

Lecture 10: Kernel ridge regression

1 / 31

Recap: Non-parametric prediction error bound

Definition (Localized (empirical) Gaussian complexity)

The localized Gaussian complexity around f^* of scale δ is

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

Lemma (Critical radius, MW 13.6.)

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

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

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

Theorem (Prediction error bound, MW Thm 13.5.)

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

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

2 / 31

Motivation for next example: kernel spaces \mathcal{F}

It's infeasible to search in generic non-parametric (say e.g. Lipschitz, or non-decreasing) function spaces since they're infinite dimensional. Kernel spaces are special in the following sense:

- Motivation 1: Computationally efficiently and tractably find global minimizer \hat{f} : *implementable* transition from linear to featurized regression via kernel trick
- Motivation 2: From research: one way to think about training NN is to view it as learning a specific kernel. Actually, convolutional neural tangent kernels (based on NN) can predict CIFAR10 with $\sim 90\%$ test accuracy
- With compute capacities nowadays, can also fit parametric classes (NN) with huge number of parameters, approximating complex function spaces well.
- However, due to being convex, can actually analyze minimizer in [Reproducing Kernel Hilbert spaces \(RKHS\)](#)

3 / 31

Plan for today

- RKHS primer (not the focus of the course and hence not covered in full detail, only mentioned to get better understand the example application of our non-parametric prediction error bound):
 - Definition
 - RKHS via kernels
 - Representer theorem
- constrained RKHS as an example for non-parametric prediction error bounds
- RKHS minimizer of regularized square loss

4 / 31

Reproducing Kernel Hilbert spaces

The reproducing property RKHS 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 $f = 0$ and $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}}$

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 $M_x < \infty$

→ can (i) design RKHS via a kernel directly, or (ii) take Hilbert space satisfying abstract definition in last slide and find kernel “in hindsight” (see “appendix”, skipped in class)

5 / 31

(i) RKHS induced by kernels

Definition (Reminder - psd kernels)

A bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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 $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, 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} .

6 / 31

(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

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

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

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

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

7 / 31

Constrained and penalized non-parametric regression

- In general non-parametric regression, the least-square estimate \hat{f} is

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 = \frac{1}{2n} \|y - f(x_1^n)\|_2^2 \text{ (possibly non-unique)}$$

- Now the function space \mathcal{F} could be an RKHS $\mathcal{F}_{\mathcal{K}}$ or even norm-bounded functions in an RKHS/metric space, with

$$\min_{\|f\|_{\mathcal{F}} \leq R} \frac{1}{2n} \|y - f(x_1^n)\|_2^2$$

- The kernel ridge regression objective is an equivalent formulation (for some appropriate λ)

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

(replaces the ℓ_2 norm of parameter vector in linear regression by RKHS norm)

8 / 31

Reparameterizing kernel (ridge) regression objective

- For \mathcal{F} being an RKHS, we can easily find (“the” or “a”, dependent on $\lambda \geq 0$) minimizer \hat{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}_K for any loss is in $\mathcal{F}_S := \text{span}\{\mathcal{K}(x_1, \cdot), \dots, \mathcal{K}(x_n, \cdot)\}$. Further the (unique) minimizer of the empirical risk (with any loss) with an additive RKHS norm penalty lies in \mathcal{F}_S . (Proof in “appendix”)

- Hence, it suffices to search for functions of the form $f(x) = \sum_{i=1}^n \alpha_i \mathcal{K}(x_i, x)$ with norm $\|f\|_{\mathcal{F}}^2 = \sum_{i=1}^n \alpha_i \alpha_j \mathcal{K}(x_i, x_j) = n\alpha^\top K\alpha$ where K is the (empirical kernel matrix) with $K_{ij} = \mathcal{K}(x_i, x_j)$

