Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 11

Yun Yang

- Uniform laws of large numbers via metric entropy

Naive discretization upper bound

We start with a crude approach to bounding the supremum of a sub-Gaussian process using a covering at a single scale.

Let $D = \sup_{\theta,\, \theta' \in \mathcal{T}} \rho_X(\theta,\, \theta')$ denote the diameter of $\mathcal{T}$.

Theorem (One-step discretization bound)

Let X_θ be a zero-mean sub-Gaussian process w.r.t. the metric ρ_X on $\mathcal{T}$. Then for any $\varepsilon \in [0, D]$,

$$\mathbb{E}[\sup_{\theta, \theta' \in \mathcal{T}} (X_\theta - X_{\theta'})] \leq 2\,\mathbb{E}[\sup_{\rho_X(\theta, \theta') \leq \varepsilon} (X_\theta - X_{\theta'})] + 2D\sqrt{\log N(\varepsilon, \mathcal{T}, \rho_X)}.$$

- The above bound always implies an upper bound on $\mathbb{E}[\sup_{\theta \in \mathcal{T}} X_\theta]$ since X_θ has zero mean. In this case, the first leading factor of 2 can be removed.
- To apply this bound, choose ε to achieve the optimal trade-off between the two terms.

Proof of the discretization upper bound

For any $\varepsilon > 0$, choose a minimal ε-cover $\{\theta^1, \ldots, \theta^N\}$ with $N = N(\varepsilon, \mathcal{T}, \rho_X)$. Then for any pair $(\theta,\, \theta') \in \mathcal{T}^2$, we can always pick $1 \leq i,\, j \leq n$ such that

$$\rho_X(\theta,\, \theta^i) \leq \varepsilon \quad \text{and} \quad \rho_X(\theta',\, \theta^j) \leq \varepsilon.$$

We have

$$\begin{aligned} X_\theta - X_{\theta'} &= (X_\theta - X_{\theta^i}) + (X_{\theta^i} - X_{\theta^j}) + (X_{\theta^j} - X_{\theta'}) \\ &\leq 2 \sup_{\rho_X(\theta_1, \theta_2) \leq \varepsilon} (X_{\theta_1} - X_{\theta_2}) + \max_{i,j} (X_{\theta^i} - X_{\theta^j}). \end{aligned}$$

Since $X_{\theta^i} - X_{\theta^j}$ is sub-Gaussian with parameter at most D^2, the Finite Lemma implies

$$\mathbb{E}[\max_{i,j} (X_{\theta^i} - X_{\theta^j})] \leq \sqrt{2D^2 \log N^2} = 2D\sqrt{2 \log N}.$$

Example: Canonical Gaussian/Rademacher process

Consider the case where $\mathcal{T} \subset \mathbb{R}^d$, and the metric is $\|\cdot\|_2$. Then

$$\mathcal{G}(\mathcal{T}) \leq \min_{\varepsilon \in [0, D]} \left\{ \mathcal{G}(\widetilde{\mathcal{T}}(\varepsilon)) + 2D\sqrt{\log N(\varepsilon, \mathcal{T}, \|\cdot\|_2)} \right\},$$

$$\widetilde{\mathcal{T}}(\varepsilon) = \left\{ \theta - \theta' : \theta, \theta' \in \mathcal{T}, \|\theta - \theta'\|_2 \leq \varepsilon \right\}.$$

The quantity $\mathcal{G}(\widetilde{\mathcal{T}}(\varepsilon))$ is called a localized Gaussian complexity.

We can upper bound it by $\varepsilon\sqrt{d}$, which leads to the naive discretization bound

$$\mathcal{G}(\mathcal{T}) \leq \min_{\varepsilon \in [0, D]} \left\{ \varepsilon\sqrt{d} + 2D\sqrt{\log N(\varepsilon, \mathcal{T}, \|\cdot\|_2)} \right\}.$$

