Stats 300b: Theory of Statistics **Winter 2019**

Lecture 11 – February 12

Lecturer: John Duchi *Scribe: Souvik Ray*

Warning: *these notes may contain factual errors*

Reading: VdV ch. 19

Outline:

- Sub-Gaussian processes.
- Entropy numbers.
- Classes with small/finite entropy numbers (non-parametric but smooth classes, VC classes).

1 Sub-Gaussian Processes

Let $\{X_t\}_{t\in T}$ be a collection of real valued random variables.
Remark As usual, all processes we deal with in this class will be separable, i.e. there exists a countable set $T' \subset T$ such that $\{X_t\}_{t\in T}$ is determined by $\{X_t\}_{t\in T'}$, i.e.,

$$\sup_{t,s\in T} |X_t - X_s| = \sup_{t,s\in T'} |X_t - X_s|.$$

Definition 1.1. *Let (T, d) be a metric space. We say $\{X_t\}_{t\in T}$ is a **Sub-Gaussian process** if*

$$\mathbb{E}\left[\exp\left(\lambda(X_s - X_t)\right)\right] \leq \exp\left(\frac{1}{2}\lambda^2 d(s,t)^2\right), \tag{1}$$

for all $\lambda \in \mathbb{R}; s, t \in T$.

Remark We say $\{X_t\}_{t\in T}$ is a σ^2-**Sub-Gaussian process** if

$$\mathbb{E}\left[\exp\left(\lambda(X_s - X_t)\right)\right] \leq \exp\left(\frac{1}{2}\lambda^2\sigma^2 d(s,t)^2\right), \tag{2}$$

for all $\lambda \in \mathbb{R}; s, t \in T$. However, the metric d can be chosen so that the sub-gaussian constant is absorbed into the metric d.

Example 1: (Gaussian process)
A Gaussian process on $\mathbb{R}^d$ is an example of a Sub-Gaussian process. To see this, let $T = \mathbb{R}^d$, and $Z \sim \mathcal{N}(0, \sigma^2 I_d)$. Then define $X_t = \langle Z, t\rangle$. Note that

$$X_s - X_t = \langle Z, s - t\rangle \sim N(0, \sigma^2||t - s||_2^2),$$

and therefore,

$$\mathbb{E}[e^{\lambda(X_s - X_t)}] \leq \exp\left(\frac{1}{2}\lambda^2\sigma^2||s - t||_2^2\right).$$

So this is σ^2-sub-Gaussian w.r.t. L_2 norm on $\mathbb{R}^d$. ♣

Example 2: (Rademacher Process)
Let $T \subset \mathbb{V}$, where $\mathbb{V}$ is a vector space equipped with a norm $||\cdot||$. Let $\ell : T \times \mathcal{X} \to \mathbb{R}$ is 1-Lipschitz in its first argument, meaning that

$$|\ell(t,x) - \ell(s,x)| \leq ||t-s|| \text{ for all } x \in \mathcal{X}; s,t \in T.$$

Let $\{\epsilon_i\}_{i=1}^n \overset{i.i.d.}{\sim} \text{Uniform}(\{\pm 1\})$. Fix $x_1, \ldots, x_n \in \mathcal{X}$. Consider the process

$$Z_t := \sum_{i=1}^n \epsilon_i \ell(t, x_i).$$

Note that for all $t, s \in T$, $\epsilon_i(\ell(t,x_i) - \ell(s,x_i))$ is bounded between $-|\ell(t,x_i) - \ell(s,x_i)|$ and $|\ell(t,x_i) - \ell(s,x_i)|$ and hence is $(\ell(t,x_i) - \ell(s,x_i))^2$-sub-Gaussian. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda(Z_t - Z_s)\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n \epsilon_i(\ell(t,x_i) - \ell(s,x_i))\right)\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[\exp\left(\lambda \epsilon_i(\ell(t,x_i) - \ell(s,x_i))\right)\right] \\
&\leq \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{\lambda^2}{2}(\ell(t,x_i) - \ell(s,x_i))^2\right)\right] \\
&\leq \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{\lambda^2}{2}||t-s||^2\right)\right] \quad (\text{ by Lipschitz condition }) \\
&= \exp\left(\frac{\lambda^2 n ||s-t||^2}{2}\right)
\end{aligned}$$

