Guarantees for Machine Learning, Fall 2023

Lecture 1: Introduction and concentration bounds

Class intro

Objective. Develop graduate students into researchers who can

- understand and criticize papers in ML theory
- conjecture and prove new theorems that with high impact

Prerequisites

- Familiar with core machine learning concepts
- Should be comfortable writing rigorous mathematical proofs (for D-MATH courses)

Course structure

- First part: classical techniques for non-asymptotic risk bounds
  - Core reference: Martin Wainwright: High-dimensional statistics (available for free online via ETH)
- Second part: projects that review and extend current papers

Logistics

- Class website sml.inf.ethz.ch/gml23/syllabus.html
- Lecture slides will be uploaded after lectures at the latest
- TAs: Konstantin Donhauser, Julia Kostin (Office hours on request)
- Internet platforms to sign up for: moodle (announcements, questions, teammate search), Gradescope (assignments)
- Important date announcements: in class and per email

Evaluation & enrollment

Evaluation

- 2 homeworks (10%), midterm (50%), project (40%)
- HWs:
  - randomly select questions graded by TAs
  - check HW release schedule on the website
- Project (in groups of two):
  - Pick a paper from list according to your interests & background on (**October 13**)
  - Discussion & extension of one theoretical paper
  - 15-20 min Presentation in last four weeks
  - $\geq$ 10 page written report (due **January 12**)

Enrollment

- Current waitlist: ~75. Admitted: 30. Limit for admissions: 30
- By experience, everybody who wants to take it, can
- Final deadline to de-register: **October 11th** else no-show
- Others welcome to audit as long as there is space

Who is here?

Which department?

1. Computer Science
2. Mathematics/Statistics
3. Data Science
4. EE & Robotics
5. Others

What stage of your studies are you?

1. Masters
2. PhD student
3. Bachelors

Plan for today

- Statistical perspective on the supervised learning pipeline
- Evaluation of an estimator using the excess risk
- Concentration bounds of empirical means

Recap: (Supervised) Machine Learning - Classification

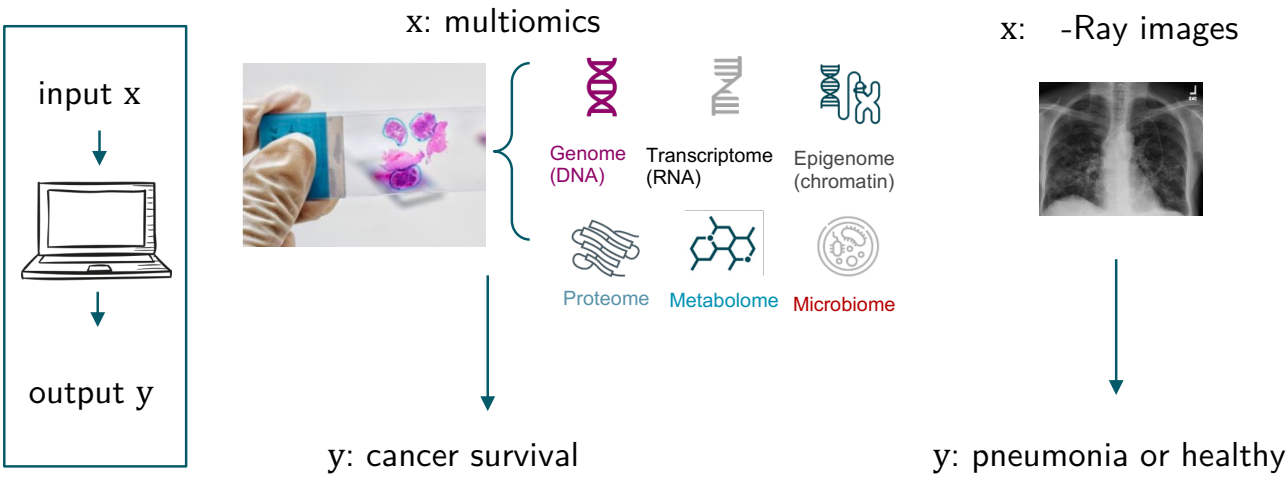


Figure 1: Classification examples

Recap: (Supervised) Machine Learning - Regression

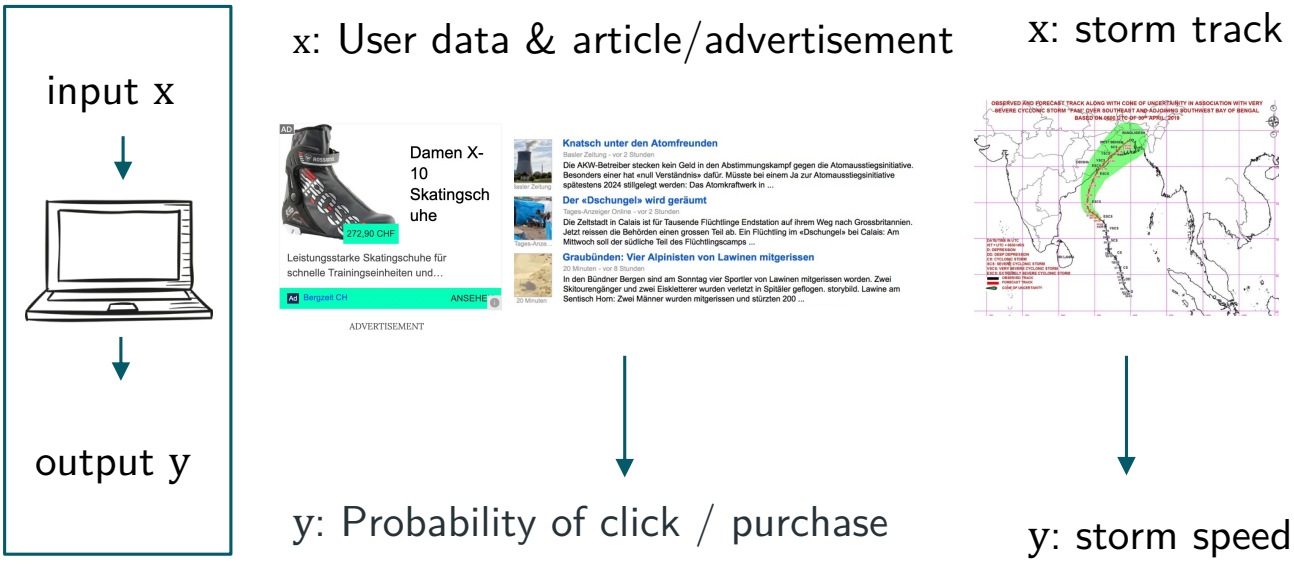


Figure 2: Regression examples

Statistical Perspective on (supervised) Machine Learning

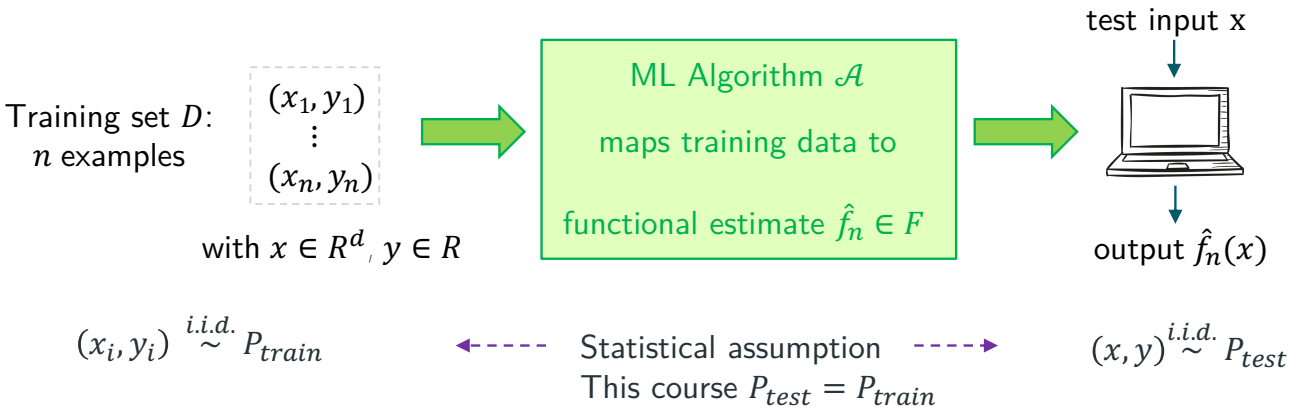


Figure 3: Supervised learning pipeline from statistical point of view

- some examples for $\mathbb{P} = \mathbb{P}_{train} = \mathbb{P}_{test}$ include
  - regression: marginal dist. over x and $y = f^\star(x) + \epsilon$ for random ϵ
  - classification: generative such as Gaussian mixture model or discriminative: marginal dist. over x and $y = \text{sign}(f^\star(x))$
- The estimate $\hat{f}_n \in \mathcal{F}$ depends on $(x_i, y_i)_{i=1}^n$ (i.e. is random) and is in some function class (e.g. linear, neural network etc.)

Evaluation of an estimator $\widehat{f}_n$

Whether $\widehat{f}_n$ is "good" is decided during test time: On average over test points (x, y), we'd like the predictions $\widehat{f}_n(x)$ to be close to y

- We measure "close" via a pointwise loss ℓ,
  e.g. $\ell((x, y), f) = (f(x) - y)^2$ for regression
  or $\ell(x, y; f) = \mathbb{1}_{f(x) \neq y}$ for classification
- We call the average loss of any function f the *population risk*
  $R(f) := R(f; \mathbb{P}) = \mathbb{E}\ell((x, y); f)$
- We further call the training loss of any f the *empirical risk*
  $R_n(f) := R(f; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \ell((x_i, y_i); f)$ estimate is
- In the next lectures we'll consider the empirical risk minimization paradigm where

$$\widehat{f}_n := \arg\min_{f \in \mathcal{F}} R_n(f)$$

Evaluation of an estimator $\widehat{f}_n$

Q: For classification, is $R(\widehat{f}_n) = 20\%$ bad or good?

A: Depends on how hard the task is! Perhaps it's not possible to achieve perfect accuracy!

We should compare population risk of $\widehat{f}_n$ with that of the best possible function *if we knew the full distribution*, i.e. evaluate the excess risk:

$$\mathcal{E}_R(n) := R(\widehat{f}_n) - \inf_f R(f) \leq UB(...)$$

Grab a neighbor: Designate a presenter. Discuss for 5 minutes.

1. How is the population risk of an estimator related to its test error?
2. Which parameters of the problem and algorithm does the excess risk depend on? What happens to the excess risk of an estimator $\widehat{f}_n$ when we vary these parameters? Categorize the phenomena
3. What are tradeoffs when we consider the *empirical risk minimizer* $\widehat{f}_n := \arg\min_{f \in \mathcal{F}} R_n(f)$

Questions on the excess risk

1. Population risk vs. test error

- Test error on n' new samples follows $R_{n'}(\widehat{f}_n) \to R_n(\widehat{f}_n)$ by law of large numbers (LLN)

2. Excess risk depends on model class $\mathcal{F}$, dimensionality of the data d, sample size n and consists of the following factors and trends

- approximation error (if $f^\star = \arg\min_f R(f)$ is complicated): larger $\mathcal{F}$, smaller d better
- optimization error (due to optimization algorithm): Lipschitz, (strong) convex loss ℓ better
- statistical error (due to finite sample and noise): larger n (usually) better (depends on $\mathcal{F}, d$ as well) of course $\leftarrow$ this course

3. Tradeoffs: Larger $\mathcal{F}$, bigger effect of noise (statistical error) but smaller approx error (variance vs. bias)

This course: Non-asymptotic take on statistical "Guarantees for Machine Learning"

We introduce general frameworks to analyze excess risk and compute concrete upper (and lower) bounds s.t. with prob. at least $1-\delta$

$$R(\widehat{f}_n) - R(f^\star) \leq UB(n, d, \mathcal{F}, f^\star)$$

where we assume $f^\star = \arg min_f R(f)$ exists.

Usually UB $\to$ Metric (CF) as $n \to \infty$

Questions we'd like to answer:

1. Does UB converge to 0 as n increases? (consistency)
1. If I collect double as much data, how much do I decrease my excess risk? $\to$ boils down to the exponent of n (statistical rate)

This course focuses on 2. We'll now discuss some probabilistic basics that give a sense for what to expect from course later.

Excess risk decomposition

- Recall the population risk $R(f) = \mathbb{E}\ell((X, Y); f)$
- Recall the empirical risk $R_n(f) = \frac{1}{n}\sum_{i=1}^n \ell((X_i, Y_i); f)$
- Remember we want to bound the excess risk

$$\begin{aligned} R(\widehat{f}_n) - R(f^\star) &= R(\widehat{f}_n) - R_n(\widehat{f}_n) + \overbrace{R_n(\widehat{f}_n) - R_n(f^\star)}^{T_3 \leq 0} + R_n(f^\star) - R(f^\star) \\ &\leq \underbrace{R(\widehat{f}_n) - R_n(\widehat{f}_n)}_{T_1} + \underbrace{R_n(f^\star) - R(f^\star)}_{T_2} \end{aligned}$$

Question: Are T_1 and T_2 qualitatively similarly hard to bound? Is $T_3 \leq 0$ always true? Briefly discuss with your neighbor.

- $T_3 \leq 0$ is only true when $f^\star \in \mathcal{F}$!
- T_1 is harder than T_2 since it's a sum of dependent variables whereas T_2 is difference between an emprical mean and its expectation.

Concentration bounds for single random variables (R.V.)

- *Markov* inequality: $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}$ for $X \geq 0$;
- Markov used on $e^{\lambda(X-\mathbb{E}X)}$ for $\lambda \geq 0$ yields the *Chernoff* bound

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]}{e^{\lambda t}}$$

  where the inf is effectively over all $\lambda \geq 0$ where the moment generating function (MGF) $\mathbb{E}e^{\lambda X}$ exists

We can use Chernoff to get tighter bounds for R.V. X with short tails

Definition (Sub-Gaussian random variables)

A random variable X with mean μ is sub-Gaussian w/ parameter σ if

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\lambda^2\sigma^2/2} \quad \text{for all } \lambda \in \mathbb{R}$$

- For σ sub-Gaussians using Chernoff we obtain the tail bound

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq \inf_{\lambda \geq 0} e^{\frac{\lambda^2\sigma^2}{2} - \lambda t} = e^{-\frac{t^2}{2\sigma^2}}$$

Examples for sub-Gaussian random variables

- Gaussians $\mathcal{N}(0, \sigma^2)$ are sub-Gaussian with parameter σ
- *Rademacher variables* $\epsilon = -1, +1$ with equal probability $1/2$ are sub-Gaussian with parameter $\sigma = 1$
  - We can directly compute and bound their MGF

$$\mathbb{E}e^{\lambda\epsilon} = \frac{1}{2}(e^{-\lambda} + e^{\lambda}) \leq e^{\lambda^2/2}$$

- Almost surely bounded in $[a, b]$ (exercise)

Empirical means of independent subgaussians

Lemma (Hoeffding's inequality)

For i.i.d sub-Gaussian R.V. X_i, it holds that

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

Neighbor-Q: Prove Hoeffding's inequality

- Recall sub-Gaussian: $\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$
- Recall Chernoff for sub-Gaussians: $\mathbb{P}(X - \mathbb{E}X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$

Proof of Hoeffding's inequality

1. We can apply Chernoff on the mean of n independent random variables with moment generating function

$$\mathbb{E}e^{\lambda(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mathbb{E}X_i))} = \prod_{i=1}^{n}\mathbb{E}e^{\frac{\lambda}{n}(X_i-\mu)} = [\mathbb{E}e^{\frac{\lambda}{n}(X_i-\mu)}]^n$$

1. Hence, the mean of n i.i.d. sub-Gaussian variables is sub-Gaussian with parameter $\frac{\sigma}{\sqrt{n}}$ since $\mathbb{E}e^{\lambda(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mathbb{E}X_i))} \leq e^{\frac{\lambda^2\sigma^2}{2n^2}n}$

1. yielding Hoeffding's inequality for the mean of iid sub-Gaussians

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

Q: How can we now use Hoeffding's inequality to bound the term T_2?

Syllabus of course

The courses focuses on bounding T_2 using so-called uniform convergence.

We'll cover

- uniform convergence using Rademacher and Gaussian complexity
- metric entropy and chaining to bound the complexity
- application to non-parametric regression (kernel methods)
- minimax lower bounds
- theory for overparameterized models

References

Concentration bounds:

- MW Chapters 2

Excess risk:

- MW Chapter 4

Lecture 2: Uniform tail bound and McDiarmid

Announcements and lecture outline

Announcements:

- HW released tonight, due in two weeks on Thursday **12.10.22 23:59** on gradescope.
- Warning: HW is *long*, start early!
- Can discuss together, but write up your *own* solution and indicate who you've worked together with
- no late HW except in medical cases (with attest from doctor)
- Post questions on HW on moodle
- Please de-register once you know you are not going to continue the course!

Plan today

1. Recap excess risk decomposition and Hoeffding's inequality
2. Concentration of functions of n dependent r.v. via bounded differences
3. McDiarmid inequality and uniform tail bound
4. Proof of McDiarmid via Doob martingales, Azuma-Hoeffding inequality

Recap last lecture: excess risk decomposition

- Recall we assume that $Z_i := (X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ with $Z_i \in \mathcal{Z}$ and evaluate a function f by the expected loss (population risk) $R(f) = \mathbb{E}\ell(Z; f)$
- The empirical risk is defined by $R_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i; f)$ and for fixed f, we have $\mathbb{E}R_n(f) = R(f)$.
- We want to bound the excess risk

$$R(\widehat{f}_n) - R(f^\star) = R(\widehat{f}_n) - R_n(\widehat{f}_n) + \overbrace{R_n(\widehat{f}_n) - R_n(f^\star)}^{\leq 0 \text{ by optimality}} + R_n(f^\star) - R(f^\star)$$
$$\leq \underbrace{R(\widehat{f}_n) - R_n(\widehat{f}_n)}_{T_1} + \underbrace{R_n(f^\star) - R(f^\star)}_{T_2}$$

- Then via Chernoff, we proved Hoeffding's inequality that holds for the mean of i.i.d. sub-Gaussians

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \geq t\right) \leq \mathrm{e}^{-\frac{nt^2}{2\sigma^2}}$$

Wakeup-Q: How can we use Hoeffdings inequality to bound T_2?

Back to term T_1

- Problem: $R_n(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i; \widehat{f}_n)$ *not* an emp. mean of i.i.d. R.V.! Can we still show some sort of concentration for $R_n(\widehat{f}_n)$?
- Crude bound: since by assumption algorithm searches in a model/function class $\mathcal{F}$, i.e. $\widehat{f}_n \in \mathcal{F}$, we can upper bound T_1 by

$$R(\widehat{f}_n) - R_n(\widehat{f}_n) \leq \sup_{f \in \mathcal{F}} R(f) - R_n(f) =: g_n(Z_1, \ldots, Z_n)$$

- Instead of averages of n i.i.d. random variables, the supremum of an *empirical process* $R(f) - R_n(f)$ is a general function $g_n : \mathcal{Z}^n \to \mathbb{R}$
- Instead of $R_n(f) \approx \mathbb{E}R_n(f) = R(f)$ for empirical means, if g_n satisfies some properties, g_n concentrates around $\mathbb{E}g_n(z)$!

Specific case: g_n satisfies bounded difference property

Definition (bounded difference property)

Define for given $z, z' \in \mathcal{Z}^n$ a new vector $z^{\backslash k}$ with the k-th element from z' and all other from z: $z_j^{\backslash k} = \begin{cases} z_j & \text{if } j \neq k \\ z'_k & \text{if } j = k \end{cases}$. We say that $g_n : \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded difference inequality if for each $k = 1, \ldots, n$ it holds that

$$|g_n(z) - g_n(z^{\backslash k})| \leq \sigma_k \quad \text{for all } z, z' \in \mathcal{Z}^n$$

Theorem (McDiarmid)

If $g_n : \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded difference condition and $Z \in \mathcal{Z}^n$ is a random vector with n independent entries, then

$$\mathbb{P}(g_n(Z) - \mathbb{E}g_n(Z) \geq t) \leq e^{-\frac{2t^2}{\sum_{k=1}^n \sigma_k^2}}$$

- Concentration with n is usually obtained via $t \sim n$ or via $\sigma_k \sim \frac{1}{n}$

Tail bound for supremum of (bounded) empirical process

- Remember for $f \in \mathcal{F}$: $R_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i, f)$
- We can now use McDiarmid on the sup. of empirical process $g_n(z_1, \ldots, z_n) = \sup_{f\in\mathcal{F}} R(f) - R_n(f)$ for bounded losses!

Theorem (Uniform tail bound)

For b-unif. bounded $\ell(\cdot, f)$, that is $\|\ell(\cdot; f)\|_\infty \leq b$ for all $f \in \mathcal{F}$, it holds that

$$\mathbb{P}(\sup_{f\in\mathcal{F}} R(f) - R_n(f) \geq \mathbb{E}[\sup_{f\in\mathcal{F}} R(f) - R_n(f)] + t) \leq e^{-\frac{nt^2}{2b^2}}$$

where the probability is over the training data.

- Note that there are other results beyond boundedness (Lipschitz functions etc.), that are tighter particularly in the context of bounding suprema of empirical process - MW Chapter 3
- This uniform tail bound can give us a (crude) high-probability bound and rate, if we can bound the expectation ($\rightarrow$ next class!)

