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

Lecture 10 – Feb 8

Lecturer: John Duchi *Scribe: Michael Feldman, Swarnadip Ghosh*

Warning: *these notes may contain factual errors*

Reading: VdV ch. 19, Vershynin ch. 1,2,8

Outline:

- Sub-Gaussian random variables
- Symmetrization
- Rademacher complexity and metric entropy

Recap: For a metric space (Θ, ρ), the covering number is $N(\Theta, \rho, \epsilon) = \min \{ N \text{ s.t. } \exists \text{ an } \epsilon\text{-cover } \{\theta_i\}_{i=1}^N \text{ of } \Theta \}$ where $\{\theta_i\}_{i=1}^N$ is an ϵ-cover if $\forall \theta \in \Theta$, $\exists\ \theta_i$ s.t. $\rho(\theta, \theta_i) \leq \epsilon$. Our goal is to prove uniform laws of large numbers, i.e.,

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{p} 0$$

1 Concentration Inequalities

Concentration inequalities are the key to proving ULLNS and are of fundamental importance in high dimensional and modern theoretical statistics and machine learning.

1.1 Sub-Gaussianity

Definition 1.1. *X is a mean-zero σ^2-sub-Gaussian RV if*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right) \quad \forall \lambda \in \mathbb{R}$$

Example: Gaussian random variables: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right) \quad \forall \lambda \in \mathbb{R}.$$

Example: Bounded random variables: If $X \in [a, b]$, then X is $\frac{(b-a)^2}{4}$ - subgaussian i.e,

$$\mathbb{E}\left[e^{\lambda\left(X - \mathbb{E}X\right)}\right] \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right) \quad \forall \lambda \in \mathbb{R}$$

Proposition 1. *Let X_i's be independent σ_i^2- sub-Gaussian random variables. Then $\sum_{i=1}^n X_i$ is a $\sum \sigma_i^2$-sub-Gaussian random variable.*

Proof W.l.o.g., let $\mathbb{E}X_i = 0$. By independence,

$$\mathbb{E}\big[e^{\lambda \sum_{i=1}^n X_i}\big] = \prod_{i=1}^n \mathbb{E}\big[e^{\lambda X_i}\big] \leq \exp\Big(\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2\Big).$$

□

We now derive two basic concentration inequalities for sub-Gaussian random variables.

1.2 Concentration inequalities

Proposition 2. *(Chernoff bound for sub-Gaussians) Let X be σ^2- sub-Gaussian. For all $t \geq 0$,*

$$\max\big(\mathbb{P}\big(X - \mathbb{E}X \geq t\big), \mathbb{P}\big(X - \mathbb{E}X \leq -t\big)\big) \leq e^{-t^2/2\sigma^2}$$

Proof Let $\mathbb{E}X = 0$ w.l.o.g. The result is proved using a standard technique, exponentiating the random variable and applying Markov' inequality:

$$\begin{aligned}
\mathbb{P}\big(X \geq t\big) &= \mathbb{P}\big(e^{\lambda X} \geq e^{\lambda t}\big) \qquad \forall \lambda \in \mathbb{R}-+ \\
&\leq \frac{\mathbb{E}\big[e^{\lambda X}\big]}{e^{\lambda t}} \\
&\leq e^{\frac{\lambda^2\sigma^2}{2} - \lambda t}.
\end{aligned}$$

The lefthand side of the above equation is minimized at $\lambda = \frac{t}{\sigma^2}$, giving

$$\mathbb{P}\big(X \geq t\big) \leq e^{t^2/2\sigma^2}$$

□

Corollary 3. *(Hoeffding bound) Let X_i be independent σ_i^2-sub-Gaussian r.v.s. Then, for $t \geq 0$,*

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

This is proved by applying the Chernoff bound to the $\sum_{i=1}^n X_i$, which is a $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian. The bound for the lower tail is identical.

Proposition 4. *(HW 1) Let $\{X_i\}_{i=1}^n$ be zero mean sub-Gaussians, possibly dependent. Then,*

$$\mathbb{E}\big(\max_{1\leq i\leq n} X_i\big) \leq \sqrt{2\sigma^2 \log n}$$

2 Symmetrization

For any class $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R}\}$,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} P_n f - Pf \geq t\Big) \leq t^{-1}\mathbb{E}\Big[\sup_{f\in\mathcal{F}} P_n f - Pf\Big]$$

If $P_n - P$ is symmetric, these expressions are much easier to deal with.

