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

Lecture 11 - CG Clases, Symmetrization, Subgaussian Processes and Chaining - 2/14/2017

Lecturer: John Duchi *Scribe: Matt Tsao*

Warning: *these notes may contain factual errors*

Reading:

Recap

For a function class $\mathcal{F}$, we defined a $\mathcal{F}$-norm

$$||P_n - P||_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P_n f - Pf|$$

We say that $\mathcal{F}$ satisfies a **uniform law of large numbers** if $\lim_{n\to\infty} ||P_n - P||_{\mathcal{F}} = 0$. Last time we discussed ϵ-covers and ϵ-brackets that allowed us to prove such ULLN statements.

Outline

- Glivenko Cantelli Classes
- Symmetrization Inequalities
- Subgaussian Processes
- Chaining and Entropy Integrals

Throughout this lecture we will be building up machinery that will allow us to get a handle on the behavior of $||P_n - P||_{\mathcal{F}}$.

1 GC Classes and Symmetrization

Definition 1.1. *$\mathcal{F}$ is a **Glivenko Cantelli Class** with respect to P if $||P_n - P||_{\mathcal{F}} \xrightarrow{p} 0$.*

Example 1: In Homework 1, we showed that for the class $\mathcal{F} = \{\mathbb{1}_{[x \le t]} : t \in \mathbb{R}\}$, $||P_n f - Pf||_{\mathcal{F}} = o_P(1)$, hence $\mathcal{F}$ is a GC class. In particular,

$$\mathbb{P}[\sup_t |P_n(X \le t) - P(X \le t)| > \epsilon] \le 2\exp(-cn\epsilon^2)$$

♣ A next natural question then, is how show that a certain function class $\mathcal{F}$ is a GC class. Certainly by Markov's Inequality we can say

$$\mathbb{P}\left[\sup_f |P_n f - Pf| \ge t\right] \le \frac{1}{t}\mathbb{E}\left[\sup_f |P_n f - Pf|\right] \tag{1}$$

$$= \frac{1}{nt}\mathbb{E}\left[\sup_f \left|\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_i)\right|\right] \tag{2}$$

We will now develop some tools to handle this expectation term.

Definition 1.2. *A* ***Rademacher*** *random variable is one which takes values in* $\{-1, 1\}$ *with equal probability.*

Theorem 1. *(Symmetrization)*
If $X_1, ..., X_n$ *are random vectors in a vector space equipped with a norm* $||\cdot||$ *and* $\epsilon_1, ..., \epsilon_n$ *are i.i.d. Rademarcher random variables which are independent of the* X_i*'s, then for* $p \geq 1$,

$$\mathbb{E}\left[\left|\left|\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]\right|\right|^p\right] \leq 2^p \mathbb{E}\left[\left|\left|\sum_{i=1}^{n} \epsilon X_i\right|\right|^p\right] \tag{3}$$

Proof Let X_i' be a random variable that has the same distribution as X_i and is independent from X_i. Then

$$\begin{aligned}\mathbb{E}\left[\left|\left|\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]\right|\right|^p\right] &= \mathbb{E}\left[\left|\left|\sum_{i=1}^{n} X_i - \mathbb{E}[X_i']\right|\right|^p\right] \\ \text{Jensen's Inequality} \rightarrow &\leq \mathbb{E}\left[\left|\left|\sum_{i=1}^{n} X_i - X_i'\right|\right|^p\right]\end{aligned}$$

Since X_i, X_i' are independent and have the same distribution, $X_i - X_i'$ is symmetric about 0, so in particular it has the same distribution as $\epsilon_i(X_i - X_i')$. Hence,

