Stats 300b: Theory of Statistics **Winter 2017**

Lecture 12 – February 15

Lecturer: John Duchi *Scribe: Emily Diana*

***Warning:** these notes may contain factual errors*

Reading: HDP Ch.8, VdV 18-19

Outline:

- Sub-Gaussian Processes
- Uniform Entropy
- VC Classes

Recap : Process $\{x_t\}_{t \in T}$ is p-sub-Gaussian if $\mathbb{E}[\exp(\lambda(x_s - x_t))] \leq \frac{\lambda^2 p(s,t)^2}{2})]$ for all $s, t \in$ T.

Example 1: : (Canonical symmetrized empirical process)
Let $x_i \overset{i.i.d}{\sim} P$ and consider $\sup_{f \in F}(P_n f - Pf)$. Then,

$$\mathbb{E}[||P_n - P||_{\mathcal{F}}] \leq 2\mathbb{E}[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{1}^{n} \epsilon_i f(x_i)] = 2\mathbb{E}[\mathbb{E}[\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i}^{n} \epsilon_i f(x_i)]|x]$$

Fix $x_{1:n} \in \mathcal{X}^n$ and consider the process $Z_f := \frac{1}{\sqrt{(n)}} \sum_{i=1}^{n} f(x_i)$. Letting $f, q \in \mathcal{F}$,

$$\mathbb{E}[exp(\lambda(Z_f - Z_g))] = \prod_{i=1}^{n} \mathbb{E}[\exp(\frac{\lambda}{\sqrt{n}} \epsilon_i (f(x_i) - g(x_i))] \leq \exp(\frac{\lambda^2}{2n} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2) = \exp(\lambda^2 2||f - g||^2_{L_1(P_n)})$$

Remark That is, $\{Z_f\}_{f \in \mathcal{F}}$ is a $|| \cdot ||_{L_2(P_n)}$-sub-Gaussian process. Note that $\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n}$ is a sub-Gaussian process with respect to the $L_2(P_n)$ norm, and

$$\mathbb{E}[\sup_{f \in \mathcal{F}} |P_n f - Pf|] \leq \frac{1}{\sqrt{n}} 2\mathbb{E}[\mathbb{E}[sup_{f \in \mathcal{F}} | \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f(x_i) |x|]]$$

Goal 1 Our first goal for this lecture is to upper bound the expected suprema of sub-gaussian processes. Recall that if $\mathcal{F}$ is bounded by B, then $\mathbb{P}(||P_n - P||_{\mathcal{F}} \geq \mathbb{E}[||P_n - P||_{\mathcal{F}}] + 1 \leq \exp(\frac{-2nt^2}{B^2})$, using Bounded Difference, which we proved last time.

New Material: Chaining (Dudley)

Let $\{X_t\}_{t \in \mathcal{T}}$ be ρ-sub-Gaussian separable and mean-zero, i.e. $\mathbb{E}[X_t] = 0$. The idea is to control $sup_{t \in \mathcal{T}} X_t$ by finer and finer approximations to the supremum. We can do this because the process is separable. Let $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, ...\mathcal{T}$ be a sequence of covers of $\mathcal{T}$, where $\mathcal{T}$ = minimal 2^{-k} $diam(\mathcal{T})$ cover of $\mathcal{T}$ in the metric (or semimetric) ρ, where $diam(\mathcal{T}) := \sup_{s,t \in \mathcal{T}} \rho(s,t)$ (assumed finite), $\mathcal{T}_0 = \{t_0\}$, and $\rho(t_0, t) \leq diam(\mathcal{T})$ $\forall t \in \mathcal{T}$.

For any $t \in \mathcal{T}$, consider sequences $t_0, t_1, ..., t_k, ... \to t$ where $t_k \in \mathcal{T}_k$ $\forall k \in \mathbb{N}$. Let $\pi_i(t) =$

$\arg\min_{t_i \in \mathcal{T}_i} \rho(t_i, t)$ be the closest point to t in $\mathcal{T}_i$. Fix any $k \in \mathbb{N}$. Then $x_i = x_{\pi_{k-1}(t)} + x_t - x_{\pi_{k-1}t)}$.

Let $\pi^i(t) := \pi_i(\pi_{i+1}(...(\pi_{k-1}(t))...)$ (a concatenation of projections). Observe that

$$x_t = \sum_{i=1}^{k} x^i_{\pi_k}(t) - x^{i-1}_{\pi_k}(t) + x^0_{\pi}(t) = \sum_{i=1}^{k} x^i_{\pi_k}(t) - x^{i-1}_{\pi_k}(t) + x_{t_0}$$

as $\pi^k_k(t) = t$. This is the "chain."

