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

Lecture 10 – Feb 9

Lecturer: John Duchi *Scribe: Qijia Jiang, Vivek Bagaria*

Warning: *these notes may contain factual errors*

Reading:

1 Sub-gaussianity

1.1 Definitions and Properties

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

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

Example: Gaussian random variables: If $X \sim \mathcal{N}(0, \sigma^2)$, then

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

Bounded random variables also fall into the category of sub-gaussian random variables:
Example: If $X \in [a, b]$, then X is $\frac{(b-a)^2}{4}$ - subgaussian i.e,

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

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

Proof $\mathbb{E}\big[\exp^{\lambda \sum_{i=1}^n X_i}\big] = \Pi_{i=1}^n \mathbb{E}\big[\exp^{\lambda X_i}\big] = \exp\left(\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}\right)$. □

Now that we've defined sub-gaussian random variables and a few simple properties, lets use them to obtain concentration inequalities similar to the Chernoff bounds.

1.2 Concentration inequalities

Lemma 2. *If X is a σ^2- subgaussian, then we have*

$$\max\big(\mathbb{P}(X - \mathbb{E}[X] \geq t), \mathbb{P}(X - \mathbb{E}[X] \leq -t)\big) \leq \exp -\left\{\frac{-t^2}{2\sigma^2}\right\}$$

Proof Let $\mathbb{E}[X] = 0$ w.l.o.g. We prove the above result using the techniques used to prove Chernoff bounds i.e, applying Markov inequality on the exponentiation of the random variable:

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

The LHS of the above equation is minimized for $\lambda = \frac{t}{\sigma^2}$, and therefore, we have

$$\mathbb{P}\big(X \geq t\big) \leq \exp -\Big\{\frac{-t^2}{2\sigma^2}\Big\}$$

$\square$

Corollary 3. *(Hoeffding inequality) Let X_i be independent σ_i^2 -subgaussian RVs. Then we have*

$$\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\big) \leq \exp -\Big\{\frac{-nt^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big\} \quad t \geq 0$$

and

$$\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^{n} X_i \leq t\big) \leq \exp -\Big\{\frac{-nt^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big\} \quad t < 0$$

This inequality is heaviliy used in proving concentration results for bounded random variables (see Example 1.1).

2 Covering Number and Metric Entropy

Let $\big(\Theta, d\big)$ be a metric space with distance measure $d : \Theta \times \Theta \to \mathbb{R}$.

Definition 2.1. *For any $\epsilon > 0$, $\big\{\theta_i\big\}_{i=1}^{N}$ is the ϵ-**cover** of Θ if*

$$\min_i d(\theta, \theta_i) < \epsilon \quad \forall \theta \in \Theta.$$

This naturally leads to the definition of covering number:

Definition 2.2. *For $\epsilon > 0$, the **covering number** of Θ for metric d is*

$$N(\Theta, d, \epsilon) = \inf\big\{N : \exists \textit{ an } \epsilon - \textit{cover } \big\{\theta_i\big\}_{i=1}^{N} \textit{ of } \Theta\big\}$$

*and $\log N(\Theta, d, \epsilon)$ is also referred as the **metric entropy.***

Covering a space is a task of covering the whole space with minimum number of balls. Extending this idea, we define packing as

Definition 2.3. *For any $\epsilon > 0$, $\big\{\theta_i\big\}_{i=1}^{M}$ is the ϵ-**packing** of Θ if*

$$\min_{i,j} d(\theta_i, \theta_j) > \epsilon.$$

Similar to covering number, we define packing number as

Definition 2.4. *For $\epsilon > 0$, the **packing number** of Θ with metric d is*

$$M(\Theta, d, \epsilon) = \sup\big\{N : \exists \textit{ an } \epsilon - \textit{packing } \big\{\theta_i\big\}_{i=1}^{N} \textit{ of } \Theta\big\}$$

*and $\log M(\Theta, d, \epsilon)$ is also referred as the **packing entropy.***

As one would suspect, covering and packing are related, and we indeed have the relation:

$$M(2\epsilon) \leq N(\epsilon) \leq M(\epsilon). \tag{1}$$

Example: Consider the balls in $\mathbb{R}^d$ with norm $||.||$, let $\mathbb{B} = \{V \in \mathbb{R}^d : ||V|| \leq 1\}$, and $\Theta = r\mathbb{B}$

