

SDS 384 11: Theoretical Statistics

Lecture 16: Uniform Law of Large Numbers- Dudley's chaining Introduction

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

Rademacher complexity of function classes

Example

Suppose \mathcal{F} is a class parametric functions $\mathcal{F} := \{f(\theta, \cdot) : \theta \in B_2\}$, where B_2 is the unit L_2 ball in \mathbb{R}^d . Assume that \mathcal{F} is closed under negation. f is L Lipschitz w.r.t. the Euclidean distance on Θ , i.e.

$$|f(\theta, \cdot) - f(\theta', \cdot)| \leq L\|\theta - \theta'\|_2.$$

$$\mathcal{R}_n(\mathcal{F}) = O\left(L\sqrt{\frac{d \log(Ln)}{n}}\right)$$

Rademacher complexity of function classes

Example

Suppose \mathcal{F} is a class parametric functions $\mathcal{F} := \{f(\theta, \cdot) : \theta \in B_2\}$, where B_2 is the unit L_2 ball in \mathbb{R}^d . Assume that \mathcal{F} is closed under negation. f is L Lipschitz w.r.t. the Euclidean distance on Θ , i.e.

$$|f(\theta, \cdot) - f(\theta', \cdot)| \leq L\|\theta - \theta'\|_2.$$

$$\mathcal{R}_n(\mathcal{F}) = O\left(L\sqrt{\frac{d \log(Ln)}{n}}\right)$$

- How do we do this?
- Using covering numbers. But we need to define a bunch of stuff first.

A Stochastic Process

- Consider a set $\mathcal{T} \subseteq \mathcal{R}^d$.
- The family of random variables $\{X_\theta : \theta \in \mathcal{T}\}$ define a Stochastic process indexed by \mathcal{T} .

A Stochastic Process

- Consider a set $\mathcal{T} \subseteq \mathcal{R}^d$.
- The family of random variables $\{X_\theta : \theta \in \mathcal{T}\}$ define a Stochastic process indexed by \mathcal{T} .
- We are often interested in the behavior of this process given its dependence on the structure of the set \mathcal{T} .
- In the other direction, we want to know the structure of \mathcal{T} given the behavior of this process.

Definition

A canonical Gaussian process indexed by \mathcal{T} is defined as:

$$G_\theta := \langle z, \theta \rangle = \sum_k z_k \theta_k,$$

where $z_k \stackrel{\text{iid}}{\sim} N(0, 1)$. The supremum $\mathcal{G}(\mathcal{T}) := E_z[\sup_{\theta \in \mathcal{T}} G_\theta]$ is the Gaussian complexity of \mathcal{T} .

Rademacher complexity

- Replacing the iid standard normal variables by iid Rademacher random variables gives a Rademacher process $\{R_\theta, \theta \in \mathcal{T}\}$, where

$$R_\theta := \langle \epsilon, \theta \rangle = \sum_k \epsilon_k \theta_k, \quad \text{where } \epsilon_k \stackrel{\text{iid}}{\sim} \text{Uniform}\{-1, 1\}$$

Rademacher complexity

- Replacing the iid standard normal variables by iid Rademacher random variables gives a Rademacher process $\{R_\theta, \theta \in \mathcal{T}\}$, where

$$R_\theta := \langle \epsilon, \theta \rangle = \sum_k \epsilon_k \theta_k, \quad \text{where } \epsilon_k \stackrel{\text{iid}}{\sim} \text{Uniform}\{-1, 1\}$$

- $\mathcal{R}(\mathcal{T}) := E_\epsilon[\sup_{\theta \in \mathcal{T}} R_\theta]$ is called the Rademacher complexity of \mathcal{T} .

How does this relate to the former notions of Rademacher complexity?

- Recall that

$$\mathcal{R}_{\mathcal{F}} := E[\sup_{f \in \mathcal{F}} |\sum_i \epsilon_i f(X_i)|] = E[E[\sup_{f \in \mathcal{F}} |\sum_i \epsilon_i f(X_i)| | X_1, \dots, X_n]]$$

How does this relate to the former notions of Rademacher complexity?

- Recall that

$$\mathcal{R}_{\mathcal{F}} := E\left[\sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i f(X_i) \right| \right] = E\left[E\left[\sup_{f \in \mathcal{F}} \left| \sum_i \epsilon_i f(X_i) \right| \middle| X_1, \dots, X_n\right]\right]$$

- Now the inner expectation can be upper bounded by

$$E_{\epsilon} \sup_{\theta \in \mathcal{T} \cup -\mathcal{T}} \sum_i \epsilon_i \theta_i, \text{ where } \mathcal{T} \subseteq \mathbb{R}^n \text{ can be written as}$$

$$\mathcal{T} = \{(f(X_1), \dots, f(X_n)) \mid f \in \mathcal{F}\}$$

Theorem

For $\mathcal{T} \in \mathbb{R}^d$,

$$\mathcal{R}(\mathcal{T}) \leq \sqrt{\frac{\pi}{2}} \mathcal{G}(\mathcal{T}) \leq c \sqrt{\log d} \mathcal{R}(\mathcal{T})$$

- This is showing that there can be there are some sets where the Gaussian complexity can be substantially larger than the Rademacher complexity.
- We will in fact give an example.

