Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 10

Yun Yang

- Uniform laws of large numbers via metric entropy

Example: Higher-order smoothness classes

For some integer α and parameter $\gamma \in (0,1]$, consider the class $\mathcal{F}_{\alpha,\gamma}$ of functions $f : [0,1] \to \mathbb{R}$ such that

$$|f^{(j)}(x)| \le C, \quad \text{for all } x \in [0,1], j = 0, 1, \ldots, \alpha, \text{ and}$$
$$|f^{(\alpha)}(x) - f^{(\alpha)}(y)| \le L\,|x-y|^{\gamma}, \quad \text{for all } x, y \in [0,1].$$

Property

The metric entropy of $\mathcal{F}_{\alpha,\gamma}$ w.r.t. the sup-norm scales as

$$\log N(\varepsilon, \mathcal{F}_L, \|\cdot\|_\infty) \asymp \big(1/\varepsilon\big)^{\frac{1}{\alpha+\gamma}}, \quad \text{as } \varepsilon \to 0.$$

More generally, we can similarly define d-dimensional class $\mathcal{F}_{\alpha,\gamma}([0,1]^d)$, and

$$\log N(\varepsilon, \mathcal{F}_L([0,1]^d), \|\cdot\|_\infty) \asymp \big(1/\varepsilon\big)^{\frac{d}{\alpha+\gamma}}, \quad \text{as } \varepsilon \to 0.$$

Example: Infinite dimensional ellipsoids in $\ell^2(\mathbb{N})$

Given a sequence of non-negative real numbers $\mu_1 \geq \mu_2 \geq \cdots$ such that $\sum_{j=1}^{\infty} \mu_j < \infty$, consider the ellipsoid

$$\mathcal{E} = \left\{ (\theta_j)_{j=1}^{\infty} \,\middle|\, \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\} \subset \ell^2(\mathbb{N}).$$

More concretely, focusing on $\mu_j = j^{-2\alpha}$ for $j = 1, 2, \ldots$ and some $\alpha > 1/2$.

Property

$$\log N(\varepsilon, \mathcal{E}, \|\cdot\|_2) \asymp \left(\frac{1}{\varepsilon}\right)^{1/\alpha} \quad \text{for sufficiently small } \varepsilon > 0.$$

Canonical Rademacher and Gaussian processes

Definition

Fix a set $\mathcal{T} \subset \mathbb{R}^n$.

1. The **canonical Gaussian process** is the stochastic process $\{G_\theta : \theta \in \mathcal{T}\}$, where

$$G_\theta = \langle g, \theta \rangle = \sum_{i=1}^{n} g_i \theta_i, \quad g_i \overset{iid}{\sim} \mathcal{N}(0, 1).$$

2. The **canonical Rademacher process** is the stochastic process $\{R_\theta : \theta \in \mathcal{T}\}$, where

$$R_\theta = \langle \varepsilon, \theta \rangle = \sum_{i=1}^{n} \varepsilon_i \theta_i, \quad g_i \overset{iid}{\sim} \text{uniform over } \{-1, +1\}.$$

Canonical Rademacher and Gaussian processes

Recall the Gaussian complexity of $\mathcal{T}$ is $\mathcal{G}(\mathcal{T}) = \mathbb{E}[\sup_{\theta\in\mathcal{T}} G_\theta]$, and the Rademacher complexity of $\mathcal{T}$ is $\mathcal{R}(\mathcal{T}) = \mathbb{E}[\sup_{\theta\in\mathcal{T}} R_\theta]$.

Properties

1. (Relation) for $\mathcal{T} \subset \mathbb{R}^d$,

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

2. (Finite Lemma) $g = (g_1, \ldots, g_d)$ has sub-Gaussian components with parameters σ^2. If $\mathcal{A} \subset \mathbb{R}^d$ has finite size, then

$$\mathbb{E}\max_{a\in\mathcal{A}}\langle g,\, a\rangle \le \sigma \max_{a\in\mathcal{A}} \|a\|_2 \sqrt{2\log|\mathcal{A}|}.$$

Proof: Left as a homework problem.