1. Since ϵ- packing is equivalent of having "disjoint" balls of radius $\epsilon/2$ we have

$$\begin{aligned} M\text{Vol}(\epsilon/2) &\leq \text{Vol}(r + \epsilon/2) \\ \implies M &\leq \left(1 + \frac{2r}{\epsilon}\right)^d \end{aligned}$$

2. Similarly ϵ - covering covers the whole space with balls of radius ϵ and hence we have

$$\begin{aligned} N\text{Vol}(\epsilon) &\geq \text{Vol}(r) \\ \implies N &\geq \left(\frac{r}{\epsilon}\right)^d \end{aligned}$$

   coupling the above inequality with equation 1, we obtain

$$\left(\frac{r}{\epsilon}\right)^d \leq N(\epsilon) \leq \left(1 + \frac{2r}{\epsilon}\right)^d$$

3 Bracketing number

When the underlying space Θ is a space of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, we can define bracketing numbers along the lines of covering , packing numbers. Formally,

Definition 3.1. *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$ be a collection of fns with measure μ. A set $\{[l_i, u_i]\}_{i=1}^N$ of functions $\mu_i, l_i : \mathcal{X} \to \mathbb{R}$ is a ϵ - bracketing set of $\mathcal{F}$ if*

$$\forall f \in \mathcal{F} \ \exists i \ s.t \ l_i \leq f(x) \leq \mu_i$$

and $\int \big(\mu_i(x) - l_i(x)\big)^p d\mu(x) \leq \epsilon^p$.

In the spirit of defining "numbers" for each notion of covering we define

Definition 3.2. *Bracketing number of $\mathcal{F}$ is*

$$N_{[]}\big(\mathcal{F}, L_p(\mu), \epsilon\big) := \inf\left\{N : \exists \ \ a \ set \ \{[l_i, u_i]\}_{i=1}^N \ which \ is \ \epsilon - bracketing \ of \ \mathcal{F}\right\}$$

Claim 4. *Let $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ where m_θ are L-Lipschitz in θ, theb $N_{[]}\big(\mathcal{F}, L_p, \epsilon L\big) \leq N(\Theta, ||.||, \epsilon/2)$.*

Proof Let $\{\theta_i\}_{i=1}^N$ be an $\epsilon/2$ -covering of Θ, then lets define

$$\begin{aligned} u_i(x) :&= m_{\theta_i}(x) + \frac{\epsilon}{2}L \\ l_i(x) :&= m_{\theta_i}(x) - \frac{\epsilon}{2}L. \end{aligned}$$

We know that for any $\theta \in \Theta$, $\exists \theta_i$ s,t $||\theta - \theta_i|| \leq \frac{\epsilon}{2}$, and from Lipschitz properties of m_θ, we have

$$\begin{aligned} \big|m_\theta(x) - m_{\theta_i}(x)\big| &\leq L||\theta - \theta_i|| \\ &\leq \frac{\epsilon}{2}L. \end{aligned}$$

□

Theorem 5. *(Uniform Convergence) Let $\mathcal{F}$ satisfy $N_{[]}\big(\mathcal{F}, L_p, \epsilon\big) < \infty$, then under i.i.d. sampling*

$$\sup_{f\in\mathcal{F}} \big|P_n f - Pf\big| \xrightarrow{p} 0.$$

Proof For any given $\epsilon > 0$ let $\big\{[l_i, u_i\big\}_{i=1}^{N}$ be ϵ-bracketing numbers then $\forall f \in \mathcal{F},\ \exists i$ s.t $l_i \leq f \leq u_i$, and therefore we have

$$\begin{aligned} P_n f - Pf &\leq P_n u_i - P l_i \\ &= P_n u_i - P u_i + P u_i - P l_i \\ &\leq o_p(1) + \epsilon. \end{aligned}$$

Since $N_{[]}$ is finite and ϵ was arbitrary, we have

$$\begin{aligned} \sup_{f\in\mathcal{F}} \big|P_n f - Pf\big| &\leq \big|N_{[]}\big| o_p(1) + \epsilon \\ &\leq 2\epsilon \to 0. \end{aligned}$$

□