9 / 31

Reparameterizing kernel (ridge) regression objective

- Therefore, kernel ridge regression reduces to

$$\begin{aligned} \min_{f \in \mathcal{F}_K} \frac{1}{2n} \|y - f(x_1^n)\|_2^2 + \lambda \|f\|_{\mathcal{F}_K}^2 &= \min_{f \in \mathcal{F}_S} \frac{1}{2n} \|y - f(x_1^n)\|_2^2 + \lambda \|f\|_{\mathcal{F}_K}^2 \\ &= \min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|y - K\alpha\|_2^2 + \lambda \alpha^\top K\alpha \end{aligned} \quad (1)$$

and similarly, the constrained version becomes

$$\min_{\|f\|_{\mathcal{F}} \leq R} \frac{1}{2n} \|y - f(x_1^n)\|_2^2 = \min_{\alpha^\top K\alpha \leq R} \frac{1}{2n} \|y - K\alpha\|_2^2 \quad (2)$$

- So the argmin of eq. 2 and eq. 1 can be written as $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i \mathcal{K}(\cdot, x_i)$ with

$$\hat{\alpha} = \arg \min_{\alpha^\top K\alpha \leq R} \frac{1}{2n} \|y - K\alpha\|_2^2$$

and correspondingly

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

10 / 31

Proof of Representer Theorem for RKHS (skipped)

- We can write $f \in \mathcal{F}_{\mathcal{K}}$ using the orthogonal decomposition of $\mathcal{F}_{\mathcal{K}} = \mathcal{F}_S \oplus \mathcal{F}_{S^\perp}$, 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

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

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 $f \in \mathcal{F}$

→ convergence in \mathcal{F} pointwise convergence

→ reduces to n -dim regression problem

11 / 31

Non-parametric regression in RKHS

We're now ready to apply our non-parametric prediction error bound on minimizers of the square loss in RKHS!

Recall the empirical square loss $\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$ (for $\lambda = 0$) and (empirical) prediction error $\frac{1}{n} \sum_{i=1}^n (f^*(x_i) - f(x_i))^2$.

Setting: $f^* \in \mathcal{F}_{\mathcal{K}}$ for some kernel \mathcal{K} and $y_i = f^*(x_i) + \sigma w_i$ w/ i.i.d. $w_i \sim \mathcal{N}(0, 1)$

Goal: instantiate the prediction error bounds for \hat{f} that are the kernel (ridge) regression minimizers above using localized complexities

Neighbor-Q:

- a) Is the kernel matrix invertible?
- b) What is the minimum value of the empirical (square) loss?
- c) How about the prediction error?

12 / 31

Constrained/Regularized regression in a metric space

If \mathcal{K} is s.t. K is pd/full-rank for all distinct inputs \rightarrow can interpolate!
 In that case the localized Gaussian complexity will be of order 1.

\mathcal{F} generally too large! \rightarrow norm-bounded $\mathcal{F}_R = \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leq R\}$

The regularized estimator \hat{f}_R is defined as (possibly non-unique)

$$\hat{f}_R \in \arg \min_{f \in \mathcal{F}_R} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 = \arg \min_{\|f\|_{\mathcal{F}} \leq R, f \in \mathcal{F}_R} \frac{1}{2n} \|y - f(x_1^n)\|_2^2$$

Theorem (Prediction error of norm-bounded functions)

Assume $f^* \in \mathcal{F}_R$ and further assume $\delta_{n;R}$ it holds that

$\frac{\sigma \tilde{\mathcal{G}}_n(\mathcal{F}_3; \delta_{n;R})}{\delta_{n;R}} \leq \delta_{n;R} R$. Then we have for the least-squares estimate

$\hat{f}_R \in \mathcal{F}_R$

$$\|\hat{f}_R - f^*\|_n^2 \leq c_0 R^2 \delta_{n;R}^2$$

with probability $\geq 1 - c_1 e^{-c' \frac{nR^2 \delta_{n;R}^2}{\sigma^2}}$.