Proof of tail bound using McDiarmid

For simplicity define $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$

Use McDiarmid by checking bounded differences assumption with $g_n(z) := \sup_{f\in\mathcal{F}} R_n(f) - R(f) = \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n h(z_i) - \mathbb{E}h$

- For b-uniformly bounded $\mathcal{H}$, we have for all $k = 1, \ldots, n$ and any $z, z' \in \mathcal{Z}^n$ that for any $h \in \mathcal{H}$

$$\frac{1}{n}\sum_i [h(z_i) - \mathbb{E}h] - \sup_{\tilde{h}\in\mathcal{H}} \frac{1}{n}\sum_i \left[\tilde{h}(z_i^{\backslash k}) - \mathbb{E}\tilde{h}\right]$$

$$\leq \frac{\sum_i h(z_i) - h(z_i^{\backslash k})}{n} \leq \frac{h(z_k) - h(z_k')}{n} \leq \frac{2b}{n}$$

- Since it holds for all $h \in \mathcal{H}$, taking the sup on both sides yields

$$g_n(z) - g_n(z^{\backslash k}) = \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_i [h(z_i) - \mathbb{E}h] - \sup_{\tilde{h}\in\mathcal{H}} \frac{1}{n}\sum_i \left[\tilde{h}(z_i^{\backslash k}) - \mathbb{E}\tilde{h}\right] \leq \frac{2b}{n}$$

- By symmetry it holds for $g_n(z^{\backslash k}) - g_n(z) \rightarrow |g_n(z) - g_n(z^{\backslash k})| \leq \frac{2b}{n}$
- Plugging in $\sigma_k = \frac{2b}{n}$ into McDiarmid then yields the result.

Proof sketch of McDiarmid

Theorem (McDiarmid)

If $g_n : \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded difference condition with $\{\sigma_k\}_{k=1}^n$ and Z is a random vector with n independent entries, then

$$\mathbb{P}(g_n(Z) - \mathbb{E}g_n(Z) \geq t) \leq e^{-\frac{2t^2}{\sum_{k=1}^n \sigma_k^2}}$$

Proof intuition:

Re-writing g_n as a sum

- For any function $g_n : \mathcal{Z}^n \to \mathbb{R}$, even though we don't have a sum per se, we can write the difference as a sum (check for yourself)

$$g_n(Z) - \mathbb{E}g_n(Z) =: \sum_{j=1}^n D_j$$

where $D_j := \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_j] - \mathbb{E}[g_n(Z)|Z_1, \ldots Z_{j-1}]$ for $j \geq 2$ and $D_1 = \mathbb{E}[g_n(Z)|Z_1] - \mathbb{E}[g_n(Z)]$

Proof intuition Part I

Discuss with your neighbor: For the special case of empirical mean $g_n(Z) = \frac{1}{n}\sum_{i=1}^n Z_i$ with Z_i independent and bounded

$\to$ D_j are independent and sub-Gaussian so that one can use Hoeffding's bound on D_j. Can we use this for general g_n?

- Indeed, for all $j = 1, \ldots, n$

$$D_j = \frac{1}{n}\sum_{i=j}^n \mathbb{E}[Z_i|Z_1, \ldots, Z_j] - \frac{1}{n}\sum_{i=j-1}^n \mathbb{E}[Z_i|Z_1, \ldots, Z_{j-1}] = \frac{Z_j}{n} - \frac{\mathbb{E}Z}{n}$$

with all D_j independent and bounded (hence sub-Gaussian)

- For general $g_n(Z)$ independence of D_j does not hold!

Proof intuition Part II

- However, we can still show that
  - D_j indepedendent $\rightarrow$ D_j martingale difference, and hope that D_j s.t.
  - D_j "conditionally" bounded (and hence still in some way subgaussian)
- (informal) Then instead of *Hoeffding* that can be used on independent **bounded** R.V., we can use *Azuma-Hoeffding*, that shows

$$\mathbb{P}(\sum_{i=1}^{n} D_i \geq t) \leq \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

  for **bounded** martingale difference sequences where $D_i \in [a_i, b_i]$ a.s.

We now formalize the proof.

"Recap": Martingale difference sequences

Let $\{Z_j\}_{j=1}^{\infty}$ be a sequence of R.V. and $\mathcal{F}_j := \sigma(Z_1, \ldots, Z_j)$,

Further, let $\{S_j\}_{j=1}^{\infty}$ be such that S_j is measurable with respect to $\mathcal{F}_j$ (i.e. we say $\{S_j\}_{j=1}^{\infty}$ is *adapted to the filtration* $\{\mathcal{F}_j\}_{j=1}^{\infty}$)

Definition (Martingale (difference))

- $\{S_j, \mathcal{F}_j\}_{j=1}^{\infty}$ is a *martingale* if for all j, $\mathbb{E}|S_j| < \infty$ and $\mathbb{E}[S_{j+1}|\mathcal{F}_j] = S_j$
- Similarly, $\{D_j, \mathcal{F}_j\}_{j=1}^{\infty}$ is a *martingale difference sequence* if for all j, $\mathbb{E}|D_j| < \infty$ and $\mathbb{E}[D_{j+1}|\mathcal{F}_j] = 0$

- For any martingale $\{S_j, \mathcal{F}_j\}_{j=0}^{\infty}$, $D_j = S_j - S_{j-1}$ for $j \geq 1$ is a martingale difference sequence.
- Doob construction: given some function $g_n : \mathcal{Z}^n \rightarrow \mathbb{R}$, for a sequence of random variables $Z_1, \ldots, Z_n$, note that $S_j = \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_j]$ fulfills exactly the above conditions if $\mathbb{E}|g_n(Z)| < \infty$. Then also $\mathbb{E}[D_{j+1}|\mathcal{F}_j] = 0$ for $D_j = S_j - S_{j-1}$
  Check with your neighbor

Formal proof of McDiarmid

Theorem ((conditional) Azuma-Hoeffding inequality)

If for martingale difference sequence $\{(D_i, \mathcal{F}_i)\}_{i=1}^n$ *it holds that* $D_i|\mathcal{F}_{i-1} \in [a_i, b_i]$ *for some* $\{(a_i, b_i)\}_{i=1}^n$ *almost surely for all* i*, then*

$$\mathbb{P}(\sum_{i=1}^n D_i \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

The proof of McDiarmid follows immediately if we can show that

- for any g_n satisfying the bounded difference property with $\{\sigma_j\}_{j=1}^n$
- we have that $g_n(Z) - \mathbb{E}g_n(Z) = \sum_{j=1}^n D_j$ with $\{D_j, \mathcal{F}_j\}_{j=1}^n$ a bounded martingale difference sequence with $b_j - a_j \leq \sigma_j$.

We now show that this fact is true.

Proof: Assumptions of Azuma-Hoeffding hold

- Define shorthand $Z_1^i = (Z_1, \ldots, Z_i) \in \mathcal{Z}^i$ for random/real vectors
- It remains to prove that: if g_n satisfies the bounded difference condition with $\{\sigma_j\}_{j=1}^n$, then for all $z_1^{j-1} \in \mathcal{Z}^{j-1}$ we have $D_j|Z_1^{j-1} = z_1^{j-1} \in [a_j, b_j]$ almost surely with $b_j - a_j \leq \sigma_j$
- We define shorthand (last equality follows by independence of Z_j): $\mathbb{E}[g_n(Z)|z_1^{j-1}] := \mathbb{E}[g_n(Z)|Z_1^{j-1} = z_1^{j-1}] = \mathbb{E}g_n(z_1^{j-1}, Z_j^n)$
- Further, by definition for all $z_1^{j-1} \in \mathcal{Z}^{j-1}$ almost surely

$$D_j|Z_1^{j-1} = z_1^{j-1} \geq \inf_{z \in \mathcal{Z}} \mathbb{E}[g_n(Z)|z_1^{j-1}, Z_j = z] - \mathbb{E}[g_n(Z)|z_1^{j-1}] =: a_j$$

$$D_j|Z_1^{j-1} = z_1^{j-1} \leq \sup_{z \in \mathcal{Z}} \mathbb{E}[g_n(Z)|z_1^{j-1}, Z_j = z] - \mathbb{E}[g_n(Z)|z_1^{j-1}] =: b_j$$

- $D_j|Z_1^{j-1} = z_1^{j-1} \in [A_j, B_j]$ and, by bounded diff. ass. on g_n, a.s:

$$\begin{aligned} b_j - a_j &= \sup_{z \in \mathcal{Z}} \mathbb{E}g_n(z_1^{j-1}, z, Z_{j+1}^n) - \inf_{z \in \mathcal{Z}} \mathbb{E}g_n(z_1^{j-1}, z, Z_{j+1}^n) \\ &\leq \sup_{z, z' \in \mathcal{Z}} \mathbb{E}|g_n(z_1^{j-1}, z, Z_{j+1}^n) - g_n(z_1^{j-1}, z', Z_{j+1}^n)| \leq \sigma_j \end{aligned}$$

Summary

- McDiarmid inequality for bounded difference
- uniform tail bound for T_1
- Proof McDiarmid: Hoeffding bound for sums of independent R.V. $\rightarrow$ martingale (difference) sequences and Azuma-Hoeffding inequality

Next up: Uniform law with symmetization and Rademacher complexity

References

Concentration bounds including Azuma-Hoeffding, McDiarmid

- MW Chapter 2
- *Boucheron, Lugosi, Massart*: Chapter 2

Martingales - any probability theory book, e.g.:

- *P. Billingsley. Probability and Measure*
- *R. Durrett. Probability: Theory and Examples (4th edition)*

(Bonus) More concentration bounds on suprema of empirical processes:

- MW Chapter 3
- *Ledoux, Talagrand: Probability for Banach spaces* for functional Bernstein

Lecture 3: Azuma-Hoeffding and uniform law

Announcements

- HW due next Thursday 23:59, write it up entirely independently yourself
- You can check paper suggestions for project work already on the project website (link to a googlesheet)
- Lec2 slides updated regarding boundedness of martingale difference sequence & Azuma-Hoeffding (will explain again today)
- Goal of in-class lecture: cannot deliver the details of each proof completely, but primarily intuition - expect to fully understand and digest after reading the book & doing homework

Plan today

- Review of proof of uniform tail bound
- Warm-up exercise: using Azuma-Hoeffding for online learning "excess risk"
- Proof of Azuma-Hoeffding
- Uniform law with Rademacher complexity
- Intuition of Rademacher complexity

Recap: Main tail bound

- $\{Z_i\}_{i=1}^n$ are training points $\overset{iid}{\sim} \mathbb{P}$, estimator $\widehat{f}_n \in \mathcal{F}$ trained on them
- We use Z both for the collection $Z = \{Z_i\}_{i=1}^n$ and a single random vector $Z \sim \mathbb{P}$ which should be clear from the context
- Goal: want to prove that $R(\widehat{f}_n) - R_n(\widehat{f}_n) \leq \sup_{f \in \mathcal{F}} \mathbb{E}\ell(Z; f) - \frac{1}{n}\sum_{i=1}^n \ell(Z_i; f) =: g_n(Z)$ small with probability at least $1 - \delta$

Theorem (Uniform tail bound)

For b-unif. bounded ℓ, it holds that

$$\mathbb{P}(\sup_{f \in \mathcal{F}} R(f) - R_n(f) \geq \mathbb{E}[\sup_{f \in \mathcal{F}} R(f) - R_n(f)] + t) \leq e^{-\frac{nt^2}{2b^2}}$$

where the probability is over the training data.

Recap: What we can do with the tail bound

Using the short-term $\mathrm{Res}(n, \mathcal{F}) := \mathbb{E}[\sup_{f \in \mathcal{F}} R(f) - R_n(f)]$ We immediately obtain

$$\mathbb{P}(\sup_{f \in \mathcal{F}} R(f) - R_n(f) \leq \mathrm{Res}(n, \mathcal{F}) + t) \geq 1 - \mathrm{e}^{-\frac{nt^2}{2b^2}}$$

This is a "high probability" bound in the sense that with probability at least $1 - \delta$ we have

$$\sup_{f \in \mathcal{F}} R(f) - R_n(f) \leq b\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} + \mathrm{Res}(n, \mathcal{F})$$

Recap: Proof of tail bound (w/o martingale speak)

Approach: Upper bound $\mathbb{P}(g_n(Z) - \mathbb{E}g_n(Z) \geq t)$ by following

1. If loss ℓ b-uniformly bounded, then $g_n = \sup_{f \in \mathcal{F}} \mathbb{E}\ell(Z, f) - \frac{1}{n}\sum_{i=1}^n \ell(Z_i, f)$ satisfies bounded difference property with $\sigma_i = \frac{2b}{n}$ for all i
2. For any g_n, we can decompose $g_n(Z) - \mathbb{E}g_n(Z) = \sum_{i=1}^n D_i$ $D_i = \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_i] - \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_{i-1}]$
3. Then, D_i satisfies that for any z_1^{i-1} there are some a_i, b_i with $b_i - a_i \leq \sigma_i$ such that $D_i|Z_1^{i-1} = z_1^{i-1} \in [a_i, b_i]$.
4. show how for such D_i (bounded martingale diff sequence) we have $\sum_{i=1}^n D_i$ concentrates around its expectation $\mathbb{E}D_i = 0$, i.e.

   $\mathbb{P}(\sum_{i=1}^n D_i > t) \leq = \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^n \sigma_i^2}} \leq \mathrm{e}^{-\frac{nt^2}{2b^2}}$ [Azuma Hoeffding]

Note: 2-4 proves McDiarmid using Azuma-Hoeffding, 2-3 prove that assumptions for Azuma-Hoeffding hold.

Not shown, will show today: Azuma-Hoeffding

Recap: Azuma-Hoeffding

- Hoeffding: Simple concentration for average of n independent sub-Gaussian (e.g bounded) Z_i

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}Z > t) \leq \mathrm{e}^{-\frac{nt^2}{2\sigma^2}}$$

- Azuma-Hoeffding: "Advanced" concentration for average of a martingale difference sequence $\{D_i\}_{i=1}^n$ bounded in intervals of length $\sigma = \frac{c}{n}$

$$\mathbb{P}(\sum_{i=1}^{n} D_i > t) \leq \mathrm{e}^{-\frac{2t^2}{n\sigma^2}} = \mathrm{e}^{-\frac{2nt^2}{c^2}}$$

Theorem (Azuma-Hoeffding inequality, MW Cor. 2.20)

If for martingale difference sequence $\{(D_i, \mathcal{F}_i)\}_{i=1}^n$ it holds that $D_i|\mathcal{F}_{i-1}$ almost surely lies in an interval of length L_i for all i, then

$$\mathbb{P}(\sum_{i=1}^{n} D_i \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n} L_i^2}}$$

Next, we gain some more intuition on Azuma-Hoeffding by applying it to a different problem related to online learning

Exercise Context I: Online learning setting

- $Z_1, \ldots, Z_n$ come in one at a time.
- At each point in time i you would like to output an estimator $\widehat{f}_{i-1}$ to predict on the next sample Z_i with small loss
- As a data scientists, we naturally consider functions that are trained using the previous examples $Z_1, \ldots, Z_{i-1}$. More formally, we assume $\widehat{f}_{i-1}$ is a *deterministic function* of the previous samples $Z_1, \ldots, Z_{i-1}$ (e.g. ERM but *does not have to be*!), i.e. measurable with respect to $\sigma(Z_1, \ldots, Z_{i-1}) = \mathcal{F}_{i-1}$.
- $\widehat{f}_0$ can be any data-independent arbitrary estimator, e.g. a randomly initialized model.
- Assume the minimizer $\widehat{f}_n := \arg\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1} \ell(Z_i; f)$ exists

Exercise Context II: Online to batch conversion

- A standard quantity people want to keep small in online learning is the regret Reg_n, the average incurred loss of the sequence $\{\widehat{f}_i\}_{i=1}^n$ with the loss of $\widehat{f}_n$

$$\mathrm{Reg}_n = \sum_{i=1}^n \ell(Z_i; \widehat{f}_{i-1}) - \sum_{i=1}^n \ell(Z_i; \widehat{f}_n)$$

- Note: Bounding the actual Reg_n is a whole area of research and in many cases, good online learning algorithms exist
- Online-to-batch conversion exploits online learning algorithms with small regret to get estimator based on batch $Z_1, \ldots, Z_n$ with good generalization. For example, one can consider a random estimator that samples from the sequence of online estimators $\{\widehat{f}_i\}_{i=0}^{n-1}$ which
  - conditioned on the data are deterministic
  - has an average (over the sampling) a risk of $\frac{1}{n}\sum_{i=1}^n R(\widehat{f}_{i-1})$
- We will now prove a high probability bound on the "average" excess risk $\frac{1}{n}\sum_{i=1}^n R(\widehat{f}_{i-1}) - R(f^\star)$

Exercise: Bound on the average excess risk

With your neighbor, prove that with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n [R(\widehat{f}_{i-1}) - R(f^\star)] \leq \frac{1}{n}\mathrm{Reg}_n + \sqrt{\frac{8\log(1/\delta)}{n}} \qquad (1)$$

with $R(f) = \mathbb{E}\ell(Z; f)$ for $\ell \in [0,1]$ using the following steps

1. Step: Prove that $D_i = [\mathbb{E}_Z \ell(Z; \widehat{f}_{i-1}) - \ell(Z_i; \widehat{f}_{i-1})] + [\ell(Z_i; f^\star) - \mathbb{E}_Z \ell(Z; f^\star)]$ is a bounded martingale difference sequence
2. Step: Decompose the risk (by including terms with $\widehat{f}$ and using its optimality) and prove

$$\frac{1}{n}\sum_{i=1}^n [R(\widehat{f}_{i-1}) - R(f^\star)] \leq \frac{1}{n}\mathrm{Reg}_n + \frac{1}{n}\sum_{i=1}^n D_i$$

3. Step: Use Step 1 and Azuma-Hoeffding to prove the bound eq. 1

Solution: Proof of average excess risk bound

We use the following shorthands for simplicity:

- $R_n(\{\widehat{f}_i\}_{i=0}^{n-1}) := \frac{1}{n}\sum_{i=1}^n \ell(Z_i; \widehat{f}_{i-1})$
- $R(\{\widehat{f}_i\}_{i=0}^{n-1}) := \frac{1}{n}\sum_{i=1}^n R(\widehat{f}_{i-1}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\ell(Z; \widehat{f}_{i-1})$

1. Risk decomposition:

$$\frac{1}{n}\sum_{i=1}^n [R(\widehat{f}_{i-1}) - R(f^\star)] \le R(\{\widehat{f}_i\}_{i=0}^{n-1}) - R_n(\{\widehat{f}_i\}_{i=0}^{n-1}) + \underbrace{R_n(\{\widehat{f}_i\}_{i=0}^{n-1}) - R_n(\widehat{f}_n}_{=\mathrm{Reg}_n}$$
$$+ \underbrace{R_n(\widehat{f}_n) - R_n(f^\star)}_{\le 0 \text{ by optimality of } \widehat{f}} + R_n(f^\star) - R(f^\star)$$

2. D_i is a martingale difference sequence because

$$\mathbb{E}D_i | \mathcal{F}_{i-1} = 0$$

as Z_i is independent of $\widehat{f}_{i-1}$ and bounded a.s. by 4.

Check: The average excess risk over $\{\widehat{f}\}_{i=1}^n$ is similar in terms of rate for large n as long as $R(\widehat{f}_n)$ is bounded

Proof of Azuma-Hoeffding

1. First of all, we have for all sequences z_1^{i-1} that for some $b_i - a_i \le L_i$

$$\mathbb{E}[e^{\lambda D_i} | Z_1^{i-1} = z_1^{i-1}] \le e^{\lambda^2 (b_i - a_i)^2/8} \le e^{\lambda^2 L_i^2/8}$$

by the fact that R.V. bounded in an interval of length L_i are $L_i/2$ subgaussian (for the right constant check MW Exercise 2.4., for an easier proof for the wrong constant check MW Example 2.4.) and hence a.s. the random variable $\mathbb{E}[e^{\lambda D_i} | Z_1^{i-1}] \le e^{\lambda^2 L_i^2/8}$

2. If D_i are independent, we have $\mathbb{E}e^{\lambda \sum_{i=1}^n D_i} = \prod_{i=1}^n \mathbb{E}e^{\lambda D_i}$

3. Note that since D_i are $\mathcal{F}_i$-measurable by definition of martingale difference sequence, we have $\mathbb{E}[e^{\lambda D_i} | G] = e^{\lambda D_i}$ for all $G \in \mathcal{F}_i$

4. Now using the tower property (TP) of conditional expectations iteratively, we see that $\sum_{i=1}^n D_i$ is $\sqrt{\sum_{i=1}^n \frac{L_i^2}{4}}$-subgaussian:

$$\mathbb{E}e^{\lambda \sum_{i=1}^n D_i} \overset{(TP)}{=} \mathbb{E}[\mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} e^{\lambda D_n} | Z_1, \ldots, Z_{n-1}]]$$
$$\overset{(3.)}{=} \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} \mathbb{E}[e^{\lambda D_n} | Z_1, \ldots, Z_{n-1}]] \le e^{\lambda^2 L_i^2/8} \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i}] = e^{\lambda^2 \sum_{i=1}^n L_i^2/8}$$

Bounding $\mathrm{Res}(n, \mathcal{F})$, Rademacher complexity

Today we use shorthand $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$ and write the uniform tail bound this way. Then we have

$$\sup_{f \in \mathcal{F}} \mathbb{E}\ell(Z, f) - \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, f) = \sup_{h \in \mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^{n} h(Z_i)$$

and it follows that

$$\mathbb{P}(\sup_{h \in \mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^{n} h(Z_i) \geq \mathrm{Res}(n, \mathcal{F}) + t) \leq \mathrm{e}^{-\frac{nt^2}{2b^2}} \qquad (2)$$

The next four sessions will be about how to bound $\mathrm{Res}(n, \mathcal{F})$!

