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

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

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 → used max of subgaussians
- For special parameterization such as linear model → 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

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

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

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

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 $\le \delta$
2. Can then take expected sup over grid points
3. Bound difference to other points again using naive bound

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

Bound using naive (1-step) covering argument

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

Let $\delta > 0$. *If a set of points* $\theta^1, \ldots, \theta^N$ *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*

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

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

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

- 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

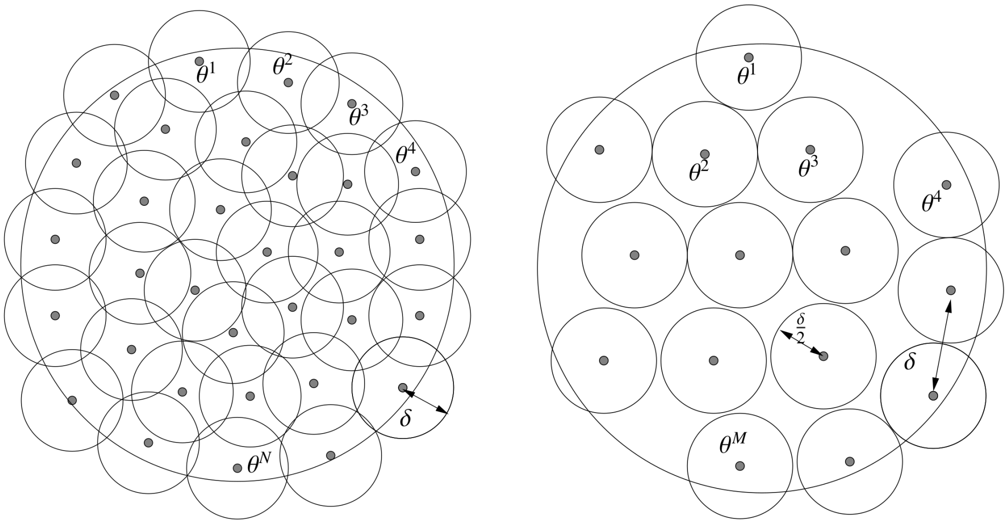

Figure 1: Left: δ-covering, Right: δ-packing

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

We now contrast the covering numbers for a parametric and non-parametric function classes $\mathcal{H} = \mathcal{F}$ (i.e. identity/no loss),

- 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.

Note that for any $\mathcal{H}$ and $f, g \in \mathcal{H}$
$\frac{\|\theta-\theta'\|_2}{\sqrt{n}} = \sqrt{\frac{1}{n}\sum_i (f(z_i) - g(z_i))^2} \leq \max_i |f(z_i) - g(z_i)| \leq \|f - g\|_\infty$

R.C. rates for function classes: Parametric example

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,

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

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

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

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 $h : [0, 1] \to \mathbb{R}$ s.t. $|h^{(\alpha)}(x) - h^{(\alpha)}(x')| \leq L|x - x'|$

- We use bounds for $\mathcal{N}(\delta; \mathcal{H}_2^\alpha, \|\cdot\|_\infty)$ 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

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

  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