$$\begin{aligned}\mathbb{E}\left[\left|\left|\sum_{i=1}^{n} X_i - X_i'\right|\right|^p\right] &= \mathbb{E}\left[\left|\left|\sum_{i=1}^{n} \epsilon_i X_i - \sum_{i=1}^{n} \epsilon_i X_i'\right|\right|^p\right] \\ &= 2^p \mathbb{E}\left[\left|\left|\frac{1}{2}\sum_{i=1}^{n} \epsilon_i X_i - \frac{1}{2}\sum_{i=1}^{n} \epsilon_i X_i'\right|\right|^p\right] \\ \text{Convexity Property} \rightarrow &\leq 2^p \left(\frac{1}{2}\left|\left|\sum_{i=1}^{n} \epsilon_i X_i\right|\right|^p + \frac{1}{2}\left|\left|\sum_{i=1}^{n} \epsilon_i X_i'\right|\right|^p\right) \\ &= 2^p \left|\left|\sum_{i=1}^{n} \epsilon_i X_i\right|\right|^p\end{aligned}$$

□

Example 2: (Rademacher Complexity)
If $\mathcal{F}$ is a function class, then by symmetrization,

$$\frac{1}{2}\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - PF|\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right|\right] \tag{4}$$

The term on the right is known as the **Rademacher Complexity** of $\mathcal{F}$. ♣

2 Subgaussian Processes

Definition 2.1. *Let* $\{X_t\}_{t \in T}$ *be a collection of real valued random variables. This is a* ***Stochastic Process*** *indexed by* T.

Remark All processes we deal with in this class will be separable, i.e. there exists a countable set T' such that $\sup_{t\in T}|X_t| = \sup_{t\in T'}|X_t|$.

Definition 2.2. *Let (T,d) be a metric space. We say $\{X_t\}_{t\in T}$ is a **subgaussian process** if*

$$\log \mathbb{E}\left[\exp\left(\lambda(X_s - X_t)\right)\right] \le \frac{\lambda^2 d(s,t)^2}{2} \tag{5}$$

for all $\lambda > 0, s, t \in T$.

Remark One might expect a subgaussian constant σ^2 to appear in (5), i.e. the upper bound should be $\frac{\lambda^2\sigma^2 d(s,t)^2}{2}$, however, the metric is chosen so that the subgaussian constant is absorbed into the metric d.

Example 3:
A gaussian process is an example of a subgaussian process. To see this, let $T = \mathbb{R}^d$, and $Z \sim \mathcal{N}(0, \sigma^2 I_d)$, define $X_t = \langle Z, t\rangle$. Note that $X_s - X_t = \langle Z, s-t\rangle$ has a normal distribution with mean zero and variance $||s-t||_2^2\sigma^2$, therefore $\log\mathbb{E}[e^{\lambda(X_s - X_t)}] \le \frac{1}{2}\lambda^2\sigma^2||s-t||_2^2$ ♣

Example 4: (Rademacher Process with a loss function) Let T be a vector space equipped with a norm $||\cdot||$, $X_i \in \mathcal{X}$ are random variables and $\ell : T\times\mathcal{X} \to \mathbb{R}$ is lipschitz in its first argument, meaning that

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

Then for $\{\epsilon_i\}_{i=1}^n$ i.i.d. Rademacher random variables, because $\epsilon_i(\ell(t,X_i) - \ell(s,X_i))$ is bounded between $-||s-t||$ and $||s-t||$, it is subgaussian, hence

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n \epsilon_i(\ell(t,X_i) - \ell(s,X_i))\right)\right] &\le \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n \epsilon_i(\ell(t,X_i) - \ell(s,X_i))\right)\right]\middle| X\right] \\
&\le \mathbb{E}\left[\exp\left(\frac{\lambda^2}{8}\sum_{i=1}^n(\ell(t,X_i) - \ell(s,X_i))^2\right)\middle| X\right] \\
&\le \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^n ||t-s||^2\right) \\
&= \exp\left(\frac{\lambda^2 n||s-t||^2}{8}\right)
\end{aligned}$$

So if $Z_t = \sum_{i=1}^n \epsilon_i \ell(t, x_i)$ then the stochastic process $\{X_t\}_{t\in T}$ is $\frac{n}{4}||\cdot||^2$-subgaussian. ♣

3 Chaining and Entropy Integrals

Recall from (1) that we are interested in $\mathbb{E}[\sup_{f\in\mathcal{F}}|P_n f - Pf|]$. By symmetrization (3) we can upper bound our desired quantity by $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right]$. Therefore, we wish to understand quantities of the form $\mathbb{E}[\sup_{t\in T} X_t]$.

