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.

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

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:

$$E \sup_{\theta \in \mathcal{T}} X_\theta \leq 8\sqrt{2} \int_0^D \sqrt{\log N(\delta; \mathcal{T}, d_X)} d\delta,$$

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.

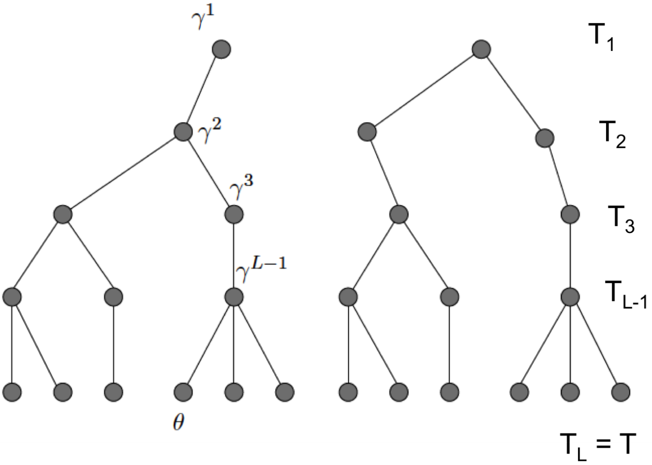
Cont.

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

Picture (Courtesy: MW's book chapter 5)



Proof

- For a member θ^i of T, obtain two sequences $\{\gamma^1, \ldots, \gamma^L\}$ where $\gamma^L = \theta^i$ and $\gamma^{m-1} := \pi_{m-1}(\gamma_m)$.
- Similarly form $\{\tilde{\gamma}^1, \ldots, \tilde{\gamma}^L\}$ for $\theta^j \in T$.
- Note that $X_\theta - X_{\gamma^1} = \sum_{i=2}^{L} (X_{\gamma^i} - X_{\gamma^{i-1}})$

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

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

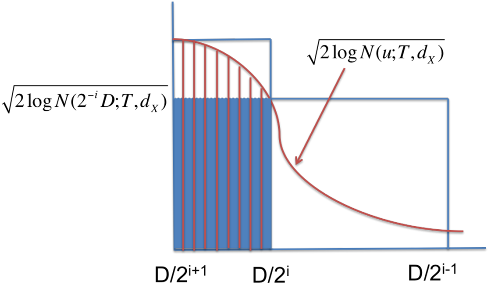
Proof Cont.

- Recall $d_X(\gamma, \pi_{i-1}(\gamma)) \leq 2^{-(i-1)}D$

$$E\left[\max_{\gamma \in T_i}(X_\gamma - X_{\pi_{i-1}(\gamma)})\right] \leq 2^{-(i-1)}D\sqrt{2\log N(2^{-i}D, \mathcal{T}, d_X)}$$

$$\leq 42^{-(i+1)}D\sqrt{2\log N(2^{-i}D, \mathcal{T}, d_X)}$$

$$\leq 4 \int_{2^{-(i+1)}D}^{2^{-i}D} \sqrt{2\log N(u; \mathcal{T}, d_X)}du$$



Done.

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

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

Example

- Recall the Rademacher complexity of the smooth parametric class?
- For $L = 1$ it was $O(\sqrt{\log n/n})$
- If you use the above integral though, you can get a sharp upper bound without the log term.