Proof (of first inequality)

$$\begin{aligned}\mathcal{G}(\mathcal{T}) &= E \sup_{\theta \in \mathcal{T}} \sum_i z_i \theta_i \\ &= E_{\epsilon} E_Z \sup_{\theta \in \mathcal{T}} \sum_i \epsilon_i |z_i| \theta_i \\ &\geq E_{\epsilon} \sup_{\theta \in \mathcal{T}} \sum_i \epsilon_i E |z_i| \theta_i \\ &= \sqrt{\frac{2}{\pi}} \mathcal{R}(\mathcal{T})\end{aligned}$$

Proof (of first inequality)

$$\begin{aligned}\mathcal{G}(\mathcal{T}) &= E \sup_{\theta \in \mathcal{T}} \sum_i z_i \theta_i \\ &= E_{\epsilon} E_Z \sup_{\theta \in \mathcal{T}} \sum_i \epsilon_i |z_i| \theta_i \\ &\geq E_{\epsilon} \sup_{\theta \in \mathcal{T}} \sum_i \epsilon_i E |z_i| \theta_i \\ &= \sqrt{\frac{2}{\pi}} \mathcal{R}(\mathcal{T})\end{aligned}$$

Example

Example

Consider the L_1 ball in \mathcal{R}^d denoted by B_1^d .

$$\mathcal{R}(B_1^d) = 1, \mathcal{G}(B_1^d) \leq \sqrt{2 \log d}$$

Example

Example

Consider the L_1 ball in \mathcal{R}^d denoted by B_1^d .

$$\mathcal{R}(B_1^d) = 1, \mathcal{G}(B_1^d) \leq \sqrt{2 \log d}$$

- $\mathcal{R}(B_1^d) = E\left[\sup_{\|\theta\|_1 \leq 1} \sum_i \theta_i \epsilon_i\right] = E[\|\epsilon\|_\infty] = 1$
- Similarly, $\mathcal{G}(B_1^d) = E[\|z\|_\infty]$

Recall the finite class lemma?

Theorem

Consider z with independent standard normal components.

$$E \max_{a \in A} \langle z, a \rangle \leq \max_{a \in A} \|a\| \sqrt{2 \log |A|}$$

Recall the finite class lemma?

Theorem

Consider z with independent standard normal components.

$$E \max_{a \in A} \langle z, a \rangle \leq \max_{a \in A} \|a\| \sqrt{2 \log |A|}$$

- In our case, $A = \{e_i, i \in [d]\}$, $e_i(j) = \pm 1(j = i)$, $|A| = 2d$ and $\max_{a \in A} \|a\| = 1$.
- This gives a weaker bound on the Gaussian complexity.

A sub-gaussian process

Definition

A stochastic process $\theta \rightarrow X_\theta$ with indexing set T is sub-Gaussian w.r.t a metric d_X if $\forall \theta, \theta' \in T$ and $\lambda \in \mathbb{R}$,

$$E \exp(\lambda(X_\theta - X_{\theta'})) \leq \exp\left(\frac{\lambda^2 d_X(\theta, \theta')^2}{2}\right)$$

- This immediately implies the following tail bound.

$$P(|X_\theta - X_{\theta'}| \geq t) \leq 2 \exp\left(-\frac{t^2}{2d_X(\theta, \theta')^2}\right)$$

Upper bound by 1 step discretization

Theorem

(1-step discretization bound). Let $\{X_\theta, \theta \in \mathcal{T}\}$ be a zero-mean sub-Gaussian process with respect to the metric d_X . Then for any $\delta > 0$, we have

$$E \left[\sup_{\theta, \theta' \in \mathcal{T}} (X_\theta - X_{\theta'}) \right] \leq 2E \left[\sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_\theta - X_{\theta'}) \right] + 2D \sqrt{\log N(\delta; \mathcal{T}, d_X)},$$

where $D := \max_{\theta, \theta' \in \Theta} d_X(\theta, \theta')$.