Remark For any $k \in \mathbb{N}$, $\max_{t \in \mathcal{T}}(x_t) \leq \max_{t \in \mathcal{T}}(x^i_{\pi_k}(t) - x^{i-1}_{\pi_k}(t)) + x^0_{\pi}(t)$. How many points are there in this maximum? $\pi^i_k(t)$ takes values in $\mathcal{T}_i$ and $\pi^{i-1}_k(t) = \pi_{i-1}({\pi_k}^i(t))$ is a deterministic function of $\pi^i_k(t)$. So this is really, at "worst", a maximum over points in a set $\mathcal{T}_i$.

We know that if $D = diam(T)$, $\rho(\pi^i_i(t), \pi^{i-1}_k(t)) \leq 2^{1-i}D$ as $\pi^{i-1}_k(t) = \pi_{i-1}(\pi^i_k(t))$, T_{i-1} is a 2^{1-i} diameter cover of T. Then,

$$\max_{t \in \mathcal{T}} x_t \leq \sum_{i=1}^{k} \max_{t \in \mathcal{T}}(x_t - x_{\pi_{i-1}}(t)) + x_0$$

where $t \in \mathcal{T} \max(x_t - x_{\pi_{i-1}}(t))$ is a finite maximum of $2^{1-i}D$-sub-Gaussian random variables. Recall that if $\{Y_i\}_{i=1}^N$ are σ^2-sub-Gaussian, then

$$\mathbb{E}[\max_i(Y_i)] \leq \sqrt{(2\sigma^2 \log(N))}$$

$$\mathbb{E}[\max_{t \in T_i}(x_t - x_{\pi_{i-1}}(t))] \leq \sqrt{4^{1-i} 2D^2 \log |T_i|}$$

where $Card(T_i) = \mathcal{N}(T, \rho, 2^{-i}D)$. Then,

$$\mathbb{E}[\max_{t \in T_k}(x_t)] \leq \sum_{i=1}^{k} \sqrt{8 \cdot 4^{-1} D^2 \log \mathcal{N}(2^{-i}D)}$$

$$= 2\sqrt(2)D \sum_{i=1}^{k} 2^{-i} \sqrt{\log \mathcal{N}(D, 2^{-i})}$$

Note tht we can think of this as a Riemann integral, so

$$\mathbb{E}[\max_{t \in T_k}(x_t)] \leq 2\sqrt(2)D \sum_{i=1}^{k} 2^{-i} \sqrt{\log \mathcal{N}(D, 2^{-i})}$$

$$\leq 4\sqrt{2}D \sum_{i=1}^{\infty} \int_{2^{-i+1}}^{2^{-i}} \sqrt{\log \mathcal{N}(D_\epsilon)} d\epsilon$$

$$= 4\sqrt{2}D \int_0^1 \sqrt{\log \mathcal{N}(D_\epsilon)} d\epsilon$$

$$= 4\sqrt{2} \int_0^{diam(T)} \sqrt{\log \mathcal{N}(T, \rho, \epsilon)} d\epsilon$$

where the last equality comes from substituting ϵ for D_ϵ and letting $D = diam(T)$.
Finally, note that $\max_{t\in T_k \cup T_0}(x_t - x_{t_0})$ is non-negative, so Fatou's lemma implies that

$$\mathbb{E}[\sup_{t\in T_k}(x_t)] \leq 4\sqrt{2}\int_0^{diam(T)} \sqrt{\log \mathcal{N}(T,\rho,\epsilon)}d\epsilon$$

Definition 0.1. *For a metric space (T,ρ) with finite ρ-diameter $J(T,\rho) := \int_0^{diam(T)} \sqrt{\log \mathcal{N}(T,\rho,\epsilon)}d\epsilon$ is Dudley's entropy integral.*

Theorem 1. *Let $\{X_t\}_{t\in T}$ be a separable ρ-sub-Gaussian process. Then $\mathbb{E}[\sup_{t\in T}(X_t)] \leq C\cdot J(T,\rho)$, where $C < \infty$ is a numerical constant.*

Examples *How do we control entropy integrals? (Hint: use $\log(1+x) \leq x$) for small x)*
Example 2: *Let $\mathcal{F} := \{l(\theta,\cdot)\}_{\theta\in\Theta}$, a collection of losses. For each $x \in X$ $l(\cdot,x)$ is $\mathcal{L}(x)$-Lipschitz with respect to $||\cdot||$ in the first argument. Assume $\log \mathcal{N}(\Theta, ||\cdot||, \epsilon) \leq d(\log(1+\frac{diam(\Theta)}{\epsilon}))$. We know by the entropy integral and symmetrization*

$$\mathbb{E}[||P_n - P||_{\mathcal{F}}] \leq C\cdot\mathbb{E}[\int_0^\infty \sqrt{\log \mathcal{N}(F, L_2(P_n),\epsilon)}d\epsilon]$$

