Stats 300b: Theory of Statistics **Winter 2018**

Lecture 11 – Feb 13

Lecturer: John Duchi *Scribe: Elena Tuzhilina, Suyash Gupta*

Warning: *these notes may contain factual errors*

Reading: VdV ch. 18-19, HDP ch. 8

Outline:

- Bounded differences and Azuma-Hoeffding inequality
- Rademacher and sub-Gaussian processes
- Entropy integrals and chaining

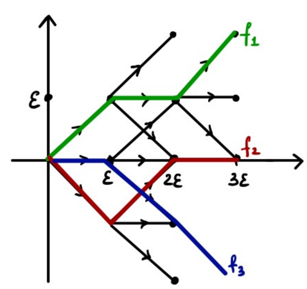
Recap Using symmetrization+covering/metric entropies to give ULLNs. Our goal is to prove $\mathbb{P}(\sup_{f\in\mathcal{F}}|P_nf - Pf| \geq t) \to 0$ as $n \to \infty$. Denote $P_n^0 = \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)$, where ϵ_i are i.i.d Redemacher random variables. Then for any $\epsilon > 0$

$$\mathbb{P}(\sup_{f\in\mathcal{F}}|P_nf - Pf| \geq t) \leq \frac{\mathbb{E}[||P_n - P||_{\mathcal{F}}]}{t} \leq \frac{2\mathbb{E}[||P_n^0||_{\mathcal{F}}]}{t} \lesssim \frac{\sqrt{\log N(\mathcal{F}, L_1(P_n), t)} + \epsilon}{\sqrt{n}t}.$$

If $\log N = o(n)$, then the RHS tends to 0.

Example: $\mathcal{F} = \{$1-Lipschitz functions on $[0,1]$ with $f(0) = 0\}$. How to calculate the covering number in sup-norm?
Fix ϵ and construct family of piecewise-linear functions with constant slope (-1, 0 or +1) in each $[0,\epsilon], [\epsilon, 2\epsilon], \ldots$. Since at each position $\{0, \epsilon, 2\epsilon, \ldots\}$ we have three choices (up, down, flat) and we have $\frac{1}{\epsilon}$ "choice" points, then we have $3^{\frac{1}{\epsilon}}$ such functions.
If $||f||_\infty = \sup_{x\in[0,1]}|f(x)|$ denotes the norm, then

$$\log N(\mathcal{F}, ||\cdot||_\infty, \epsilon) \asymp \tfrac{1}{\epsilon}\log 3 \text{ and } \log N(\mathcal{F}, L_1(P_n), \epsilon) \lesssim \tfrac{1}{\epsilon}.$$



Remark If $\mathcal{F} = \{$1-Lipschitz functions on $[0,1]^d\}$ then $\log N(\mathcal{F}, ||\cdot||_\infty, \epsilon) \sim \left(\frac{1}{\epsilon}\right)^d$ and we still get uniform law but exponentially in d(slower).

1 Concentration inequalities(revisited)

Goal: Often we want to understand concentration of more sophisticated things than averages, e.g. $\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(f(X_i) - Pf)$.

Definition 1.1. *A sequence $\{X_i\}$ adapted to a filtration $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ (increasing sequence of σ-fields) is a **Martingale difference sequence** if*

- *$X_i \in \mathcal{F}_i$ for any $i \in \mathbb{N}$*

- $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$ *for any* $i \in \mathbb{N}$.

Recall $M_n = \sum_{i=1}^n X_i$ *is associated martingale* ($X_i = M_i - M_{i-1}$).

Definition 1.2. *Let* X_i *be a MGD, it is* σ_i^2***-sub-Gaussian MGD*** *if* $\mathbb{E}[\exp(\lambda X_i)|\mathcal{F}_{i-1}] \leq \exp\left(\frac{\lambda^2\sigma_i^2}{2}\right)$ *for any* $i \in \mathbb{N}$.

Example: If $|X_i| \leq c_i$, then $\{X_i\}$ is c_i^2-sub-Gaussian MGD.

Theorem 1. *(Azuma-Hoeffding) If* $\{X_i\}$ *is* σ_i^2*-sub-Gaussian MGD, then for* $t \geq 0$

$$\mathbb{P}(\textstyle\sum_{i=1}^n X_i \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right) \text{ and } \mathbb{P}(\textstyle\sum_{i=1}^n X_i \leq -t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$$

Proof Note that $\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian, as

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n X_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda X_i}\mathbb{E}\left[e^{\lambda X_n}|\mathcal{F}_{n-1}\right]\right] \leq \mathbb{E}\left[\prod_{i=1}^{n-1} e^{\lambda X_i}\right]\exp\left(\frac{\lambda^2\sigma_n^2}{2}\right) \leq \exp\left(\frac{\lambda}{2}\sum_{i=1}^n \sigma_i^2\right)$$

□

2 Arbitrary function of independent random variables

Let $\{X_i\}_{i=1}^n$ be independent, $X_i \in \mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$. Can we control $f(X_{1:n}) - \mathbb{E}[f]$?