So $\{Z_t\}_{t \in T}$ is n-sub-Gaussian for norm $||\cdot||$. ♣

Remark Let's use sub-Gaussianity and Radamacher symmetrization to control ULLN. All we want to argue is that

$$\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf| = o(1).$$

2 Uniform laws via Entropy/Covering numbers

Definition 2.1. *For empirical distribution P_n, let $L_p(P_n)$ norm be defined as*

$$||f-g||^p_{L_p(P_n)} := \int |f(x) - g(x)|^p dP_n(x) = \frac{1}{n}\sum_{i=1}^n |f(X_i) - g(X_i)|^p.$$

We shall use this for symmetrized process.

Theorem 1. *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \longrightarrow \mathbb{R}$, with envelope $F : \mathcal{X} \longrightarrow \mathbb{R}^+$ (i.e. $|f(x)| \leq F(x)$, $\forall\, x \in \mathcal{X}, f \in \mathcal{F}$). Suppose $F \in L_1(P)$. Let $\mathcal{F}_M := \{f_M = fI(|f| \leq M) : f \in \mathcal{F}\}$. If*

$$\log N(\mathcal{F}_M, L_1(P_n), \epsilon) = o_p(n), \;\; \epsilon > 0, M < \infty,$$

then

$$\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| = ||P_n - P||_{\mathcal{F}} \xrightarrow{P} 0, \quad \text{as } n \longrightarrow \infty.$$

Proof Let $P_n^0 f := \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)$ be the symmetrized process. Note that $|P_n^0 f| \leq P_n|f| = ||f||_{L_1(P_n)}$.

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} |P_n f - Pf|\right] &\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} |P_n^0 f|\right] \\
&\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - f_M(X_i))|\right] + 2\mathbb{E}\left[\sup_{f\in\mathcal{F}_M} |P_n^0 f|\right].
\end{aligned}$$

Note that, $|f - f_M| = |f|I(|f| > M) \leq FI(F > M)$, and therefore

$$\text{RHS} \leq 2\mathbb{E}\left[F(X)I(F(X) > M\right] + 2\mathbb{E}\left[||P_n^0||_{\mathcal{F}_M}\right].$$

Fix $\epsilon > 0$ and $\mathcal{G}$ be a minimal ϵ-cover of $\mathcal{F}_M$ w.r.t. norm $L_1(P_n)$. Then

$$\sup_{f\in\mathcal{F}_M} |P_n^0 f| \leq \sup_{g\in\mathcal{G}} |P_n^0 g| + \epsilon. \tag{3}$$

Now conditional on $X_1^n := (X_1, \ldots, X_n)$, we have $nP_n^0 g$ is $\sum_{i=1}^n g(X_i)^2 = n||g||^2_{L_2(P_n)}$ sub-Gaussian, and therefore,

$$\sqrt{n}\mathbb{E}\left[\sup_{g\in\mathcal{G}} |P_n^0 g| \middle| X_1^n\right] \leq \sqrt{2\sigma^2(X_1^n)\log(2|\mathcal{G}|)} = \sqrt{2\sigma^2(X_1^n)\log(2N(\mathcal{F}_M, L_1(P_n), \epsilon))}, \tag{4}$$

where, $\sigma^2(X_1^n) := \sup_{g\in\mathcal{G}} ||g||^2_{L_2(P_n)} \leq M^2$. By assumption, $\sqrt{\log(2N(\mathcal{F}_M, L_1(P_n), \epsilon))} = o_p(\sqrt{n})$, which gives,

