

Lecture 7: Coverings and metric entropy

1 / 19

Announcements and plan

- Homework was due, everybody made it in the course!
- Project choices due next Tuesday **14.10.** 2pm, sign up on same googlesheet.
 - Find project partners! Groups of two (alone is also ok but not recommended). Looks like we'll have 12 groups.
 - Think about 2-3 options that you'd be excited about in case multiple groups want to pick the same.
 - If you're suggesting your own one and have discussed with us, add it to googlesheets

Plan today

- Rademacher complexity for general classes via covering numbers and metric entropy

2 / 19

About project choice

1. Identify and motivate problem - why should I / the community care?
Including literature review
2. “Detective hat”: Intuitive (not just technical level) understanding of proof, assumptions, statement in depth
3. “Reviewer hat”: Which relevant questions does it aim to address and does the paper indeed 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? Which experimental questions were not conclusively answered and are important?
5. Break down the identified follow-up problem into feasible chunks (e.g. lemmas, experiments) and show your attempts to tackle the first few steps.

3 / 19

Recap: Bounding Rademacher complexities so far

Generally, for vector $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ we write for emp. R.C.

$$n\tilde{\mathcal{R}}_n(\mathbb{T}) = \mathbb{E} \sup_{\tilde{\theta} \in \mathbb{T}} \langle \epsilon, \tilde{\theta} \rangle \text{ where } \mathbb{T} = \mathcal{H}(z_1^n) \text{ (*)}$$

So far, we bounded the R.C. of two types of classes \mathcal{H}

- discrete: i.e. $\mathcal{H}(z_1^n) \subset \{h : \mathcal{X} \rightarrow \{0, 1\}\}$ where the bound depended on the cardinality $|\mathcal{H}(z_1^n)| \leq 2^n$ using Massart’s Lemma
- continuous but specific form: linear functions with bounded $\|w\| \leq B$ for domain $\|x\| \leq D$ where $\mathcal{H}(z_1^n) = (\ell_\gamma \circ \mathcal{F})(z_1^n)$
 - Here where since the ramp loss $\ell_\gamma : \mathbb{R} \rightarrow [0, 1]$ takes on continuous values, the set $(\ell_\gamma \circ \mathcal{F})(z_1^n) = \{(\ell_\gamma(f(z_1)), \dots, \ell_\gamma(f(z_n)))\}$ is a subset in \mathbb{R}^n containing **infinitely, uncountably many elements**.
 - Using Rademacher contraction: $\tilde{\mathcal{R}}_n((\ell_\gamma \circ \mathcal{F})(z_1^n)) \leq \frac{1}{\gamma} \tilde{\mathcal{R}}_n(\mathcal{F}(z_1^n))$
 - Now in (*), because of the specific structure of $\mathbb{T} = \mathcal{F}(z_1^n)$ for linear functions we can write
$$n\tilde{\mathcal{R}}_n(\mathcal{F}(z_1^n)) = \mathbb{E} \sup_{\|\theta\| \leq B} \langle \sum_i \epsilon_i x_i, \theta \rangle = B \mathbb{E} \|\sum_i \epsilon_i x_i\| = O(\sqrt{n})$$

In general however, since $\sup_{\tilde{\theta} \in \mathbb{T}} \|\tilde{\theta}\|_2$ grows with n , without specific structure, just using Cauchy Schwartz would give vacuous bound.

4 / 19

R.C. rates for different function classes

- Today we'll see examples for general parametric and non-parametric infinite-dimensional and real-valued \mathcal{H} where $\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n)) \leq O(\frac{1}{n^\beta})$ for some $\beta \leq 1/2$, for every z_1^n

- Then with probability at least $1 - \delta$, we have by uniform law

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

- 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"

5 / 19

A general approach to bound the R.C.

