

SDS 384 11: Theoretical Statistics

Lecture 17: Uniform Law of Large Numbers- Chaining

Purnamrita Sarkar

Department of Statistics and Data Science

The University of Texas at Austin

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:

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

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

Theorem

Let X_θ be zero mean sub-Gaussian process w.r.t. a metric d_X on \mathcal{T} .

We have:

$$E \sup_{\theta \in \mathcal{T}} X_\theta \leq K \int_0^D \sqrt{\log N(\delta; \mathcal{T}, d_X)} d\delta,$$

where $D := \sup_{\gamma, \gamma' \in \mathcal{T}} d_X(\gamma, \gamma')$.

- From before: $E \sup_{\theta \in \mathcal{T}} X_{\theta} = E \sup_{\theta, \theta' \in \mathcal{T}} (X_{\theta} - X_{\theta'})$
- Recall that we first choose a δ cover \mathcal{T} and two points θ^1, θ^2 from \mathcal{T} which are δ close to θ and θ' .

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

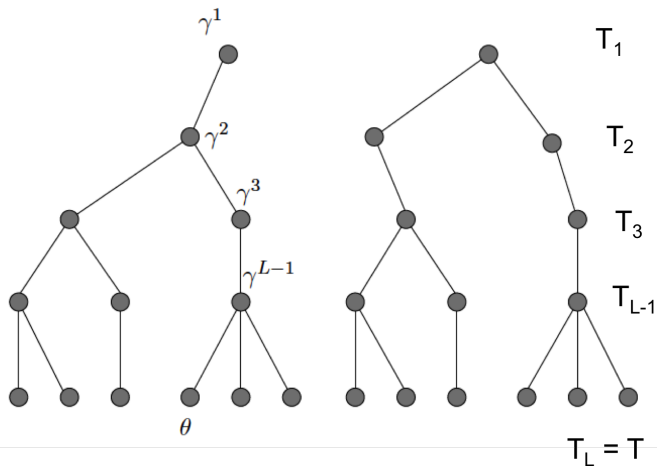
- For the expectation of the last part we used the finite class lemma.
- Now we will take a series of finer covers of smaller diameters.

- For each integer $i = 1, \dots, L$,
 - Let $\epsilon_m = D2^{-m}$
 - Form the minimal ϵ_m cover T_m of T .
 - Since $T \subseteq \mathcal{T}$, $N_m := |T_m| \leq N(\epsilon_m; \mathcal{T}, d_X)$
 - When $L = \log_2(D/\delta)$, we have $T_L = T$
 - Let

$$\pi_m(\theta) := \arg \min_{\beta \in T_m} d_X(\theta, \beta)$$

- $\pi_m(\theta)$ is the best approximation of θ from T_m
- Also, $d_X(\gamma, \pi_m(\gamma)) \leq 2^{-m}D$ when $\gamma \in T$

Picture (Courtesy: MW's book chapter 5)



- For a member θ^i of T , obtain two sequences $\{\gamma^0, \dots, \gamma^L\}$ where $\gamma^L = \theta^i$ and $\gamma^{m-1} := \pi_{m-1}(\gamma^m)$.
- Similarly form $\{\tilde{\gamma}^0, \dots, \tilde{\gamma}^L\}$ for $\theta^j \in T$.
- Note that $X_{\theta} - X_{\gamma^0} = \sum_{i=1}^L (X_{\gamma^i} - X_{\gamma^{i-1}})$

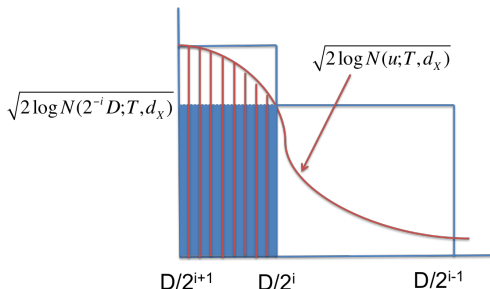
$$X_{\theta^i} - X_{\theta^j} = \sum_{i=1}^L (X_{\gamma^i} - X_{\gamma^{i-1}}) - \sum_{i=1}^L (X_{\tilde{\gamma}^i} - X_{\tilde{\gamma}^{i-1}})$$