(This theorem actually holds for any normed function space \mathcal{F} as long as $\mathcal{F} - \mathcal{F}$ is star-shaped)

13 / 31

Proof for Theorem (prediction error of $\hat{f} \in \mathcal{F}_R$)

- Scaling basic inequality $\|\hat{\Delta}\|_n^2 \leq \frac{2\sigma}{n} \langle w, \hat{\Delta} \rangle$ by $\frac{1}{R^2}$, using $\tilde{f}^* = \frac{f^*}{R}$, $\tilde{f} = \frac{\hat{f}}{R}$, $\tilde{\Delta} = \tilde{f} - \tilde{f}^*$, $\tilde{\sigma} = \frac{\sigma}{R}$, we obtain $\|\tilde{\Delta}\|_n^2 \leq 2\frac{\tilde{\sigma}}{n} \sum_{i=1}^n w_i \tilde{\Delta}(x_i)$
- Since $\tilde{f}^*, \tilde{f} \in \mathcal{F}_1$, $\tilde{\Delta} \in \mathcal{F}_1^* = \mathcal{F}_1 - \tilde{f}^* \subset \mathcal{F}_3$ (\mathcal{F}_2 suffices for norm-bounded RKHS, but use \mathcal{F}_3 for penalized later)
- Now argue similar to last lecture
 - Want $\frac{\tilde{\sigma}}{n} \sum_i w_i \tilde{\Delta}(x_i) \leq 2\|\tilde{\Delta}\|_n \delta_{n;R}$ for all $\|\tilde{\Delta}\|_n \geq \delta_{n;R}$ for some $\delta_{n;R}$
 - Since \mathcal{F} is a Hilbert space, \mathcal{F}_R is convex and $\mathcal{F}_R - \mathcal{F}_R$ is convex \rightarrow star-shaped for any R . Then, like last time we get

$$\sup_{\|\tilde{\Delta}\|_n \geq \delta_{n;R}, \tilde{\Delta} \in \mathcal{F}_3} \frac{\sigma}{n} \sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\|\tilde{\Delta}\|_n} \leq \sup_{\|\tilde{\Delta}\|_n \leq \delta_{n;R}, \tilde{\Delta} \in \mathcal{F}_3} \frac{\sigma}{n} \sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\delta_{n;R}}$$
 - In particular, by the concentration of Lipschitz functions of Gaussians, for each R we again have

$$\sup_{\|\tilde{\Delta}\|_n \leq \delta_{n;R}, \tilde{\Delta} \in \mathcal{F}_3} \frac{\tilde{\sigma}}{n} \sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\delta_{n;R}} \leq \delta_{n;R}$$
 where we use the modified critical inequality

$$\mathbb{E}_w \sup_{\tilde{\Delta} \in \mathcal{F}_3, \|\tilde{\Delta}\|_n \leq \delta_{n;R}} \frac{\tilde{\sigma}}{n} \sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\delta_{n;R}} = \frac{\tilde{\sigma}}{\delta_{n;R}} \tilde{\mathcal{G}}_n(\mathcal{F}_3; \delta_{n;R}) \leq \delta_{n;R}$$
 in tail bound
- Observing $\|\hat{f} - f^*\|_n^2 = R^2 \|\tilde{\Delta}\|_n^2$ yields the theorem. □ 14 / 31

Instantiation for norm-bounded RKHS

- Key: compute the localized G.C. to find $\delta_{n;R}$ satisfying the assumptions of the theorem
- Instead of using Dudley's integral, for RKHS we can directly use the special structure given by the spectrum of the kernel matrix.
- By the representer theorem we have $\hat{f}_R(\cdot) = \sum_{i=1}^n \hat{\alpha}_i \mathcal{K}(\cdot, x_i)$ with

$$\hat{\alpha} = \arg \min_{\alpha^\top K \alpha \leq R} \frac{1}{2n} \|y - K\alpha\|_2^2$$

- Denote the eigenvalues of the rescaled kernel matrix $\frac{K}{n}$ as $\hat{\mu}_j$

Lemma (R -modified critical radius $\delta_{n;R}$, MW Cor. 13.18)

