Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 12

Yun Yang

- Uniform laws of large numbers via metric entropy

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

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

Dudley's entropy integral

Theorem (Dudley's entropy integral bound)

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

$$\mathbb{E}[\sup_{\theta, \theta' \in \mathcal{T}} (X_\theta - X_{\theta'})] \leq 2\,\mathbb{E}[\sup_{\rho_X(\theta, \theta') \leq \varepsilon} (X_\theta - X_{\theta'})] + 8\sqrt{2}\,J(\varepsilon/2,\, D)$$

where

$$J(\varepsilon,\, D) = \int_\varepsilon^D \sqrt{\log N(t,\, \mathcal{T},\, \rho_X)}\, dt.$$

- $J(0, D) = \int_0^D \sqrt{\log N(t,\, \mathcal{T},\, \rho_X)}\, dt$ is known as the Dudley's entropy integral.
- The proof is based on the chaining method, a substantial refinement of the one step discretization method.

Proof: Dudley's entropy integral

Previously, we established

$$X_\theta - X_{\theta'} = (X_\theta - X_{\theta^i}) + (X_{\theta^i} - X_{\theta^j}) + (X_{\theta^j} - X_{\theta'})$$
$$\leq 2 \sup_{\rho_X(\gamma,\gamma')\leq\varepsilon} (X_\gamma - X_{\gamma'}) + \max_{i,j} |X_{\theta^i} - X_{\theta^j}|.$$

Now we use a more refined method to bound the second supremum over the ε-cover $\hat{T} = \{\theta^j\}_{j=1}^N$. We consider a sequence of progressively better approximations to elements of $\hat{T}$ (which leads to sets with progressively smaller diameters).

Suppose the diameter of $\hat{T}$ is D. We first define $\hat{T}_L = \hat{T}$, and think of it as a $2^{-L}D$-cover of $\hat{T}$, where $L = \lceil \log_2(D/\varepsilon) \rceil$ ensures that $2^{-LD} \leq \varepsilon$.

Then we define $\hat{T}_{m-1} =$ a minimal $2^{-(m-1)}D$-cover of $\hat{T}_m$, for m going from $L-1$ down to 0. Notice that $\hat{T}_0$ is a minimal D-cover of $\hat{T}$, so $|\hat{T}_0| = 1$. Denote $\hat{T}_0 = \{\theta_0\}$.

Proof: Dudley's entropy integral

For each $m = 0, \ldots, L-1$, define the mapping $\pi_m : \hat{T} \to \hat{T}_m$ via

$$\pi_m(\theta) = \underset{\gamma \in \hat{T}_m}{\operatorname{argmin}} \ \rho_X(\theta, \gamma),$$

so that $\pi_m(\theta)$ is the best approximation of $\theta \in \hat{T}$ from the set $\hat{T}_m$.
For each $\theta \in \hat{T}$, define the sequence $(\gamma^0, \ldots, \gamma^L)$ recursively via $\gamma^L = \theta$, $\gamma^{m-1} = \pi_{m-1}(\gamma^m)$ for $m = L, L-1, \ldots, 1$. By construction, we always have $\gamma^0 = \theta_0$, and the chaining relation:

$$X_{\gamma^L} - X_{\theta_0} = \sum_{m=2}^{L} (X_{\gamma^m} - X_{\gamma^{m-1}}).$$

Consider another $\tilde{\theta} \in \hat{T}$ with associated sequence $(\tilde{\gamma}^0, \ldots, \tilde{\gamma}^L)$.

$$X_\theta - X_{\tilde{\theta}} \leq 2 \sum_{m=1}^{L} \max_{\beta \in \hat{T}_m} \left(X_\beta - X_{\pi_{m-1}(\beta)} \right),$$

Proof: Dudley's entropy integral

$$X_\theta - X_{\tilde{\theta}} \leq 2 \sum_{m=1}^{L} \max_{\beta \in \hat{T}_m} \big(X_\beta - X_{\pi_{m-1}(\beta)}\big).$$

