IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 16 & 17
Apr. 7 & 9, 2020

Goals: We will study suprema of stochastic processes with a certain metric structure. We will develop a single-scale covering argument and then improve it through a chaining technique.

1. SUPREMA OF GAUSSIAN AND SUBGAUSSIAN PROCESSES

Definition: Stochastic process $(U_\theta)_{\theta\in\Theta}$, indexed by $\theta \in \Theta$, is a collection of random variables on a common probability space.

The index θ can be "time," but we will be primarily interested in cases where Θ has some metric structure.

We will be interested in the behavior of the supremum of the stochastic process, and in particular

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta.$$

To understand this object, we need to have a sense of the dependence structure of U_θ and $U_{\theta'}$ for a pair of parameters, but also about the metric structure of Θ.

Gaussian process is a collection of random variables such that any finite collection $U_{\theta_1}, \dots, U_{\theta_n}$, for any $n \geq 1$, is zero-mean and jointly Gaussian. In this case

$$\mathbb{E}\exp\{\lambda(U_\theta - U_{\theta'})\} = \exp\{\lambda^2 d(\theta,\theta')^2/2\}$$

with $d(\theta,\theta')^2 = \mathbb{E}(U_\theta - U'_\theta)^2$. Hence, there is a natural metric for Gaussian process.

1.1 SubGaussian Processes

Definition: Stochastic process $(U_\theta)_{\theta\in\Theta}$ is sub-Gaussian with respect to a metric d on Θ if U_θ is zero-mean and

$$\forall \theta,\theta' \in \Theta, \lambda \in \mathbb{R}, \qquad \mathbb{E}\exp\{\lambda(U_\theta - U_{\theta'})\} \leq \exp\{\lambda^2 d(\theta,\theta')^2/2\}$$

The main examples we will be studying have a particular linearly parametrized form:

Gaussian process: Let $G_\theta = \langle g, \theta\rangle$, $g = (g_1, \dots, g_n)$, $g_i \sim N(0,1)$ i.i.d. Take $d(\theta,\theta') = \|\theta - \theta'\|$. Then

$$G_\theta - G'_\theta = \langle g, \theta - \theta'\rangle \sim N(0, \left\|\theta - \theta'\right\|^2)$$

In particular, this Gaussian process is also, trivially, sub-Gaussian with respect to the Euclidean distance on Θ.

Rademacher process: Let $R_\theta = \langle \epsilon, \theta \rangle$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, ϵ i.i.d. Rademacher. Again, take $d(\theta, \theta') = \|\theta - \theta'\|$. Then

$$R_\theta - R'_\theta = \langle \epsilon, \theta - \theta' \rangle$$

is subGaussian with parameter $\|\theta - \theta'\|^2$.

Note that in this linear parametrization of U_θ, the expected supremum can be seen as a kind of average 'width' of the set Θ.

Definition: We will call $\widehat{\mathcal{R}}(\Theta) = \mathbb{E} \sup_{\theta \in \Theta} \langle \epsilon, \theta \rangle$ the (empirical) Rademacher averages of Θ. The corresponding expected supremum of the Gaussian process will be called the Gaussian averages or the Gaussian width of Θ and denoted by $\widehat{\mathcal{G}}(\Theta)$.

1.1.1 A few examples

Let $U_\theta = \langle \epsilon, \theta \rangle$, $\Theta \subset \mathbb{R}^n$, and take Euclidean distance as the metric. We have

$$\widehat{\mathcal{R}}(\mathsf{B}_\infty^n) = \mathbb{E} \sup_{\theta \in \mathsf{B}_\infty^n} U_\theta = \mathbb{E} \sup_{\theta \in \mathsf{B}_\infty^n} \langle \epsilon, \theta \rangle = n.$$

To get a sublinear growth in n, we have to make sure Θ is significantly smaller than B_∞^n.

