36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 5: January 28

Lecturer: Siva Balakrishnan *Scribe: Zhili Feng*

5.1 Recap

1. Sub-gaussian process:
$$\mathbb{E}[e^{t(X_\theta - X_{\theta'})}] \le \exp\left(\frac{t^2\rho^2(\theta,\theta')}{2}\right)$$
For the canonical Gaussian process we discussed so far $\rho(x,y) = \|x-y\|_2$.

2. One-step discretization bound:
$$\mathbb{E}[\sup_\theta X_\theta] \le \mathbb{E}\sup_\theta(X_\theta - X_{\theta'}) \le c\mathbb{E}_{\rho(\theta,\theta')\le\delta}(X_\theta - X_{\theta'}) + \sqrt{D^2\log\mathcal{N}(\delta,\mathbb{T},\rho)}$$
where $D = \sup_{\theta,\theta'\in\mathbb{T}}\rho(\theta,\theta')$

3. Application: $\mathbb{E}\|W\|_2 \le C(\sqrt{n}+d)$, where $W \in \mathbb{R}^{n\times d}$ with W_{ij} is a zero mean, one subgaussian RV.

5.2 Apply Naive Discretization Bound to Regression

Recall the setup, we observe $(x_i, y_i)_{i=1}^n$, where $x_i \sim P_X[0,1]$,
$$y_i = f^*(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,1).$$

The goal is to estimate f^*. We restrict $\mathcal{F} = \{f : f(0) = 0, \text{supp}(f) = [0,1], L\text{-Lipschitz}\}$.

A naive estimator is $\widehat{f} = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_i (y_i - f(x_i))^2$. Using the basic inequality, we have:
$$\frac{1}{n}\sum_i(\widehat{f}(x_i) - f^*(x_i))^2 \le -\frac{2}{\sqrt{n}}\langle\epsilon, \frac{\widehat{f}-f^*}{\sqrt{n}}\rangle$$

Note that:
$$\langle\epsilon, \frac{f_1 - f^*}{\sqrt{n}}\rangle - \langle\epsilon, \frac{f_2 - f^*}{\sqrt{n}}\rangle = \langle\epsilon, \frac{\widehat{f_1} - f_2}{\sqrt{n}}\rangle \sim \mathcal{N}(0, \frac{1}{n}\sum_i(f_1(x_i) - f_x(x_1))^2)$$

which gives a natural metric

$$\rho(f_1, f_2) = \frac{1}{n}\sum_i (f_1(x_i) - f_x(x_1))^2 \triangleq \|f_1 - f_2\|_n \leq \|f_1 - f_2\|_\infty.$$

Since the natural metric is data-dependent, which is not-ideal, it suffices to cover the space with $\|\cdot\|_\infty$ for upper bounds. The naive discretization bound then gives:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\left(\mathbb{E}\sup_{\|f_1-f_2\|_n\leq\delta}\langle\epsilon, \frac{f_1-f_2}{\sqrt{n}}\rangle + \sqrt{L^2\log\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)}\right)\\
\leq&\frac{c}{\sqrt{n}}\left(\mathbb{E}\|\epsilon\|_2\|\frac{f_1-f_2}{\sqrt{n}}\|_2 + \sqrt{L^2\times L/\delta}\right)\\
\lesssim& n^{-1/3} \qquad \text{by picking } \delta^3 = L^2/3.
\end{aligned}$$

5.3 Dudley's Bound

This section gives a tighter upper bound than the naive discretization bound.

Definition 5.1 *Dudley's entropy integral*

$$\mathcal{J}(\delta) = \int_\delta^D \sqrt{\log\mathcal{N}(u, \mathbb{T}, \rho)}du$$

Lemma 5.2

$$\mathbb{E}\sup_\theta X_\theta \leq c\left(\mathbb{E}\sup_{\rho(\theta,\theta')<\delta}(X_\theta - X_{\theta'}) + \mathcal{J}(\delta)\right)$$

Under mild regularity conditions we can take $\delta \to 0$ and obtain

$$\mathbb{E}\sup_\theta X_\theta \leq c\mathcal{J}(0).$$

Remark 5.3 $\mathcal{J}(\delta) \leq \sqrt{\mathcal{N}(\delta)}(D-\delta)$ *since $\mathcal{N}$ is non-decreasing.*

Example 5.4 *We use Dudley's bound for non-parametric regression with Lipschitz functions:*

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbb{E}\sup_f\langle\epsilon, \frac{f-f^*}{\sqrt{n}}\rangle\\
\leq&\frac{1}{\sqrt{n}}\int_0^L\sqrt{L/u}du\\
\lesssim&\, n^{-1/2}.
\end{aligned}$$

Note this gives a better bound than the naive discretization.

Example 5.5 *Let $\mathcal{A}$ be a collection of sets with VC-dimension $d < \infty$. We want to bound* $\mathbb{E}\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_i \mathbb{1}_{x_i\in A} - P(A)\right|$.

Write $\mathcal{F} = \{f = \mathbb{1}_A, A \in \mathcal{A}\}$, we have:

$$
\begin{aligned}
&\mathbb{E}\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_i \mathbb{1}_{x_i\in A} - P(A)\right| \\
=&\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_i f(x_i) - \mathbb{E}f\right| \\
\le&\mathbb{E}\mathcal{R}(\mathcal{F}, x_1^n) \\
\le&\sqrt{\frac{d\log n}{n}}
\end{aligned}
$$

where $\mathcal{R}$ is the Rademacher complexity, $x_1^n = \{x_1, \ldots, x_n\}$. Here the last step follows from Massart's lemma.

