two slides are for reference, as a recap of kernel trick:

Feature maps are motivated by search in nonlinear function spaces

- Instead of linear function $w^\top x$ with $w \in \mathbb{R}^d$, we want $w^\top \phi(x)$ with $w \in \mathbb{R}^p$ where ϕ is feature vector with p elements $\phi_j : X \to \mathbb{R}$
- In fact this includes feature maps that satisfy $\phi : X \to \ell_2(\mathbb{N})$ where ℓ_2 is the space of square summable sequences
- Define $\mathcal{F} = \{f : X \to \mathbb{R} : f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}_0} \text{ with } w \in \ell_2(\mathbb{N})\}$ and consider loss $l((x, y); f) = l(f(x), y)$

Lemma (dependence only on inner products)

There exists a global empirical risk minimizer $\widehat{f} = \min_{f \in \mathcal{F}} \sum_{i=1}^{n} l(y_i, f(x_i))$ *such that for any test sample* $x \in X$, $\widehat{f}(x)$ *only depends on* x, x_i *via inner products* $\langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}_0}$ *and* $\langle \phi(x_i), \phi(x) \rangle_{\mathcal{H}_0}$

Recap: Proof of Lemma (skipped in class)

Define $S = \text{span}\{\phi(x_1), \ldots, \phi(x_n)\}$

1. Note that because $f(x_i) = w^\top \phi(x_i)$, the value of the empirical risk only depends on $w_S := \prod_S w$, we can limit search space to $w \in S$. This is because you can decompose $w = w_S + w_{S^\perp}$ with $S^\perp$ the orthogonal complement of S and hence $w_{S^\perp}^\top \phi(x_i) = 0$ for all i
2. To search in $\mathcal{F}_S = \{f : f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}_0}\ w \in S\}$ we can parameterize $w = \sum_{i=1}^n \alpha_i \phi(x_i)$ and hence $f(x_j) = \sum_{i=1}^n \alpha_i \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}_0}$ and
3. The ERM $\widehat{f}$ can then be obtained by minimizing over α obtaining $\hat{\alpha}$ which depends on training points x_i only via $\langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}_0}$
4. Observing that $\widehat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i \langle \phi(x_i), \phi(x) \rangle_{\mathcal{H}_0}$ the proof is complete

□