Examples: balls in $\mathbb{R}^d$

- Euclidean ball of unit norm $\mathbb{B}_2^d = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$:

$$\mathcal{R}(\mathbb{B}_2^d) = \sqrt{d}, \quad \mathcal{G}(\mathbb{B}_2^d) \leq \sqrt{d}, \quad \mathcal{G}(\mathbb{B}_2^d)/\sqrt{d} \to 1 \text{ as } d \to \infty.$$

- Unit ℓ_1-ball in d dimensions
  $\mathbb{B}_1^d = \{\theta \in \mathbb{R}^d : \|\theta\|_1 = \sum_{j=1}^d |\theta_j| \leq 1\}$:

$$\mathcal{R}(\mathbb{B}_1^d) = 1, \; \mathcal{G}(\mathbb{B}_1^d) \leq \sqrt{2\log d}, \; \mathcal{G}(\mathbb{B}_1^d)/\sqrt{2\log d} \to 1 \text{ as } d \to \infty.$$

- ℓ_0-ball in d dimensions
  $\mathbb{B}_0^d(s) = \{\theta \in \mathbb{R}^d : \|\theta\|_0 = \sum_{j=1}^d \mathbb{I}(\theta_j \neq 0) \leq s\}$. Consider the set $\mathcal{S}^d(s) = \mathbb{B}_0^d \cap \mathbb{B}_2^d$: for some universal constants $c, \; C > 0$,

$$c\sqrt{s \log \frac{e\,d}{s}} \leq \mathcal{G}(\mathcal{S}^d(s)) \leq C\sqrt{s \log \frac{e\,d}{s}}.$$

Example: Gaussian complexity of function class

For a function class $\mathcal{F}$, we have defined, for any fixed collection $x_1^n = (x_1, \ldots, x_n)$ of points, the subset of $\mathbb{R}^n$

$$\mathcal{F}(x_1^n) = \Big\{ \big(f(x_1), \ldots, f(x_n)\big) \,\Big|\, f \in \mathcal{F} \Big\}.$$

Define the Gaussian complexity of this set (rescaled by n^{-1}) as

$$\mathcal{G}\big(\mathcal{F}(x_1^n)/n\big) = \mathbb{E}_w \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n w_i \, f(x_i) \right| \right],$$

where w_i are i.i.d. $\mathcal{N}(0,1)$. Define the empirical $\mathcal{L}^2(\mathbb{P}_n)$ norm on $\mathcal{F}$ as $\|f - g\|_n = \sqrt{n^{-1} \sum_{i=1}^n \big(f(x_i) - g(x_i)\big)^2}$. Suppose all functions in $\mathcal{F}$ have $\|\cdot\|_n$ norm bounded by $b > 0$, then

$$\mathcal{G}\big(\mathcal{F}(x_1^n)/n\big) \leq b \frac{\mathbb{E}[\|w\|_2]}{\sqrt{n}} \leq b.$$

Sub-Gaussian process

Definition

A stochastic process $\theta \mapsto X_\theta$ with indexing set $\mathcal{T}$ is said to be sub-Gaussian with respect to a metric ρ_X on $\mathcal{T}$ if for all $\theta, \theta' \in \mathcal{T}$ and $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[\exp\{\lambda(X_\theta - X_{\theta'})\}\big] \leq \exp\left(\frac{\lambda^2 \rho_X^2(\theta, \theta')}{2}\right).$$

- Imposing a sub-Gaussian tail bound is an equivalent way for defining a sub-Gaussian process.
- The canonical Rademacher and Gaussian processes are sub-Gaussian w.r.t. the Euclidean metric $\|\theta - \theta'\|_2$.

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

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

Recall that we have showed the concentration of $\|W\|_{\mathsf{op}}$ around its expectation $\mathbb{E}[\|W\|_{\mathsf{op}}]$, when its entries are i.i.d. $\mathcal{N}(0,1)$. In this example, by viewing $\mathbb{E}[\|W\|_{\mathsf{op}}]$ 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

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

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

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

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

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