Step I (this week): we first use eq. 2 & that $\mathrm{Res}(n, \mathcal{F})$ is bounded by

Definition (Rademacher complexity)

Given a function class $\mathcal{H}$ and distribution $\mathbb{P}$ on its domain $\mathcal{Z}$, for i.i.d. Rademacher R.V. ϵ_i, we define the Rademacher complexity as

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon, Z} \sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(Z_i)$$

Step II (next 2 weeks): We'll discuss how to bound $\mathcal{R}_n(\mathcal{H})$ as a function of $n, \mathcal{H}$

Step I: Uniform law with Rademacher complexity

Theorem (Uniform law for the risk, MW Thm 4.10.)

For b-unif. bounded $\mathcal{H}$, with prob. over the training data

$$\mathbb{P}(\sup_{h \in \mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^{n} h(Z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t) \leq e^{-\frac{nt^2}{2b^2}}$$

- By using $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$ we get

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon, Z} \sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(Z_i) = \mathbb{E}_{\epsilon, Z} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \ell(Z_i, f)$$

  and after showing $\mathrm{Res}(n, \mathcal{F}) \leq 2\mathcal{R}_n(\mathcal{H})$, directly obtain our desired bound on $\sup_{f \in \mathcal{F}} R(f) - R_n(f)$

- Note if $\mathcal{R}_n(\mathcal{H}) = o(1)$, then $\sup_{f \in \mathcal{F}} R(f) - R_n(f) \overset{a.s.}{\to} 0$.
- Before the proof, we aim to gain some intuition for the quantity $\mathcal{R}_n(\mathcal{H})$ and how it may behave with different n and $\mathcal{H}$

References

Azuma-Hoeffding

- MW Chapter 2

Online to batch conversion with Azuma-Hoeffding

- https://home.ttic.edu/~tewari/lectures/lecture13.pdf

Uniform law and Rademacher complexity

- MW Chapter 4

Lecture 4: Uniform law and Rademacher complexity

FAQ for muddiest point

Online learning

- added more motivation and explanation, also lecture note from Tewari, Kakade.

Azuma-Hoeffding

- How martingale properties allow the AH bound (will discuss now)

Questions on the uniform law - will discuss today

Plans for today

- Recap Azuma-Hoeffding proof
- Intuition for Rademacher complexity
- Proof of uniform law with symmetrization
- Application of Rademacher complexity: VC bound for binary classification
  - Proof of VC bound using uniform law
  - Proof of Massart's lemma

Recap: From Hoeffding to Azuma-Hoeffding

Why use $D_i = \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_i] - \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_{i-1}]$ to decompose g_n? Azuma-Hoeffding is a *generalization* of Hoeffding (i.e. Azuma-Hoeffding implies Hoeffding), for functions of n independent R.V. instead of sum of n independent RV.

Decomposition of g_n

- For Hoeffding, in $\frac{1}{n}\sum_{i=1}^n X_i$ for X_i independent, each R.V. adds fresh randomness $\rightarrow$
- For AH, in decomposition $\sum_{i=1}^n D_i$, each D_i has the additional randomness that is due to addition of Z_i only. This is why we chose the particular D_i (property 1 next slide)

In addition the D_i are in some sense **bounded** (for McDiarmid, generally subgaussian is fine), so sth "like Hoeffding" should work:

- For Hoeffding, each summand is subgaussian $\rightarrow$
- For AH (for proving McDiarmid), each summand is conditionally a.s. bounded and hence also conditionally subgaussian (property 2)

Recap: Martingale properties to prove Azuma-Hoeffding

The following properties of this choice are what we need in the proof (these are the properties of martingale differences)

1. D_i is $\mathcal{F}_i$ measurable, i.e. D_i is a deterministic function given specific values for $Z_1, \ldots, Z_i$
2. For any values $z_1, \ldots, z_{i-1}$, for some a_i, b_i
   - the random variable $D_i | Z_1^{i-1} = z_1^{i-1}$ is bounded in an interval $[a_i, b_i]$ of length L_i and
   - $\mathbb{E}[D_i | Z_1^{i-1} = z_1^{i-1}] = 0$ and hence together we use the fact that r.v. bounded a.s. in $[a_i, b_i]$ are $\frac{b_i - a_i}{2}$ subgaussian to get

$$\mathbb{E}[e^{\lambda(D_i - \mathbb{E}[D_i | Z_1^{i-1} = z_1^{i-1}])} | Z_1^{i-1} = z_1^{i-1}] \leq e^{\lambda^2 (b_i - a_i)^2/8} \leq e^{\lambda^2 L_i^2/8}$$

Further we use the tower property (TP): $\mathbb{E}[\mathbb{E}[X|Y,Z]|Y] = \mathbb{E}[X|Y]$

$$\mathbb{E} e^{\lambda \sum_{i=1}^n D_i} \overset{(TP)}{=} \mathbb{E}[\mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} e^{\lambda D_n} | Z_1, \ldots, Z_{n-1}]]$$

$$\overset{(1.)}{=} \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} \mathbb{E}[e^{\lambda D_n} | Z_1, \ldots, Z_{n-1}]] \overset{(2.)}{\leq} e^{\lambda^2 L_i^2/8} \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i}] = e^{\lambda^2 \sum_{i=1}^n L_i^2/8}$$

Recap: Uniform tail bound via Rademacher complexity

- Define $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$
- ϵ_i are i.i.d. Rademacher R.V.
- $Z = \{Z_i\}_{i=1}^n$ are training points $\overset{iid}{\sim} \mathbb{P}$

Definition (Rademacher complexity)

Given a function class $\mathcal{H}$ and distribution $\mathbb{P}$ on its domain $\mathcal{Z}$, we define the Rademacher complexity as

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon, z} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(z_i)$$

Theorem (Uniform law for the risk, MW Thm 4.10.)

For b-unif. bounded $\mathcal{H}$, with prob. over training data,

$$\mathbb{P}(\sup_{h \in \mathcal{H}} [\mathbb{E}h - \frac{1}{n} \sum_{i=1}^n h(Z_i)] \geq 2\mathcal{R}_n(\mathcal{H}) + t) \leq e^{-\frac{nt^2}{2b^2}}$$

Intuition for Rademacher complexity

Consider binary classification setting $\ell(z_i; f) = \mathbb{1}(f(x_i)y_i < 0)$.

1. How does the empirical Rademacher complexity

$$\tilde{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \ell(z_i, f)$$
$$\text{with } \mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$$

   depend on the factors $\mathcal{F}, \ell, n$ to control excess risk?

2. What is the connection between R.C. and VC dimension?

3. (Why) is it easier to reason about than the original $\text{Res}(n, \mathcal{H}) = \mathbb{E}g_n(Z)$

Some answers

- If $\mathcal{F}$ larger $\to$ $\mathcal{H}$ larger $\to$ $\tilde{\mathcal{R}}_n(\mathcal{H})$ larger (VC dim)
- Similarly if ℓ has small variance $\to$ $\tilde{\mathcal{R}}_n(\mathcal{H})$ is smaller (Lipschitz)
- As n grows, harder to fit $\to$ $\tilde{\mathcal{R}}_n(\mathcal{H})$ smaller

Intuition (see figures in handwritten notes)

- Let's look $\tilde{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i h(z_i)$ for fixed z_i and $h(z_i) = \ell(z_i; f)$ and see how it might decrease with n
- For simplicity, let $\mathcal{Z} = \mathbb{R}$, use e.g. $h(z) = \text{sgn} f(z)$ (you can do it more generally for ℓ)
- Let $\mathcal{F}$ be "smooth" functions, given a draw/sample $\epsilon_1, \ldots, \epsilon_n$

Which $f \in \mathcal{F}$ can achieve large $\tilde{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i \ell(z_i, f)$?

- Maximizing $\tilde{\mathcal{R}}_n(\mathcal{H})$ requires for each $\{\epsilon_i\}_{i=1}^n$ matching "induced labeling" of f $(\{f(z_i)\}_{i=1}^n)$
- For small n, you can find a f for each sample of $\{\epsilon_i\}_{i=1}^n$ that matches in sign, i.e. $|\{(h(z_1), \ldots, h(z_n)) : h \in \mathcal{H}\}| = 2^n$, then $\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i h(z_i) = 1$
- For large n, points are too dense, if $\mathcal{F}$ need to be smooth, not that possible for some very "wiggly" $\{\epsilon_i\}_{i=1}^n \to \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i h(z_i)$

Caveats of the uniform law

- Requires boundedness of ℓ (for bounded differences)
  - for regression you also bound suprema of empirical processes, can use Gaussian complexity and Lipschitz-of-Gaussians rule (see MW 3)
  - or argue that ℓ bounded with high probability, cause X and hence $f(X)$ bounded for continuous f
- Super loose bound $\rightarrow \mathcal{F}$ needs to be algorithm / data dependent
  - we will see for regularized optimizers
  - structural risk minimization
- in second half of lectures we'll discuss a different way to bound for regression $\rightarrow$ however even there concentration of suprema of empirical processes will be needed

Proof of uniform law - Step I: Tail bound

Theorem (Uniform tail bound)

For b-unif. bounded ℓ, it holds that

$$\mathbb{P}(\sup_{f\in\mathcal{F}} R(f) - R_n(f) \geq \mathbb{E}[\sup_{f\in\mathcal{F}} R(f) - R_n(f)] + t) \leq e^{-\frac{nt^2}{2b^2}}$$

where the probability is over the training data.

We recapped the proof last lecture, using McDiarmid.

In particular, by the uniform tail bound, if we can prove that $\mathbb{E}[\sup_{f\in\mathcal{F}} R(f) - R_n(f)] \leq 2\mathcal{R}_n(\mathcal{H})$ then it immediately follows that

$$\mathbb{P}(\sup_{h\in\mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^{n} h(Z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t)$$

$$\leq \mathbb{P}(\sup_{f\in\mathcal{F}} R(f) - R_n(f) \geq \mathbb{E}[\sup_{f\in\mathcal{F}} R(f) - R_n(f)] + t) \leq \mathrm{e}^{-\frac{nt^2}{2b^2}}$$

This proof step is called symmetrization

Proof of uniform law - Step II: Symmetrization

(i) For any H, $\sup_H \mathbb{E}H(Z) \leq \mathbb{E}\sup_H H(Z)$ (Exercise)

(ii) $h(Z_i) - h(\tilde{Z}_i)$ is symmetric $\rightarrow$ multiplying by ϵ_i preserves distr.

$$\begin{aligned}
\mathbb{E}_Z g_n(Z) &= \mathbb{E}_Z \sup_{h\in\mathcal{H}} \mathbb{E}h - \frac{1}{n}\sum_i h(Z_i) \\
&= \mathbb{E}_Z \sup_{h\in\mathcal{H}} \mathbb{E}_{\tilde{Z}} \frac{1}{n}\sum_{i=1}^n h(\tilde{Z}_i) - \frac{1}{n}\sum_{i=1}^n h(Z_i) \\
&\overset{(i)}{\leq} \mathbb{E}_{Z,\tilde{Z}} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n [h(Z_i) - h(\tilde{Z}_i)] \\
&\overset{(ii)}{=} \mathbb{E}_{Z,\tilde{Z},\epsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i [h(Z_i) - h(\tilde{Z}_i)] \\
&\leq 2\mathbb{E}_{Z,\epsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(Z_i) =: 2\mathcal{R}_n(\mathcal{H})\square
\end{aligned}$$

- Tight: $\frac{\mathcal{R}_n(\mathcal{H})}{2} \leq \mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_i h - \mathbb{E}h \leq 2\mathcal{R}_n(\mathcal{H})$ (MW Prop 4.11.)

Classification setup

- Labels are now in discrete domain $y \in \{-1, +1\}$
- Given f, we predict the label of some x using $\hat{y} = sign(f(x))$
- Evaluation metric: $\ell((x, y); f) = \mathbb{1}_{\{yf(x)<0\}}$ and hence population risk: $R(f) = \mathbb{E}\ell((x, y); f) = \mathbb{P}(y \neq sign(f(x))$
- A fixed $f \in \mathcal{F}$ defines a labeling from domain $\mathcal{X} \rightarrow \{-1, +1\}$. For a *given set* $Z^n = \{Z_i = (x_i, y_i)\}_{i=1}^n$, the function space $\mathcal{F}$ induces a set in $\{-1, 1\}^n$ that reads $\mathcal{F}(Z^n) = \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}$
- We again use notation $h(z) = \ell(z, f)$ and define

$$\mathcal{H}(Z^n) = \{(\ell(Z_1; f), \ldots, \ell(Z_n; f)) : f \in \mathcal{F}\}$$

Notice that $|\mathcal{F}(Z^n)| = |\mathcal{H}(Z^n)|$

Classification generalization bound

Recap **definition VC dimension** for binary classification: Biggest $n \in \mathbb{N}$ s.t. there exists $Z^n \in \mathcal{Z}^n$ with $\mathcal{H}(Z^n) = \{0,1\}^n$

Finite VC dimension can make $\mathcal{H}$ Glivenko-Cantelli, i.e. $\mathcal{R}_n(\mathcal{H}) = o(1)$. With your neighbor, use that

$$\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) \leq \sqrt{\frac{2d_{\text{VC}} \log(n+1)}{n}}$$

to prove the following bound

Theorem (uniform VC bound)

If $\mathcal{H}$ has VC dimension d_{VC}, w/ prob $\geq 1-\delta$ for any estimator $f \in \mathcal{F}$

$$\mathbb{P}(yf(X) < 0) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i f(x_i)<0} + 4\sqrt{\frac{d_{VC}\log(n+1)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

Proof of VC bound

Lemma (Massart)

For n points $Z^n := \{Z_1, \ldots, Z_n\}$, let all $h : \mathcal{Z} \to \{0,1\}$ and $\mathcal{H}(Z^n) := \{(h(Z_1), \ldots, h(Z_n)) : h \in \mathcal{H}\}$ with cardinality $|\mathcal{H}(Z^n)|$.

$$\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) := \mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(Z_i) \leq \sqrt{\frac{2\log|\mathcal{H}(Z^n)|}{n}}$$

- $|\mathcal{H}(Z^n)|$ corresponds to # labelings for Z^n induced by $\mathcal{H}$
- if $|\mathcal{H}(Z^n)|$ grows exponentially $\to \tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) = O(1)$

Lemma (Sauer-Shelah, MW Prop 4.18.)

If $\mathcal{F}$ has VC dimension d_{VC}, then for any $Z_1, \ldots, Z_n$ we have growth function $N_{\mathcal{H}}(n) := \sup_{Z^n\in\mathcal{Z}^n} |\mathcal{H}(Z^n)| \leq (n+1)^{d_{VC}}$ for all $n \geq d_{VC}$.

Hence can use $\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{Z^n}\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) \leq \sup_{Z^n\in\mathcal{Z}^n} \tilde{\mathcal{R}}_n(\mathcal{H}(Z^n))$ and Massart with Sauer-Shelah (loose since distribution independent!) in the uniform law to yield result

Proof of Massart

Lemma (Massart)

For n points $Z^n := \{Z_1, \dots, Z_n\}$, let all $h : \mathcal{Z} \to \{0, 1\}$ and $\mathcal{H}(Z^n) := \{(h(Z_1), \dots, h(Z_n)) : h \in \mathcal{H}\}$ with cardinality $|\mathcal{H}(Z^n)|$.

$$\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) := \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(Z_i) \leq \sqrt{\frac{2 \log |\mathcal{H}(Z^n)|}{n}}$$

- Step 1: For Rademacher ϵ_i and any Z_1^n we have that $\theta_i := h(Z_i) \in \{0, 1\}$, show $\frac{1}{n} \epsilon^\top \theta$ is zero-mean and $\frac{1}{\sqrt{n}}$ sub-gaussian (similar to Hoeffding proof). This follows from the fact that $[a_i, b_i]$ bounded r.v. are $[b_i - a_i]/2$ subgaussian
- Step 2: Use the fact from HW 1 that, for N zero-mean subgaussians $X_1, \dots, X_N$ with sub-gaussian parameter σ

$$\mathbb{E} \max_{i=1..N} X_i \leq \sqrt{2\sigma^2 \log N}$$

  Here, $N = \mathcal{H}(Z^n)$ the number of different vectors $(h(Z_1), \dots, h(Z_n))$

References

Uniform law

- MW Chapter 4
- "Understanding machine learning" by Shalev-Shwartz, Ben-David, Chapter 26

Lecture 5: VC bound and margin bound

Announcements

- Homework 1 due Thursday 23:59
- Moodle finally has forums to ask questions re HW or lecture (just realized yesterday)
- Project sign-ups Monday 14:00 - find your partner on moodle If you want to present a paper not on the list, please double check with us.

Feedback compilation

- Good: interactivity, intuition
- can be improved: handwriting, references to some results that are not explicitly noted in MW (adding some from SS), more intuition before proof but also more proof details

About project choice

1. Identify and motivate problem - why should I / the community care? Including literature review (done-ish)
2. "Detective hat": Intuitive (not just technical level) understanding of proof, assumptions, statement in depth
3. "Reviewer hat": Which relevant questions does it shed light on and does the paper answer/shed light on it? How significant is the addition of this paper compared to existing literature? This is a key step towards Step 4.
4. "Researcher hat": What are **interesting, impactful** follow-up questions they did not answer and would be interesting and perhaps feasible to pursue?
5. Break down the identified follow-up problem into feasible chunks (e.g. lemmas, experiments) and optionally show your attempts to tackle the first few steps.

Outline for today

- VC bound and proof
- Rademacher contraction
- Interactive: Proof using the ramp loss and contraction (students)

Recap: Massart's lemma

Note: in this lecture, we often write $z^n := z_1^n$ and the same for x.

Last time, we bounded the Rademacher for function classes $\mathcal{F}$ that induce a finite set $\mathcal{H}(Z^n) = \{(\ell(Z_1; f), \dots, \ell(Z_n; f)) : f \in \mathcal{F}\}$ using Massart's lemma

Lemma (Massart, SS Lemma 26.8)

For n points $Z^n := \{Z_1, \dots, Z_n\}$, let all $h : \mathcal{Z} \to \{0, 1\}$ and $\mathcal{H}(Z^n) := \{(h(Z_1), \dots, h(Z_n)) : h \in \mathcal{H}\}$ with cardinality $|\mathcal{H}(Z^n)|$.

$$\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) := \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(Z_i) \leq \sqrt{\frac{2 \log |\mathcal{H}(Z^n)|}{n}}$$

- $|\mathcal{H}(Z^n)|$ corresponds to # labelings for Z^n induced by $\mathcal{H}$
- if $|\mathcal{H}(Z^n)|$ grows exponentially $\to \tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) = O(1)$

VC bound

We now use Massart to upper bound the generalization gap $R(f) - R_n(f)$ for function classes of finite VC dimension, where $|\mathcal{H}(Z^n)|$ does not grow exponentially in n for any Z^n.

Recap **definition VC dimension** for binary classification:

Definition (VC dimension)

Biggest $n \in \mathbb{N}$ s.t. there exists $Z^n \in \mathcal{Z}^n$ with $\mathcal{H}(Z^n) = \{0, 1\}^n$

Function classes $\mathcal{F}$ with finite VC dimension can make $\mathcal{H}$ Glivenko-Cantelli, i.e. $\mathcal{R}_n(\mathcal{H}) = o(1)$. More specifically:

Theorem (uniform VC bound)

If $\mathcal{H}$ has VC dimension d_{VC}, w/ prob $\geq 1 - \delta$ for any estimator $f \in \mathcal{F}$

$$\mathbb{P}(yf(X) < 0) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{y_i f(x_i) < 0} + 4\sqrt{\frac{d_{VC} \log(n+1)}{n}} + \sqrt{\frac{2 \log(1/\delta)}{n}}$$

Proof of VC bound 1

Now we first prove a high-probability upper bound for the population 0-1 loss $\ell((x,y);f) = \mathbb{1}_{yf(x)<0}$ for finite function classes $\mathcal{F}$.

Plugging in the definition of the loss, using the uniform law, we get

$$\mathbb{P}(Yf(X) < 0) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i f(x_i)<0} + 2\mathcal{R}_n(\mathcal{H}) + c\sqrt{\frac{\log(1/\delta)}{n}} \qquad (1)$$

for some universal constant c. The proof uses the uniform law (U.L.)

$$\begin{aligned} R(f) - R_n(f) &= \mathbb{E}\ell((x,y);f) - \frac{1}{n}\sum_{i=1}^{n}\ell((x,y);f) \\ &= \mathbb{P}(yf(x) < 0) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{y_i f(x_i)<0} \\ &\leq \sup_{f\in\mathcal{F}} R(f) - R_n(f) \overset{U.L.}{\leq} 2\mathcal{R}_n(\mathcal{H}) + c\sqrt{\frac{\log(1/\delta)}{n}} \end{aligned}$$

Proof of VC bound 2

- Note that $\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{Z^n}\tilde{\mathcal{R}}_n(\mathcal{H}) \leq \sup_{Z^n}\tilde{\mathcal{R}}_n(\mathcal{H})$ (this is crude!)
- Further by Massart, $\sup_{Z^n}\tilde{\mathcal{R}}_n(\mathcal{H}) \leq \sup_{Z^n}\sqrt{\frac{2\log|\mathcal{H}(Z^n)|}{n}}$ yielding

$$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2\log\sup_{Z^n}|\mathcal{H}(Z^n)|}{n}} \qquad (2)$$

(loose since distribution independent!)