Let $\delta_{n;R}$ be the smallest $\delta > 0$ satisfying

$$\frac{4}{\sqrt{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}} \leq \frac{\delta^2 R}{\sigma}$$

Then it also holds that $\frac{\sigma \tilde{\mathcal{G}}_n(\mathcal{F}_3; \delta_{n;R})}{\delta_{n;R}} \leq \delta_{n;R} R$

15 / 31

Localized G.C. for RKHS with bounded norm

The key for the proof of the above lemma is the following bound

Lemma (local G.C. for norm-bounded RKHS, MW Lem. 13.22)

Let \mathcal{F} be an RKHS with kernel function \mathcal{K} and $\hat{\mu}_j$ be the eigenvalues of the rescaled kernel matrix $\frac{K}{n}$. Let \mathcal{F}_r be the norm-bounded space of functions in \mathcal{F} with RKHS norm r . It holds for all $\delta > 0$ that

$$\tilde{\mathcal{G}}_n(\mathcal{F}_1; \delta) \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}}.$$

In fact, more generally $\tilde{\mathcal{G}}_n(\mathcal{F}_r; \delta) \leq \sqrt{\frac{r^2+1}{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}}.$

In what follows, we denote as K the rescaled version of the empirical matrix for convenience, i.e. $K_{ij} = \frac{1}{n} \mathcal{K}(x_i, x_j)$.

16 / 31

Proof of Lemma (skipped during class)

- By representer theorem, can take sup over \mathcal{F}_S by parameterizing $\Delta(\cdot) = \frac{1}{\sqrt{n}} \sum_i \alpha_i \mathcal{K}(\cdot, x_i) \in \mathcal{F}_S \subset \mathcal{F}$ and hence $\Delta(x_1^n) = \sqrt{n} K \alpha$, s.t.

$$\begin{aligned} \tilde{\mathcal{G}}_n(\mathcal{F}_r; \delta) &= \mathbb{E}_w \sup_{\|\Delta\|_{\mathcal{F}} \leq r, \|\Delta\|_n \leq \delta} \frac{1}{n} \sum_i w_i \Delta(x_i) \\ &= \frac{1}{\sqrt{n}} \mathbb{E}_w \sup_{\alpha^\top K \alpha \leq r^2, \alpha^\top K^2 \alpha \leq \delta^2} w^\top K \alpha \end{aligned}$$

- Let $K = U^\top \Lambda U$ and $\theta := \Lambda U \alpha \rightarrow \tilde{\mathcal{G}}_n(\mathcal{F}_r; \delta) = \frac{1}{\sqrt{n}} \mathbb{E}_w \max_{\theta \in \mathbb{T}} w^\top \theta$

$$\text{with } \mathbb{T} = \left\{ \theta \in \mathbb{R}^n \mid \sum_i \theta_i^2 \leq \delta^2, \sum_{i=1}^n \frac{\theta_i^2}{\hat{\mu}_i} \leq r^2 \right\}$$

- Let $\mathcal{E} := \{ \theta \in \mathbb{R}^n \mid \sum_i \eta_i \theta_i^2 \leq 1 + r^2 \} \supset \mathbb{T}$ w/ $\eta_i = \max\{\delta^{-2}, \hat{\mu}_i^{-1}\}$

$$\max_{\theta \in \mathcal{E}} \langle w, \theta \rangle \iff \max_{\theta^\top \text{diag}(\eta_i) \theta \leq 1+r^2} \langle w, \theta \rangle \iff \max_{\|\beta\|_2 \leq \sqrt{1+r^2}} \langle \text{diag}^{-1/2}(\eta_i) w, \beta \rangle$$

$$\rightarrow \tilde{\mathcal{G}}_n(\mathcal{F}_r; \delta) \leq \sqrt{\frac{1+r^2}{n}} \mathbb{E}_w \sqrt{\sum_i \frac{w_i^2}{\eta_i}} \leq \sqrt{\frac{1+r^2}{n}} \sqrt{\sum_i \frac{1}{\eta_i}} \text{ via Jensen's } \square$$