Example: Gaussian complexity of unit ball

- Consider the canonical Gaussian process with $\mathcal{T}$ the unit ball in $\mathbb{R}^d$.
- We have $D = 2$ and $\log N(\varepsilon, \mathcal{T}, \|\cdot\|_2) \leq d \log(1 + 2/\varepsilon)$.
- The previous argument leads to

$$\mathcal{G}(\mathcal{T}) \leq \min_{\varepsilon \in [0,2]} \left\{ \varepsilon \sqrt{d} + 2D \sqrt{\log N(\varepsilon, \mathcal{T}, \|\cdot\|_2)} \right\}.$$

- Choose $\varepsilon = 1/2$, we obtain

$$\mathcal{G}(\mathcal{T}) \leq \sqrt{d} \left(\frac{1}{2} + 4\sqrt{\log 5} \right).$$

- Using direct method, we proved $\mathcal{G}(\mathcal{T}) = \sqrt{d}(1 - o(1))$.

Example: Maximum singular value of sub-Gaussian random matrix

Let $W \in \mathbb{R}^{n \times d}$ be a random matrix with i.i.d. 1-sub-Gaussian entries. The ℓ_2-operator norm of W is its largest singular value, which has the variational characterization

$$\|W\|_{\text{op}} = \sup_{v \in \mathbb{S}^{d-1}} \|Wv\|_2, \quad \text{where } \mathbb{S}^{d-1} \text{ is the unit sphere in } \mathbb{R}^d.$$

Recall that we have showed the concentration of $\|W\|_{\text{op}}$ around its expectation $\mathbb{E}[\|W\|_{\text{op}}]$, when its entries are i.i.d. $\mathcal{N}(0,1)$. In this example, by viewing $\mathbb{E}[\|W\|_{\text{op}}]$ as the Gaussian complexity of certain subset of $\mathbb{R}^{n \times d}$, we will show:

Property

There is some universal constant $c > 0$ such that

$$\frac{\mathbb{E}[\|W\|_{\text{op}}]}{\sqrt{n}} \leq c\left(1 + \sqrt{\frac{d}{n}}\right).$$

Example: Empirical Gaussian complexity of parametric function class

Recall that when $\mathcal{F}$ be a parameterized class of functions

$$\mathcal{F} = \left\{ f_\theta(\cdot) :\ \theta \in \mathbb{R}^d \right\},$$

and the mapping $\theta \mapsto f_\theta(\cdot)$ is L-Lipschitz, then

$$N(\varepsilon,\, \mathcal{F}(x_1^n)/\sqrt{n},\, \|\cdot\|_2) \le N(\varepsilon,\, \mathcal{F},\, \|\cdot\|_\infty) \le d \log(L/\varepsilon).$$

Assume $\|f\|_\infty \le 1$ for each $f \in \mathcal{F}$, then

$$\mathcal{G}(\mathcal{F}(x_1^n)/n) \le \frac{1}{\sqrt{n}} \min_{\varepsilon \in [0,2]} \left\{ \varepsilon \sqrt{n} + 4\sqrt{d \log(L/\varepsilon)} \right\}.$$

Choose $\varepsilon = 1/\sqrt{n}$, we obtain

$$\mathcal{G}(\mathcal{F}(x_1^n)/n) \le c \sqrt{\frac{\log n}{n}}.$$

Example: Gaussian complexity of Lipschitz function class

For L-Lipschitz function class

$$\mathcal{F}_L = \left\{ g : [0, 1] \to \mathbb{R} \,\middle|\, g(0) = 0,\ g \text{ is } L\text{-Lipschitz} \right\}.$$

We derived its metric entropy w.r.t. the sup-norm scales as bounded by

$$\log N(\varepsilon,\ \mathcal{F}_L,\ \|\cdot\|_\infty) \asymp L/\varepsilon.$$

Therefore, we have

$$\mathcal{G}(\mathcal{F}_L(x_1^n)/n) \leq \frac{c}{\sqrt{n}} \min_{\varepsilon \in [0, 1]} \left\{ \varepsilon \sqrt{n} + \sqrt{\frac{L}{\varepsilon}} \right\}.$$

Choosing $\varepsilon = (L/n)^{1/3}$ leads to

$$\mathcal{G}(\mathcal{F}_L(x_1^n)/n) \leq c \left(\frac{L}{n} \right)^{1/3}.$$