- $E \left[\max_{\theta, \theta' \in T} X_{\theta^i} - X_{\theta^j} \right] \leq 2 \sum_{i=1}^L E \left[\max_{\gamma \in T_i} |X_{\gamma} - X_{\pi_{i-1}(\gamma)}| \right]$

Proof Cont.

- Recall $d_X(\gamma, \pi_{i-1}(\gamma)) \leq 2^{-(i-1)}D$. Now the finite class lemma gives:

$$\begin{aligned}
 E \left[\max_{\gamma \in T_i} |X_\gamma - X_{\pi_{i-1}(\gamma)}| \right] &\leq 2^{-(i-1)}D \sqrt{2 \log N(2^{-i}D, \mathcal{T}, d_X)} \\
 &\leq 42^{-(i+1)}D \sqrt{2 \log N(2^{-i}D, \mathcal{T}, d_X)} \\
 &\leq 4 \int_{2^{-(i+1)}D}^{2^{-i}D} \sqrt{2 \log N(u; \mathcal{T}, d_X)} du
 \end{aligned}$$



Done.

$$\begin{aligned} E \sup_{\theta \in \mathcal{T}} X_{\theta} &= E \sup_{\theta, \theta' \in \mathcal{T}} (X_{\theta} - X_{\theta'}) \\ &\leq 2E \left[\sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_{\theta} - X_{\theta'}) \right] + E \left[\sup_{\theta^i, \theta^j \in \mathcal{T}} (X_{\theta^i} - X_{\theta^j}) \right] \\ &\leq 2E \left[\sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_{\theta} - X_{\theta'}) \right] + 2 \sum_{i=1}^L E \left[\max_{\gamma \in T_i} |X_{\gamma} - X_{\pi_{i-1}(\gamma)}| \right] \\ &\leq 2E \left[\sup_{\substack{\theta, \theta' \in \mathcal{T} \\ d_X(\theta, \theta') \leq \delta}} (X_{\theta} - X_{\theta'}) \right] + 8 \int_{\delta/2}^D \sqrt{2 \log N(u; T, d_X)} du \end{aligned}$$

Taking $\delta = 0$ gives the desired bound.

Example

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.

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

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

- We computed this just using the discretization bound.
- It was $O(L\sqrt{d \log(nL)/n})$
- Using chaining takes the logarithmic term away.

- 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 $Y_\theta - Y_{\theta'} \sim \text{Subgaussian}(d_X(\theta, \theta'))$
- We have:

$$d_X(\theta, \theta') = \|f(\theta, X_1^n) - f(\theta', X_1^n)\| \leq \sqrt{n}L\|\theta - \theta'\|_2$$

- So it is $L\sqrt{n}$ Lipschitz.
- $D = 2L\sqrt{n}$

Example

- $N(\delta, f(\Theta, X_1^n), d_X) \leq N(\delta/(L\sqrt{n}), \Theta, \|\cdot\|_2) \leq (1 + 2L\sqrt{n}/\delta)^d$

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}) &\leq \frac{K}{n} \int_0^D \sqrt{\log N(\delta/(L\sqrt{n}), \Theta, \|\cdot\|_2)} d\delta \\ &\leq \frac{K}{n} \int_0^D \sqrt{d \log(1 + 2L\sqrt{n}/\delta)} d\delta \\ &= \frac{K}{n} \int_0^D \sqrt{d \log(1 + D/\delta)} d\delta \\ &\leq \frac{K\sqrt{D}\sqrt{d}}{n} \int_0^D \delta^{-1/2} d\delta \\ &= O\left(L\sqrt{\frac{d}{n}}\right)\end{aligned}$$

Example- VC class

Example

For a function class \mathcal{F} of $\{0, 1\}$ valued functions with VC dimension d ,

$$\mathcal{R}_{\mathcal{F}} = O\left(\sqrt{\frac{d}{n}}\right)$$

- First derive with the finite class lemma.
- Then try chaining.