Since for each $\beta \in \hat{T}_m$, $X_\beta - X_{\pi_{m-1}(\beta)}$ is sub-Gaussian with parameter at most $\rho(\beta, \pi_{m-1}(\beta)) \leq 2^{-(m-1)}D$, and $\hat{T}_m$ has at most $N(2^{-m}D, \mathcal{T}, \rho_X)$ elements, the Finite Lemma implies

$$\begin{aligned}\mathbb{E}\Big[\max_{\beta \in \hat{T}_m} \big(X_\beta - X_{\pi_{m-1}(\beta)}\big)\Big] &\leq 2^{-(m-1)}D\sqrt{2\log N(2^{-m}D, \mathcal{T}, \rho_X)} \\ &\leq 4\int_{2^{-(m+1)}D}^{2^{-m}D} \sqrt{2\log N(t, \mathcal{T}, \rho_X)}\, dt.\end{aligned}$$

Putting pieces together, we obtain

$$\mathbb{E}\Big[\max_{\theta, \tilde{\theta} \in \hat{T}} (X_\theta - X_{\tilde{\theta}})\Big] \leq 8\sqrt{2}\int_{\varepsilon/2}^{D} \sqrt{2\log N(t, \mathcal{T}, \rho_X)}\, dt.$$

Example: Empirical Gaussian complexity of parametric function class

Previously, we applied the naive discretization bound to get

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

Here, we show that the Dudley entropy integral yields a sharper upper bound without the $\log n$ factor. Recall that $N(\varepsilon, \mathcal{F}(x_1^n)/\sqrt{n}, \|\cdot\|_\infty) \leq c\log(1+\varepsilon^{-1})$, implying

$$\mathcal{G}(\mathcal{F}(x_1^n)/n) \leq \frac{c'}{\sqrt{n}}\int_0^2 \sqrt{\log(1+1/t)}\,dt = \frac{c''}{\sqrt{n}}.$$

Example: Bounds for Vapnik-Chervonenkis classes

Let $\mathcal{F}$ be a b-uniformly bounded class of functions with finite VC dimension d. We are interested in controlling the random variable

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) - \mathbb{P} f \right) \right|.$$

Define the zero-mean Rademacher process

$$Z_f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(x_i).$$

Then $Z_f - Z_g$ is sub-Gaussian with parameter

$$\|f - g\|_n^2 = n^{-1} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2.$$

Example: Bounds for Vapnik-Chervonenkis classes

Therefore, Dudley's entropy integral bound implies

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\big[\sup_{f\in\mathcal{F}} |Z_f|\big] \leq \frac{8\sqrt{2}}{\sqrt{n}} \int_0^{2b} \sqrt{\log N(t,\, \mathcal{F},\, \|\cdot\|_n)}\, dt.$$

Use the known fact that

$$N(\varepsilon,\, \mathcal{F},\, \|\cdot\|_n) \leq C\, d(16e)^d \left(\frac{b}{\varepsilon}\right)^{2d},$$

we obtain

$$\mathcal{R}_n(\mathcal{F}) \leq c' \sqrt{\frac{d}{n}}, \qquad \text{and for all } \delta > 0,$$

$$\mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq c\sqrt{\frac{d}{n}} + \delta\right] \leq 2e^{-\frac{n\delta^2}{8}}.$$

Gaussian comparison inequality

Suppose that we are given a pair of Gaussian vectors $\{X_j, j = 1, \ldots, N\}$ and $\{Y_j, j = 1, \ldots, N\}$ of the same dimension. Gaussian comparison inequalities compare the two Gaussian vectors in terms of the expected value of some real-valued function F defined on $\mathbb{R}^n$.

Theorem (Sudakov-Fernique)

Given a pair of centered Gaussian vectors $\{X_j, j = 1, \ldots, N\}$ *and* $\{Y_j, j = 1, \ldots, N\}$*, suppose that*

$$\mathbb{E}(X_i - X_j)^2 \le \mathbb{E}(Y_i - Y_j)^2 \quad \textit{for all pair } (i,j) \in N^2.$$

Then $\mathbb{E}[\max_{j=1,\ldots,N} X_j] \le \mathbb{E}[\max_{j=1,\ldots,N} Y_j]$*.*

The results can be extended for comparing two Gaussian processes, by taking limits of maxima over finite subsets.

