THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

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

Dudley's chaining

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

Proof

- 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 T and two points θ^1, θ^2 from T which are δ close to θ and θ'.

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

- 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.

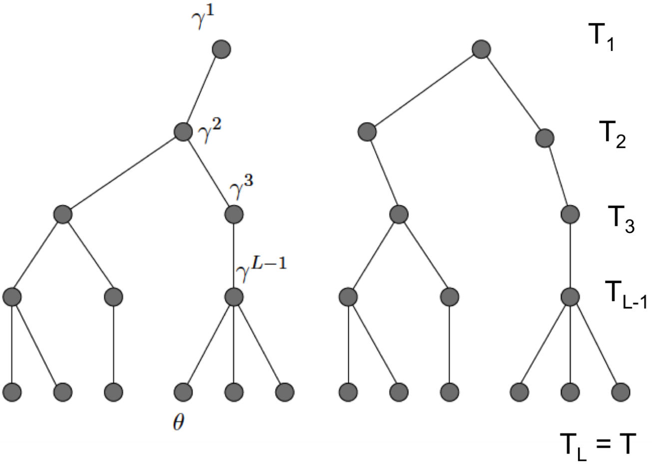
Cont.

- For each integer $i = 1, \ldots L$,
  - 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, $d_X(\gamma, \pi_m(\gamma)) \leq 2^{-m} D$

Picture (Courtesy: MW's book chapter 5)



Proof

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

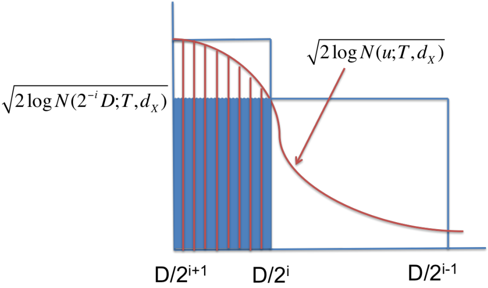
$$X_{\theta^i} - X_{\theta^j} = \sum_{i=1}^{L} (X_{\gamma^i} - X_{\gamma^{i-1}}) - \sum_{i=2}^{L} (X_{\tilde{\gamma}^i} - X_{\tilde{\gamma}^{i-1}})$$

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

Proof Cont.

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

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



Done.

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

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

Example

Example

Suppose $\mathcal{F}$ is a class parametric functions $\mathcal{F} := \{f(\theta, .) : \theta \in B_2\}$, 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. $|f(\theta, .) - f(\theta', .)| \leq L\|\theta - \theta'\|_2$.

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

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

Proof

- Denote $f(\theta, X_1^n)$ as the vector $(f(\theta, X_1), \ldots, f(\theta, X_n))$.
- 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 $f(\theta, X_1^n) \to \langle \epsilon, f(\theta, X_1^n)\rangle =: Y_\theta$ is mean zero subgaussian.
- Note that $Y_\theta - Y_\theta' \sim Subgaussian(d_X(\theta, \theta'))$
- 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.

Example

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

$$
\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &\leq \frac{K}{n} \int_0^D \sqrt{\log N(\delta/(L\sqrt{n}), \Theta, \|.\|_2)}d\delta \\
&\leq \frac{K}{n} \int_0^2 \sqrt{d \log(1 + 2L\sqrt{n}/\delta)}d\delta \\
&\leq \frac{C_1 L\sqrt{2d}}{\sqrt{n}} \int_0^2 \sqrt{\log(2/u)}du \\
&\leq \frac{C_2 L\sqrt{d}}{\sqrt{n}} \int_0^\infty v^2 e^{-v^2/2} dv \\
&= \frac{CL\sqrt{d}}{\sqrt{n}} E[Z^2] \qquad \text{where } Z \sim N(0,1) \\
&= O\left(\sqrt{\frac{d}{n}}\right)
\end{aligned}
$$

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 $\|f-g\|^2_{L_2(\hat{F}_n)} = \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-g(X_i))^2$
- Haussler et al show that (You did something similar in your homework)

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

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

Example VC class with chaining

- Using chaining we get:

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

- 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

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

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

Acknowledgement

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