- For finite classes or when labels are finite and discrete \rightarrow can use max of subgaussians
- For special parameterization such as linear model \rightarrow used boundedness of parameters and inputs

Today, we present a generic approach by

- viewing the R.C. as the expected supremum of a subgaussian process
- 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 $\theta, \tilde{\theta} \in \mathbb{T}$, the random variable $X_\theta - X_{\tilde{\theta}}$ is subgaussian w/ parameter $\rho(\theta, \tilde{\theta})$ for some metric ρ and $\mathbb{E}X_\theta = 0$

6 / 19

From R.C. to supremum of subgaussian processes

First note that, for Rademacher variables ϵ_i , we can write for any $\mathbb{T} \subset \mathbb{R}^n$

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

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 following

Proposition (Rademacher as a sup of subgaussian processes)

For any \mathbb{T} , the collection of zero-mean variables $\{X_\theta = \frac{1}{\sqrt{n}} \langle \epsilon, \theta \rangle\}_{\theta \in \mathbb{T}}$ 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

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

7 / 19

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

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

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

8 / 19

Can we still 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

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

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 $\rho(\theta, \theta') = \frac{\|\theta - \theta'\|_2}{\sqrt{n}}$, then we have

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

9 / 19

Proof using naive (1-step) covering argument

- Define $i = \arg \min_j \rho(\theta^j, \theta)$ with $\rho(\theta^i, \theta) \leq \delta$ and correspondingly \tilde{i} for $\tilde{\theta}^*$
- For general ρ we can rewrite for any arbitrary $\theta, \tilde{\theta} \in \mathbb{T}$

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

- 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 definition of ρ and 3. of previous slide \square

Using the proposition to bound R.C. by $N(\delta)$

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

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

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 $\delta \sim \frac{1}{\sqrt{n}} \rightarrow \tilde{\mathcal{R}}_n(\mathbb{T}) \leq O(\sqrt{\frac{\log n}{n}})$

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

11 / 19

Covering number and entropy

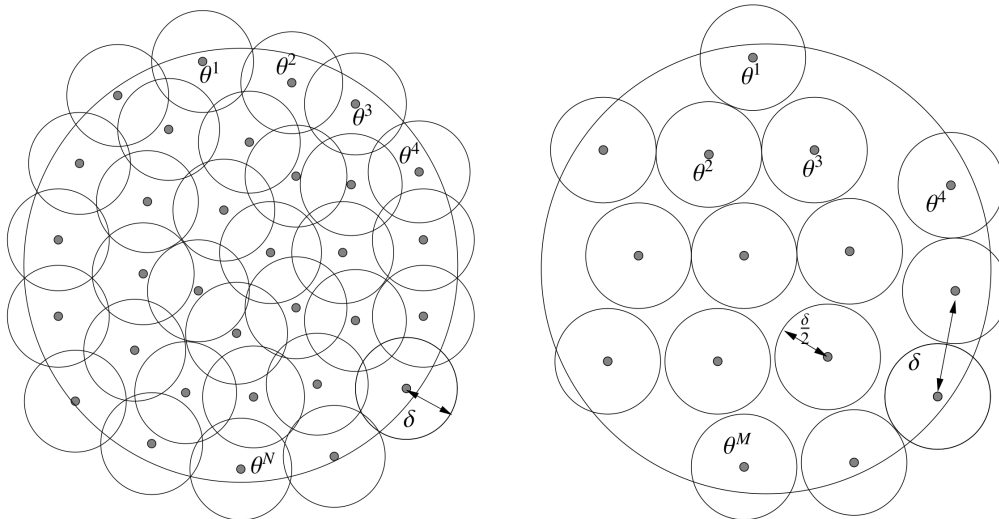


Figure 1: Left: δ -covering, Right: δ -packing

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 $\max_{\theta \in \mathbb{T}} \min_i \rho(\theta_i, \theta) \leq \epsilon$ (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

12 / 19

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!)

13 / 19

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 scaled norm $\rho(\theta, \theta') = \|\theta - \theta'\|_n := \frac{\|\theta - \theta'\|_2}{\sqrt{n}}$ as the metric.

Note that for any \mathcal{H} and $f, g \in \mathcal{H}$

$$\frac{\|\theta - \theta'\|_2}{\sqrt{n}} = \sqrt{\frac{1}{n} \sum_i (f(z_i) - g(z_i))^2} \leq \max_i |f(z_i) - g(z_i)| \leq \|f - g\|_\infty$$

from which it follows that $\mathcal{N}(\delta; \mathcal{H}(z_1^n), \|\cdot\|_n) \leq \mathcal{N}(\delta; \mathcal{H}, \|\cdot\|_\infty)$

14 / 19

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. The class of linear functions in a bounded domain falls into this category!