17/31

More discussion on the error bound

Note: Can easily generalize to $f^* \notin \mathcal{F}_R$ (more technical, without new core insights) with additional approx. error $\inf_{\|f\|_{\mathcal{F}} \leq R} \|f - f^*\|_n^2$

Rates for actual kernel spaces \mathcal{F}

- Ex. 1: α -smooth functions (derivatives up to order α are square-integrable) w/ $\hat{\mu}_j \sim j^{-2\alpha} \rightarrow \|\hat{f} - f^*\|_n^2 \leq \left(\frac{R\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$ (you'll prove it in HW2)
- Ex. 2: Gaussian kernel w/ $\hat{\mu}_j \sim e^{-cj \log j} \rightarrow \|\hat{f} - f^*\|_n^2 \leq \frac{\sigma^2 \log(\frac{Rn}{\sigma})}{n}$ (see MW example 13.21)
- Bonus: For \mathcal{K} on compact \mathcal{X} empirical matrix eigenvalues $\hat{\mu}_j \sim \mu_j$ for big n where μ_j are integral operator eigenvalues of the operator $T_{\mathcal{F}}$ defined by $T_{\mathcal{F}}(f)(y) = \int \mathcal{K}(x, y) f(x) d\mathbb{P}(x)$ (Koltchinskii, Gine '00)

Penalized regression guarantees for metric spaces

- Recall kernel ridge regression solution

$$\hat{f}_{\lambda_n} = \arg \min_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{F}}^2$$

- With the same definition of the critical quantity $\delta_{n;R}$ as before

Theorem (Prediction error for reg. estimators - MW Thm 13.17.)

For any convex function class \mathcal{F} with a norm and \mathcal{F}^* star-shaped, when $\lambda_n \geq 2\delta_{n;R}^2$, there is a universal constant such that for $f^* \in \mathcal{F}_R$

$$\|\hat{f}_{\lambda_n} - f^*\|_n^2 \leq cR^2(\delta_{n;R}^2 + \lambda_n) \text{ w/ prob. } \geq 1 - c_0 e^{-c_1 \frac{nR^2 \delta_{n;R}^2}{\sigma^2}}.$$

- Again, if $f^* \notin \mathcal{F}_R$ yields add. approx. error $\inf_{\|f\|_{\mathcal{F}} \leq R} \|f - f^*\|_n^2$
- if additional term $\lambda_n \sim \delta_{n;R}^2$, same order as constrained
- in practice, don't know $\delta_n \rightarrow$ choose λ_n via cross-validation

19 / 31

Proof of bound for regularized regression estimate

For simplicity we write \hat{f} for \hat{f}_{λ_n}

1. By optimality we have

$$\frac{1}{2n} \sum_{i=1}^n (f^*(x_i) + \sigma w_i - \hat{f}(x_i))^2 + \lambda_n \|\hat{f}\|_{\mathcal{F}}^2 \leq \frac{\sigma^2}{2n} \sum_{i=1}^n w_i^2 + \lambda_n \|f^*\|_{\mathcal{F}}^2$$

which yields different **basic inequality** after rearranging terms

$$\frac{1}{2} \|\hat{\Delta}\|_n^2 \leq \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) + \lambda_n (\|f^*\|_{\mathcal{F}}^2 - \|\hat{f}\|_{\mathcal{F}}^2)$$

2. Divide both sides by R^2 and use normalized f^*, \hat{f}, σ by $\frac{1}{R}$ like for norm-bounded $\rightarrow \tilde{f}^*, \tilde{f}, \tilde{\sigma}, \tilde{\Delta} = \tilde{f} - \tilde{f}^*$ (\tilde{f} different than in MW!)

$$\frac{1}{2} \|\tilde{\Delta}\|_n^2 \leq \underbrace{\frac{\tilde{\sigma}}{n} \sum_{i=1}^n w_i \tilde{\Delta}(x_i)}_{T_1} + \underbrace{\lambda_n (\|\tilde{f}^*\|_{\mathcal{F}}^2 - \|\tilde{f}\|_{\mathcal{F}}^2)}_{T_2}$$

