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

Lecture 4: January 23

Lecturer: Siva Balakrishnan *Scribe: Audrey Huang*

Sub-Gaussian Processes

The Gaussian processes that we have described previously are particular examples of sub-Gaussian processes.

Definition 4.1 *A random variable X with mean $\mu = \mathbb{E}[X]$ is **sub-Gaussian** if there is a positive number σ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2} \quad \forall \lambda \in \mathbb{R}.$$

Definition 4.2 *A collection of zero-mean random variables $\{X_\theta, \theta \in \mathbb{T}\}$ is a **sub-Gaussian process** with respect to a metric ρ on $\mathbb{T}$ if*

$$\mathbb{E}[e^{\lambda(X_\theta - X'_\theta)}] \leq e^{\lambda^2\rho^2(\theta,\theta')/2} \quad \forall \theta, \theta' \in \mathbb{T}, \lambda \in \mathbb{R}.$$

In the case of Rademacher and Gaussian complexities, the metric is a norm $\rho(\theta, \theta') = \Theta(\|\theta - \theta'\|)$.

Let $X_\theta = \langle \theta, \epsilon \rangle$ where $\theta \in \mathbb{T}$ and $\epsilon \sim N(0, I_d)$. Then $X_\theta - X_{\theta'} = \langle \theta - \theta', \epsilon \rangle \sim N(0, \|\theta - \theta'\|_2^2)$. From the zero-mean assumption, we have that

$$\begin{aligned}\mathbb{E}[\sup_{\theta\in\mathbb{T}} X_\theta] &= \mathbb{E}[\sup_{\theta\in\mathbb{T}}(X_\theta - X_{\theta_0})] \\ &\leq \mathbb{E}[\sup_{\theta,\theta'\in\mathbb{T}} (X_\theta - X_{\theta'})].\end{aligned}$$

If we upper bound the last term, we can then upper bound the first term $\mathbb{E}[\sup_{\theta\in\mathbb{T}} X_\theta]$, which is the Gaussian complexity.

When $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-Gaussian process, we can then use the **1-step discretization method** to upper bound $\mathbb{E}[\sup_{\theta\in\mathbb{T}}(X_\theta - X_{\theta_0})]$. The discretization method uses the metric entropy of the set and is much easier to use than direct methods to bound the Gaussian complexity, while achieving the same rate of decay.

1-Step Discretization Method

The general outline of forming the naive discretization bound is as follows. We construct a δ-covering of the set $\mathbb{T}$, which allows us to replace the supremum over $\mathbb{T}$ by a finite maximum over the δ-cover. We must also then add an approximation error that scales with δ, which forms the upper bound.

The bound obtained is

$$\mathbb{E}[\sup_{\theta,\theta' \in \mathbb{T}} (X_\theta - X_{\theta'})] \leq 2\mathbb{E} \sup_{\theta,\theta \in \mathbb{T}, \|\theta - \theta\| \leq \delta} (X_\theta - X_{\theta'}) + 2\sqrt{D^2 \log N(\delta; \mathbb{T}, \rho)},$$

where D is the diameter, given by $D = \sup_{\theta,\theta' \in \mathbb{T}} \rho(\theta, \theta')$.

Gaussian Complexity of the Unit Ball

Recall from previous lectures that the metric entropy of a unit ball is

$$d \log(\frac{1}{\delta}) \leq \log N(\delta; \mathbb{B}, \|\cdot\|) \leq d \log(1 + \frac{2}{\delta}).$$

We can use the discretization bound and what we know about the covering number already to bound the Gaussian complexity of the L2 unit ball. Let $B_2(1) = \{\theta : \|\theta\|_2 \leq 1\}$. Then

$$\begin{aligned}
\mathcal{G}(\mathbb{B}_2(1)) &= \mathbb{E}[\sup_{\theta \in \mathbb{B}_2(1)} X_\theta] \\
&\leq \mathbb{E}[\sup_{\theta,\theta' \in \mathbb{B}_2(1)} (X_\theta - X_{\theta'})] \\
&\leq 2\mathbb{E} \sup_{\theta,\theta \in \mathbb{B}_2(1), \|\theta - \theta]\|_2 \leq \delta} \langle \theta - \theta', \epsilon \rangle + 4\sqrt{4d \log(1 + \frac{2}{\delta})}
\end{aligned}$$

Using Cauchy-Schwarz, $\mathbb{E} \sup_{\theta,\theta \in \mathbb{B}_2(1), \|\theta - \theta]\|_2 \leq \delta} \langle \theta - \theta', \epsilon \rangle \leq \delta \mathbb{E}\|w\|_2 \leq \delta\sqrt{d}$. Plugging this in,

$$\leq 2\sqrt{d}\delta + 8\sqrt{d \log(1 + \frac{2}{\delta})}$$

If we choose $\delta = \frac{1}{2}$

$$\begin{aligned}
&= \sqrt{d}(\frac{1}{2} + 2\sqrt{2 \log 5}) \\
&\lesssim C\sqrt{d}
\end{aligned}$$

In the previous lecture we used direct methods to prove that $\mathcal{G}(\mathbb{B}_2(1)) = \sqrt{d}(1 - o(1))$. Using the discretization bound achieves the same $\sqrt{d}$ scaling, but with worse control of the constant prefactor.