Definition 2.1. *ε is a Rademacher random variable if $\varepsilon \in \{-1, 1\}$ and $\mathbb{E}(\varepsilon) = 0$.*

Theorem 5. *(Symmetrization) Let $X_1, ..., X_n$ be independent random vectors in a Banach space equipped with a norm $||\cdot||$ and let $\varepsilon_1, ..., \varepsilon_n$ be i.i.d. Rademacher variables which are independent of the X_i's. For $p \geq 1$,*

$$\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\Bigg\|^p\Bigg] \leq 2^p\,\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}\varepsilon_i X_i\Bigg\|^p\Bigg]$$

Proof Let X_i' be an independent copy of X_i. Then,

$$\begin{aligned}\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\Bigg\|^p\Bigg] &= \mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i')\Bigg\|^p\Bigg] \\ &\leq \mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}(X_i - X_i')\Bigg\|^p\Bigg]\end{aligned}$$

by Jensen's inequality ($\|\cdot\|^p$ is convex as $p \geq 1$). Notice that $X_i - X_i'$ is symmetric about 0, so $X_i - X_i' \stackrel{d}{=} \varepsilon_i(X_i - X_i')$. Therefore,

$$\begin{aligned}\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\Bigg\|^p\Bigg] &\leq \mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}\epsilon_i(X_i - X_i')\Bigg\|^p\Bigg] \\ &= 2^p\,\mathbb{E}\Bigg[\Bigg\|\frac{1}{2}\sum_{i=1}^{n}\epsilon_i X_i - \frac{1}{2}\sum_{i=1}^{n}\epsilon_i X_i'\Bigg\|^p\Bigg] \\ &\leq 2^{p-1}\,\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}\epsilon_i X_i\Bigg\|^p\Bigg] + 2^{p-1}\,\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}\epsilon_i X_i'\Bigg\|^p\Bigg] \\ &= 2^p\cdot\mathbb{E}\Bigg[\Bigg\|\sum_{i=1}^{n}\varepsilon_i X_i\Bigg\|^p\Bigg]\end{aligned}$$

The second inequality follows from the convexity of $\|\cdot\|^p$. □

This result is useful for several reasons:

1. symmetric r.v.s are often easier to work with
2. we can find more precise bounds for symmetric sums
3. proofs of ULLNS will be easier
4. Conditional on $\{X_i\}_{i=1}^n$, $\sum_{i=1}^n \varepsilon_i X_i$ is $\sum_{i=1}^n X_i^2$-sub-Gaussian.

By symmetrization,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} P_n f - Pf \geq \varepsilon\Big) \leq \frac{1}{\varepsilon}\mathbb{E}\Big[\sup_{f\in\mathcal{F}} P_n f - Pf\Big] \leq \frac{2}{n\varepsilon}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^{n}\varepsilon_i f(x_i)\Big|\Big]$$

Definition 2.2. *The Rademacher complexity $R_n(\mathcal{F})$ is defined as*

$$R_n(\mathcal{F}) = \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^{n}\varepsilon_i f(x_i)\Big|\Big]$$

If $R_n(\mathcal{F}) = o(n)$, then we have a ULLN. Typically we require an envelope function F, a function that satisifies $F(x) \geq |f(x)|$, for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$. For $M \in \mathbb{R}_+$, let

$$f_M(x) = \begin{cases} f(x) & |f(x)| \leq M \\ 0 & |f(x)| > M \end{cases}$$

and $\mathcal{F}_M = \{f_m : f \in \mathcal{F}\}$.

Theorem 6. *Let $\mathcal{F}$ be a class of functions with envelope $F \in L_1(P)$. If $\log N(\mathcal{F}_M, L_1(P_n), \varepsilon) = o_p(n)$ for all $M < \infty$ and $\varepsilon > 0$, then $\|P_n - P\|_{\mathcal{F}} \xrightarrow{p} 0$.*

Proof Let $P_n^0 f = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)$ where the ϵ_i are i.i.d. Rademachers. By symmetrization,

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

Call the first term above T_1 and the second T_2. $T_1 \leq 2\mathbb{E}\big[F(X)\mathbf{1}_{F(X)\geq M}\big] \to 0$ as $M \to \infty$. Let $\mathcal{G}$ be minimal ε-cover of $\mathcal{F}_M$ in $L_1(P_n)$ norm. Then,

$$\sup_{f\in\mathcal{F}_M}\Big\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\Big\| \leq \max_{g\in\mathcal{G}}\Big\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\Big\| + \epsilon$$

Conditional on X_i , $\sum_{i=1}^{n}\varepsilon_i g(X_i)$ is $n\sigma_n^2 := \sum_{i=1}^{n} g^2(X_i)$ sub-Gaussian. Since $\sum_{i=1}^{n} g^2(X_i) \leq nM^2$, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i g(X_i)$ is M^2 sub-Gaussian.

$$\begin{aligned}\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i g(X_i)\Big\|\Big|X\Big] &\leq \sqrt{2\sigma_n^2\log(2|\mathcal{G}|)} \\ &\leq \sqrt{2M^2 log(2N(\mathcal{F}_M, L_1(P_n), \epsilon))} \\ &= o_p(\sqrt{n})\end{aligned}$$

Therefore we get, $\mathbb{E}\big[\|P_n - P\|_{\mathcal{F}}\big] \leq 2\mathbb{E}\big[F\mathbf{1}_{F\geq M}\big] + 2\mathbb{E}\big[M \wedge o_p(1)\big] + 2\epsilon$. Now, let $M \to \infty$, $n \to \infty$, and $\varepsilon \downarrow 0$. The righthand side goes converges to 0, concluding the proof. □