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 $O(\left(\frac{\log n}{n}\right)^{1/3})$ vs. Dudley: $O(\left(\frac{\log n}{n}\right)^{1/2})$ (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*

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

- 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:

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

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

b. eq. 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