Furthermore, we have the following upper bound on the size of $\mathcal{H}(Z^n)$

Lemma (Sauer-Shelah, MW Prop 4.18.)

If $\mathcal{F}$ has VC dimension d_{VC}, then for any $Z^n = Z_1,\ldots,Z_n$ we have growth function $N_{\mathcal{H}}(n) := \sup_{Z^n\in\mathcal{Z}^n}|\mathcal{H}(Z^n)| \leq (n+1)^{d_{VC}}$ for all $n \geq d_{VC}$.

Plugging Sauer-Shelah into eq. 2, and that into eq. 1 in the uniform law to yield result

Empirical Rademacher complexity - notation

In the following, we will slightly abuse notation and write the more general empirical Rademacher complexity for $\mathbb{T} \subset \mathbb{R}^n$ as

$$\tilde{\mathcal{R}}_n(\mathbb{T}) = \mathbb{E} \sup_{\theta \in \mathbb{T}} \sum_{i=1}^{n} \epsilon_i \theta_i.$$

Note that hence we can write $\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n))$ for $\tilde{\mathcal{R}}_n(\mathcal{H})$.

The following lemma can connect the empirical Rademacher comp. of a function class $\widetilde{\mathcal{F}}$ to the empirical Rademacher comp. of a specific loss $\ell : \mathbb{R} \to \mathbb{R}$ acting on a function class, specifically when $\mathcal{H} = \ell \circ \widetilde{\mathcal{F}}$

First note that for $\mathbb{T} = \widetilde{\mathcal{F}}(Z^n)$ we can write the empirical Rademacher complexity in two ways (abusing notation)

$$\tilde{\mathcal{R}}_n(\ell \circ \widetilde{\mathcal{F}}) = \mathbb{E} \sup_{\tilde{f} \in \widetilde{\mathcal{F}}} \sum_{i=1}^{n} \epsilon_i \ell(\tilde{f}(Z_i)) \text{ same as}$$

$$\tilde{\mathcal{R}}_n(\ell \circ \mathbb{T}) = \mathbb{E} \sup_{\theta \in \mathbb{T}} \sum_{i=1}^{n} \epsilon_i \ell(\theta_i)$$

Rademacher contraction

In the case of classification, we often have a loss of the form (again, slightly abusing notation) $\ell(Z_i, f) = \ell(Y_i f(X_i))$ and can define $\tilde{f}(Z_i) = Y_i f(X_i)$.

The following lemma holds for general losses $\boldsymbol{\ell} : \mathbb{R}^n \to \mathbb{R}^n$ (again, abuse of notation) where the loss may differ for each element, with $\boldsymbol{\ell}(\theta) = (\ell_1(\theta_1), \ldots, \ell_n(\theta_n))$ with $L-$Lipschitz $\ell_j : \mathbb{R} \to \mathbb{R}$, i.e.

$$|\ell_j(a) - \ell_j(b)| \leq L|a - b| \text{ for all } a, b \in \mathbb{R}.$$

Lemma (Rademacher contraction, SS Lemma 26.9)

For any $\mathbb{T} \subset \mathbb{R}^n$ and $\boldsymbol{\ell} : \mathbb{R}^n \to \mathbb{R}^n$ with univariate L-Lipschitz functions it holds that

$$\tilde{\mathcal{R}}_n(\boldsymbol{\ell} \circ \mathbb{T}) \leq L \tilde{\mathcal{R}}_n(\mathbb{T})$$

In the following when $\ell_i = \ell$ for all i, then $\tilde{\mathcal{R}}_n(\boldsymbol{\ell} \circ \mathbb{T}) = \tilde{\mathcal{R}}_n(\ell \circ \mathbb{T})$ as in the previous slide.

Skipped during lecture: Proof ingredients

Let ϵ be the vector of n i.i.d. Rademacher r.v. and define the shorthand $\epsilon_{2:n} = (\epsilon_2, \ldots, \epsilon_n)$ and same for θ.

The following holds for all n

- Key 1: de-symmetrize using the tower property: For any g we have $\mathbb{E}_\epsilon g(\epsilon) = \mathbb{E}_{\epsilon_1}[\mathbb{E}[g(\epsilon)|\epsilon_1]] = \frac{1}{2}\mathbb{E}[g(\epsilon)|\epsilon_1 = 1] + \frac{1}{2}\mathbb{E}g(\epsilon)|\epsilon = -1]$
- Key 2: Lipschitz property $\ell_i(\theta_i) - \ell_i(\tilde{\theta}_i) \leq L|\theta_i - \tilde{\theta}_i|$ for all i
- Key 3: For each ϵ we can define $h(\theta_{2:n}) = \sum_{i=2}^n \epsilon_i \ell_i(\theta_i)$. One can prove via contradiction that

$$\sup_{\theta,\tilde{\theta}\in\mathbb{T}} |\theta_1 - \tilde{\theta}_1| + h(\theta_{2:n}) + h(\tilde{\theta}_{2:n}) = \sup_{\substack{\theta,\tilde{\theta}\in\mathbb{T} \\ \theta_1 \geq \tilde{\theta}_1}} \theta_1 - \tilde{\theta}_1 + h(\theta_{2:n}) + h(\tilde{\theta}_{2:n})$$

Skipped during lecture: R.C. contraction proof

$$n\tilde{\mathcal{R}}_n(\ell \circ \mathbb{T}) = \mathbb{E}_\epsilon \sup_{\theta\in\mathbb{T}} \sum_{i=1}^n \epsilon_i \ell_i(\theta_i)$$

$$\overset{1.}{=} \frac{1}{2}\left[\mathbb{E}_{\epsilon_{2:n}} \sup_{\theta\in\mathbb{T}} \ell_1(\theta_1) + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sup_{\tilde{\theta}\in\mathbb{T}} -\ell_1(\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i)\right]$$

$$= \frac{1}{2}\left[\mathbb{E}_{\epsilon_{2:n}} \sup_{\theta,\tilde{\theta}\in\mathbb{T}} \ell_1(\theta_1) - \ell_1(\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i)\right]$$

$$\overset{2.}{\leq} \frac{1}{2}\left[\mathbb{E}_{\epsilon_{2:n}} \sup_{\theta,\tilde{\theta}\in\mathbb{T}} L|\theta_1 - \tilde{\theta}_1| + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i)\right]$$

$$\overset{3.}{=} \frac{1}{2}\left[\mathbb{E}_{\epsilon_{2:n}} \sup_{\theta\in\mathbb{T}} L\theta_1 + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i) + \sup_{\tilde{\theta}\in\mathbb{T}}(-L\tilde{\theta}_1) + \sum_{i=2}^n \epsilon_i \ell_i(\tilde{\theta}_i)\right]$$

$$\overset{1.}{=} \mathbb{E}_\epsilon \sup_{\theta\in\mathbb{T}} L\epsilon_1\theta_1 + \sum_{i=2}^n \epsilon_i \ell_i(\theta_i)$$

Use the same argument for the RHS inductively on each coordinate. □

Mimicking proof-based research in collaboration

- Learning objectives: Both for actual guarantees and presentation, collaboration
  1. Get intuition why a problem / conjecture should be true
  2. Break down a proof to parts
  3. Prove individual parts
- Matching questions in the interactive session today
  1. Intuitively why should enforcing a large margin yield better generalization? Show graphically (no right or wrong)
  2. Given contraction inequality, ramp loss and Rademacher complexity for linear functions, prove the margin bound
  3. Prove Rademacher complexity for linear function class

Instructions

- Groups:
  - We will divide the class into three groups of ≈ 4 people each.
  - Each group will solve one of the three questions jointly.
  - Once you know your group, choose a representative to present later
- Group work:
  - 15 minutes of discussion to solve the question - if done early, feel free to solve another groups' question
  - Another 5 minutes to prepare the representative's blackboard presentation
- Final presentation
  - 30 minutes of 3 short presentations (7 min presentation, 3 min Q&A)
  - Introduce yourself and group members by names
  - Present your results.

Primer on margins for linear classifiers

- Class of linear classifiers $\mathcal{F} = \{f : f(x) = w^\top x \; w \in \mathbb{R}^d\}$
- Intuition in introductory lectures for linearly separable data: large minimum distance to the boundary is good that can be computed as

$$d_{\min} = \min_i y_i \frac{w^\top x_i}{\|w\|_2}$$

  where $\min_i y_i \langle w, x_i \rangle$ is called the margin
- Can obtain set of maximizing directions by solving

$$\max_{\gamma, w} \gamma \text{ s.t. } y_i \langle \frac{w}{\|w\|_2}, x_i \rangle \geq \gamma$$

  which for bounded $\|w\|_2 \leq B$ is the same as solving

$$\max_{\gamma', \|w\|_2 \leq B} \gamma' \text{ s.t. } y_i \langle w, x_i \rangle \geq \gamma'$$

- We will look the generalization performance of feasible w with $\|w\|_2 \leq B$ which achieve a margin of at least some γ

Margin bound for binary classification

Key ingredient of proof (in interactive session)

Definition (ramp loss)

The ramp loss ℓ_γ is defined as

$$\ell_\gamma(u) = \begin{cases} 1 & u \in (-\infty, 0) \\ 1 - \frac{u}{\gamma} & u \in [0, \gamma] \\ 0 & u \in (\gamma, \infty) \end{cases}$$

and $\frac{1}{\gamma}$-Lipschitz.

Margin bound for linear classifiers

Definitions

- Set of linear functions $\mathcal{F}_B = \{f(x) = \langle w, x\rangle : \|w\|_2 \leq B\}$
- Define the risk $R_n^\gamma(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{y_i f(x_i)\leq\gamma}$ and $R^\gamma(f) = \mathbb{E}_{X,Y}\mathbb{1}_{Yf(X)\leq\gamma}$

Assumption (A): Boundedness of covariates $\mathbb{P}(\|x\|_2 \leq D) = 1$

Theorem (margin bound for linear classifiers)

If the assumptions are valid for any fixed γ, w/ prob. at least $1-\delta$, for any $f \in \mathcal{F}_B$ we have

$$R^0(f) = \mathbb{P}[y \neq sign(f(x))] \leq R_n^\gamma(f) + \frac{2DB}{\gamma\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{n}}$$

for some constant $c > 0$.

Solution: Proof of margin bound for linear classifiers

1. First we prove the following lemma

Lemma (uniform law with margin loss)

For $\mathcal{F}$ symmetric, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} R^0(f) - R_n^\gamma(f) \geq \frac{2}{\gamma}\mathcal{R}_n(\mathcal{F}) + t\right) \leq e^{-cnt^2}$$

2. Then we note that the class of linear functions $\mathcal{F}_B$ is symmetric and

Lemma (Rademacher complexity of bounded linear function class)

For $\mathcal{F}_B$ the empirical Rademacher complexity for specific $x_1, \ldots x_n$ is

$$\tilde{\mathcal{R}}_n(\mathcal{F}_B(x_1^n)) \leq \frac{B\max_i \|x_i\|_2}{\sqrt{n}}$$

so that $\mathcal{R}_n(\mathcal{F}_B) \leq \sup_{x_1^n\in\mathcal{X}_1^n} \tilde{\mathcal{R}}_n(\mathcal{F}_B(x_1^n)) \leq \frac{BD}{\sqrt{n}}$

3. Plugging in $t = c\sqrt{\frac{\log(1/\delta)}{n}}$ then yields the theorem. □

Solution: Proof of uniform law with margin loss

Define $R_{\ell_\gamma}(f) := \mathbb{E}_{(X,Y)}\ell_\gamma(Yf(X))$ and $R_{\ell_\gamma,n}(f)$ its empirical version. We first use the uniform law to bound $R_{\ell_\gamma}(f)$.

1. In particular, given $z_i = (x_i, y_i)$, define $\widetilde{\mathcal{F}}(z_1^n)$ by $\tilde{f}(z_i) = y_i f(x_i)$ for $f \in \mathcal{F}$. Because $\mathcal{F}$ is symmetric, we have $\widetilde{\mathcal{F}}(z_1^n) = \mathcal{F}(x_1^n)$
2. Defining $\mathcal{H}(z_1^n) = \{\ell_\gamma(\cdot, f) : f \in \mathcal{F}\}$ the Rademacher complexity reads
$$\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n)) = \tilde{\mathcal{R}}_n(\ell_\gamma \circ \mathcal{F}(x_1^n)).$$
3. The contraction inequality implies $\tilde{\mathcal{R}}_n(\ell_\gamma \circ \mathcal{F}(x_1^n)) \leq \frac{1}{\gamma}\tilde{\mathcal{R}}_n(\mathcal{F}(x_1^n))$ and the same holds when taking expectations
4. The uniform law then yields that w.p. $\geq 1 - \mathrm{e}^{-cnt^2}$
$$\sup_{f\in\mathcal{F}} R_{\ell_\gamma}(f) - R_{\ell_\gamma,n}(f) \leq \frac{2}{\gamma}\mathcal{R}_n(\mathcal{F}) + t$$
5. The lemma follows by noting that for every $\gamma > 0$ and any f it holds that $R^0(f) \leq R_{\ell_\gamma}(f)$ and $R_{\ell_\gamma,n}(f) \leq R_n^\gamma(f)$.

Solution: Rademacher complexity for linear classes

Proof of lemma via direct calculation

We utilize the fact that $\|x\|_2 = \sqrt{\|x\|_2^2}$ and that $\sqrt{\cdot}$ is a concave function whence Jensen's inequality yields

$$n\tilde{\mathcal{R}}_n(\mathcal{F}_B(x_1^n)) = \mathbb{E}_\epsilon \sup_w \sum_i \epsilon_i w^\top x_i \leq B\mathbb{E}_\epsilon \|\sum_i \epsilon_i x_i\|$$
$$= B\sqrt{\mathbb{E}_\epsilon \|\sum_i \epsilon_i x_i\|^2} = B\sqrt{\sum_i \|x_i\|^2} \leq B\sqrt{n}\max_i \|x_i\|_2$$

In contrast: Rade. Comp. via VC Dimension

1. VC dimension of a class of linear classifiers (without bias term!) in R^d is d ($d_{\mathrm{VC}} \geq d$ is clear, $d_{\mathrm{VC}} \leq d$ via construction using linear dependence for $d+1$ points)
2. Then, using the VC bound we would obtain a bound of the order $\sqrt{\frac{d\log(n+1)}{n}}$, which is generally much larger then the dimension independent B.

References

- Massart, Rademacher for classification: Shalev-Schwartz & Ben-David Chapter 26

Lecture 6: Covering and metric entropy

Announcements

- HW was due, thanks for handing in
- HW solutions will be up end of this week. HW2 will be up in 1.5 weeks, i.e. **27.10.**
- Thanks for signing up for projects - a few have not yet signed up
- Project proposals due Friday, **24.10. 23:59** - send to konstantin.donhauser at inf.ethz.ch via email

Plan today

- Rademacher complexity as supremum of subgaussian process
- Bounding the supremum using max of subgaussian result and covering argument (metric entropy)
- Examples beyond linear functions

Recap: Uniform law

Recap $\mathcal{H} = \ell \circ \mathcal{F}$

Theorem (Uniform law for the risk)

For b-unif. bounded $\mathcal{H}$, with prob. over the training data

$$\mathbb{P}(\sup_{h\in\mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^{n} h(Z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t) \leq e^{-\frac{nt^2}{2b^2}}$$

Our task was then to bound $\overbrace{\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n))}$

$$\mathcal{R}_n(\mathcal{H}) := \mathbb{E}_z \overbrace{\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_i \epsilon_i h(z_i)}^{\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n))} =: \mathcal{R}_n(\mathcal{H})$$

Here, we write $\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n))$ (where we stress dependence on samples) for $\tilde{\mathcal{R}}_n(\mathcal{H})$ with a slight abuse of notation. More generally, for any set $\mathbb{T} \subset \mathbb{R}^n$ we define

$$\tilde{\mathcal{R}}_n(\mathbb{T}) = \mathbb{E}_\epsilon \sup_{\theta\in\mathbb{T}} \sum_{i=1}^{n} \epsilon_i \theta_i.$$

Recap: VC bound vs. margin bound

Last lecture, we obtained a completely distribution independent VC bound of the Rademacher complexity via

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{Z^n}\tilde{\mathcal{R}}_n(\mathcal{H}) \leq \sup_{Z^n} \tilde{\mathcal{R}}_n(\mathcal{H}(Z_1^n))$$

by bounding the RHS via the VC dimension.

Q: How about the margin bound for linear functions? Is it to distribution dependent?

A: It depended on $D := \sup_{x\in\mathcal{X}} \|x\|_2$. When using the upper bound for the 0-1 loss (for some empirically trained $\widehat{f}$), it implicitly also depends on the margin of the distribution γ as that affects how small $R_n^\gamma(\widehat{f})$ can be.

Recap: Margin bound proof and Rademacher contraction

Assume that for some function class $\mathcal{F}$ all samples z_1^n from the distribution $\mathbb{P}$ can achieve a margin of γ

1. Define the proxy function class $\widetilde{\mathcal{F}}(z_1^n) = \{y_i f(x_i) : f \in \mathcal{F}\}$ function class. Then $\mathcal{H} := \{h : h(z) = \ell(z; f), f \in \mathcal{F}\} = \ell \circ \widetilde{\mathcal{F}}$
2. Rademacher contraction implies that (via uniform law) that L-Lipschitz loss functions would generalize better.
3. Then we can use the uniform law on the $\mathcal{H} = \ell_\gamma \circ \mathcal{F}$ with ramp loss ℓ_γ and obtain that with probability at least $1 - \delta$

$$R^0(f) \le R_{\ell_\gamma}(f) \le R_{\ell_\gamma,n}(f) + 2\mathcal{R}_n(\ell_\gamma \circ \widetilde{\mathcal{F}}) + \sqrt{\frac{c\log(1/\delta)}{n}}$$

$$\le R_n^\gamma(f) + \frac{2}{\gamma} \underbrace{\mathcal{R}_n(\widetilde{\mathcal{F}})}_{\le \sup_{x_1^n} \tilde{\mathcal{R}}_n(\widetilde{\mathcal{F}}(x_1^n))} + \sqrt{\frac{c\log(1/\delta)}{n}}$$

Intuition for Rademacher contraction on the board.

R.C. rates for different function classes

So far we bounded R.C. of finite VC classes, of linear (parametric) function classes by $O(\frac{1}{\sqrt{n}})$.

- Today we'll see examples for infinite-dimensional $\mathcal{F}$ where $\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n)) \le O(\frac{1}{n^\beta})$ for some $\beta \le 1/2$, for every z_1^n
- Then with probability at least $1 - \delta$, the generalization gap

$$\sup_{f \in \mathcal{F}} R(f) - R_n(f) \le O(\frac{1}{n^\beta}) + O(\sqrt{\frac{\log 1/\delta}{n}})$$

- For $\beta < 1/2$ the Rademacher term always dominates the excess risk since we have fast concentration for the sup of empirical process $\rightarrow$ the parametric $\sqrt{n}$ rate is "best one can hope for"

A general approach to bound the R.C.

- For finite classes $\to$ used max of subgaussians
- For special parameterization such as linear model $\to$ used boundedness of parameters and inputs

Today, we present a generic approach by

1. viewing the R.C. as the expecteed supremum of a subgaussian process
2. bounding the expected supremum of subgaussian processes via metric entropy

Definition (subgaussian process)

$\{X_\theta, \theta \in \mathbb{T}\}$ is a zero-mean subgaussian process if for all $\theta, \tilde{\theta} \in \mathbb{T}$, random variable $X_\theta - X_{\tilde{\theta}}$ is subgaussian w/ parameter $\rho(\theta, \tilde{\theta})$ for some metric ρ and $\mathbb{E}X_\theta = 0$

From R.C. to supremum of subgaussian processes

First note that we can write $\mathbb{T} \subset \mathbb{R}^n$

$$\tilde{\mathcal{R}}_n(\mathbb{T}) = \mathbb{E}_\epsilon \sup_{\theta \in \mathbb{T}} \frac{1}{n} \sum_i \epsilon_i \theta_i =: \frac{1}{\sqrt{n}} \mathbb{E}_\epsilon \sup_{\theta \in \mathbb{T}} X_\theta$$

where $X_\theta := \frac{1}{\sqrt{n}} \langle \epsilon, \theta \rangle$ and the scaling is chosen for later convenience

Then X_θ is a subgaussian process as per the next

Proposition (Rademacher as a sup of subgaussian processes)

For any $\mathbb{T}$, X_θ is a σ-subgaussian process with parameter $\sigma = \sup_{\theta, \tilde{\theta} \in \mathbb{T}} \rho(\theta, \tilde{\theta})$ where $\rho(\theta, \tilde{\theta}) = \frac{\|\theta - \tilde{\theta}\|_2}{\sqrt{n}}$ and it holds that

$$\sqrt{n} \tilde{\mathcal{R}}_n(\mathbb{T}) \leq \mathbb{E} \sup_{\theta, \theta' \in \mathbb{T}} X_\theta - X_{\theta'}$$

Proof of proposition

1. First $\mathbb{E}X_\theta = 0$ for all θ
2. $X_\theta - X_{\tilde{\theta}}$ is subgaussian wrt $\rho(\theta, \tilde{\theta}) := \frac{1}{\sqrt{n}}\|\theta - \tilde{\theta}\|_2 =: \|\theta - \tilde{\theta}\|_n$ since

$$\mathbb{E}e^{\lambda(X_\theta - X_{\tilde{\theta}})} = \mathbb{E}e^{\frac{\lambda}{\sqrt{n}}\sum_i \epsilon_i(\theta_i - \tilde{\theta}_i)} \leq \prod_i \mathbb{E}e^{\frac{\lambda(\theta_i - \tilde{\theta}_i)}{\sqrt{n}}\epsilon_i} \leq e^{\frac{\lambda^2 \frac{1}{n}\|\theta - \tilde{\theta}\|_2^2}{2}}$$

3. Because $\mathbb{E}X_{\tilde{\theta}} = 0$ for all $\tilde{\theta} \in \mathbb{T}$, we can then write empirical Rademacher complexity

$$\begin{aligned}\sqrt{n}\tilde{\mathcal{R}}_n(\mathbb{T}) &= \mathbb{E}_\epsilon \sup_{\theta \in \mathbb{T}} \frac{1}{\sqrt{n}}\langle \epsilon, \theta \rangle = \mathbb{E} \sup_{\theta \in \mathbb{T}} X_\theta - \mathbb{E}X_{\tilde{\theta}} \\ &\overset{(i)}{=} \mathbb{E} \sup_{\theta \in \mathbb{T}} X_\theta - X_{\tilde{\theta}} \leq \mathbb{E} \sup_{\theta, \tilde{\theta} \in \mathbb{T}} X_\theta - X_{\tilde{\theta}}\end{aligned}$$

where (i) holds because of linearity of expectation and for any $\tilde{\theta}$, which is smaller than sup-ing the difference over $\tilde{\theta}$

How can we leverage max of subgaussian lemma now?