A few other sets:

$$\widehat{\mathcal{R}}(\mathsf{B}_2^n) = \mathbb{E} \sup_{\theta \in \mathsf{B}_2^n} \langle \epsilon, \theta \rangle = \mathbb{E} \|\epsilon\|_2 = \sqrt{n}$$

and

$$\widehat{\mathcal{G}}(\mathsf{B}_2^n) \leq \sqrt{n}.$$

However, we observe that

$$\widehat{\mathcal{R}}(\mathsf{B}_1^n) = \mathbb{E} \sup_{\theta \in \mathsf{B}_1^n} \langle \epsilon, \theta \rangle = \mathbb{E} \|\epsilon\|_\infty = 1.$$

and yet for the Gaussian process,

$$\widehat{\mathcal{G}}(\mathsf{B}_1^n) = \mathbb{E} \sup_{\theta \in \mathsf{B}_1^n} \langle g, \theta \rangle = \mathbb{E} \max_{i \in [n]} |g_i| \leq \sqrt{2 \log(2n)}.$$

In fact, this discrepancy between the Rademacher and Gaussian averages for B_1^n is the worst that can happen and for any Θ

$$\widehat{\mathcal{R}}(\Theta) \lesssim \widehat{\mathcal{G}}(\Theta) \lesssim \sqrt{\log n} \cdot \widehat{\mathcal{R}}(\Theta). \tag{1.1}$$

Furthermore, the discrepancy is only there because B_1^n has a small ℓ_1 diameter, and for many of the applications in statistics, we will work with a function class that will not have such a small ℓ_1 diameter.

For a singleton,

$$\widehat{\mathcal{R}}(\{\theta\}) = 0$$

while for the vector $\mathbf{1}_n = (1, \ldots, 1)$,

$$\widehat{\mathcal{R}}(\{-\mathbf{1}_n, \mathbf{1}_n\}) = \mathbb{E} \max\{\langle \epsilon, \mathbf{1}_n \rangle, -\langle \epsilon, \mathbf{1}_n \rangle\} = \mathbb{E} \left| \sum_{i=1}^n \epsilon_i \right| \leq \sqrt{n}.$$

Some further properties of both Rademacher and Gaussian averages:

$$\widehat{\mathcal{R}}(\Theta) \lesssim \operatorname{diam}(\Theta)\sqrt{\log \operatorname{card}(\Theta)},$$

$$\widehat{\mathcal{R}}(\operatorname{conv}(\Theta)) = \widehat{\mathcal{R}}(\Theta),$$

$$\widehat{\mathcal{R}}(c\Theta) = |c|\widehat{\mathcal{R}}(\Theta) \quad \text{for constant } \ c$$

1.2 Finite-class lemma and a single-scale covering argument

Lemma: Let d be a metric on Θ and assume (U_θ) is a subGaussian process. Then for any finite subset $A \subseteq \Theta \times \Theta$,

$$\mathbb{E} \max_{(\theta,\theta')\in A} U_\theta - U_{\theta'} \leq \max_{(\theta,\theta')\in A} d(\theta,\theta') \cdot \sqrt{2\log \operatorname{card}(A)} \tag{1.2}$$

How do we go beyond finite cover?

Definition: Let (Θ, d) be a metric space. A set $\theta_1, \ldots, \theta_N \in \Theta$ is a (proper) cover of Θ at scale ϵ if for any θ there exists $j \in [N]$ such that $d(\theta, \theta_j) \leq \epsilon$. The covering number of Θ at scale ϵ is the size of the smallest cover, denoted by $\mathcal{N}(\Theta, d, \epsilon)$.