Example - VC class with finite class lemma

- The finite class lemma says

$$\begin{aligned}\mathcal{R}_{\mathcal{F}} &\leq \frac{\sup_{f \in \mathcal{F}} \|f(X_1^n)\|_2 \sqrt{2 \log |\mathcal{F}|}}{n} \\ &\leq \frac{\sqrt{2 \log(ne/d)^d}}{\sqrt{n}} \\ &\leq \frac{\sqrt{2d \log(ne/d)}}{\sqrt{n}} \\ &= O\left(\sqrt{\frac{d \log(n/d)}{n}}\right)\end{aligned}$$

Example - VC class with chaining

- To use chaining we first need the covering number in terms of the VC dimension.
- First define the $\|f - g\|_{L_2(\hat{F}_n)}^2 = \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2$
- Haussler et al show that (You did something similar in your homework)

$$N(\delta; \mathcal{F}, \|\cdot\|_{L_2(P_n)}) \leq c_1 d \left(\frac{c_2}{\delta^2} \right)^d$$

- Note that the map $\langle \epsilon, f(X_1^n) \rangle / \sqrt{n}$ is subGaussian w.r.t. the $d_X = L_2(\hat{F}_n)$ norm.

Example VC class with chaining

- Using chaining we get:

$$\begin{aligned}\mathcal{R}_{\mathcal{F}} &\leq \frac{K}{\sqrt{n}} \int_0^1 \sqrt{\log N\left(\delta, \mathcal{F}, \|\cdot\|_{L_2(\hat{F}_n)}\right)} d\delta \\ &\leq \frac{c_3}{\sqrt{n}} \int_0^1 \sqrt{\log(c_1 d) + d \log(c_2/\delta^2)} d\delta \\ &\leq \frac{c_3}{\sqrt{n}} \int_0^1 \left(\sqrt{\log(c_1 d)} + \sqrt{d \log(c_2/\delta^2)} \right) d\delta \\ &= O\left(\sqrt{\frac{d}{n}}\right)\end{aligned}$$

- We have again lost the $\log(n/d)$ term.

Why use chaining?

- Recall the Glivenko Cantelli lemma?
- We have $\|\hat{F}_n - F\|_\infty \leq 2\mathcal{R}_{\mathcal{F}} + \delta$ with probability at least $1 - e^{-n\delta^2/2}$
- For the function class $\mathcal{F} := \{1(-\infty, t] : t \in \mathbb{R}\}$, we used the finite class lemma in lecture 12 to show that, $\mathcal{R}_{\mathcal{F}} = O\left(\sqrt{\frac{\log(n)}{n}}\right)$.
- But, now we can use chaining to show that, in fact,
 $\|\hat{F}_n - F\|_\infty \leq \frac{c}{\sqrt{n}} + \delta$ with probability at least $1 - e^{-n\delta^2/2}$ for some constant c . This bound is un-improvable in terms of the rate.

When does the entropy integral exist?

- Suppose \mathcal{T} has diameter D w.r.t d_X , and $\log N(\delta; \mathcal{T}, d) = O(\epsilon^{-d})$.
Then

$$\begin{aligned}\int_0^D \sqrt{\log N(\delta; \mathcal{T}, d_X)} d\delta &\leq C \int_0^D \delta^{-D/2} d\delta \\ &= O\left(\frac{D^{1-d/2}}{1-d/2}\right)\end{aligned}$$

- The integral only exists when $d = 1$.

Acknowledgement

- The slides were primarily made using Martin Wainwright's book and Peter Bartlett's lectures.