- The mean zero condition gives us:

$$E[\sup_{\theta \in \mathcal{T}} X_\theta] = E[\sup_{\theta \in \mathcal{T}} (X_\theta - X_{\theta_0})] \leq E[\sup_{\theta, \theta' \in \mathcal{T}} (X_\theta - X_{\theta'})]$$

$$E \left[\sup_{\theta, \theta' \in \mathcal{T}} (X_\theta - X_{\theta'}) \right] \leq \underbrace{2 E \left[\sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_\theta - X_{\theta'}) \right]}_{\text{Approximation error}} + \underbrace{4 \sqrt{D^2 \log N(\delta; \mathcal{T}, d_X)}}_{\text{Estimation error}}$$

- As $\delta \rightarrow 0$, the cover becomes more refined, and so the approximation error decays to zero.
- But the estimation error grows.
- In practice the δ can be chosen to achieve the optimal trade-off between two terms.

- Choose a δ cover T .
- For $\theta, \theta' \in \mathcal{T}$, let $\theta^1, \theta^2 \in T$ such that $d_X(\theta, \theta^1) \leq \delta$ and $d_X(\theta', \theta^2) \leq \delta$.

$$\begin{aligned} X_\theta - X_{\theta'} &= (X_\theta - X_{\theta^1}) + (X_{\theta^1} - X_{\theta^2}) + (X_{\theta^2} - X_{\theta'}) \\ &\leq 2 \sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_\theta - X_{\theta'}) + \sup_{\theta^i, \theta^j \in T} (X_{\theta^i} - X_{\theta^j}) \end{aligned}$$

- But note that $X_{\theta^1} - X_{\theta^2} \sim \text{Subgaussian with variance proxy } (d_X(\theta^1, \theta^2))$.

Finite class lemma for subgaussian processes

Theorem

Consider X_θ sub-gaussian w.r.t d on \mathcal{T} and A is a set of pairs from \mathcal{T} .

$$E \max_{(\theta, \theta') \in A} (X_\theta - X_{\theta'}) \leq D \sqrt{2 \log |A|},$$

where $D := \max_{(\theta, \theta') \in A} d_X(\theta, \theta')$.

Finite class lemma

$$\begin{aligned}\exp\left(\lambda E \max_{(\theta, \theta') \in A} (X_\theta - X_{\theta'})\right) &\leq E \exp\left(\lambda \max_{(\theta, \theta') \in A} (X_\theta - X_{\theta'})\right) \\ &= \max_{(\theta, \theta') \in A} E \exp(\lambda(X_\theta - X_{\theta'})) \\ &\leq \sum_{(\theta, \theta') \in A} \exp\left(\frac{\lambda^2 d_X(\theta, \theta')^2}{2}\right) \\ &\leq |A| \exp\left(\frac{\lambda^2 D^2}{2}\right)\end{aligned}$$

- Now optimize over λ .

Finishing the proof

$$X_\theta - X_{\theta'} \leq 2 \sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_\theta - X_{\theta'}) + \sup_{\theta^i, \theta^j \in \mathcal{T}} (X_{\theta^1} - X_{\theta^2})$$

$$\begin{aligned} E \left[\sup_{\theta, \theta' \in \mathcal{T}} (X_\theta - X_{\theta'}) \right] &\leq 2E \left[\sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_\theta - X_{\theta'}) \right] + E \left[\sup_{\theta^i, \theta^j \in \mathcal{T}} (X_{\theta^1} - X_{\theta^2}) \right] \\ &\leq 2E \left[\sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_\theta - X_{\theta'}) \right] + D \sqrt{2 \log N(\delta; \mathcal{T}, d_X)^2} \end{aligned}$$

Revisiting: smoothly parametrized class

Example

Suppose \mathcal{F} is a class parametric functions $\mathcal{F} := \{f(\theta, \cdot) : \theta \in B_2\}$, where B_2 is the unit L_2 ball in \mathbb{R}^d . Assume that \mathcal{F} is closed under negation. f is L Lipschitz w.r.t. the Euclidean distance on Θ , i.e.

$$|f(\theta, \cdot) - f(\theta', \cdot)| \leq L\|\theta - \theta'\|_2.$$

$$\mathcal{R}_n(\mathcal{F}) = O\left(L\sqrt{\frac{d \log(Ln)}{n}}\right)$$

- Denote $f(\theta, X_1^n)$ as the vector $(f(\theta, X_1), \dots, f(\theta, X_n))$.
- Recall that $n\mathcal{R}_n(\mathcal{F}) = E \left[\sup_{f \in \mathcal{F}} \langle \epsilon, f(\theta, X_1^n) \rangle \right] = E \left[\sup_{\theta \in \Theta} \langle \epsilon, f(\theta, X_1^n) \rangle \right]$
- The process $f(\theta, X_1^n) \rightarrow \langle \epsilon, f(\theta, X_1^n) \rangle =: Y_\theta$ is mean zero subgaussian.
- Note that $Y_\theta - Y_{\theta'} \sim \text{Subgaussian with var proxy } d_X(\theta, \theta')^2$
- We have:

$$d_X(\theta, \theta') = \|f(\theta, X_1^n) - f(\theta', X_1^n)\|^2 \leq nL^2 \|\theta - \theta'\|_2^2$$

- So it is $L\sqrt{n}$ Lipschitz.