For general function classes, the set e.g. $\mathbb{T} = \mathcal{H}(z_1^n)$ is infinite (even when it's bounded). How to get to a finite set to use max of subgaussians like in Massarts Lemma?

Main idea (high-level):

1. Cover $\mathbb{T}$ with a finite set of N points such that for any $\theta \in \mathbb{T}$, there is a point in the cover with distance $\leq \delta$
2. Can then take expected sup over grid points
3. Bound difference to other points again using naive bound

$$\frac{1}{\sqrt{n}}\mathbb{E}_\epsilon \sup_{\frac{\|\theta\|}{\sqrt{n}} \leq \delta} \frac{1}{\sqrt{n}} \sum_i \epsilon_i \theta_i \leq \delta \mathbb{E}_\epsilon \frac{\|\epsilon\|_2}{\sqrt{n}} \leq \delta$$

Bound using naive (1-step) covering argument

Proposition (using Pollard's bound - MW Prop 5.17)

Let $\delta > 0$. If a set of points $\theta^1, \dots, \theta^N$ satisfies $\min_j \rho(\theta, \theta^j) \leq \delta$ for all $\theta \in \mathbb{T}$ and $\sup_{\theta,\theta' \in \mathbb{T}} \rho(\theta, \theta') \leq \sigma$ with $\rho = \frac{\|\cdot\|_2}{\sqrt{n}}$, then we have

$$\tilde{\mathcal{R}}_n(\mathbb{T}) \leq 2[\delta + 2\sigma\sqrt{\frac{\log N(\delta)}{n}}]$$

Proof: For general ρ we can rewrite for any arbitrary $\theta, \tilde{\theta} \in \mathbb{T}$

$$\begin{aligned} X_\theta - X_{\tilde{\theta}} &= X_\theta - X_{\theta^\star} + X_{\theta^\star} - X_{\tilde{\theta}^\star} + X_{\tilde{\theta}^\star} - X_{\tilde{\theta}} \\ &= 2 \sup_{\rho(\theta,\theta') \leq \delta} X_\theta - X_{\theta'} + \max_{i,j \in [N]} X_{\theta^i} - X_{\theta^j} \end{aligned}$$

- Taking expectations, we obtain Pollard's bound for general ρ

$$\mathbb{E} \sup_{\theta,\tilde{\theta} \in \mathbb{T}} X_\theta - X_{\tilde{\theta}} \leq 2\mathbb{E} \sup_{\rho(\theta,\theta') \leq \delta} X_\theta - X_{\theta'} + 2\sqrt{2\sigma^2 \log N(\delta)}$$

  using the max of subgaussians upper bound you proved in HW1.

- Proposition follows by using specific ρ and 3. of previous slide $\square$.

How large is $N(\delta)$ for a given δ?

- **For a given** δ we'd like to find the **smallest number** N for which the condition in the proposition holds, depends δ and call this $N(\delta)$ (covering number, next slide).
- Then, we can choose δ to minimize $\delta + 2\sigma\sqrt{\frac{\log N(\delta)}{n}}$, i.e.

$$\tilde{\mathcal{R}}_n(\mathbb{T}) \leq 2 \inf_{\delta > 0} [\delta + 2D\sqrt{\frac{\log N(\delta)}{n}}]$$

In order for this term to decrease with n we require

- δ to decrease with n
- $N(\delta)$ not increase exponentially with decreasing δ.

Good example: $N(\delta) \sim 1/\delta$ and $\delta \sim \frac{1}{\sqrt{n}} \to \tilde{\mathcal{R}}_n(\mathbb{T}) \leq O(\sqrt{\frac{\log n}{n}})$

The minimum $N(\delta)$ for a given δ can be found using the covering number (next slide).

Covering number and entropy

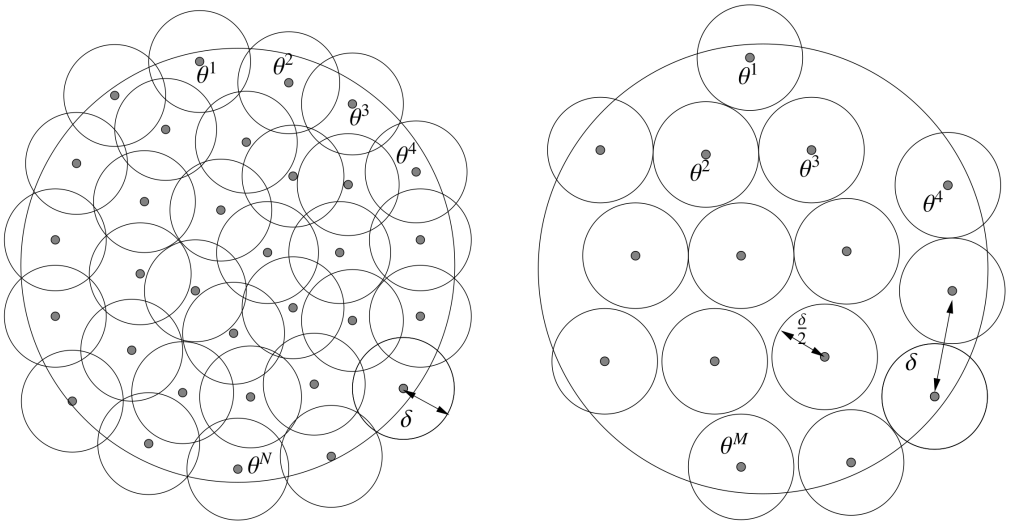


Figure 1: Left: δ-covering, Right: δ-packing

Definition (covering number, metric entropy)

For a metric ρ let the ϵ-*covering number* $\mathcal{N}(\epsilon; \mathbb{T}, \rho)$ be the smallest N such that a set of N points $S = \{\theta_i\}_{i=1}^N$ satisfies $\max_{\theta \in S} \min_i \rho(\theta_i, \theta) \leq \epsilon$ (S is ϵ-cover). *The metric entropy* is $\log \mathcal{N}(\epsilon; \mathbb{T}, \rho)$. Usually in our course $\mathcal{N} < \infty$ for any ϵ

Packing number

Definition (packing number)

The ϵ-*packing number* $\mathcal{M}(\epsilon; \mathbb{T}, \rho)$ is the biggest M such that a set of M points $S = \{\theta_i\}_{i=1}^M$ satisfies $\min_{i \neq j} \rho(\theta_i, \theta_j) \geq \epsilon$ (S is ϵ-packing).

Lemma (Packing vs. covering number - MW Lemma 5.5)

The following sandwich relationship holds
$\mathcal{M}(2\epsilon; \mathbb{T}, \rho) \leq \mathcal{N}(\epsilon; \mathbb{T}, \rho) \leq \mathcal{M}(\epsilon; \mathbb{T}, \rho)$

- Growth of $\mathcal{N}$ depends on
  - metric ρ on $\mathbb{T}$
  - for abstract $\mathbb{T}$: geometry of the set
  - for $\mathbb{T} = \mathcal{H}(z_1^n)$: covering/complexity of $\mathcal{H}$ (very loose!)

R.C. rates for function classes: Parametric example

We now contrast the covering numbers for a parametric and non-parametric function classes $\mathcal{H} = \mathcal{F}$ (i.e. identity/no loss), i.e. setting $\mathbb{T} = \mathcal{H}(z_1^n)$ and using the empirical error $\rho = \|\cdot\|_n := \frac{\|\theta-\theta'\|_2}{\sqrt{n}}$ as the metric.

Example I: Smoothly parameterized function class $\mathcal{H}_1$ with h s.t.

$$\sup_z |h(z;u) - h(z;u')| \leq L\|u - u'\|_2$$

where $u \in \mathbb{B}_2(1) \subset \mathbb{R}^d$ is the 2-norm ball of radius 1. For any z_1^n,

$$\mathcal{N}(\delta; \mathcal{H}(z_1^n), \|\cdot\|_n) \leq (1+\frac{2L}{\delta})^d \to \log \mathcal{N}(\delta; \mathcal{H}(z_1^n), \|\cdot\|_n) \asymp d \log(1+\frac{L}{\delta})$$

Further the set is bounded as

$$\|h(z_1^n; u) - h(z_1^n; u')\|_n \leq \|h(z;u) - h(z;u')\|_\infty \leq L\|u - u'\|_2$$

Finally plugging in $\delta = \sqrt{\frac{d \log n}{n}}$ yields $\mathcal{R}_n(\mathcal{H}_1) \leq O(\sqrt{\frac{d \log n}{n}})$.

Proof of covering number of $\mathcal{H}_1$ (skipped in class)

1. By assumption on h we have

$$\|h(z_1^n; u) - h(z_1^n; u')\|_n \leq \|h(z;u) - h(z;u')\|_\infty \leq L\|u - u'\|_2$$

2. Any δ/L-cover for $\mathbb{B}_2(1) \subset \mathbb{R}^d$ is also an δ-cover for $\mathcal{H}(z_1^n)$
3. (MW Lem. 5.7.) Covering of a ball of metric ρ wrt metric ρ has $\mathcal{N}(\delta; \mathbb{B}_\rho, \rho) = (1 + \frac{2}{\delta})^d$ using volume ratio bound

$$\to \quad \mathcal{N}(\delta; \mathcal{H}(z_1^n), \|\cdot\|_n) \leq \mathcal{N}(\frac{\delta}{L}; \mathbb{B}_2(1), \|\cdot\|_2) \leq (1 + \frac{2L}{\delta})^d$$

R.C. rates for function classes: Nonparametric example

We now move on to an infinite-dimensional function class

Example II: Smooth non-parametric function classes $\mathcal{H}_2^\alpha$ with $h : [0,1] \to [0,1]$ s.t. $|h^{(\alpha)}(x) - h^{(\alpha)}(x')| \leq L|x - x'|$

- We use bounds for $\mathcal{N}(\delta; \mathcal{H}_2^\alpha, \|\cdot\|_\infty)$ since for any $\mathcal{H}$ and $f, g \in \mathcal{H}$ $\frac{\|\theta-\theta'\|_2}{\sqrt{n}} = \sqrt{\frac{1}{n}\sum_i (f(z_i) - g(z_i))^2} \leq \max_i |f(z_i) - g(z_i)| \leq \|f - g\|_\infty$ and thus $\mathcal{N}(\delta; \mathcal{H}(z_1^n), \|\cdot\|_n) \leq \mathcal{N}(\delta; \mathcal{H}, \|\cdot\|_\infty)$
- For $\alpha = 0$, using the sandwich inequality and constructing a packing, we get for any z_1^n

$$\mathcal{N}(\delta; \mathcal{H}_2^0, \|\cdot\|_\infty) = O(e^{L/\delta}) \to \log \mathcal{N}(\delta; \mathcal{H}_2^0, \|\cdot\|_\infty) \asymp \frac{1}{\delta}$$

  and hence we have $\mathcal{R}_n(\mathcal{H}_2^0) \leq O(n^{-1/3})$ (see MW Example 5.10.).
- For general α, we have $\log \mathcal{N}(\delta; \mathcal{H}_2^\alpha, \|\cdot\|_\infty) \asymp (\frac{1}{\delta})^{\frac{1}{\alpha+1}}$ and hence obtain rates of $\mathcal{R}_n(\mathcal{H}_2^\alpha) \leq O(n^{-\frac{1}{2}\frac{(2\alpha+2)}{(2\alpha+3)}})$ (MW Ex. 5.11.).

References

Metric entropy

- MW Chapter 5

Lecture 7: Chaining, non-parametric regression and localized complexity

Announcements and plan

- Project proposals due next Tuesday **24.10.**, send to Konstantin and supervisor
- One page is enough, instructions on project website (plan how you split up work among the group)

Plan today

- Pollard: One-step discretization $\rightarrow$ Finer argument via Dudley's integral: Chaining
- Moving from classification to (non-parametric) regression

Recap: Metric entropy to bound excess risk

- Excess risk $R\widehat{f}_n) - R(f^\star)$ bounded by generalization gap and standard concentration terms.
- For bounded losses, generalization gap $R(\widehat{f}_n) - R_n(\widehat{f}_n)$ is bounded by Rademacher complexity w.h.p.
- Can bound (population) R.C. via sup of empirical R.C.
- View the empirical R.C. as expected supremum of subgaussian process $X_\theta := \frac{1}{\sqrt{n}}\langle \epsilon, \theta \rangle$ for Rademacher vector ϵ and $\theta \in \mathcal{H}(x_1^n) = \{(h(x_1), \ldots, h(x_n)) | h \in \mathcal{H}\}$
- Bounded this expectation using the covering number (Pollard's bound)

Recap: Covering number

Proposition (using Pollard's bound - MW Prop 5.17)

Let $\delta > 0$. If a set of points $\theta^1, \ldots, \theta^N$ is a covering of $\mathbb{T}$ in the metric $\rho = \frac{\|\cdot\|_2}{\sqrt{n}}$, i.e. it satisfies $\min_j \rho(\theta, \theta^j) \leq \delta$ for all $\theta \in \mathbb{T}$ and $\sup_{\theta,\theta' \in \mathbb{T}} \rho(\theta, \theta') \leq \sigma$, then we have

$$\tilde{\mathcal{R}}_n(\mathbb{T}) \leq \mathbb{E} \sup_{\theta,\theta' \in \mathbb{T}} X_\theta - X_{\theta'} \leq 2[\delta + 2\sigma\sqrt{\frac{\log N(\delta)}{n}}]$$

This bound holds in particular for the covering number

Definition (covering number, metric entropy)

For a metric ρ let the *ϵ-covering number* $\mathcal{N}(\epsilon; \mathbb{T}, \rho)$ be the smallest N such that a set of N points $S = \{\theta_i\}_{i=1}^N$ satisfies $\max_{\theta \in S} \min_i \rho(\theta_i, \theta) \leq \epsilon$ (S is ϵ-cover). *The metric entropy* is $\log \mathcal{N}(\epsilon; \mathbb{T}, \rho)$.

Recap: Examples

Example I: Smoothly parameterized function class $\mathcal{H}_1$ with h s.t.

$$\sup_z |h(z;u) - h(z;u')| \leq L\|u - u'\|_2$$

where $u \in \mathbb{B}_2(1) \subset \mathbb{R}^d$ is the 2-norm ball of radius 1.

Covering number: order $\log(1 + \frac{L}{\delta})$ and $\mathcal{R}_n(\mathcal{H}_1) \leq O(\sqrt{\frac{d \log n}{n}})$.

Example II: Smooth non-parametric function classes $\mathcal{H}_2^\alpha$ with $h : [0,1] \to \mathbb{R}$ s.t. $|h^{(\alpha)}(x) - h^{(\alpha)}(x')| \leq L|x - x'|$

For $\alpha = 0$, covering number: order $\frac{L}{\delta}$ and $\mathcal{R}_n(\mathcal{H}_2^0) \leq O(n^{-1/3})$.

For general α we have $\mathcal{R}_n(\mathcal{H}_2^\alpha) \leq O(n^{-\frac{1}{2}\frac{(2\alpha+2)}{(2\alpha+3)}})$ (MW Ex. 5.10., 5.11. and 5.21).

Can check for yourself in both cases that the diameter $\sup_{\theta,\theta' \in \mathbb{T}} \frac{\|\theta - \theta'\|_2}{\sqrt{n}}$ is bounded by a constant

Metric entropy refinement: chaining

- Remember Pollard's bound with $D = \sup_{\theta,\tilde{\theta} \in \mathbb{T}} \rho(\theta, \tilde{\theta})$

$$\tilde{\mathcal{R}}_n(\mathbb{T}) \leq \frac{2}{\sqrt{n}} \inf_{\delta > 0} [\delta\sqrt{n} + 2D\sqrt{\log N(\delta)}]$$

- For the last term we're combining a large D with a small δ (hence big $N(\delta)$) $\to$ lose lose.
- Intuitive question: can we use a finer argument such that small δ is paired with big $N(\delta)$?

Theorem (Dudley's entropy integral - MW Thm 5.22.)

Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean subgaussian process wrt some metric ρ. Define $D = \sup_{\theta,\tilde{\theta} \in \mathbb{T}} \rho(\theta, \tilde{\theta})$. Then for any $\delta \in [0, D]$ we have

$$\mathbb{E} \max_{\theta,\tilde{\theta} \in \mathbb{T}} X_\theta - X_{\tilde{\theta}} \leq 2\mathbb{E} \sup_{\gamma,\gamma' : \rho(\gamma,\gamma') \leq \delta} X_\gamma - X_{\gamma'} + 16 \int_{\delta/4}^{D} \sqrt{\log \mathcal{N}(t; \mathbb{T}, \rho)} dt$$

Re Tightness: for non-decreasing functions Pollard's bound yields $O\left(\left(\frac{\log n}{n}\right)^{1/3}\right)$ vs. Dudley: $O\left(\left(\frac{\log n}{n}\right)^{1/2}\right)$ (exercise, nontrivial)

Example of using Dudley for Lipschitz functions

Remember the examples of the parametric and non-parametric function classes.

Example I: Smoothly parameterized function class $\mathcal{H}_1$ with h s.t.

$$\sup_z |h(z;u) - h(z;u')| \leq \|u - u'\|_2$$

where $u \in \mathbb{B}_2(1) \subset \mathbb{R}^d$ is the 2-norm ball of radius 1.

The covering number is of order $d \log(\frac{1}{\delta})$.

Example II: Smooth non-parametric function classes $\mathcal{H}_2^0$ with $h : [0,1]^d \to \mathbb{R}$ s.t. $|h(x) - h(x')| \leq \|x - x'\|_\infty$.

The covering number is of order $(\frac{1}{\delta})^d$.

With your neighbor: Use these approximate covering numbers to compute an upper bound for the Rademacher complexity using Dudley's entropy integral and compare the rates obtained using Pollard's bound (focus on $d = 1$ first)

Proof of Dudley's integral: Part I

Define shorthand $N_{\mathbb{T}}(\delta) := \mathcal{N}(\delta; \mathbb{T}, \rho)$

- Define $L = \lceil \log_2 \frac{D}{\delta} \rceil$ sets of $\delta_i = D2^{-i}$ covers $\mathcal{C}_i$ of $\mathbb{T}$ with $|\mathcal{C}_i| = N_{\mathbb{T}}(\delta_i)$. The finest cover (original/smallest δ) is $\mathcal{C}_L$.
- Remember the one-step discretization for Pollard's bound:

$$\begin{aligned} X_\theta - X_{\tilde{\theta}} &= X_\theta - X_{\theta_\star^{(L)}} + X_{\theta_\star^{(L)}} - X_{\tilde{\theta}_\star^{(L)}} + X_{\tilde{\theta}^\star} - X_{\tilde{\theta}} \\ &= 2 \sup_{\rho(\gamma,\gamma') \leq \delta} X_\gamma - X_{\gamma'} + \max_{\theta,\theta' \in \mathcal{C}_L} X_\theta - X_{\theta'} \end{aligned}$$

where $\theta_\star^{(i)}$ denotes closest point of θ in $\mathcal{C}_i$.

- We can now "recursively" act on $\max_{\theta,\theta' \in \mathcal{C}_L} X_\theta - X_{\theta'}$ by using the same argument on the set $\mathcal{C}_L$ with the coarser cover $\mathcal{C}_{L-1}$.