Note that T_2 is a new term **and** $\tilde{\Delta}, \tilde{f}$ are not necessarily \mathcal{F} -norm-bounded which enters in localized G.C. for \mathcal{F}_R to bound T_1

20 / 31

Proof of bound for regularized regression estimate

3. Either $\|\tilde{\Delta}\|_{n;R} \leq \delta_n$ and we are done, or $\|\tilde{\Delta}\|_n > \delta_{n;R}$ on which event we further analyze two events differing by the \mathcal{F} -norm of $\tilde{\Delta}$ and show that in both events it holds that

$$c' \|\tilde{\Delta}\|_n^2 \leq c\delta_{n;R} \|\tilde{\Delta}\|_n + \lambda_n$$

for different constants c', c (details in next slide)

- a) Event 1 $\|\tilde{f}\|_{\mathcal{F}} \leq 2$: then $\|\tilde{\Delta}\|_{\mathcal{F}} \leq 3$ and using previous arguments as for the prediction error for norm-bounded RKHS using the critical inequality and tail bound we obtain $T_1 \leq c\delta_{n;R} \|\tilde{\Delta}\|_n$, as well as the fact that $T_2 \leq \|\tilde{f}^*\|_{\mathcal{F}}^2 \leq 1$.
- b) Event 2 $\|\tilde{f}\|_{\mathcal{F}} > 2$: using a new (peeling) lemma for all $\|\tilde{\Delta}\|_{\mathcal{F}} \geq 1$ see next slide. There we use T_2 to “cancel” large norms
4. Solving the quadratic yields $\|\tilde{\Delta}\|_n^2 \leq c(\delta_{n;R}^2 + \lambda_n)$ \square

21 / 31

Proof of 3b. - regularization plays role of norm-bounding

We use the shorthand δ_n for $\delta_{n;R}$. We now show that on both events 1 & 2, $c' \|\tilde{\Delta}\|_n^2 \leq c\delta_n \|\tilde{\Delta}\|_n + \lambda_n$ for some (different) constants c', c

- b) Event 2: $\|\tilde{f}\|_{\mathcal{F}} > 2 > 1 \geq \|\tilde{f}^*\|_{\mathcal{F}} \rightarrow \|\tilde{\Delta}\|_{\mathcal{F}} \geq 1$

- T_1 : can still bound T_1 using similar idea as before, but iteratively (peeling lemma) on event $\|\tilde{\Delta}\|_{\mathcal{F}} \geq 1$ (MW Lem. 13.23) yields with probability at least $\geq 1 - c_1 e^{-\frac{n\delta_n^2}{c_2\sigma^2}}$

$$\sup_{\tilde{\Delta} \in \mathcal{F}^*, \|\tilde{\Delta}\|_{\mathcal{F}} \geq 1} \frac{\tilde{\sigma}}{n} \sum_i w_i \tilde{\Delta}(x_i) \leq 2\delta_n \|\tilde{\Delta}\|_n + 2\delta_n^2 \|\tilde{\Delta}\|_{\mathcal{F}} + \frac{\|\tilde{\Delta}\|_n^2}{16} \quad (3)$$

- T_2 : $\lambda_n (\|\tilde{f}^*\|_{\mathcal{F}}^2 - \|\tilde{f}\|_{\mathcal{F}}^2) \leq 2\lambda_n - \lambda_n \|\tilde{\Delta}\|_{\mathcal{F}}$ using $\|\tilde{\Delta}\|_{\mathcal{F}} \leq \|\tilde{f}\|_{\mathcal{F}} + \|\tilde{f}^*\|_{\mathcal{F}}$ and $\|\tilde{f}^*\|_{\mathcal{F}}^2 - \|\tilde{f}\|_{\mathcal{F}}^2 \leq \|\tilde{f}^*\|_{\mathcal{F}} - \|\tilde{f}\|_{\mathcal{F}}$
 \rightarrow green “swallows” red term for large enough $\lambda_n \geq 2\delta_n^2$
 \rightarrow regularization takes care of not having explicit norm bound!
- Putting things together yields $\frac{1}{2} \|\tilde{\Delta}\|_n^2 \leq c\delta_n \|\tilde{\Delta}\|_n + \frac{1}{16} \|\tilde{\Delta}\|_n^2 + 2\lambda_n$