Expected Operator Norm of a Random Matrix

Let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be a random 1-sub-Gaussian matrix with i.i.d entries $W_{ij} \sim N(0,1)$. The expected operator norm of this matrix is then

$$\mathbb{E}\|W\|_{op} = \mathbb{E} \sup_{u,v,\|u\|\|v\|=1} u^T W v = \mathbb{E} \sup_{\|u\|_2=1} \|Wu\|_2$$

Note that

$$\mathbb{E} \sup_u \|Wu\|_2 = \mathbb{E} \sup_{u,v} u^T W v = \mathbb{E} \sup_{u,v} \langle W, u^T v\rangle = \mathbb{E} \sup_{u,v} tr(W^T(uv^T))$$

Then let $\{M : M = uv^T, \|u\|_2 = \|v\|_2 = 1\}$ which is the same as $\{M : rank(M) = 1, \|M\|_F = 1\}$. We can rewrite $\mathbb{E}\|W\|_{op}$ as

$$\mathbb{E}\|W\|_{op} = \mathbb{E} \sup_{M\in\mathcal{M}} \langle W, M\rangle \leq \mathbb{E} \sup_{M,M'\in\mathcal{M}} \langle W, M - M'\rangle.$$

Let us denote by $\|M\|_*$ the nuclear norm of M (i.e. the sum of its singular values).

Note that $\langle W, M - M'\rangle \sim N(0, \|M - M'\|_F^2)$ and $\rho(M, M') = \|M - M'\|_F$. Using the 1-step discretization bound,

$$\begin{aligned}
\mathbb{E} \sup_{M,M'\in\mathcal{M}} \langle W, M - M'\rangle &\leq 2\mathbb{E} \sup_{M,M'\in\mathcal{M}, \|M-M'\|_F\leq\delta} \langle W, M - M'\rangle + C\sqrt{\log N(\delta; M\| \cdot \|_F)} \\
&\leq 2\mathbb{E}\|W\|_{op}\|M - M'\|_* + C\sqrt{\log N(\delta; M\| \cdot \|_F)} \\
&\leq 2\mathbb{E}\|W\|_{op}\sqrt{rank(M - M')}\|M - M'\|_F + C\sqrt{\log N(\delta; M\| \cdot \|_F)} \\
&\leq 2\sqrt{2}\mathbb{E}\|W\|_{op}\delta + C\sqrt{\log N(\delta; M\| \cdot \|_F)} \\
&\leq 2\sqrt{2}\mathbb{E}\|W\|_{op}\delta + C\sqrt{(n + d)\log(1 + \frac{2}{\delta})}.
\end{aligned}$$

If we set δ to get a tight bound and absorb constants, for a constant $c > 0$ we get

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

Lipschitz Functions

The class of Lipschitz function is given by

$$\mathcal{F}_L = \{g : [0, 1] \to \mathbb{R} | g(0) = 0, |g(x) - g(x')| \leq L|x - x'| \forall x, x' \in [0, 1]\}$$

Suppose we have a dataset of $(X_1, Y_1), ..., (X_n, Y_n)$ where $X_i \sim U[0,1]$ and $Y_i = f(X_i) + \epsilon_i$ with $\epsilon_i \sim N(0,1)$. We construct an estimate $\hat{f} = \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$. In this nonparametric setting we will instead consider $\mathcal{F}_L(X_1^n)$, and the metric is $\|\cdot\|_n$ the empirical L_2 distance $\|f - f'\|_n = \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} (f(x_i) - f'(x_i))^2}$. We want to bound the Gaussian complexity of $\mathcal{F}_L(X_1^n)$.

Since $\|f - f'\|_n \leq \|f - f'\|_\infty$, $\log N(\delta; \mathcal{F}_L(X_1^n), \|\cdot\|_n) \leq \log N(\delta; \mathcal{F}_L, \|\cdot\|_\infty)$. We can then upper bound using the discretization bound with the L_∞ metric,

$$\mathcal{G}(\frac{\mathcal{F}_L(X_1^n)}{n}) \leq \frac{1}{\sqrt{n}} [\text{naive } \delta \text{ bound} + C\sqrt{D^2 \log N(\delta; \mathcal{F}_L, \|\cdot\|_\infty)}].$$

Recall for suitably small $\delta > 0$ we know that the metric entropy bound on this class of functions is $\log N_\infty(\delta; \mathcal{F}_L) \asymp \frac{L}{\delta}, \forall \delta < \delta_0$. We can then plug this into the second term. Further, since the functions in $\mathcal{F}_L$ are uniformly bounded by 1, and the Cauchy-Schwarz inequality gives a naive δ bound of $\leq \delta\sqrt{n}$,

$$\mathcal{G}(\frac{\mathcal{F}_L(X_1^n)}{n}) \leq \frac{1}{\sqrt{n}} (\delta\sqrt{n} + C\sqrt{\frac{L}{\delta}}).$$

Choosing $\delta = n^{-1/3}$ to obtain the tightest upper bound,

$$\mathcal{G}(\frac{\mathcal{F}_L(X_1^n)}{n}) \lesssim n^{-1/3}$$

which is much slower than the parametric case. We will see in the next lecture that we can however sharpen this bound using a sharper bound on the metric entropy obtained using Dudley's chaining.