More generally for any two $\theta, \tilde{\theta} \in \mathcal{C}_i$ we have:

$$\begin{aligned} X_\theta - X_{\tilde{\theta}} &\leq X_\theta - X_{\theta_\star^{(i-1)}} + X_{\theta_\star^{(i-1)}} - X_{\tilde{\theta}_\star^{(i-1)}} + X_{\tilde{\theta}_\star^{(i-1)}} - X_{\tilde{\theta}} \\ &\leq 2 \max_{\theta \in \mathcal{C}_i} X_\theta - X_{\theta_\star^{(i-1)}} + \max_{\theta,\theta' \in \mathcal{C}_{i-1}} X_\theta - X_{\theta'} \end{aligned}$$

Proof of Dudley's integral: Part II

- note that in $max_{\theta\in\mathcal{C}_i} X_\theta - X_{\theta_\star^{(i-1)}}$, for each $\theta \in \mathcal{C}_i$ we have $\theta_\star^{(i-1)}$ be **its** closest point, not of the "original" θ *in* $\mathbb{T}$
- "Rolling out" the induction, we obtain

$$\max_{\theta,\tilde{\theta}\in\mathcal{C}_L} X_\theta - X_{\tilde{\theta}} \leq 2\sum_{i=2}^{L} \max_{\theta\in\mathcal{C}_i} X_\theta - X_{\theta_\star^{(i-1)}} + \max_{\theta,\theta'\in\mathcal{C}_1} X_\theta - X_{\theta'}$$

Rolling out from $L \to 1$ or going from $\mathcal{C}_L$ to $\mathcal{C}_1$, we iteratively

- reduced the cover cardinality until only one element is left (with large diameter),
- while all the intermediate terms (in sum) are δ_{i-1}-subgaussian (instead of fixed D)
- with increasing δ but decreasing corresponding cover cardinality

Proof of Dudley's integral: Part III

In order to compute the final expectation observe that

1. max of subgaussians: $X_\theta - X_{\theta_\star^{(i-1)}}$ is a δ_{i-1}-subgaussian process $\to$

$$\mathbb{E}\max_{\theta\in\mathcal{C}_i} X_\theta - X_{\theta_\star^{(i-1)}} \leq 2\delta_{i-1}\sqrt{\log|\mathcal{C}_i|}$$

2. Covering number non-increasing as δ increases and interval $[D2^{-(i+1)}, D2^{-i}]$ is of length $D2^{-(i+1)} = D2^{-(i-1)}\frac{1}{4}$:

$$\delta_{i-1}\sqrt{\log|\mathcal{C}_i|} = D2^{-(i-1)}\sqrt{\log N_{\mathbb{T}}(D2^{-i})} \leq 4\int_{D2^{-(i+1)}}^{D2^{-i}} \sqrt{\log N_{\mathbb{T}}(t)}\mathrm{d}t$$

3. Putting things together and because $\delta_L = D2^{-L} \leq \delta$

$$\mathbb{E}\max_{\theta,\tilde{\theta}\in\mathcal{C}_L} X_\theta - X_{\tilde{\theta}} \leq 4\sum_{i=2}^{L} D2^{-(i-1)}\sqrt{\log N_{\mathbb{T}}(D2^{-i})} + 2D\sqrt{\log N_{\mathbb{T}}(D/2)}$$

$$\leq 16\int_{\delta/4}^{D} \sqrt{\log N_{\mathbb{T}}(t)}\mathrm{d}t$$

□

Short navigation slide

Whole topic of this class: For each $\mathcal{F}$ define $f^\star = \arg\min_{f\in\mathcal{F}} R(f)$. Interested in bounding excess risk w.h.p.

$$R(\widehat{f}_n) - R(f^\star) = R(\widehat{f}_n) - R_n(\widehat{f}_n) + \overbrace{R_n(\widehat{f}_n) - R_n(f^\star)}^{\leq 0 \text{ by optimality}} + R_n(f^\star) - R(f^\star)$$

- so far: via uniform convergence and Rademacher complexity using

$$\mathbb{P}(\sup_{h\in\mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum_{i=1}^n h(Z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t) \leq \mathrm{e}^{-\frac{nt^2}{2b^2}}$$

  for $\mathcal{H} = \ell \circ \mathcal{F}$ and bounding empirical Rademacher complexity for finite classes, more generally w/ metric entropy and chaining (today)

This line of reasoning was useful for **classification**, for the second half of lectures, we'll switch to **regression**. Can we just continue to use this uniform convergence technique to obtain bounds?

(Non-)parametric regression setting - fixed design

- Square loss and constrained regression
- Fixed design, i.e. only care about prediction on training inputs $x_1, \ldots, x_n$
- Gaussian observation noise, i.e. $W = Y - f^\star(X) \in \mathcal{N}(0, \sigma^2)$
- Analyze minimizer $\widehat{f} = \arg\min_{f\in\mathcal{F}} R_n(f) := \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2$ or 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

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

Partner-Q: Derive a h.p. upper bound for $\|f - f^\star\|_n^2$ for linear functions $f(x) = \langle w, x\rangle$ with $\|x\|_2 \leq D, \|w\|_2 \leq B$. Compare a closed-form vs. a uniform law approach - where might the difference come from?

Warm-up using closed-form solution - linear regression

For linear/kernel regression, can directly analyze closed-form solution of both ridge and min-norm interpolator. For linear:

- first recall $y = X\theta^\star + w$ and solution $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|_2^2$
- minimizer $\widehat{f}(x) = \hat{\theta}^\top x$ with $\hat{\theta} = (X^\top X)^{-1} X^\top (X\theta^\star + w)$
- $\|\widehat{f} - f^\star\|_n^2 = \frac{1}{n}\|X(\hat{\theta} - \theta^\star)\|^2 = \frac{1}{n} w^\top X (X^\top X)^{-1} X^\top w$
- only need to bound $\frac{1}{n} w^\top X (X^\top X)^{-1} X^\top w \to$ use that the norm of a Gaussian is a Lipschitz function of Gaussian for concentration (here with Lipschitz constant $\sqrt{\frac{\text{rank}(X)}{n}}$ via SVD) and MW Thm 2.26
- Further $\mathbb{E}\frac{1}{n} w^\top X (X^\top X)^{-1} X^\top w = \sigma^2 \frac{\text{rank}(X)}{n}$

This stands in contrast to the uniform law approach where you can use contraction to obtain a bound using Rademacher complexity of linear function classes and at most get a $\frac{1}{\sqrt{n}}$ bound

Beyond closed-form solutions

- First of all, notice the "slow" uniform excess risk bound holds for any $\mathcal{F}$, including ones for which $f^\star \notin \mathcal{F}$!
- Further, in our argument using uniform law, we used optimality of $\widehat{f}_n$ only once

$$R(\widehat{f}_n) - R(f^\star) = R(\widehat{f}_n) - R_n(\widehat{f}_n) + \overbrace{R_n(\widehat{f}_n) - R_n(f^\star)}^{\leq 0 \text{ by optimality}} + R_n(f^\star) - R(f^\star)$$

Next few classes: using *localized complexities* to prove tighter bounds for particular estimator: global minimizer of square loss for regression!

- Idea: By using **optimality of** $\widehat{f}$ instead of uniform bound
  1. circumvent uniform boundedness
  2. can get more restricted function space

Basic inequality circumventing boundedness and more

Optimality of $\widehat{f}$ yields the *basic inequality*

$$R_n(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(y_i - f^\star(x_i))^2 = R_n(f^\star)$$
$$\|\widehat{f} - f^\star\|_n^2 \leq \frac{2\sigma}{n}\sum_{i=1}^{n} w_i(\widehat{f}(x_i) - f^\star(x_i)) \tag{1}$$

- Taking expectations defining $\mathcal{F}^\star = \mathcal{F} - f^\star$
  $\rightarrow \mathbb{E}\|\widehat{f} - f^\star\|_n^2 \leq 2\sigma\widetilde{\mathcal{G}}_n(\mathcal{F}^\star(x_1^n)) := \mathbb{E}_w \sup_{g\in\mathcal{F}^\star} \frac{2\sigma}{n}\sum_{i=1}^{n} w_i g(x_i)$
- Gaussian complexity popped out without needing uniform boundedness (same "order" as Radmacher, satisfies sandwich relationship, porved in HW 2, for each $\mathbb{T}$)
  $\frac{1}{2\log n}\widetilde{\mathcal{G}}_n(\mathbb{T}) \leq \tilde{\mathcal{R}}_n(\mathbb{T}) \leq \sqrt{\frac{\pi}{2}}\widetilde{\mathcal{G}}_n(\mathbb{T})$
- But still stuck with a huge function space $\mathcal{F}$!

The trick is to notice eq. 1 restricts function space!

Non-parametric regression prediction error bound

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.

We can then use this quantity to bound

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 $t \geq 1$

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

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.

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

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:

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

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

b. eq. 2 follows from h.p. bound of this (locally uniform!) quantity

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

and if *localized (empirical) Gaussian complexity* is bounded

$$\sigma \widetilde{\mathcal{G}}_n(\mathcal{F}^\star; \delta_n) := \sigma \widetilde{\mathcal{G}}_n(\mathcal{F}^\star(x_1^n) \cap \mathbb{B}_n(\delta_n)) = \mathbb{E} \sup_{\substack{\|\hat{\Delta}\|_n \leq \delta_n \\ \hat{\Delta} \in \mathcal{F}^\star}} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \leq \delta_n^2$$

References

Dudley's integral

- MW Chapter 5

Non-parametric regression

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 $x_1, \ldots, x_n$
- 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:

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

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

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

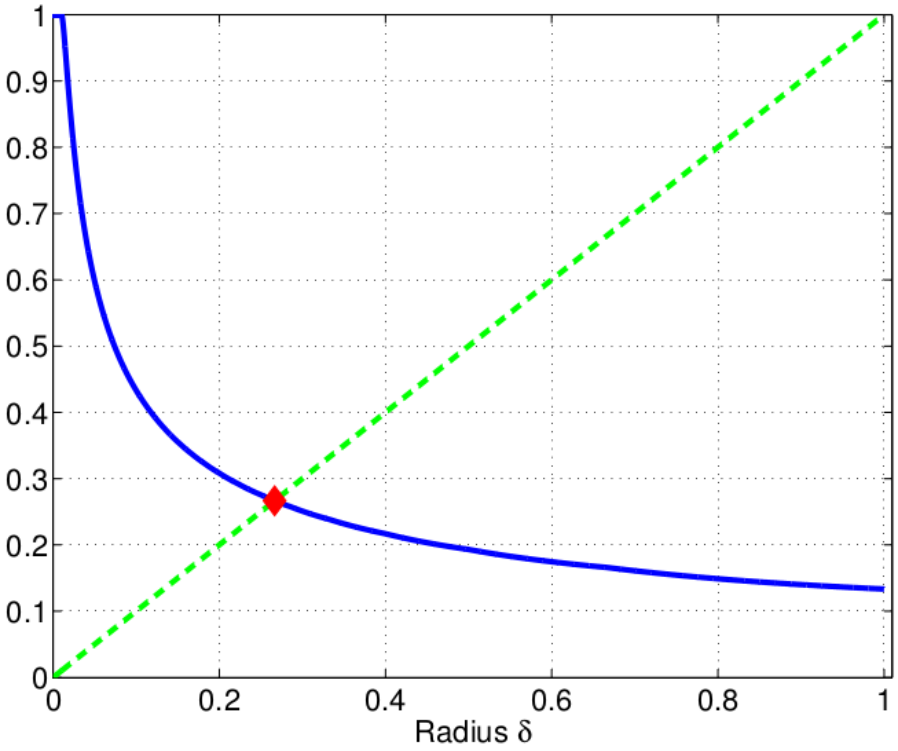


Figure 1: Blue solid: $f(\delta) = \frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\delta)}{\delta}$, Green dashed: $f(\delta) = \delta$

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 $t \geq 1$

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

- 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 $\mathbb{P}(g_n(w) \geq \mathbb{E}g_n(w) + s) \leq \mathrm{e}^{-\frac{ns^2}{2\sigma^2 t\delta_n^2}}$ (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

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

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

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 $\int_0^D \sqrt{\log \mathcal{N}(t; \mathcal{F}(x_1^n), \|\cdot\|_n)} \mathrm{d}t$ 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 $f : \mathcal{X} \to \mathbb{R}$ 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 $f : \mathcal{X} \to \mathbb{R}$ with evaluation functional that is bounded and linear, i.e. for all $x \in \mathcal{X}$ there exists $L_x : \mathcal{F} \to \mathbb{R}$ 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 $\mathcal{F}_S := span\{\mathcal{K}(x_1, \cdot), \ldots, \mathcal{K}(x_n, \cdot)\}$. 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 $f \in \mathcal{F}$
$\rightarrow$ convergence in $\mathcal{F}$ pointwise convergence
$\rightarrow$ 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 $\mathcal{F} = \{f : X \to \mathbb{R} : f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}_0}$ 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

□

Lecture 9: Kernel ridge regression

Announcements

- HW 2 out tonight, due 9.11. 23:59
- Proofs skipped in class / exercise for home: You are supposed to fully understand those steps, also of the exercises in class and in the homework - the oral exam will primarily test your understanding of how different proof steps fit together

Plan for today

- Another example of prediction error of square-loss minimizer: Prediction error bound for ERM of norm-bounded RKHS
- Prediction error bound for *regularized* regression

Recap: Non-parametric prediction error bound

Definition (Localized (empirical) Gaussian complexity)

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

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

Lemma (Critical radius, MW 13.6.)

For any star-shaped $\mathcal{F}$, it holds that $\frac{\widetilde{\mathcal{G}}_n(\mathcal{F};\delta)}{\delta}$ 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.

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 $t \geq 1$

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

Recap: Reproducing Kernel Hilbert Spaces (RKHS)

- Recap motivation of kernel trick and kernel spaces
- abstract definition of reproducing kernel Hilbert spaces $\rightarrow$ can be associated uniquely with a kernel $\mathcal{K}$ and equal to its induced (unique) Hilbert space which is the completion of
- $\mathcal{F}_{\text{pre}} = \{\sum_{i=1}^N \alpha_i \mathcal{K}(\cdot, x_i) : N \in \mathbb{N}, \alpha \in \mathbb{R}^N, x_1, \ldots, x_N \in \mathcal{X}\}$ with inner product $\langle \mathcal{K}(\cdot, x), \mathcal{K}(\cdot, y) \rangle_{\mathcal{F}_\mathcal{K}} = \mathcal{K}(x, y)$

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

- $\mathcal{F}_{lin} = \{f : f(x) = \langle w, x \rangle, w \in \mathbb{R}^d\}$ is an RKHS with $\mathcal{K}(x, y) = \langle x, z \rangle$ as a reproducing kernel as a reproducing kernel $f = \langle w_f, \cdot, \rangle$ and $g = \langle w_g, \cdot \rangle$ the inner product $\langle f, g \rangle = w_f^\top w_g$

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

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

Non-parametric regression in RKHS

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

- Recall the non-parametric (unpenalized) estimate $\widehat{f}$ is defined as

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

Today:

- compute generalization bound for $\widehat{f}$ in a particular RKHS
- Minimization of square loss in constrained space
  $\mathcal{F}_R = \{f \in \mathcal{F} : \|f\|_\mathcal{F} \le R\}$ (ommitting subscript $\mathcal{K}$) or kernel ridge regression (regularized square loss) using localized complexities

Unregularized kernel regression

- Given empirical loss $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ and (empirical) prediction error $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$.
- Define the *empirical kernel matrix* K with $K_{ij} := \frac{\mathcal{K}(x_i,x_j)}{n}$ (*this is the normalized kernel matrix, more interpretable since eigenvalues converge to operator eigenvalues*)
- Now assume that the empirical kernel matrix is invertible.

Neighbor-Q:
a) What is the minimum value of the empirical loss?
b) How about the prediction error?
c) How about the localized Gaussian complexity?
d) For which kernels is the kernel matrix invertible?

Remember how to rewrite the empirical loss in matrix vector notation.
Compute the localized complexity and critical radius

Regularized kernel regression

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.

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

So we defined the regularized estimator $\widehat{f}_R$ is defined as

$$\widehat{f}_R \in \arg\min_{f \in \mathcal{F}_R} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \text{ (possibly non-unique)}$$

By the representer theorem we can then write it as

$$\min_{f \in \mathcal{F}_R} \frac{1}{2n}\|y - f(x_1^n)\|_2^2 = \min_{\alpha \in \mathbb{R}^n} \frac{1}{2n}\|y - K\alpha\|_2^2$$

- We now see eigenvalues of the kernel matrix can be used to bound prediction error of $\widehat{f}_R$ w.h.p. via the critical inequality!

Localized G.C. for RKHS with bounded norm

Lemma (local G.C. for norm-bounded RKHS, MW Cor. 13.18)

Defining $\hat{\mu}_j$ as eigenvalues of the kernel matrix K we have

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

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

Definition (R-modified critical quantity $\delta_{n;R}$)

We define $\delta_{n;R}$ to be the smallest $\delta > 0$ satisfying

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

- By Lemma it then holds that $\frac{\sigma\widetilde{\mathcal{G}}_n(\mathcal{F}_3;\delta_{n;R})}{\delta_{n;R}} \leq \delta_{n;R}R$

Prediction error bound for RKHS with bounded norm

Theorem (Prediction error of norm-bounded RKHS)

Assume $f^\star \in \mathcal{F}_R$. Then we have for least-squares estimate $\widehat{f}_R \in \mathcal{F}_R$

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

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

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

Rates for actual kernel spaces $\mathcal{F}$

- Ex. 1: α-smooth functions w/ $\hat{\mu}_j \sim j^{-2\alpha} \rightarrow \|\widehat{f} - f^\star\|_n^2 \leq (\frac{R\sigma^2}{n})^{2/3}$
- Ex. 2: Gaussian kernel w/ $\hat{\mu}_j \sim e^{-cj\log j} \rightarrow \|\widehat{f} - f^\star\|_n^2 \leq \frac{\sigma^2\log(\frac{Rn}{\sigma})}{n}$
- 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)

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

- Scale basic inequality by R to obtain $\widetilde{f^\star} = \frac{f^\star}{R}$, $\widetilde{f} = \frac{\widehat{f}}{R}$, $\tilde{\sigma} = \frac{\sigma}{R}$

$$\frac{1}{nR^2}\sum_{i=1}^n (y_i - \widehat{f}(x_i))^2 \leq \frac{1}{nR^2}\sum_{i=1}^n (y_i - f^\star(x_i))^2$$

$$\|\widetilde{f} - \widetilde{f^\star}\|_n^2 \leq 2\frac{\tilde{\sigma}}{n}\sum_{i=1}^n w_i(\widetilde{f}(x_i) - \widetilde{f^\star}(x_i))$$

- Since $\widetilde{f^\star}, \widetilde{f} \in \mathcal{F}_1$, $\tilde{\Delta} \in \mathcal{F}_1^\star = \mathcal{F}_1 - \widetilde{f^\star} \subset \mathcal{F}_3$ ($\mathcal{F}_2$ 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 $\delta_{n;R}$
  - Using $\mathbb{E}_w \sup_{\tilde{\Delta}\in\mathcal{F}_3, \|\tilde{\Delta}\|_n \leq \delta} \frac{\tilde{\sigma}}{n}\sum_{i=1}^n w_i\tilde{\Delta}(x_i) = \tilde{\sigma}\widetilde{\mathcal{G}}_n(\mathcal{F}_3;\delta)$
  - It's sufficient that $\sup_{\|\tilde{\Delta}\|_n \leq \delta_{n;R}, \tilde{\Delta}\in\mathcal{F}_3} \frac{\tilde{\sigma}}{n}\sum_{i=1}^n w_i\frac{\tilde{\Delta}(x_i)}{\delta_{n;R}} \leq \delta_{n;R}$ where we ned modified critical inequality $\tilde{\sigma}\widetilde{\mathcal{G}}_n(\mathcal{F}_3;\delta_{n;R}) \leq \delta_{n;R}^2$ in tail bound
- Observing $\|\widehat{f} - f^\star\|_n^2 = R^2\|\tilde{\Delta}\|_n^2$ yields the theorem. □

Proof of Lemma (local. compl. for norm-bounded RKHS)

- 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.

$$\widetilde{\mathcal{G}}_n(\mathcal{F}_r;\delta) = \mathbb{E}_w \sup_{\|\Delta\|_{\mathcal{F}}\leq r, \|\Delta\|_n \leq \delta} \frac{1}{n}\sum_i w_i\Delta(x_i)$$

$$= \frac{1}{\sqrt{n}}\mathbb{E}_w \sup_{\alpha^\top K\alpha \leq r^2, \alpha^\top K^2\alpha \leq \delta^2} w^\top K\alpha$$

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

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

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

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

- Hence $\widetilde{\mathcal{G}}_n(\mathcal{F}_r;\delta) \leq \sqrt{\frac{1+r^2}{n}}\mathbb{E}_w\sqrt{\sum_i \frac{w_i^2}{\eta_i}} \leq \sqrt{\frac{1+r^2}{n}}\sqrt{\sum_i \frac{1}{\eta_i}}$ via

Regularized regression guarantees for metric spaces

- So far looked at empirical risk minimizers for the square loss of type $\widehat{f} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$
- But often type we minimize a loss with an additive penalty such as in ridge regression

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

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

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

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

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

- Again, if $f^\star \notin \mathcal{F}_R$ yields add. approx. error $\inf_{\|f\|_{\mathcal{F}} \leq R} \|f - f^\star\|_n^2$
- if additional term $\lambda_n \sim \delta_{n;R}^2$, same order as constrained

Proof of bound for regularized regression estimate

For simplicity we write $\widehat{f}$ for $\widehat{f}_{\lambda_n}$

1. By optimality we have

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

which yields basic inequality after rearranging terms

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

2. Normalize $f^\star, \widehat{f}, \sigma$ by $\frac{1}{R}$ like for norm-bounded $\rightarrow$ $\widetilde{f^\star}, \widetilde{f}, \tilde{\sigma}, \tilde{\Delta} = \widetilde{f} - \widetilde{f^\star}$ ($\widetilde{f}$ different than in MW!)