22 / 31

Peeling lemma idea - MW Lem. 13.23 (skipped in class)

- The idea is to make T_1 depend on the \mathcal{F} -norm which we can then “kill” via regularization (large enough λ_n)
- By star-shapedness of \mathcal{F} it suffices to show inequality with sup over $\|\tilde{\Delta}\|_{\mathcal{F}} = 1$ (it that held, then for any $\|\tilde{\Delta}\|_{\mathcal{F}} > 1$, we could plug in $\frac{\tilde{\Delta}}{\|\tilde{\Delta}\|_{\mathcal{F}}}$ to obtain the same bound)
- However then, we no longer have $\|\tilde{\Delta}\|_n \geq \delta_n$ (can essentially only use the star-shaped argument on one of the norms)
- Then we do something like in chaining - split up event where eq. 3 does not hold and $\|\tilde{\Delta}\|_{\mathcal{F}} = 1$ (without boundedness of $\|\tilde{\Delta}\|_n$) into subevents where $\|\tilde{\Delta}\|_n \in [t_m, t_{m+1}]$ with $t_m = 2^m \delta_n$
- Lastly, union bounding with this choice of t_m with the usual concentration bound (Lipschitz function of Gaussians in MW Thm 2.26) yields the result.

For a detailed proof we refer to the book.

23 / 31

References

Reproducing Kernel Hilbert spaces:

- MW Chapter 12
- SC Chapter 4

(Penalized and constrained) Non-parametric regression

- MW Chapter 3

24 / 31

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 (bonus):

- 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 $\mathcal{K}_{\mathcal{F}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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)

25 / 31

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 $\mathcal{K}(x, y) = \langle x, y \rangle$ as a reproducing kernel
- Following discussion about \mathcal{F}_{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 $f_n \rightarrow f$ 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 $[0, 1]$

$\mathcal{W}_2^1([0, 1]) = \{f : [0, 1] \rightarrow \mathbb{R} \mid f(0) = 0, f' \in \mathcal{L}^2([0, 1])\}$ where derivative exists almost everywhere

- IP $\langle f, g \rangle = \int_0^1 f'(x)g'(z)dz$ (interpretable)
- Sobolev kernel: $\mathcal{K}(x, y) = \min\{x, y\}$
- Checking it's reproducing:
 $\langle f(\cdot), \min\{\cdot, z\} \rangle = \int_0^1 f'(x)\mathbb{1}_{x \leq z} dx = \int_0^z f'(x) dx = f(z)$
- can extend to higher order derivatives / smoothness (HW 3)

26 / 31

From function class (RKHS) to kernel: Sobolev spaces

$\mathcal{L}^2([0, 1])$ is not an RKHS because convergence not point-wise

Some restrictions on $\mathcal{L}^2([0, 1])$ can fix that: Sobolev space on $[0, 1]$

$\mathcal{W}_2^1([0, 1]) = \{f : [0, 1] \rightarrow \mathbb{R} \mid f(0) = 0, f' \in \mathcal{L}^2([0, 1])\}$ where derivative exists almost everywhere

- IP $\langle f, g \rangle = \int_0^1 f'(x)g'(z)dz$ (interpretable)
- Sobolev kernel: $\mathcal{K}(x, y) = \min\{x, y\}$
- Reproducing prop.:
 $\langle f(\cdot), \min\{\cdot, z\} \rangle = \int_0^1 f'(x)\mathbb{1}_{x \leq z} dx = \int_0^z f'(x)dx = f(z)$
- can extend to higher order derivatives / smoothness (HW 2)
 $\mathcal{W}_2^\alpha([0, 1]) = \{f : [0, 1] \rightarrow \mathbb{R} \mid f^{(\alpha)}(0) = 0, f^{(\alpha)} \in \mathcal{L}^2([0, 1])\}$