Let $\{X_t\}_{t\in T}$ be a $d^2(\cdot,\cdot)$ subgaussian process. We will approximate X_t by finier and finer discretizations in the following way: Let $D = diam(T) = \sup_{s,t\in T} d(s,t)$, and assume $D < \infty$. Let $T_0 \subset T_1 \subset T_2 \subset ... \subset T$ be a sequence of minimal covers of T where T_k is a minimal $2^{-k}D$ cover of T.

For $t \in T$, consider the "best" sequence $t_0, t_1, ...$ converging to t so that $t_k \in T_k$. Let $\pi_i(t) := \arg\min_{t_i\in T_i} d(t_i, t) \leq 2^{-i}D$. For any $k \in \mathbb{N}$, for $t \in T_k$ define $\pi^{(i)}(t) = \pi_i(\pi^{(i+1)}(t))$. In other words, you are projecting $k - i$ times. Now for any $t \in T_k$,

$$\begin{aligned}
X_t &= X_{\pi_{k(t)}} \\
&= (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) + X_{\pi_{k-1}(t)} \\
&= (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) + (X_{\pi_{k-1}(t)} - X_{\pi^{(k-2)}(t)}) + X_{\pi^{(k-2)}(t)} \\
&\vdots \\
&= X_{t_0} + \sum_{i=1}^{k} X_{\pi^{(i)}(t)} - X_{\pi^{(i-1)}(t)}
\end{aligned}$$

So if we take a maximum over all $t \in T_k$, and noting that $X_{t_0} = 0$, we see that:

$$\begin{aligned}
\max_{t\in T_k} X_t &= \max_{t\in T_k} \sum_{i=1}^{k} X_{\pi^{(i)}(t)} - X_{\pi^{(i-1)}(t)} \\
&\leq \sum_{i=1}^{k} \max_{t\in T_k} X_{\pi^{(i)}(t)} - X_{\pi^{(i-1)}(t)} \\
&= \sum_{i=1}^{k} \max_{\tau\in T_i} X_\tau - X_{\pi_{i-1}(\tau)}
\end{aligned}$$

Since T_i is a $2^{-i}D$ cover of T, $d(\tau, \pi_{i-1}(\tau)) \leq 2^{1-i}D$. Therefore by the subgaussianity assumption, $X_\tau - X_{\pi_{i-1}(\tau)}$ is $2^{2-2i}D$ subgaussian. Next we will use the following fact:

Fact 2. *If $X_1, ..., X_n$ are independent σ^2-subgaussian random variables,* $\mathbb{E}[\max_k X_k] \leq \sqrt{2\sigma^2 \log n}$

Because there are $N(T, 2^{1-i}D)$ elements in T_{i-1}, applying the fact gives:

$$\mathbb{E}\left[\max_{t\in T_i} \left(X_\tau - X_{\pi_{i-1}(\tau)}\right)\right] \leq \sqrt{8D^2 4^{-i} \log N(T, 2^{-i}D)}$$

Therefore by linearity of expectation, we have

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

By separability, $\lim_{k\to\infty} \max_{t\in T_k} X_t = \sup_{t\in T} X_t$. Since the sets $\{T_k\}_{k=1}^{\infty}$ are nested, $\max_{t\in T_k} X_t$ is an increasing sequence in k. Thus by the Monotone Convergence Theorem,

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in T} X_t\right] &\leq 2\sqrt{2}D \sum_{i=1}^{\infty} 2^{-i} \sqrt{\log N(T, 2^{-i}D)} \\
\text{via the integral test} \to &\leq 2\sqrt{2}D \int_0^1 \epsilon \sqrt{\log N(T, D\epsilon)}\, d\epsilon \\
\text{Change of variables} \to &= 4\sqrt{2} \int_0^{diam(T)} \sqrt{\log N(T, \epsilon)}\, d\epsilon
\end{aligned}$$

The integral on the right is known as the **Entropy Integral**.