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

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

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 based on 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) on Event 1 $\|\widetilde{f}\|_{\mathcal{F}} \leq 2$ using previous arguments on T_1 as for the prediction error for norm-bounded RKHS using the critical inequality and tail bound, as well as the fact that $T_2 \leq \|\widetilde{f^\star}\|_{\mathcal{F}}^2 \leq 1$.

b) on Event 2 $\|\widetilde{f}\|_{\mathcal{F}} > 2$ using a new (peeling) lemma for all $\|\tilde{\Delta}\|_{\mathcal{F}} \geq 1$. There we use T_2 to "cancel" large norms

4. Solving the quadratic yields $\|\tilde{\Delta}\|_n^2 \leq c(\delta_{n;R}^2 + \lambda_n)$ □

Proof of 4. - 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

a) Event 1: $\|\widetilde{f}\|_{\mathcal{F}} \leq 2$, then $\|\tilde{\Delta}\|_{\mathcal{F}} \leq 3$ and we can use slide 10 and the fact that $T_2 \leq 1$: $\rightarrow$ yields $\frac{1}{2}\|\tilde{\Delta}\|_n^2 \leq c\delta_{n;R}\|\tilde{\Delta}\|_n + \lambda_n$,

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

- T_1: can still bound T_1 using similar idea as in sl. 10, but iteratively (peeling lemma) on event $\|\tilde{\Delta}\|_{\mathcal{F}} \geq 1$ (MW Lem. 13.23) yields with probability at least $\geq 1 - c_1 \mathrm{e}^{-\frac{n\delta_n^2}{c_2\tilde{\sigma}^2}}$

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

- T_2: $\lambda_n(\|\widetilde{f^\star}\|_{\mathcal{F}}^2 - \|\widetilde{f}\|_{\mathcal{F}}^2) \leq 2\lambda_n - \lambda_n\|\tilde{\Delta}\|_{\mathcal{F}}$ using $\|\tilde{\Delta}\|_{\mathcal{F}} \leq \|\widetilde{f}\|_{\mathcal{F}} + \|\widetilde{f^\star}\|_{\mathcal{F}}$ and $\|\widetilde{f^\star}\|_{\mathcal{F}}^2 - \|\widetilde{f}\|_{\mathcal{F}}^2 \leq \|\widetilde{f^\star}\|_{\mathcal{F}} - \|\widetilde{f}\|_{\mathcal{F}}$
  $\rightarrow$ 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 $\frac{1}{2}\|\tilde{\Delta}\|_n^2 \leq c\delta_n\|\tilde{\Delta}\|_n + \frac{1}{16}\|\tilde{\Delta}\|_n^2 + 2\lambda_n$

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 $\mathcal{F}$ we only need to show inequality with sup over $\|\tilde{\Delta}\|_{\mathcal{F}} = 1$
- 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. 1 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 $t_m = 2^m \delta_n$ and union bound.
- Union bounding with this choice of t_m with the usual concentration bound (Lipschitz function of Gaussians in MW Thm 2.26)

For a detailed proof we refer to the book.

References

Reproducing Kernel Hilbert spaces:

- MW Chapter 12
- SC Chapter 4

Non-parametric regression:

- MW Chapter 13

Kernel eigenvalues (skipped in class)

- 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 $T_{\mathcal{K}} : \mathcal{L}^2 \to \mathcal{L}^2$ with $T_{\mathcal{K}} f = \int \mathcal{K}(x, y) f(y) \mathrm{d}y$, 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 eigenfcuntions 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 $\mathbb{P}$!

Proof of Mercer's Theorem (skipped in class)

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

Lecture 10: NTK and random design

Announcements

- HW 2 released, due 9.11. 23:59
- HW 1 grades released these days via gradescope

Plan for today

- Prediction error bound for random design
- Add-on: Random features and NTK

Random design

- So far, we only controlled $\|\widehat{f} - f^\star\|_n^2$ w.h.p. over observation noise w

$$\|\widehat{f} - f^\star\|_n^2 = R(\widehat{f}) - R(f^\star) = \mathbb{E}_w \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{f}(x_i))^2 - \mathbb{E}_w \frac{1}{n}\sum_{i=1}^n w_i^2$$
$$= \frac{1}{n}\sum_{i=1}^n (\widehat{f}(x_i) - f^\star(x_i))^2$$

- can be bounded using empirical Gaussian complexities via basic inequality $\rightarrow$ basic inequality

How does the error look like on the whole domain $\mathcal{X}$?

Now we view X as random and take expectation also over X, i.e. assuming $f \in \mathcal{L}^2(\mathbb{P})$, want to bound

$$\|\widehat{f} - f^\star\|_2^2 = R(f) - R(f^\star) = \mathbb{E}_{X,W}(Y - \widehat{f}(X))^2 - \mathbb{E}W^2$$
$$= \mathbb{E}_X(\widehat{f}(X) - f^\star(X))^2 = \mathbb{E}_{x_1,\dots,x_n}\|\widehat{f} - f^\star\|_n^2$$

Prediction error bound for random design - uniform law?

Definition (Rademacher complexity - recap)

Given a function class $\mathcal{H}$ and distribution $\mathbb{P}$ on its domain $\mathcal{Z}$, we define the Rademacher complexity as

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon,z} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(z_i)$$

Theorem (Uniform law - recap)

For b-unif. bounded $\mathcal{H}$ with $\mathcal{R}_n(\mathcal{H}) = \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(z_i)$

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \mathbb{E}h - \frac{1}{n}\sum_{i=1}^n h(z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t\right) \leq e^{-\frac{nt^2}{2b^2}}$$

w/ prob. over the training data. If $\mathcal{R}_n(\mathcal{H}) = o(1)$, then $\sup_{h\in\mathcal{H}} R(h) - R_n(h) \overset{a.s.}{\to} 0$.

Using the uniform law for (uniformly bounded) regression

Partner-Q: Using the uniform law, derive a h.p. upper bound for $\|f - f^\star\|_2^2$ for linear functions $f(x) = \langle w, x \rangle$ with $\|x\|_2 \leq D, \|w\|_2 \leq B$, bounded noise. Use Rademacher contraction

First of all, in this setting, by assumption, the loss is uniformly bounded since $|y_i - f(x_i)| \leq D'$ is bounded by some constant D'.

Further, we can show that in the given setting, the loss is L-Lipschitz for some constant L independent of the specific probability measure.

- Define $\widetilde{\mathcal{F}} = \{(y_1 - f(x_1), \ldots, y_n - f(x_n)) : f \in \mathcal{F}\}$
- Then for the square function $\ell(u) = u^2$ we have $|\ell(u) - \ell(u')| \leq |u^2 - u'^2| \leq |u - u'||u + u'| \leq 2D'|u - u'|$.

Using the Lipschitz property, we can then use Rademacher contraction to bound the Rademacher complexity of $\ell \circ \widetilde{\mathcal{F}}$ using the Rade. comp. on bounded linear functions like for the SVM. Analogously to the SVM excess risk bound, the uniform law yields a squared error bound of order $O(1/\sqrt{n})$.

Precise statement of localized uniform law

- Now how about just using the uniform law to bound $\sup_{\hat{\Delta} \in \mathcal{F}^\star} \|\hat{\Delta}\|_2^2 - \|\hat{\Delta}\|_n^2$ using (population) Rademacher complexity? (from now on we write g instead of $\hat{\Delta}$ for simplicity)
- In fact, we can *localize* the uniform law as well!
- Indeed, for b-uniformly bounded $\mathcal{F}^\star$, we can define the critical inequality on the *population* localized Rademacher complexity

$$\mathcal{R}_n(\mathcal{F}^\star; \delta) = \frac{1}{n} \mathbb{E}_{X,\epsilon} \sup_{g \in \mathcal{F}, \|g\|_2 \leq \delta} \sum_{i=1}^n \epsilon_i g(x_i) \leq \frac{\delta^2}{16b}$$

- Let $\overline{\delta}_n$ be a δ that satisfies this inequality.

Now what?

- Can't directly use our localization / basic inequality approach, since that only holds for finite samples!
- We also only have $\|\hat{\Delta}\|_n \leq \delta_n$ *with high probability* over x

Precise statement of localized uniform law

Theorem (Localized uniform law, MW Thm 14.1)

For star-shaped and $b-$uniformly bounded $\mathcal{F}^\star$, let $\overline{\delta}_n$ as defined above. Then if $\overline{\delta}_n^2 > c\frac{\log[4\log(1/\overline{\delta}_n)]}{n}$ then w.p. at least $1 - c_1 e^{-c_2 \frac{n\overline{\delta}_n^2}{b^2}}$ we have

$$\sup_{g \in \mathcal{F}^\star} \|g\|_2 - \|g\|_n \leq c\overline{\delta}_n$$

Proof idea:

- For localization we used the basic inequality for the empirical error
- There we had LHS $\|g\|_n^2$ with $g \in \mathcal{F}^\star$ which we self-bounded by $\delta_n \|g\|_n$
- We can do something similar here: we choose $\|g\|_2^2 - \|g\|_n^2$ as our RHS and will also "self-upper-bound" it

Proof idea for localized uniform law

- Observe that the binomial formula yields for any $g \in \mathcal{F}^\star$

$$\|g\|_2 - \|g\|_n = \frac{\|g\|_2^2 - \|g\|_n^2}{\|g\|_2 + \|g\|_n}$$

- Hence the proof goes through either with
  a) $\frac{\|g\|_2^2 - \|g\|_n^2}{\|g\|_2 + \|g\|_n} \leq \overline{\delta}_n$ if $\|g\|_2 \leq \overline{\delta}_n$
  b) $\|g\|_2 \overline{\delta}_n$ with high prob. if $\|g\|_2 \geq \overline{\delta}_n$ (uniformly for all $g \in \mathcal{F}^\star$) yields

We give intuition for the proof of b)

Proof steps - case $\|g\|_2 \geq \overline{\delta}_n$

For simplicity of the proof, assume $b = 1$ and hence $\|g\|_2 \leq 1$ (general case follows from scaling arguments as last time)

We can show $\sup_{g \in \mathcal{F}^\star, \|g\|_2 \geq \overline{\delta}_n} \|g\|_2^2 - \|g\|_n^2 \leq \|g\|_2 \overline{\delta}_n$ w.h.p.:

1. Step: For fixed $r \geq \overline{\delta}_n$, bounding $\sup_{g \in \mathcal{F}^\star, \|g\|_2 \leq r} \|g\|_2^2 - \|g\|_n^2$ (MW Lemma 14.9.)

- symmetrization and Rademacher contraction for $r \geq \overline{\delta}_n$

$$\mathbb{E} \sup_{g \in \mathcal{F}^\star, \|g\|_2 \leq r} \|g\|_2^2 - \|g\|_n^2 \leq 2\mathbb{E} \sup_{g \in \mathcal{F}^\star, \|g\|_2 \leq r} \frac{1}{n} \sum_{i=1}^n \epsilon_i g^2(x_i)$$
$$\leq 4\mathbb{E} \sup_{g \in \mathcal{F}^\star, \|g\|_2 \leq r} \frac{1}{n} \sum_{i=1}^n \epsilon_i g(x_i) \leq r\overline{\delta}_n$$

  where the last inequality follows from definition of $\overline{\delta}_n$

- we then use Talagrand concentration (MW Thm 3.27) that states that w.p $\geq 1 - \mathrm{e}^{-cn\overline{\delta}_n^2}$ we have $\sup_{g \in \mathcal{F}^\star, \|g\|_2 \leq r} \|g\|_2^2 - \|g\|_n^2 \leq \frac{r\overline{\delta}_n}{2}$

Proof steps - case $\|g\|_2 \geq \overline{\delta}_n$

2. Step: If we could plug in $r = \|g\|_2$ we'd be done, but above h.p. bound only holds for fixed r!

- Use peeling argument like before and split $S := \{\sup_{g \in \mathcal{F}^\star, \|g\|_2 \geq \overline{\delta}_n} \|g\|_2^2 - \|g\|_n^2 \geq \|g\|_2 \overline{\delta}_n\}$ into sub-events: $S_m = \{\|g\|_2 \in [t_{m-1}, t_m]\}$ where $t_m = 2^m \overline{\delta}_n$. In particular, by uniform boundedness $\|g\|_2 \leq 1$, we have that $S \subset \bigcup_{m=1}^M \{S \cap S_m\}$ with $M = 4 \log(1/\overline{\delta}_n)$

- using $\sup_{g \in \mathcal{F}^\star, \|g\|_2 \leq r} \|g\|_2^2 - \|g\|_n^2 \leq \frac{r\overline{\delta}_n}{2}$ with $r = t_m$ and using union bound gives

$$\mathbb{P}(S) \leq \sum_{m=1}^M \mathbb{P}(S \cap S_m) \leq \sum_{m=1}^M \mathbb{P}\Big(\sup_{g \in \mathcal{F}^\star, \|g\|_2 \leq t^m} \|g\|_2^2 - \|g\|_n^2 \geq \frac{t_m \overline{\delta}_n}{2} \Big)$$
$$\leq \sum_{m=1}^M \mathrm{e}^{-cn\overline{\delta}_n^2} \leq \mathrm{e}^{-cn\overline{\delta}_n^2 + \log M} \leq \mathrm{e}^{-cn\overline{\delta}_n^2}$$

Kernel $\to$ feature maps (unbounded, translation-invariant)

We saw some examples for RKHS and their kernels with **compact supports** (e.g Sobolev spaces). What if domain is non-compact?

Consider RBF kernels $\mathcal{K}(x, y) = h(\|x - y\|_2)$

Theorem (Bochner: feature maps for translation-invariant kernels)

If $\mathcal{K}(x, y) = h(x - y)$ with h continuous and $x, y \in \mathbb{R}^d$, then there is a unique, finite, non-negative measure μ on $\mathbb{R}^d$ such that

$$h(t) = \int_{\mathbb{R}^d} e^{-i\langle t,\omega\rangle} \mu(d\omega)$$

Reminiscent of the Fourier basis, we call μ spectral measure, and if it has a density, we call $s(\omega)d\omega = \mu(d\omega)$ the spectral density

Kernels as expectations

For Gaussian kernels $\mathcal{K}(x, y) = \mathrm{e}^{-\frac{\|x-y\|_2^2}{2\sigma^2}}$ on $\mathbb{R}^d$ where Bochner holds with $s(\omega) = \left(\frac{2\pi}{\sigma^2}\right)^{-d/2} \mathrm{e}^{-\frac{\sigma^2\|\omega\|_2^2}{2}}$ (Fourier transform)

- For feature maps $\phi(\omega; x) = \mathrm{e}^{-\mathrm{i}\langle x,\omega\rangle}$, we can rewrite the kernel as an expectation over measure $\mu(\mathrm{d}\omega) = s(\omega)\mathrm{d}\omega$, i.e.

$$\mathcal{K}(x, y) = \mathbb{E}_{\omega\sim\mu}\phi(\omega; x)\phi(\omega; y) = \langle\phi(\cdot; x), \phi(\cdot; y)\rangle_{\mathcal{L}^2(\mu)}$$

  proof by completing the square

The corresponding kernel space $\mathcal{F}_\mathcal{K}$ can be described as follows:

- kernel space
  $\mathcal{F}_\mathcal{K} = \{f : f(x) = \int \tilde{f}(\omega)\mathrm{e}^{-\mathrm{i}\langle x,\omega\rangle}\mu(\mathrm{d}x) = \langle\tilde{f}, \phi\rangle_{\mathcal{L}^2(\mu)}, \tilde{f} \in \mathcal{L}^2(\mu)\}$

Kernels as expectations $\to$ random features

- Instead of the true expectation, can approximate/unbiased estimate $\mathcal{K}$ via empirical expectation $\hat{\mathbb{P}}_m$ over m samples of ω_j from μ

$$\hat{\mathcal{K}}(x,y) = \mathbb{E}_{\omega\sim\hat{\mathbb{P}}_m}\phi(\omega;x)\phi(\omega;y) := \frac{1}{m}\sum_{j=1}^{m}\phi(\omega_j;x)\phi(\omega_j;y)$$

- w/ (approx) m-dim feature map $\widehat{\phi}(x) = \frac{1}{\sqrt{m}}(\phi(\omega_1;x),\ldots,\phi(\omega_m;x))$
- can then again define the induced RKHS $\mathcal{F}_{\hat{\mathcal{K}}} = \{f : f = \frac{1}{m}\sum_{j=1}^{m}\tilde{f}(\omega_j)\phi(\omega_j;x), \tilde{f}\in\mathcal{H}\}$

Random features 'ctd

Theorem (Approximation for random features, Rahimi Recht '08)

For $f = \mathbb{E}_{\omega\sim\mu}\tilde{f}(\omega)\phi(\omega;\cdot)\in\mathcal{F}_{\mathcal{K}}$ with $\|\tilde{f}\|_\infty \le C$, define $\widehat{f} = \mathbb{E}_{\omega\sim\hat{\mathbb{P}}_m}\tilde{f}(\omega)\phi(\omega;\cdot)\in\mathcal{F}_{\hat{\mathcal{K}}}$. Then w/ prob. $\ge 1-\delta$ we have

$$\|\widehat{f} - f\|^2_{\mathcal{L}^2(\mathbb{P})} \le \frac{C}{\sqrt{m}}(1+\sqrt{2\log 1/\delta}).$$

- Proof via McDiarmid + Jensen's (on the expectation of norms) (see Percy Liang's notes)
- ∞-dim to n-dim to m-dim problem, since we can just solve linear problem by expressing $f(x_1^n) = \Phi\alpha$ with $\alpha\in\mathbb{R}^m$

$\to$ choosing m too small gets bad approx. error. In practice would choose $\sim n$ (statistical error), so no real computational gain if no additional structural assumptions are made on $\mathcal{F}_{\mathcal{K}}$

Example: two-layer fully-connected NN

- Taylor "linearization" around initialization of width-m 2-layer NN

$$f_{NN}(x) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m} a_j\sigma(\langle w_j, x\rangle) \approx \frac{1}{\sqrt{m}}\sum_{j=1}^{m} a_{0,j}\sigma(\langle w_{0,j}, x\rangle)$$
$$+ \underbrace{\sum_j \frac{(a_j - a_{0,j})}{\sqrt{m}}\sigma(\langle w_{0,j}, x\rangle)}_{T_1(x)} + \underbrace{\sum_j (w_j - w_{0,j})^\top (a_{0,j} x\sigma'(\langle w_{0,j}, x\rangle))}_{T_2(x)}$$

  where $w_{0,j} \overset{i.i.d.}{\sim} \mu_w$, $a_{0,j} \overset{i.i.d.}{\sim} \mu_a$ at initialization, w/ non-linearity σ

- $T_1 \in \mathcal{F}_{RF} := \{f_1 : f_1(x) = \frac{1}{\sqrt{m}}\sum_{j=1}^m s_j\sigma(\langle w_{0,j}, x\rangle), s \in \mathbb{R}^m\}$ with feature maps $\phi_j(x) = \sigma(\langle w_{0,j}, x\rangle) \to \mathcal{F}_{RF}$ has kernel $\hat{\mathcal{K}}(x,y) = \frac{1}{m}\sum_{j=1}^m \sigma(\langle w_{0,j}, x\rangle)\sigma(\langle w_{0,j}, y\rangle)$ that approximates $\mathcal{K}(x,y) = \mathbb{E}_{\mu_w}\sigma(\langle w_{0,j}, x\rangle)\sigma(\langle w_{0,j}, y\rangle)$ as the layer width $m \to \infty$

- $T_2 \in \mathcal{F}_{NTK} := \{f_2 : f_2(x) = \frac{1}{\sqrt{m}}\sum_{j=1}^m v_j^\top (a_{0,j} x\sigma'(\langle w_{0,j}, x\rangle)), v_j \in \mathbb{R}^d\}$ with feature maps $\phi_{ij} = x_i a_{0,j}\sigma'(\langle w_{0,j}, x\rangle)$, $i \in [d], j \in [m]$ $\to \mathcal{F}_{NTK}$ has kernel $\hat{\mathcal{K}}(x,y) = \frac{1}{m}\sum_{j=1}^m x^\top y\sigma'(\langle w_{0,j}, x\rangle)\sigma'(\langle w_{0,j}, y\rangle)$ that approximates $\mathcal{K}(x,y) = \mathbb{E}_{\mu_w} x^\top y\sigma'(\langle w_{0,j}, x\rangle)\sigma'(\langle w_{0,j}, y\rangle)$

Idea:

- $\mathcal{F}_{RF}$ corresponds to class where first layer stays fixed at initialized value, second layer trainable, and $\mathcal{F}_{NTK}$ vice versa
- sum of both kernels yields another kernel and hence forms a "new" RKHS $\mathcal{F} = \mathcal{F}_{RF} \oplus \mathcal{F}_{NTK}$

$\to$ You could say, optimizing 2-layer NN $\approx$ optimizing loss in RKHS ($\to$ analyzable!)