2.1 Doob martingale

Idea: Turn $f - \mathbb{E}[f]$ into n summands with Martingale difference structure.
Let $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$ is a σ-field generated by $X_1, \ldots, X_n$. Define

$$D_i = \mathbb{E}[f(X_{1:n})|\mathcal{F}_i] - \mathbb{E}[f(X_{1:n})|\mathcal{F}_{i-1}].$$

Note that $\mathbb{E}[f(X_{1:n})|\mathcal{F}_n] = f(X_{1:n})$ and $\mathbb{E}[f(X_{1:n})|\mathcal{F}_0] = \mathbb{E}[f]$. Therefore,

$$\sum_{i=1}^n D_i = f(X_{1:n}) - \mathbb{E}[f(X_{1:n})].$$

Also $\mathbb{E}[D_i|\mathcal{F}_{i-1}] = \mathbb{E}[|\mathbb{E}[f|\mathcal{F}_i]\mathcal{F}_{i-1}] - \mathbb{E}[f|\mathcal{F}_{i-1}] = 0$.

Observation D_i is a MD sequence adapted to $\{\mathcal{F}_i\}$, where $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$.

2.2 Bounded differences

Theorem 2. *Let all f satisfy* c_i *bounded differences*($|f(X_{1:(i-1)}, x_i, X_{i+1:n}) - f(X_{1:(i-1)}, x_i^{'}, X_{i+1:n})| \leq c_i$)*. Then* $f - Pf$ *is* $\frac{1}{4}\sum_{i=1}^n c_i^2$*subgaussian..*
Proof *Apply Azuma-Hoeffding inequality to associated Doob martingale.*

$$D_i = \mathbb{E}[f(X_{1:n})|\mathcal{F}_i] - \mathbb{E}[f(X_{1:n})|\mathcal{F}_{i-1}].$$

Let

$$U_i = sup_{x_i'} \left[\int f(X_{1:(i-1)}, x_i^{'}, X_{i+1:n}) \,\mathrm{d}P(X_{i+1:n}) - \int f(X_{1:(i-1)}, x_i, X_{i+1:n}) \,\mathrm{d}P(x_i) \,\mathrm{d}P(x_{i+1:n}) \right]$$
$$L_i = inf_{x_i'} \left[\int f(X_{1:(i-1)}, x_i^{'}, X_{i+1:n}) \,\mathrm{d}P(X_{i+1:n}) - \int f(X_{1:(i-1)}, x_i, X_{i+1:n}) \,\mathrm{d}P(x_i) \,\mathrm{d}P(x_{i+1:n}) \right]$$

Observe that

$$L_i \leq D_i \leq U_i$$

and

$$U_i - L_i \leq c_i$$

so, D_i *is* $\sigma_i^2 = \frac{c_i^2}{4}$ *sub gaussian.*

□

Corollary 3. *(McDiarmid's inequality) If* $f : \chi^n \mapsto \mathbb{R}$ *satisfies* c_i *bounded differences then for* $t \geq 0$,

$$P(f(X_{1:n}) - \mathbb{E}(f) \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).(similar \;\; for \;\; lower \;\;\; tail)$$

Idea Processes/functions satisfying bounded differences reduce problem of controlling tails to controlling expectations. Let $\mathcal{F} \subseteq \chi \mapsto \mathbb{R}$. Assume that

$$|f(x) - f(x^{'})| \leq B < \infty \forall x, x^{'} \in \chi.$$

Proposition 4. *Both* $sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i) - Pf$ *and* $sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n f(X_i) - Pf|$ *satisfy* $\frac{B}{n}$ *bounded differences.*

Proof Fix any $x_1, x_2, ..., x_n, x_i^{'} \in [n]$. Then

$$\begin{aligned} sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{j=1}^n f(x_j) - Pf\right) - sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{j=1, j\neq i}^n f(x_j) + f(x_i^{'}) - Pf\right) \\ \leq sup_{f\in\mathcal{F}} \left[\left(\frac{1}{n}\sum_{j=1}^n f(x_j) - Pf\right) - \left(\frac{1}{n}\sum_{j=1, j\neq i}^n f(x_j) + f(x_i^{'}) - Pf\right)\right] \\ = sup_{f\in\mathcal{F}} \frac{1}{n}\left[f(x_i) - f(x_i^{'})\right] \\ \leq \frac{B}{n}. \end{aligned}$$

□

Corollary 5. *Let $\mathcal{F} \subseteq \chi \mapsto \mathbb{R}$, $|f(x) - f(x^{'})| \leq B < \infty \forall x, x^{'} \in \chi$,then*

$$P\left(sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - Pf \right| \geq \mathbb{E}[||P_n - P||_{\mathcal{F}}] + t \right) \\ \leq \exp\left(-\frac{2nt^2}{B^2} \right)$$

Consequence To prove ULLN or even concentration/high probability version everything boils down to controlling $\mathbb{E}[||P_n - P||_{\mathcal{F}}] \leq 2\mathbb{E}[||P_n^o||_{\mathcal{F}}] = 2R_n((F))(Rademacher\ \ complexity)$.

3 Subgaussian Processes

Definition 3.1. *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 3.2. *Let (T, d) be a metric space. We say $\{X_t\}_{t\in T}$ is a **subgaussian process** if*

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

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

Remark One might expect a subgaussian constant σ^2 to appear in (1), 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 1:
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 $\log \mathbb{E}[e^{\lambda(X_s - X_t)}] \leq \frac{1}{2}\lambda^2\sigma^2||s - t||_2^2$ ♣

Example 2: (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

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

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

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

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. ♣