If instead we use Dudley's bound, and use Hassler's bound that

$$\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|) \le Cd \times 2^d \times (1/\delta)^d,$$

we have

$$
\begin{aligned}
&\mathbb{E}\mathcal{R}(\mathcal{F}, x_1^n) \\
\le&\frac{C}{n}\int_0^C \sqrt{\log(cd \times 2^d \times \log(1/\delta))}d\delta \\
\le&\frac{C}{n}\int_0^C \sqrt{\log(d\log(1/\delta))}d\delta \\
\le&C\sqrt{\frac{d}{n}}
\end{aligned}
$$

which gets rid of the $\log d$ factor of Massart's lemma.

As an application of this we can now recover something closer to the DKW inequality for CDF functions. The DKW inequality states that:

$$P(\sup_x |\widehat{F}(x) - F(X)| \ge t) \le 2e^{-nt^2}.$$

Since this corresponds to uniform convergence over the class of left intervals (i.e. intervals of the form $(-\infty, t]$, $t \in \mathbb{R}$) which has VC dimension 1, the above result together with the Azuma-Hoeffding bound for concentration yield,

$$P(\sup_x |\widehat{F}(x) - F(X)| \ge t) \le Ce^{-cnt^2},$$

for some constants $c, C > 0$.

5.3.1 Useful Inequalities

Theorem 5.6 *Sudakov-Fernique Inequality*

Given two sequences of random variables $\{X_1, \ldots\}$ *and* $\{Y_1, \ldots\}$ *and* $F : \mathbb{R}^n \to \mathbb{R}$*. Suppose that* $\mathbb{E}(X_i - X_j)^2 \leq \mathbb{E}(Y_i - Y_j)^2$ *for all* $(i,j) \in \mathbb{N}^2$*, then*

$$\mathbb{E} \sup_i X_i \leq \mathbb{E} \sup_i Y_i.$$

Lemma 5.7 *Gaussian Contraction Inequality:*

Let $\epsilon \sim \mathcal{N}(0, I_d)$*,* $\theta \in \Theta^d$*,* $\psi = \{\psi_1, \ldots, \psi_d\}$ *where each* $\psi_i : \Theta \to \mathbb{R}$ *and* $\|\psi\| \leq 1$*, then:*

$$\mathbb{E} \sup_\theta \langle \epsilon, \theta \rangle \geq \mathbb{E} \sup_\theta \langle \epsilon, \psi(\theta) \rangle,$$

where $\psi(\theta) = \{\psi_1(\theta_1), \ldots, \psi_d(\theta_d)\}$*.*

Proof: Since ψ is a contraction, we have $\mathbb{E}(\theta_i - \theta_j)^2 \geq \mathbb{E}(\psi(\theta_i) - \psi(\theta_j))^2$ for all $i, j \in \mathbb{N}$. This allows us to use Sudakov-Fernique. ∎

Example 5.8 *Let* $\mathcal{F}^2(x_1, \ldots, x_n) = \{f^2(x_1), \ldots, f^2(x_n), f \in \mathcal{F}\}$*. We want* $\mathcal{G}(\mathcal{F}^2) \leq 2b\mathcal{G}(\mathcal{F})$ *if* $\|f\|_\infty \leq b$*.*

Let $\psi(t) = t^2/(2b)$*, then if we can show that* ψ *is contraction we obtain by the Gaussian contraction inequality that*

$$\mathcal{G}(\mathcal{F}^2) = \mathbb{E} \sup_f \langle \epsilon, f^2 \rangle = 2b\mathbb{E} \sup_f \langle \epsilon, \psi(f) \rangle \leq 2b\mathbb{E} \sup_f \langle \epsilon, f \rangle = 2b\mathcal{G}(\mathcal{F}).$$

We are left to show that ψ *is a contraction:*

$$|\psi(f_1) - \psi(f_2)| \leq |\frac{f_1^2}{2b} - \frac{f_2^2}{2b}| \leq |\frac{(f_1 + f_2)(f_1 - f_2)}{2b}| \leq |f_1 - f_2|.$$

5.3.2 Tightness of Dudley's Bound

Theorem 5.9 *Sudakov Minoration*

Let $\{X_\theta\}$ *be a Gaussian process,* $\rho(\theta_1, \theta_2) = \sqrt{\mathbb{E}(X_{\theta_1} - X_{\theta_2})^2}$ *(usually called the "intrinsic metric" of* X_θ*), then*

$$\mathbb{E} \sup_\theta X_\theta \geq \sup_{\delta > 0} \left(\frac{\delta}{2} \sqrt{\log \mathcal{M}(\delta, \mathbb{T}, \rho)} \right)$$

Proof: Let $\{\theta^1, \dots, \theta^{\mathcal{M}}\}$ be a packing wrt ρ. Since it's a packing we have $\mathbb{E}(X_\theta - X_{\theta'})^2 \geq \delta^2$. Now let $Y_{\theta^1}, \dots, Y_{\theta^{\mathcal{M}}} \sim \mathcal{N}(0, \delta^2/2)$ be i.i.d, hence $\mathbb{E}(Y_\theta - Y_{\theta'})^2 \leq \mathbb{E}(X_\theta - X_{\theta'})^2$. Now we can apply Sudakov-Fernique:

$$\mathbb{E} \sup_\theta X_\theta \geq \mathbb{E} \max_{i \in [\mathcal{M}]} X_{\theta^i} \geq \mathbb{E} \max_{i \in [\mathcal{M}]} Y_{\theta^i} = c\delta \sqrt{\log \mathcal{M}(\delta, \mathbb{T}, \rho)}$$

■