- linear expansion is only good when $\|w_j - w_{0,j}\|$ small $\to$ people show for large enough width changes are indeed small
- just showed that infinite-width limit kernels "make sense" (check out arc-cosine kernel)
- infinite width is far from what we use $\to$ people are trying to show optimization and generalization results for poly or logarithmic in n, d

References

Translation-invariant kernels and Random features

- *Percy Liang Lecture Notes*: Lectures 11, 12
- *Rahimi and Recht '08: Random Features for Large-Scale Kernel Machines* (Neurips)

Neural networks and kernels

- Matus Telgarsky's deep learning theory lectures: https://mjt.cs.illinois.edu/courses/dlt-f19~/files/lec5-handout.pdf
- *Cho, Saul '09: Kernel methods for deep learning* (Neurips): arc-cosine kernel
- NTK related: e.g. Jacot, Gabriel, Hongler '18, Chizat, Bach '19
- Approximation properties of $\mathcal{F}_{NTK}, \mathcal{F}_{RF}$ and the infinite width limit: Ghorbani, Misiakiewicz, Mei, Montanari '19, Mei, Montanari '19

Random design

- MW Chapter 14lat

Lecture 11: Minimax lower bounds

Announcements

- Homework 2 was due last night, solutions out today
- Please fill out your oral exam availabilities sent out in email, taking place 20.11./21.11. 9 am - 5 pm
  - mark *all slots* where you do not have a strict conflict
  - exams are 20 minutes long

Recap: Upper bound for random design

We considered the non-parametric regression setting $Y = f^\star(X) + w$

We view X as random and take expectation also over X, i.e. for any $f \in \mathcal{L}^2(\mathbb{P})$, we have

$$\|f - f^\star\|_2^2 = R(f) - R(f^\star) = \mathbb{E}_{X,W}(Y - f(X))^2 - \mathbb{E}W^2$$
$$= \mathbb{E}_X(f(X) - f^\star(X))^2 = \mathbb{E}_{x_1,\dots,x_n}\|f - f^\star\|_n^2$$

and want to bound $\|\widehat{f} - f^\star\|_2^2$ for an estimator $\widehat{f}$

Theorem (Localized uniform law, MW Thm 14.1)

For star-shaped and $b-$uniformly bounded $\mathcal{F}^\star$, let $\overline{\delta}_n$ be population critical radius. Then if $\overline{\delta}_n^2 > c\frac{\log[4\log(1/\overline{\delta}_n)]}{n}$ then w.p. at least $1 - c_1 e^{-c_2 \frac{n\overline{\delta}_n^2}{b^2}}$ we have $\sup_{g\in\mathcal{F}^\star} \|g\|_2 - \|g\|_n \le c\overline{\delta}_n$

For bounded domains, we can then plug in $g = \widehat{f} - f^\star$, use the h.p. upper bound for the empirical error $\|\widehat{f} - f^\star\|_n^2 \le U(n)$ and obtain w.h.p

$$\|\widehat{f} - f^\star\|_2^2 \le U(n) + c\overline{\delta}_n$$

Estimation task

- Let $\mathcal{P}$ be a set of probability distributions on $(\mathcal{X}, \mathcal{Y})$, can then view a quantity of interest to be a mapping F acting on a probability distribution (outputting a function or parameter)
- For today, we consider each $\mathbb{P}_\mathcal{F} \in \mathcal{P}$ defined via $y = f^\star(x) + w$ (either y or both x, y random), for different $f^\star \in \mathcal{F}$ but fixed distributions over x and noise w and the object of interest could be $F(\mathbb{P})(x) = \mathbb{E}[Y|x] = f^\star(x)$.
- View estimating procedure/algorithm for $F(\mathbb{P})$ as a mapping $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$ from dataset to space of functions, where $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with $(x_i, y_i) \sim \mathbb{P}$, outputting $\widehat{f}_\mathcal{D} = \mathcal{A}(\mathcal{D})$
- So far we've seen: Error bounds of the type $\|\widehat{f}_\mathcal{D} - f^\star\|_2^2 \le O(n^{-\alpha})$

Pair-Q: Discuss with your neighbor: What is a reasonable notion of optimality of an algorithm that a practitioner might care about?
Today: Compare to what's the best possible (*optimal*) given the data?

Minimax risk

Definition (Minimax risk)

The minimax risk or error of estimating the mapping $F : \mathcal{P}_{\mathcal{F}} \to \mathcal{F}$ in some squared metric $\|\cdot\|^2$ is defined as

$$\mathfrak{M}(F(\mathcal{P}), \|\cdot\|^2) = \inf_{\mathcal{A}} \sup_{\mathbb{P} \in \mathcal{P}_{\mathcal{F}}} \mathbb{E}_{\mathcal{D} \sim \mathbb{P}^n} \|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2$$

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ has i.i.d. samples from $\mathbb{P}^n \to \mathcal{A}(\mathcal{D})$ is random
- Note that more generally $\mathcal{F}$ can also be a parameter space for parameterized function classes (as we will see next lecture)
- Here $\mathcal{A}$ is not constrained to any particular procedure (could be minimization of risk but also something else) but "knows" to search in set $\mathcal{F}$ that induces $\mathcal{P}_{\mathcal{F}}$
- Here we consider deterministic (i.e. not random) algorithms $\mathcal{A}$
- could use as $\|\cdot\|$ standard metric of $\mathcal{F}$ (see MW Chapter 15)

Minimax lower bounds

What do we learn if we could obtain $\mathfrak{M}(F(\mathcal{P}), \|\cdot\|^2) \geq O(n^{-\alpha})$?

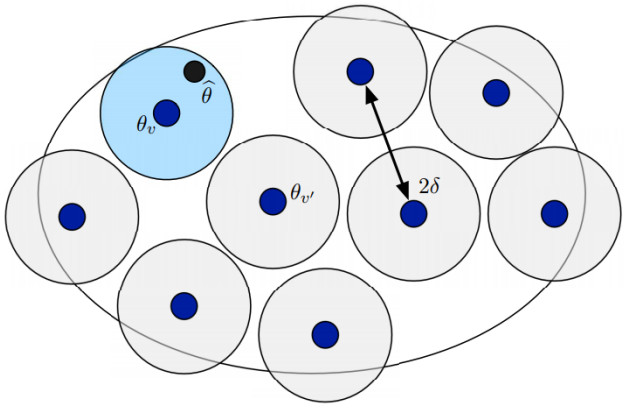
- no estimator (knowing $\mathcal{P}_{\mathcal{F}}$ or, equivalently, $\mathcal{F}$ and) can achieve smaller risk (for their resp. hardest case)
- if upper bound of an estimation procedure matches lower bound:
  - practically we don't need to waste time looking for "better"
  - if we want to do better in the worst case

This class: Find **lower bounds** for the minimax risk as large as possible for **given** $\mathcal{P}$, F

- From estimation to "testing" / classification
- Fano's method: bounding the probability of testing error via mutual information (MI)
- Upper bounding MI using Yang-Barron
- Examples: non-parametric regression on Sobolev functions

Main idea: From estimation to testing (intuition)

- Consider M finite functions f^i spread across $\mathcal{F}$ s.t. pairwise distances $> 2\delta$ (e.g. in a packing set of $\mathcal{F}$)
- If $\mathcal{A}$ can find $\widehat{f}$ (black dot) that is δ close to any true $f^\star \in \mathcal{F}$
  $\rightarrow$ if data is drawn from f^j, $\mathcal{A}$ induces a test that correctly identifies f^j by choosing the closest f^i (blue dot) to the estimated $\widehat{f}$
  $\rightarrow$ no "testing" error



- As we want a lower bound on estimation, can reverse the argument

$\rightarrow$ Problem reduces to: given n points, what's the smallest possible δ so that we can distinguish from which f^i the data was drawn?

Main idea: from estimation to testing

We sometimes write $\widehat{f}_{\mathcal{D}} = \mathcal{A}(\mathcal{D})$, omitting $\mathcal{A}$ subscript. Define

- For any M let $\{f^i\}_{i=1}^M$ be a set of functions in $\mathcal{F}$
- For each $\tilde{f} \in \mathcal{F}$, define $\mathbb{P}_{\tilde{f}}$ as a unique distribution with $F(\mathbb{P}_{\tilde{f}}) = \tilde{f}$
- Define the mixture distribution $\mathbb{Q}_M$ for $\mathcal{D}, J$ by defining
  1. J a uniform R.V. (flat "prior") with values in $[M] = \{1, \ldots, M\}$, i.e. $\mathbb{Q}_M(J = j) = \frac{1}{M}$ for all j
  2. and drawing random i.i.d. datapoints $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ from $\mathbb{P}_{f^j}^n$, i.e. $\mathbb{Q}_M(\mathcal{D}|J = j) = \mathbb{P}_{f^j}^n$
- Decision / Testing functions of form $\psi : (\mathcal{X} \times \mathcal{Y})^n \rightarrow [M]$

Lemma (Estimation vs. testing, MW Prop 15.1)

Choose $\{f^i\}_{i=1}^{M(2\delta)}$ to be a 2δ-packing of $\mathcal{F}$ in the $\|\cdot\|$ metric so that $M(2\delta) \leq \mathcal{M}(2\delta; \mathcal{F}, \|\cdot\|)$, then

$$\inf_{\mathcal{A}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathcal{D} \sim \mathbb{P}} \|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 \geq \delta^2 \inf_{\psi} \mathbb{Q}_{M(2\delta)}(\psi(\mathcal{D}) \neq J)$$

Proof of Lemma

Omitting $\mathbb{Q}_M$ subscript, define $\psi_{\mathcal{A}}(\mathcal{D}) := \arg\min_{i\in[M]} \|\mathcal{A}(\mathcal{D}) - f^i\|$

1. Markov's inequality yields

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 &\geq \delta^2\mathbb{P}(\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 \geq \delta^2) \\ &= \delta^2\mathbb{P}(\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\| > \delta)\end{aligned}$$

2. Key link between estimation and "testing" (via intuition sl. 8):

$$\mathbb{Q}(\{\|\mathcal{A}(\mathcal{D}) - f^i)\| \leq \delta\}|J = i) \leq \mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) = i\}|J = i)$$

   because for any $f \in \mathcal{F}$ such that $\|f - f^i\| < \delta$, for any $j \neq i$ we have $\|f - f^j\| > \|f^j - f^i\| - \|f - f^i\| > \delta \to \psi_{\mathcal{A}}(\mathcal{D}) = i$

3. Then the Lemma follows by the distribution of J

$$\delta^{-2}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathcal{D}\sim\mathbb{P}}\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 \overset{1.}{\geq} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n(\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\| > \delta)$$

$$\geq \frac{1}{M}\sum_{i\in[M]} \mathbb{P}^n_{f^i}(\|\mathcal{A}(\mathcal{D}) - f^i\| > \delta) = \sum_{i\in[M]} \mathbb{Q}(J = i)\mathbb{Q}(\|\mathcal{A}(\mathcal{D}) - f^i\| > \delta|J =$$

$$\overset{2.}{\geq} \sum_{i\in[M]} \mathbb{Q}(J = i)\mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) \neq i\}|J = i) = \mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) \neq J\})$$

Lower bounding $\mathbb{Q}(\psi(\mathcal{D}) \neq J)$ with Fano's method

For simplicity assuming densities of joint and conditional distributions:

Definitions (Entropy and mutual information)

For any two R.V. X, Y with joint probability distribution $\mathbb{P}$ define

- the *entropy* $H(X, Y) = -\mathbb{E}_{\mathbb{P}} \log p(X, Y)$
- the *conditional entropy* $H(X|Y) = -\mathbb{E}_{\mathbb{P}} \log p(X|Y)$
- the *mutual information* $I(X, Y) = H(X) - H(X|Y)$

Intuitively (imprecise):

- $H(X|Y)$: uncertainty "left" about X if value of Y were known
- $I(X, Y)$: information of X in Y and vice versa

Theorem (Fano's method, MW Sec 15.4.)

For some $M \in \mathbb{N}$ and $\{f^i\}_{i=1}^M$, let $\mathbb{Q}_M$ be a mixture distribution as in slide 9. Then for any decision/testing function ψ, it holds that

$$\mathbb{Q}_M(\psi(\mathcal{D}) \neq J) \geq 1 - \frac{I(\mathcal{D}, J) + \log 2}{\log M}$$

Proof of Theorem (Fano's method)

Define Bernoulli $E_\psi = \mathbb{1}_{\psi(\mathcal{D}) \neq J}$ with $\mathbb{Q}_M(E_\psi = 1) = \mathbb{Q}_M(\psi(\mathcal{D}) \neq J)$

1. We first establish *Fano's inequality* after which the proof is trivial

$$H(J|\mathcal{D}) \leq H(E_\psi) + \mathbb{Q}_M(\psi(\mathcal{D}) \neq J) \log(M-1)$$

- Proof: First, by Bayes' theorem and def. of conditional expectations

$$\underbrace{H(E_\psi|J,\mathcal{D})}_{=0} + H(J|\mathcal{D}) = H(J, E_\psi|\mathcal{D}) = H(J|E_\psi,\mathcal{D}) + \underbrace{H(E_\psi|\mathcal{D})}_{\leq H(E_\psi)}$$

- Proof then follows from

$$H(J|E_\psi,\mathcal{D}) = \underbrace{H(J|E_\psi = 0,\mathcal{D})}_{=0}\mathbb{Q}(E_\psi = 0) + \underbrace{H(J|E_\psi = 1,\mathcal{D})}_{\leq \log(M-1)}\mathbb{Q}(E_\psi = 1)$$

2. Since E_ψ Bernoulli $H(E_\psi) \leq \log 2$ for all ψ
   and since J uniform $H(J) = \log M$
3. Using Fano's inequality and $H(J|\mathcal{D}) = H(J) - I(\mathcal{D}, J)$ yields Thm.

Fano's method to lower bound minimax risk

- We would like to ultimately plug in Fano's lower bound into the lemma.
- If we choose $\{f^i\}_{i=1}^{M(2\delta)}$ to be a 2δ-packing as in Lemma we can plug in $M = M(2\delta) \leq \mathcal{M}(2\delta; \mathcal{F}, \|\cdot\|)$ to get

$$\mathbb{Q}_{M(2\delta)}(\psi(\mathcal{D}) \neq J) \geq 1 - \frac{I(\mathcal{D}, J) + \log 2}{\log M(2\delta)}$$

- If δ is chosen such that $I(\mathcal{D}, J) \sim \log M(2\delta)$ then the Lemma implies a lower bound of order δ^2
- This might or might not be a tight lower bound (if it matches some algorithm dependent upper bound, you're in luck)

Upper bounding the mutual information

- To bound the mutual information we recall the

Definition (Kullback-Leibler divergence)

The KL divergence between any two probability distributions $\mathbb{P}, \mathbb{Q}$

$$KL(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}} \log \frac{d\mathbb{P}}{d\mathbb{Q}}$$

- We can write $I(\mathcal{D}, J) = KL(\mathbb{Q} \,\|\, \mathbb{Q}_{\mathcal{D}}\mathbb{Q}_J)$ and then for q densities of $\mathbb{Q}$, we have

$$\mathbb{E}_J \mathbb{E}_{\mathcal{D}} \log \frac{q_{\mathcal{D}|J}}{q_{\mathcal{D}}} = \mathbb{E}_J KL(\mathbb{Q}_{\mathcal{D}|J} \,\|\, \mathbb{Q}_{\mathcal{D}})$$

$$= \frac{1}{M} \sum_{i=1}^{M} KL(\mathbb{P}_{fi}^n \,\|\, \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{fj}^n)$$

- The next theorem bounds the mutual information in Fano's method.

Theorem (Yang-Barron, MW Lemma 15.21)

$$I(\mathcal{D}, J) \leq \inf_{\epsilon > 0} \epsilon^2 + \log \mathcal{N}(\epsilon^2; \mathcal{P}^n, KL)$$

Summary: One recipe for minimax lower bounds

Recipe for using Yang-Barron + Fano to get lower bounds:

1. Choose ϵ such that $\epsilon^2 \geq \log \mathcal{N}(\epsilon^2; \mathcal{P}^n, KL)$
2. Choose δ such that $\log \mathcal{M}(2\delta; \mathcal{F}, \|\cdot\|) \geq 4\epsilon^2 + 2\log 2$
3. Hence $1 - \frac{I(\mathcal{D},J) + \log 2}{\log M(2\delta)} \geq \frac{1}{2}$ and via Fano's method

$$\inf_{\mathcal{A}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 \geq \frac{1}{2}\delta^2$$

Minimax prediction error for estimating Sobolev functions

Example: Sobolev functions $\mathcal{F} = \mathcal{W}_2^\alpha([0,1])$ with

- Consider the family of distributions $\mathcal{P}_\mathcal{F}$ generated via: $X \sim U([0,1])$ and $y = f^\star(x) + w$ with standard normal w and $f^\star \in \mathcal{W}_2^\alpha([0,1])$ so that conditional distribution $Y|x \sim \mathcal{N}(f(x), \sigma^2)$ (our non-parametric regression setting)
- We're interested in estimating $f^\star = \mathbb{E}_\mathbb{P}[Y|x]$ and evaluate it via the $\mathcal{L}^2([0,1])$ norm
- Recall *upper bounds* for constrained kernel regression
  - w.h.p. $\|\widehat{f} - f^\star\|_n^2 \leq O\big(\frac{\sigma^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}$ (HW 2)
  - $\widehat{f} - f^\star$ is uniformly bounded by reproducing property and Hilbert norm constraint $\rightarrow$ MW Thm 14.1. and MW Prop 14.25 yields $\|\widehat{f} - f^\star\|_{\mathcal{L}^2([0,1])}^2 \leq O\big(\frac{\sigma^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}$

Minimax prediction error for estimating Sobolev functions

Corollary (Minimax error for Sobolev function estimation)

Writing $\|\cdot\|_2 := \|\cdot\|_{\mathcal{L}^2([0,1])}^2$, *we have for* $\frac{n}{\sigma^2}$ *larger than a constant*

$$\mathfrak{M}(F(\mathcal{P}), \|\cdot\|_2^2) \geq O\Big(\frac{\sigma^2}{n}\Big)^{\frac{2\alpha}{2\alpha+1}}$$

Proof of Corollary

a) Writing out the conditional distribution we have for $n = 1$

$$KL(\mathbb{P}_f \parallel \mathbb{P}_g) = \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{P}_f} g^2(X) - f^2(X) + 2(f(X) - g(X))Y$$

$$= \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{P}_f} g^2(X) - f^2(X) + 2(f(X) - g(X))f(X) = \frac{\|f-g\|_2^2}{2\sigma^2}$$

b) For n samples we have an extra factor of n, since for $z_i = (x_i, y_i)$

$$KL(\mathbb{P}_f^n \parallel \mathbb{P}_g^n) = \int \prod_{i=1}^n p_f(z_i) \log \prod_{i=1}^n \frac{p_f(z_i)}{p_g(z_i)} \mu(dz^n)$$

$$= \sum_{i=1}^n \int p_f(z_i) \log \frac{p_f(z_i)}{p_g(z_i)} \mu(\mathrm{d}z_i) = n\frac{\|f-g\|_2^2}{2\sigma^2}$$

Proof ctd'

c) Hence $\mathcal{N}(\epsilon^2; \mathcal{P}^n, KL) = \mathcal{N}(\frac{\epsilon\sqrt{2\sigma^2}}{\sqrt{n}}; \mathcal{W}_2^\alpha([0,1]), \|\cdot\|_2)$

d) Using the result in next slide about covering number of Sobolev spaces

- Using $\log \mathcal{N}(\delta; \mathcal{W}_2^\alpha([0,1]), \|\cdot\|_2^2) = O(\frac{1}{\delta})^{1/\alpha}$ and 1. in slide 15 we require

$$\epsilon^2 \geq (\frac{n}{2\sigma^2})^{\frac{1}{2\alpha}} \epsilon^{-1/\alpha} \quad \rightarrow \quad \epsilon^2 = O(\frac{n}{\sigma^2})^{\frac{1}{2\alpha+1}}$$

- Recalling that $\mathcal{M}(2\delta) \geq \mathcal{N}(2\delta)$ and using 2. in slide 15, it suffices to require

$$(\frac{1}{\delta})^{\frac{1}{\alpha}} \geq c\big[(\frac{n}{\sigma^2})^{\frac{1}{2\alpha+1}} + 2\log 2\big] \quad \rightarrow \quad \delta^2 = O(\frac{\sigma^2}{n})^{\frac{2\alpha}{2\alpha+1}}$$

for $\frac{\sigma^2}{n}$ smaller than a universal constant.

e) Hence by 3. (Fano's method) $\|\widehat{f} - f^\star\|^2_{\mathcal{L}^2([0,1])} \geq O\Big(\frac{\sigma^2}{n}\Big)^{\frac{2\alpha}{2\alpha+1}}$ □

Metric entropy for higher order Sobolev spaces (bonus)

Lemma (Metric entropy for α-order compact Sobolev spaces)

It holds that $\log \mathcal{N}(\delta; \mathcal{W}_2^\alpha([0,1]), \|\cdot\|_2^2) = O(\frac{1}{\delta})^{\frac{1}{\alpha}}$.

Proof steps

Define $\mathcal{E}_\alpha = \{\theta \in \ell_2(\mathbb{N}) : \sum_{j=1}^\infty j^{2\alpha}\theta_j^2 \leq 1\}$

1. First observation: $\mathcal{N}(\delta; \mathcal{W}_2^\alpha([0,1]), \|\cdot\|_2^2) = \mathcal{N}(\delta; \mathcal{E}_\alpha, \|\cdot\|_{\ell^2(\mathbb{N})})$
   - Note that by Mercer's Theorem, we can write for some orthonormal basis in $\|\cdot\|_2$ $\mathcal{W}_2^\alpha([0,1]) = \{f : f = \sum_{j=1}^\infty \theta_j\phi_j \text{ for } \theta \in \mathcal{E}_\alpha\}$
   - Kernel operator eigenvalues decay as $j^{2\alpha}$ (hinges on spectra of differential operators that we won't prove)
   - Because ϕ_j are orthonormal in $\|\cdot\|_2$ norm we have $\|f\|_2^2 = \|\theta_f\|^2_{\ell^2(\mathbb{N})}$
2. MW Example 5.12. proves $\log \mathcal{N}(\delta; \mathcal{E}_\alpha, \|\cdot\|_{\ell^2(\mathbb{N})}) \leq O\Big(\frac{1}{\delta}\Big)^{\frac{1}{\alpha}}$ □

References

Main source

- MW Chapter 15

Additional reading

- *John Duchi Information Theory (Stats 311) Lecture Notes*: Lectures 3, 5, 6
- *Bin Yu '97*: Assouad, Fano and LeCam, "Festschrift for Lucien LeCam" - overview of different minimax methods (including two we did not talk about)
- *Yang, Barron '99*: Information theoretic determination of minimax rates of convergence.