As a simple consequence,

Lemma: If $(U_\theta)_{\theta\in\Theta}$ is subGaussian with respect to d on Θ, then for any $\delta > 0$,

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta \leq 2\mathbb{E} \sup_{d(\theta,\theta')\leq\delta} (U_\theta - U_{\theta'}) + 2\operatorname{diam}(\Theta)\sqrt{\log \mathcal{N}(\Theta, d, \delta)}$$

Proof. Observe that

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta = \mathbb{E}\sup_{\theta\in\Theta} U_\theta - U_{\theta'} \leq \mathbb{E} \sup_{\theta,\theta'\in\Theta} U_\theta - U_{\theta'}$$

Let $\widehat{\Theta}$ be a δ-cover of Θ. Then

$$\begin{aligned} U_\theta - U_{\theta'} &= U_\theta - U_{\hat{\theta}} + U_{\hat{\theta}} - U_{\hat{\theta}'} + U_{\hat{\theta}'} - U_{\theta'} && (1.3)\\ &\leq 2 \sup_{d(\theta,\theta')\leq\delta} (U_\theta - U_{\theta'}) + \sup_{\hat{\theta},\hat{\theta}'\in\widehat{\Theta}} (U_{\hat{\theta}} - U_{\hat{\theta}'}) && (1.4)\end{aligned}$$

The last term is

$$\mathbb{E} \sup_{\hat{\theta},\hat{\theta}'\in\widehat{\Theta}} U_{\hat{\theta}} - U_{\hat{\theta}'} \leq \operatorname{diam}(\Theta)\sqrt{2\log(\operatorname{card}(\widehat{\Theta})^2)}$$

□

1.3 Example: Rademacher/Gaussian processes

Let $U_\theta = \langle g, \theta \rangle$ or $\langle \epsilon, \theta \rangle$, $\Theta \subset \mathbb{R}^n$, and take Euclidean distance as the metric. Then

$$\mathbb{E} \sup_{d(\theta,\theta') \le \delta} U_\theta - U_{\theta'} \le \mathbb{E} \sup_{\|\theta\| \le \delta} \langle g, \theta \rangle \le \delta \mathbb{E} \|g\| \le \delta \sqrt{n}$$

Hence,

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \le 2\delta\sqrt{n} + 2\mathrm{diam}(\Theta) \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|_2, \delta)} \tag{1.5}$$

Roughly speaking, the supremum over Θ can be upper bounded by the supremum within a ball of radius δ ("local complexity") and the maximum over a finite collection of centers of δ-balls. We will see this decomposition/idea again within the context of optimal estimators with general (possibly nonparametric) classes of functions.

Let's step back and ask what kind of generic statement we can say about a d-dimensional subset of a Euclidean ball. Suppose that $\Theta \subseteq \mathsf{B}_2^n$ and assume that Θ lives in a d-dimensional subspace. Then

$$\mathcal{N}(\Theta, \|\cdot\|_2, \delta) \le \left(1 + \frac{2}{\delta}\right)^d$$

and by taking $\delta = \sqrt{d/n}$ the estimate in (1.5) becomes

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \le 2\sqrt{d} + 4\sqrt{d \log\left(1 + 2\sqrt{n/d}\right)} \lesssim \sqrt{d \log(n/d)}. \tag{1.6}$$

Here we tacitly assumed $d < n$. Recall that in Lecture 5 we obtained an upper bound of $O(\sqrt{d})$ in this setup by having a cover at scale $1/2$ and comparing the supremum to the maximum *multiplicatively*. Another way to see it is

$$\mathbb{E} \sup_{\theta \in \mathsf{B}_2^d} \langle \epsilon, \theta \rangle = \mathbb{E} \|\epsilon\| = \sqrt{d}$$

and similarly

$$\mathbb{E} \sup_{\theta \in \mathsf{B}_2^d} \langle g, \theta \rangle = \mathbb{E} \|g\| \le \sqrt{\sum_{i=1}^n \mathbb{E} g_i^2} \le \sqrt{d}$$

Hence, we lost a logarithmic factor by appealing to the general machinery of the previous section. We will also see that we can remove the extraneous logarithm by looking at a cover at multiple scales.