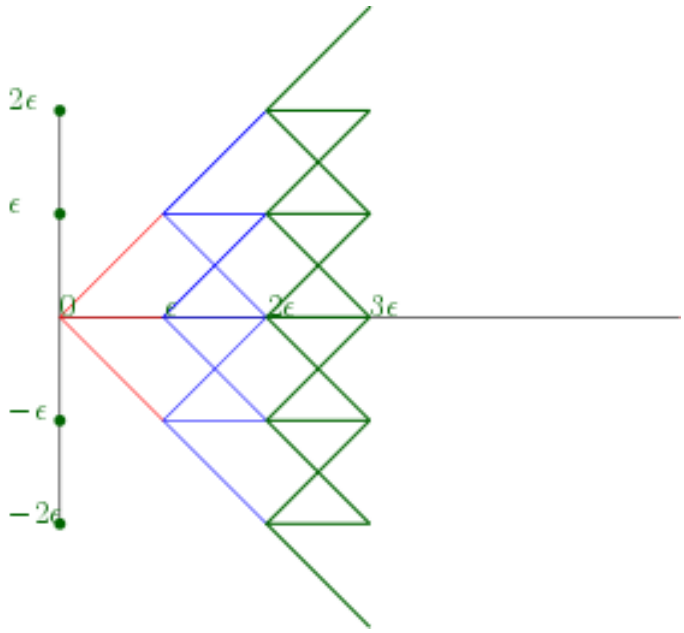
$$\mathbb{E}\left[||P_n - P||_{\mathcal{F}}\right] \leq 2\mathbb{E}\left[F(X)I(F(X) > M\right] + 2\mathbb{E}\left[M \wedge o_p(1)\right] + 2\epsilon.$$

Now the result follows by taking $n \longrightarrow \infty$ first and then $M \uparrow \infty$ and $\epsilon \downarrow 0$ along with the fact that $\mathbb{E}F(X) < \infty$. □

Remark Goal : Describe classes of functions $\mathcal{F}$ such that the metric entropy $\log N(\mathcal{F}, L_1(P), \epsilon)$ are (uniformly) small for all distributions P.

Example 3: Let $\mathcal{F}$ be a collection of 1-Lipschitz functions on $[0, 1]$ with $f(0) = 0$. Fix $\epsilon > 0$. What is the ϵ-covering number in sup-norm?

Take $x_0 = 0 < x_1 < \cdots < x_n = 1$ such that $x_{i+1} - x_i = \epsilon$, for all $i = 0, \cdots, n-2$, and $1 - x_{n-1} \leq \epsilon$. Construct family of piecewise-linear functions with constant slope (-1, 0 or +1) in each $[x_i, x_{i+1}]$, for all $i = 0, \ldots, n-1$. Since at each position $x_0, x_1, \ldots, x_{n-1}$ we have three choices (up, down, flat) and we have $\lceil \frac{1}{\epsilon} \rceil$ many choice points, then we have $3^{\lceil \frac{1}{\epsilon} \rceil}$ such functions.



Any $f \in \mathcal{F}$ has some function g among these such that $||f - g||_\infty \leq \epsilon$. Also there is a subcollection of at least $3^{\lfloor \frac{1}{\epsilon} \rfloor}$ of these functions s.t. $||g - g'||_\infty \geq \epsilon$, for all $g, g' \in$ subcollection. So,

$$\log N(\mathcal{F}, ||\cdot||_\infty, \epsilon) \asymp \frac{1}{\epsilon} \log 3.$$

Consequently, as $L_1(P_n) \leq ||\cdot||_\infty$, we have

$$\frac{1}{n} \log N(\mathcal{F}, L_1(P_n), \epsilon) \leq \frac{1}{n} \log N(\mathcal{F}, ||\cdot||_\infty, \epsilon) \leq \frac{C}{n\epsilon}.$$

Therefore, using (3) and (4), we get

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n^0 f|\right] \leq \mathbb{E}\left[\sup_{g \in \mathcal{G}} |P_n^0 g|\right] + \epsilon \leq \frac{C'}{\sqrt{n\epsilon}} + \epsilon, \quad \forall\, \epsilon > 0.$$

Therefore,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n^0 f|\right] \leq C_1 n^{-\frac{1}{3}}.$$

♣

Remark This rate of convergence is somewhat tight.