Remark $||l(t,\cdot)-l(s,\cdot)||_{L_2(P_n)} \leq \sqrt{(\log \mathcal{N}(\mathcal{F}, L_2(P_n),\epsilon)}d\epsilon$ *by $L(x)$-Lipschitz. Thus,* $\log \mathcal{N}(\mathcal{F}, L_2(P_n),\epsilon) = 0$ *if $\epsilon \geq diam(\Theta)\sqrt{P_n L^2}$. Also,* $\log \mathcal{N}(\mathcal{F}, ||\cdot||_{P_n L^2}, \epsilon) \leq \log \mathcal{N}(\Theta, ||\cdot||, \frac{\epsilon}{\sqrt{P_n L^2}})$
So, we have

$$\mathbb{E}[\int_0^\infty \sqrt{\log \mathcal{N}(F, L_2(P_n),\epsilon)}d\epsilon] \leq \mathbb{E}[\int_0^{\sqrt{P_n L^2}diam(\Theta)} \sqrt{\log \mathcal{N}(\Theta, ||\cdot||, \frac{\epsilon}{\sqrt{P_n L^2}})}d\epsilon]$$

$$= diam(\Theta)\mathbb{E}[\sqrt{P_n L^2}\int_0^1 \sqrt{\log \mathcal{N}(\Theta, P_u)}du]$$

where $u = \frac{\epsilon}{diam\sqrt{P_n L^2}}$

$$\leq diam(\Theta)\mathbb{E}[L(x)^2]^{\frac{1}{2}}\int_0^1 \sqrt{d\log(1+\frac{1}{u})}du$$

$$\leq diam(\Theta)\mathbb{E}[L(x)^2]^{\frac{1}{2}}\int_0^1 \sqrt{\frac{d}{u}}du$$

$$\leq diam(\Theta)\mathbb{E}[L(x)^2]^{\frac{1}{2}}\sqrt{d}$$

Next Goal *Give classes $\mathcal{F}$ for which we can bound $\sup_Q \mathcal{N}(\mathcal{F}, L_2(Q),\epsilon)$.*

VC Classes *Big example of classes allowing uniform bounds on entropy numbers.*

Definition 0.2. *Let $\mathcal{C}$ be a collection of sets and $X = x_1, ...x_n$. A vector $y \in \{\pm 1\}^n$ is a labeling of X. We say $\mathcal{C}$ shatters X if for all labelings $y \in \{\pm 1\}^n$, $\exists$ a set $A \in \mathcal{C}$ such that $x_i \in A$ ir $y_i = +1$ and $x_i \notin A$ if $y_i = -1$.*
Example 3: *Let $x_1, x_2, x_3 \in \mathbb{R}^3$ not collinear. $\mathcal{C}$=Half-spaces in $\mathbb{R}^2$. For any labeling, these points can be shattered.*

Definition 0.3. *Given* $\mathcal{C} \subset 2^{\mathcal{X}}$*, the shattering coefficient of* $\mathcal{C}$ *on* $x_1, x_2, ...x_n$ *is* $\Delta_n(\mathcal{C}, x_{1:n}) :=$ *card*$\{A \cap x_1, ...x_n : A \in \mathcal{C}\}$ = *the number of labelings of* $x_{1:n}$ *that* $\mathcal{C}$ *gives.*

The VC-dimension (Vapnik-Chervonenkis) of $\mathcal{C}$ *is*
$VC(\mathcal{C}) := \sup\{n \in \mathbb{N} : \max_{X_{1:n} \in \mathcal{X}^n} \Delta_n(\mathcal{C}, x_{1:n}) = 2^n\}$ = *the size of the largest set of points that* $\mathcal{C}$ *ca shatter.*

Lemma 2. *Sauer-Shelah lemma For any class* $\mathcal{C}$ *of sets,*

$$\max_{x_{1:n} \in X^n} \Delta_n(\mathcal{C}, x_{1:n}) \leq \sum_{j=0}^{VC(\mathcal{C})} \binom{n}{j} = O(n^{VC(\mathcal{C})})$$

Consequence: *If* $\max_{x_{1:n} \in X^n} \Delta_n(\mathcal{C}, x_{1:n}) < 2^n$*, then* $VC(\mathcal{C}) < n$ *and*

$$\Delta_n(\mathcal{C}, x_{1:n}) \leq O(1) \cdot n^{VC(\mathcal{C})}$$

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