- Also,

$$n\mathcal{R}_n(\mathcal{F}) = E[\sup_{\theta \in \Theta} (Y_\theta - Y_{\theta'})] \leq E[\sup_{\theta, \theta' \in \Theta} (Y_\theta - Y_{\theta'})]$$

-

$$n\mathcal{R}_n(\mathcal{F}) \leq \underbrace{2E \sup_{\substack{d_X(\theta, \theta') \leq \delta \\ \theta, \theta' \in \Theta}} (Y_\theta - Y_{\theta'})}_A + 2D\sqrt{\log N(\delta; \mathcal{F}(\Theta, X_1^n), d_X)}$$

- $A \leq \delta E \left[\sup_{\|v\|_2=1} \langle \epsilon, v \rangle \right] \leq \delta \sqrt{n}$
- $D = \sup_{\theta, \theta'} d_X(\theta, \theta') = 2L\sqrt{n}$

- $N(\delta; \mathcal{F}, d_X) \leq N(\delta/L\sqrt{n}, \Theta, \|\cdot\|_2) \leq \left(1 + \frac{L\sqrt{n}}{\delta}\right)^d$
- Finally,

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{4\delta}{\sqrt{n}} + 4L\sqrt{\frac{d \log(1 + L\sqrt{n}/\delta)}{n}}$$

- Setting $\delta = 1$ gives:

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{4L}{\sqrt{n}} + 4L\sqrt{\frac{d \log(1 + L\sqrt{n})}{n}}$$

Examples: Nonparametric functions

Example

Suppose \mathcal{F} is a class of L Lipschitz functions which are supported on $[0, 1]$ and $f(0) = 0$. Note that \mathcal{F} is closed under negation. f is L Lipschitz i.e. $|f(x) - f(x')| \leq L|x - x'| \forall x, x' \in [0, 1]$.

$$\mathcal{R}_n(\mathcal{F}) = O\left(\frac{L}{n}\right)^{1/3}$$

Examples: Nonparametric functions

Example

Suppose \mathcal{F} is a class of L Lipschitz functions which are supported on $[0, 1]$ and $f(0) = 0$. Note that \mathcal{F} is closed under negation. f is L Lipschitz i.e. $|f(x) - f(x')| \leq L|x - x'| \forall x, x' \in [0, 1]$.

$$\mathcal{R}_n(\mathcal{F}) = O\left(\frac{L}{n}\right)^{1/3}$$

Examples: Nonparametric functions

- Consider the process $f(X_1^n) \rightarrow \langle \epsilon, f(X_1^n) \rangle = Y_f$
- $Y_f - Y_{f'} \sim \text{subGaussian}$ with var proxy $\|f(X_1^n) - f'(X_1^n)\|_2^2$
- So $d_Y(f, f') = \|f(X_1^n) - f'(X_1^n)\|_2 \leq \sqrt{n} \|f - f'\|_\infty$
- The diameter is $D = \sup_{f, f' \in \mathcal{F}(X_1^n)} d_X(f, f') \leq 2L\sqrt{n}$
- So, $N(\delta, \mathcal{F}(X_1^n), \|\cdot\|_2) \leq N(\delta/\sqrt{n}, \mathcal{F}, \|\cdot\|_\infty)$

$$\begin{aligned} n\mathcal{R}_n(\mathcal{F}) &\leq E\left[\sup_{f \in \mathcal{F}(X_1^n)} Y_f\right] \leq E\left[\sup_{f, f' \in \mathcal{F}(X_1^n)} (Y_f - Y_{f'})\right] \\ &\leq 2E\left[\sup_{d_Y(f, f') \leq \delta} (Y_f - Y_{f'})\right] + 2D\sqrt{\log N(\delta/\sqrt{n}, \mathcal{F}, \|\cdot\|_\infty)} \\ &\leq 2\delta\sqrt{n} + 4L\sqrt{n(L\sqrt{n})/\delta} \\ &\leq 2\delta\sqrt{n} + 4L^{3/2}\sqrt{n^{3/2}/\delta} \end{aligned}$$

- Set $\delta^{3/2} = CL^{3/2}n^{1/4}$, i.e. $\delta = C'Ln^{1/6}$ to get $\mathcal{R}_n = O(n^{-1/3})$