3 VC Classes

VC classes are big examples of classes allowing uniform laws and uniform entropy numbers. i.e., $\log N(\mathcal{F}, L_p(Q), \epsilon)$ is bounded independent of Q.

Definition 3.1. (Vapnik-Chervonenkis classes) *Let $\mathcal{C}$ be a collection of subsets of $\mathcal{X}$ and $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ be a collection of points. A labelling of $x_1^n := (x_1, \ldots, x_n)$ is a vector $\mathbf{y} \in \{\pm 1\}^n$. We say that $\mathcal{C}$ shatters x_1^n if for all labelling $\mathbf{y} \in \{\pm 1\}^n$ of x_1^n, $\exists A_{\mathbf{y}} \in \mathcal{C}$, s.t.*

$$\begin{cases} x_i \in A_{\mathbf{y}}, & \text{if } y_i = 1, \\ x_i \notin A_{\mathbf{y}}, & \text{if } y_i = -1. \end{cases}$$

Example 4: Let $x_1, x_2, x_3 \in \mathbb{R}^2$ and they are not collinear. $\mathcal{C}$=Collection of half-spaces in $\mathbb{R}^2$. Then $\mathcal{C}$ shatters $\{x_1, x_2, x_3\}$. ♣

Definition 3.2. Shattering number *Given $\mathcal{C} \subset 2^{\mathcal{X}}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the shattering number of $\mathcal{C}$ on $x_1^n = (x_1, x_2, ...x_n)$ is the number of labellings of x_n^1 that $\mathcal{C}$ realizes, i.e.,*

$$\Delta_n(\mathcal{C}, x_1^n) := Card\{A \cap x_1, ...x_n : A \in \mathcal{C}\}.$$

Definition 3.3. (VC dimension) *The VC dimension is the size of the largest shattered set, i.e.,*

$$VC(\mathcal{C}) := \sup\{n \in \mathbb{N} : \exists\, x_1, \ldots, x_n \in \mathcal{X} \text{ s.t. } \Delta_n(\mathcal{C}, x_1^n) = 2^n\}.$$

Example 5: Let $\mathcal{C}$ is the collection of half spaces on $\mathbb{R}^2$. We have shown $VC(\mathcal{C}) \geq 3$. Consider any 4 points on $\mathbb{R}^2$. It is not possible to shatter these four points by $\mathcal{C}$ (for example, if you can form a convex quadrilateral joining these four points, label two diagonally opposite points by 1 and other two by -1. It is not possible to shatter this set for this labelling). So $VC(\mathcal{C}) = 3$. ♣

Example 6: Let $\mathcal{C}$ is the collection of half spaces on $\mathbb{R}^d$. Then $VC(\mathcal{C}) = d + 1$. ♣

Lemma 2. Sauer-Shelah lemma *For any collection of sets $\mathcal{C} \subset 2^{\mathcal{X}}$,*

$$\max_{x_1^n \in \mathcal{X}^n} \Delta_n(\mathcal{C}, x_1^n) \leq \sum_{j=0}^{VC(\mathcal{C})} \binom{n}{j} = O(n^{VC(\mathcal{C})}).$$

Remark Consequence of Sauer-Shelah Lemma

If $\max_{x_1^n \in \mathcal{X}^n} \Delta_n(\mathcal{C}, x_1^n) < 2^n$, then $VC(\mathcal{C}) < n$ and hence

$$\Delta_m(\mathcal{C}, x_1^m) \leq O(m^{VC(\mathcal{C})}) << 2^m.$$

Additional lectures notes on the course website provide a further reference on this topic.