27 / 31

Kernel trick (you should know)

The following two slides are for reference, as a recap of kernel trick (which you should know):

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 $\phi_j : X \rightarrow \mathbb{R}$
- In fact this includes feature maps that satisfy $\phi : X \rightarrow \ell_2(\mathbb{N})$ where ℓ_2 is the space of square summable sequences
- Define $\mathcal{F} = \{f : X \rightarrow \mathbb{R} : f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}_0} \text{ with } w \in \ell_2(\mathbb{N})\}$ and consider loss $l((x, y); f) = l(f(x), y)$

Lemma (dependence only on inner products)

There exists a global empirical risk minimizer

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

28 / 31

Proof of Lemma (you should know)

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

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 $\mathcal{F}_S = \{f : f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}_0} \mid w \in S\}$ 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 \hat{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 $\hat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i \langle \phi(x_i), \phi(x) \rangle_{\mathcal{H}_0}$ the proof is complete

□

29 / 31

Kernel eigenvalues (bonus)

- The empirical and population Gaussian complexities are close within constants MW Prop 14.25.
- population Gaussian compl. depends on kernel operator eigenvalues
- For \mathcal{K} on compact \mathcal{X} empirical matrix eigenvalues $\hat{\mu}_j \sim \mu_j$ for big n where μ_j are integral operator eigenvalues (Koltchinskii, Gine '00)

Define bounded, linear Hilbert-Schmidt integral operator

$T_{\mathcal{K}} : \mathcal{L}^2 \rightarrow \mathcal{L}^2$ with $T_{\mathcal{K}} f = \int \mathcal{K}(x, y) f(y) dy$, and we call μ_j eigenvalues and ψ_j eigenfunctions if $T_{\mathcal{K}} \psi_j = \mu_j \psi_j$

Theorem (Mercer's) (SC Thm 4.49, 4.51, MW Thm 12.20)

For \mathcal{K} psd with RKHS $\mathcal{F}_{\mathcal{K}}$, there exist eigenfunctions and eigenvalues $\psi_j, \mu_j \geq 0$ of $T_{\mathcal{K}}$ that satisfy

1. ψ_j form an ONB in $\mathcal{L}^2(\mathbb{P})$ and $\phi_j = \sqrt{\mu_j} \psi_j$ is an ONS in $\mathcal{F}_{\mathcal{K}}$.
2. $\mathcal{K}(x, y) = \sum_j \mu_j \psi_j(x) \psi_j(y)$ converges in $\mathcal{L}^2(\mathbb{P})$
3. If \mathcal{K} also continuous, above sum converges absolutely and uniformly

Crucial: μ_j, ψ_j depends on distribution \mathbb{P} !

30 / 31

Proof of Mercer's Theorem (bonus)

1. Main component: Hilbert-Schmidt Theorem (spectral theorem)
(e.g. Knapp Thm 2.5., any functional analysis book)
 - For any kernel, $T_{\mathcal{K}}$ is compact, self-adjoint, has eigenspaces
 - decomposition of image of $T_{\mathcal{K}}$ into ψ_j (countable) ONB of \mathcal{L}_2 that are eigenvectors of $T_{\mathcal{K}}$
 - sum converges in \mathcal{L}^2 .
2. Positivity by definition of the operator and kernel psd
3. Why $T_{\mathcal{K}}$ maps to $\mathcal{F}_{\mathcal{K}}$ SC 4.26.: Hoelder ineq, Bochner integrability
4. Absolute uniform convergence of sum for continuous kernel:
Non-decreasing sequences of continuous functions with a continuous limit converge uniformly (e.g. Rudin 7.13).

Notes in S.C. they define it $T_{\mathcal{K}}$ more rigorously