1.4 Function class

In particular, we will be interested in the following indexing set Θ. Let $x_1, \ldots, x_n$ be fixed, and let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$. We call

$$\Theta = \frac{1}{\sqrt{n}} \mathcal{F}|_{x_1,\ldots,x_n} = \left\{ \frac{1}{\sqrt{n}} (f(x_1), \ldots, f(x_n)) : f \in \mathcal{F} \right\} \subseteq \mathbb{R}^n$$

a (scaled by $1/\sqrt{n}$) *projection* of $\mathcal{F}$ onto $x_1, \ldots, x_n$. Take

$$d(\theta, \theta')^2 = \left\|\theta - \theta'\right\|^2 = \left\|f - f'\right\|_n^2 = \frac{1}{n} \sum_{i=1}^n (f(x_i) - f'(x_i))^2$$

where $\theta = (f(x_1), \ldots, f(x_n))$ and $\theta' = (f'(x_1), \ldots, f'(x_n))$, $f, f' \in \mathcal{F}$. With these definitions, we can define a Gaussian or Rademacher process with respect to Θ and d.

Important point: the symmetrization lemma allows us to relate supremum of the empirical process to supremum of a Rademacher process.

1.4.1 Example: Linear Function Class

We now focus on a specific example of linear functions

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle : w \in \mathsf{B}_2^d\}.$$

Then for fixed $x_1, \ldots, x_n \in \mathsf{B}_2^d$, a direct calculation yields

$$\mathbb{E} \sup_{w \in \mathsf{B}_2^d} \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i \langle w, x_i \rangle = \frac{1}{\sqrt{n}} \mathbb{E} \left\| \sum_{i=1}^n \epsilon_i x_i \right\| \leq \frac{1}{\sqrt{n}} \sqrt{\mathbb{E} \left\| \sum_{i=1}^n \epsilon_i x_i \right\|^2} \leq 1. \tag{1.7}$$

Let's see if we can recover this via our machinery. After all, the above object is precisely a supremum of a subgaussian process. Observe that

$$\Theta = \frac{1}{\sqrt{n}} \mathcal{F}|_{x_1, \ldots, x_n} \subseteq \frac{1}{\sqrt{n}} \mathsf{B}_\infty^n \subset \mathsf{B}_2^n \tag{1.8}$$

and that

$$\mathcal{F}|_{x_1, \ldots, x_n} = \left\{ (\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle) \in \mathbb{R}^n : w \in \mathsf{B}_2^d \right\} = \{Xw : w \in \mathsf{B}_2^d\}$$

is a subset of a d-dimensional subspace. Hence, appealing to the previous example (1.6), we get an upper bound of $O(\sqrt{d \log(n/d)})$.

Looking back at (1.7), however, we see that we also gained an extra $\sqrt{d}$ factor, which can be a big loss in high-dimensional situations. Where did we gain it? We can see that the set $\frac{1}{\sqrt{n}} \mathcal{F}|_{x_1, \ldots, x_n}$ in (1.8) is, in fact, much smaller than a d-dimensional Euclidean ball.

2. CHAINING

Theorem: Let $(U_\theta)_{\theta \in \Theta}$ be a (mean-zero) subGaussian stochastic process with respect to a metric d. Let $D = \operatorname{diam}(\Theta)$. Then for any $\delta \in [0, D]$,

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 2\mathbb{E} \sup_{d(\theta, \theta') \leq \delta} (U_\theta - U'_\theta) + 8\sqrt{2} \int_{\delta/4}^{D/2} \sqrt{\log \mathcal{N}(\Theta, d, \varepsilon)} d\varepsilon \tag{2.9}$$

Proof. Let Θ_j be a cover of Θ at scale $2^{-j} D$. We have $\operatorname{card}(\Theta_0) = 1$. Let

$$N = \min \left\{ j : 2^{-j} D \leq \delta \right\}$$

(which means $2^{-N} D \leq \delta \leq 2^{-(N-1)} D$) and $\operatorname{card}(\Theta_N) = \mathcal{N}(\Theta, d, 2^{-N} D) \geq \mathcal{N}(\Theta, d, \delta)$. As before, we start with a single (finest-scale) cover:

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 2\mathbb{E} \sup_{d(\theta, \theta') \leq \delta} (U_\theta - U_{\theta'}) + \mathbb{E} \sup_{\theta_N, \theta'_N \in \Theta_N} (U_{\theta_N} - U_{\theta'_N}).$$

For $\theta_N \in \Theta_N$,

$$U_{\theta_N} = \sum_{i=1}^{N} U_{\theta_i} - U_{\pi_{i-1}(\theta_i)} + U_{\theta_0} \tag{2.10}$$

where, recursively, we define $\theta_{i-1} = \pi_{i-1}(\theta_i)$ to be the element of Θ_{i-1} closest to θ_i. The sequence $\theta_0, \theta_1, \ldots, \theta_N$ is a "chain" linking an element of the covering to the corresponding closest element at the coarser scale.

Let the corresponding chain for $\theta'_N \in \Theta_N$ be denoted by $\theta'_0, \theta'_1, \ldots, \theta'_N$. Then

$$U_{\theta_N} - U_{\theta'_N} = \left(\sum_{i=1}^{N} U_{\theta_i} - U_{\pi_{i-1}(\theta_i)} \right) - \left(\sum_{i=1}^{N} U_{\theta'_i} - U_{\pi_{i-1}(\theta'_i)} \right)$$

and

$$\mathbb{E} \max_{\theta,\theta' \in \Theta_N} U_\theta - U_{\theta'} \leq \sum_{i=1}^{N} \mathbb{E} \max_{\theta_i \in \Theta_i} (U_{\theta_i} - U_{\pi_{i-1}(\theta_i)}) + \sum_{i=1}^{N} \mathbb{E} \max_{\theta'_i \in \Theta_i} (U_{\pi_{i-1}(\theta'_i)} - U_{\theta'_i}) \tag{2.11}$$

$$\leq 2 \sum_{i=1}^{N} D 2^{-(i-1)} \sqrt{2 \log \mathcal{N}(\Theta, d, 2^{-i} D)} \tag{2.12}$$

$$= 8 \sum_{i=1}^{N} D 2^{-(i+1)} \sqrt{2 \log \mathcal{N}(\Theta, d, 2^{-i} D)} \tag{2.13}$$

$$\leq 8 \sum_{i=1}^{N} \int_{2^{-(i+1)} D}^{2^{-i} D} \sqrt{2 \log \mathcal{N}(\Theta, d, \varepsilon)} d\varepsilon \tag{2.14}$$

Observe that $2^{-(N+1)} D \geq \delta/4$, which concludes the proof. □

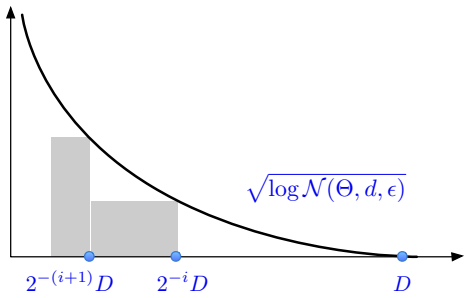

Figure 1: Illustration of the Dudley integral upper bound

Sudakov's theorem gives a single-scale lower bound:

Theorem: For a Gaussian process $(U_\theta)_{\theta \in \Theta}$,

$$C \sup_{\alpha \geq 0} \alpha \sqrt{\log \mathcal{N}(\Theta, d, \alpha)} \leq \mathbb{E} \sup_{\theta \in \Theta} U_\theta$$

for some constant C.

We can interpret this lower bound as the largest rectangle under the curve in Figure 1. This lower bound can be tight in the applications we consider (whenever the sum of the areas of rectangles Figure 1 is of the same order as the largest one).

Summary: We now have tools to analyze suprema of subGaussian processes in terms of the geometric descriptions (e.g. covering numbers) of the indexing set. These techniques will be applied to a number of parametric and nonparametric regression and classification problems in the subsequent lectures, after we introduce a few more tools such as combinatorial parameters.