We prove in next slide: For any z_1^n

$$\mathcal{N}(\delta; \mathcal{H}_1(z_1^n), \|\cdot\|_n) \leq \left(1 + \frac{2L}{\delta}\right)^d \rightarrow \log \mathcal{N}(\delta; \mathcal{H}_1(z_1^n), \|\cdot\|_n) \asymp d \log\left(1 + \frac{L}{\delta}\right)$$

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 $\mathcal{R}_n(\mathcal{H}_1) \leq O\left(\sqrt{\frac{d \log n}{n}}\right)$.

15 / 19

Proof of covering number of \mathcal{H}_1

1. By assumption on h we have for any z_1^n

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

2. Any δ/L -cover $S_{\mathbb{B}} = \{\theta^i\}_{i=1}^N$ for $\mathbb{B}_2(1) \subset \mathbb{R}^d$ automatically induces a δ -cover for $\mathcal{H}_1(z_1^n)$: if for all θ there exists i s.t. $\|\theta - \theta^i\|_2 \leq \frac{\delta}{L}$, then we have by above that $\|h(z_1^n; \theta) - h(z_1^n; \theta^i)\|_n \leq \delta$. Hence $S = \{h(\cdot; \theta^i)\}_{i=1}^N$ would be a δ -cover for $\mathcal{H}_1(z_1^n)$

$$\rightarrow \mathcal{N}(\delta; \mathcal{H}_1(z_1^n), \|\cdot\|_n) \leq \mathcal{N}\left(\frac{\delta}{L}; \mathbb{B}_2(1), \|\cdot\|_2\right)$$

3. (MW Lem. 5.7.) Covering of a ball of metric ρ wrt metric ρ has $\mathcal{N}(\delta; \mathbb{B}_\rho, \rho) = \left(1 + \frac{2}{\delta}\right)^d$ using volume ratio bound

$$\rightarrow \mathcal{N}\left(\frac{\delta}{L}; \mathbb{B}_2(1), \|\cdot\|_2\right) \leq \left(1 + \frac{2L}{\delta}\right)^d$$

16 / 19

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^α with $h : [0, 1] \rightarrow \mathbb{R}$ 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 the fact that $\mathcal{N}(\delta; \mathcal{H}_2^\alpha(z_1^n), \|\cdot\|_n) \leq \mathcal{N}(\delta; \mathcal{H}_2^\alpha, \|\cdot\|_\infty)$
- For $\alpha = 0$, using the sandwich inequality between packing and covering numbers, and constructing a 2δ -packing that is also a δ -covering (see next slide), 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 $\log \mathcal{N}(\delta; \mathcal{H}_2^\alpha, \|\cdot\|_\infty) \asymp \left(\frac{1}{\delta}\right)^{\frac{1}{\alpha+1}}$ 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.).

17 / 19

2δ -packing and δ -cover for nonparametric example

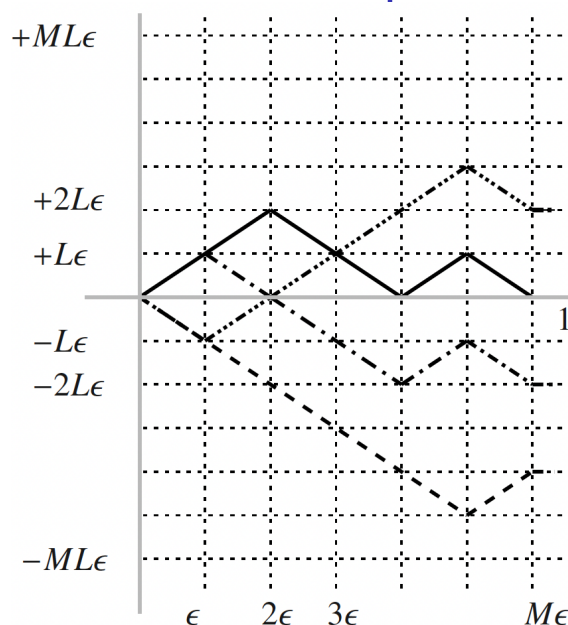


Figure 2: Cover for L -Lipschitz functions with $h(0) = 0$

- See Ex 5.10. for mathematical expression of this set of functions
- One can see visually that picking $\epsilon = \frac{\delta}{L}$ yields
 - a 2δ -packing since biggest difference between any two functions is at least 2δ and
 - a δ -cover because any Lipschitz function with $h(0) = 0$ at any point is $\leq \delta$ close to the next function in this set

18 / 19

References

Metric entropy

- MW Chapter 5