Sudakov's lower bound

The following theorem provides a lower bound on the expected supremum of Gaussian process.

Theorem (Sudakov minoration)

Let X_θ be a zero-mean Gaussian process defined on non-empty set $\mathcal{T}$. Then

$$\mathbb{E}\big[\sup_{\theta\in\mathcal{T}} X_\theta\big] \geq \sup_{\varepsilon>0} \frac{\varepsilon}{2}\sqrt{\log M(\varepsilon,\, \mathcal{T},\, \rho_X)},$$

where $\rho_X(\theta,\, \theta') = \sqrt{\mathit{Var}\,(X_\theta - X_{\theta'})}$.

Proof: For any $\varepsilon > 0$, let $\{\theta^1, \ldots, \theta^M\}$ be an ε-packing of $\mathcal{T}$. Let $Y_i = X_{\theta^i}$. Define $X_i \overset{iid}{\sim} \mathcal{N}(0, \varepsilon^2/2)$. Then

$$\mathbb{E}[(Y_i - Y_j)^2] \geq \varepsilon^2 = \mathbb{E}[(X_i - X_j)^2].$$

Therefore, $\mathbb{E}\big[\sup_{\theta\in\mathcal{T}} X_\theta\big] \geq \mathbb{E}\big[\max_i Y_i\big] \geq E\big[\max_i X_i\big] \geq \frac{\varepsilon}{2}\sqrt{\log M}$.

Example: Gaussian complexity of ℓ_2-ball

We have proved previously that

$$\mathcal{G}(\mathbb{B}_2^d) \leq \sqrt{d}.$$

Now we apply the Sudakov minoration to capture a $\mathcal{O}(\sqrt{d})$ lower bound. We proved that

$$\log N(\varepsilon, \mathbb{B}_2^d, \|\cdot\|_2) \geq d \log(1/\varepsilon).$$

Therefore, the Sudakov bound implies

$$\mathcal{G}(\mathbb{B}_2^d) \geq \sup_{\varepsilon > 0} \left\{ \frac{\varepsilon}{2} \sqrt{d \log(1/\varepsilon)} \right\} \geq \frac{\sqrt{\log 2}}{4} \sqrt{d},$$

by choosing $\varepsilon = 1/2$.

Example: Metric entropy of ℓ_1-ball

Recall that we have the Gaussian complexity upper bound

$$\mathcal{G}(\mathbb{B}_1^d) \leq \sqrt{2 \log d}.$$

Now we apply Sudakov's minoration to get an upper bound on the metric entropy,

$$\log N(\varepsilon, \mathbb{B}_1^d, \|\cdot\|_2) \leq c\,(1/\varepsilon)^2 \log d.$$

This bound is tight in ε and d, suggesting that the ℓ_1-ball is much smaller than the ℓ_2 ball when d is large.

Example: Lower bounds on maximum singular value

Recall that for a standard Gaussian random matrix $W \in \mathbb{R}^{n \times d}$, we can write

$$\mathbb{E}[\|W\|_{\mathsf{op}}] = \mathbb{E}\Big[\sup_{\Theta \in \mathbb{M}} \langle\!\langle W,\, \Theta \rangle\!\rangle\Big],$$

where $\mathbb{M} = \big\{\Theta \in \mathbb{R}^{n \times d} :\, \mathrm{Tr}(\Theta) = 1,\, \mathrm{rank}(\Theta) = 1\big\}$.
It can be shown that there exists some universal constant $c > 0$ such that

$$\log N(\varepsilon,\, \mathbb{M},\, \|\cdot\|_{\mathsf{F}}) \geq c\,(n + d)\, \log(1/\varepsilon).$$

This implies

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