36-709: Advanced Statistical Theory I **Spring 2019**

Lecture 6: February 7

Lecturer: Alessandro Rinaldo *Scribes: Mikaela Meyer*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

We begin this lecture by talking about how concentration inequalities can be used to give confidence intervals and moment bounds.

6.1 Moment Bounds and Confidence Intervals

Moment bound: If $\mathbb{P}(|X| \geq t) \leq ce^{-cnt^\alpha}, \alpha \in \{1, 2\}$, then $\mathbb{E}(|X_n|) \leq C' n^{-1/\alpha}$.
Confidence interval: If $\mathbb{P}(|X_n| \geq t) \leq ae^{\frac{-nbt^2}{c+dt}}, a, b, c, d > 0$, then $\forall \delta \in (0, 1)$, $|X_n| \leq \sqrt{\frac{c}{nb}} + \frac{d}{nb} \log(\frac{a}{\delta})$ with probability at least $1 - \delta, \delta = \mathcal{O}(\frac{1}{n^p}), p > 0$

The downside of using this confidence interval is that it might not be the sharpest interval. Knowing the likelihood function would give better values of the constants, which would produce a tighter bound. However, if you are not concerned about this downside, the benefit is that you have a finite sample confidence interval that holds for all n.

6.1.1 Application: Maxima

We know how to use the union bound to bound probabilities. It turns out that bounding the expected value of the maximum of a random variable can use a comparable trick. Being able to bound the expected value of the maximum of a random variable is one of the most important bounds we can get in high dimensional statistics.

Theorem 6.1 *Let $X_1, ..., X_n$ be centered random variables that are not necessarily independent such that for all $\lambda \in [0, b), b \leq \infty$, $\mathbb{E}[e^{\lambda X_i}] \leq \psi(\lambda)$, where $\psi(.)$ is convex on $[0, b)$. Then*

$$\log \mathbb{E}[\max_i X_i] \leq \inf_{\lambda \in [0,b)} \left\{ \frac{\log(n) + \psi(\lambda)}{\lambda} \right\}$$

Proof: By Jensen's inequality, for any $\lambda \in [0, b)$,

$$\begin{aligned}
\exp\{\lambda \mathbb{E}[\max_i X_i]\} &\leq \mathbb{E}\left[e^{\lambda \max_i X_i}\right] && f(x) = e^x \text{ is convex} \\
&= \mathbb{E}\left[\max_i e^{\lambda X_i}\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right] \\
&\leq n e^{\psi(\lambda)}
\end{aligned}$$

Taking the log of both sides, we get $\lambda \mathbb{E}[\max_i X_i] \leq \log(n) + \psi(\lambda)$, so $\mathbb{E}[\max_i X_i] \leq \frac{\log(n)+\psi(\lambda)}{\lambda}$ ■

We didn't see the convexity of $\psi(.)$ used in this proof.

Result: If ψ is convex, continuously differentiable on $[0, b)$, and $\psi(0) = \psi'(0) = 0$, then $\forall \mu > 0$ $\inf_{\lambda \in [0,b)} \left\{\frac{\mu+\psi(\lambda)}{\lambda}\right\} = \inf\{t \geq 0 : \psi^*(t) \geq \mu\}$ where $\psi^*(t) = \sup_{\lambda \in [0,b)} \lambda t - \psi(\lambda)$ (Proof found in [SB12])

Example (Sub-Gaussian Random Variables): Given all $X_i \in \text{SG}(\sigma^2)$, then $\psi(\lambda) = \frac{\lambda^2\sigma^2}{2}$. The bound is $\inf_{\lambda>0} \frac{\log(n)}{\lambda} + \frac{\lambda^2\sigma^2}{2\lambda}$. Set $\lambda = \sqrt{\frac{2\log(n)}{\sigma^2}}$ to obtain $\mathbb{E}[\max_i X_i] \leq \sqrt{2\sigma^2 \log(n)}$. This yields an important result: the expected value of the maximum of sub-Gaussian random variables with the same parameter grows at a rate of $\sqrt{\log(n)}$.

Example (Sub-Exponential Random Variables): If $\psi(\lambda) = \frac{\lambda^2\nu^2}{2(1-\lambda b)}$ for $\lambda \in (0, \frac{1}{b})$, then it is possible (though not fun) to show $\mathbb{E}[\max_i X_i] \leq \sqrt{2\nu^2 \log(n)} + b\log(n)$. If you are really interested in showing this, see [SB12], page 29. Note that this bound looks similar to the one for sub-Gaussian random variables, except that it includes an additional $b\log(n)$ term. So, if $X_i \sim \chi^2_p$, $\mathbb{E}[\max X_i] \leq 2\sqrt{p\log(n)} + 2\log(n)$.

Going forward, we will see many examples of how to deal with this bound on the maximum. By the way, you already know how to bound $\mathbb{P}(\max X_i \geq t) \leq n e^{\frac{-t^2}{2\sigma^2}}$ by the union bound.

6.2 Bounded Differences Inequality (a.k.a. McDiarmid's Inequality or Azuma-Hoeffding Inequality)

Most of the time, we see bounds on averages. But averages are just one type of function of random variables we can be interested in bounding. Picture an arbitrary function of independent random variables. Can we create a concentration inequality for this arbitrary function?

As an aside, it should be noted that $\mathbb{E}(Y|X)$ is a random variable because it is a function of the random variable, X. However, $\mathbb{E}(Y|X = x)$ is not a random variable because it is a function of the fixed x.

Let $Z = f(X_1, ..., X_n)$ $f : \mathbb{R}^n \to \mathbb{R}$. We are interested in the concentration inequality for $Z - \mathbb{E}(Z)$. Set $Y_0 = \mathbb{E}(Z)$, making it a degenerate random variable. For $k = 1, ..., n$, set $Y_k = \mathbb{E}(Z|X_1, ..., X_k)$ so $Y_n = Z$ (also a random variable). Then $Z - \mathbb{E}(Z) = Y_n - Y_0 = \sum_{k=1}^n Y_k - Y_{k-1} = \sum_{k=1}^n D_k$. This is a sum of n random variables, though these random variables are not independent. If we want to go about making this bound, we will need to first spend time making an aside about Martingales.

6.2.1 Martingales

A sequence $Y_0, Y_1, Y_2, ...$ of random variables is a **Martingale** if:

1. $\mathbb{E}[|Y_i|] < \infty$ (integrable)
2. For every $k \geq 1$, $\mathbb{E}(Y_k|Y_0, ..., Y_{k-1}) = Y_{k-1}$

Doob Construction: Let $Z = f(X_1, ..., X_n)$. Let $Y_k = \mathbb{E}(Z|X_1, ..., X_k)$. Then $Y_1, ..., Y_n$ is a Martingale. If $Y_0, ..., Y_n$ is a Martingale, then $D_1, ..., D_n$ where $D_k = Y_k - Y_{k-1}$ for all k is called a **Martingale difference**. $\mathbb{E}[D_k] = 0$ for all k.

Theorem 6.2 *Let $D_1, ..., D_n$ be Martingale differences such that $\forall |\lambda| < \frac{1}{\alpha_k}, \nu_k, \alpha_k > 0$ for all k, $\mathbb{E}[e^{\lambda D_x}|D_1, ..., D_{k-1}] \leq e^{\frac{\lambda^2 \nu_k^2}{2}}$ (the differences are sub-exponential). Then:*

1. $\sum_{k=1}^n D_k \in SE\left(\sum_{k=1}^n \nu_k^2, \max_k \alpha_k\right)$ *(same as if they were independent)*
2. *For $t > 0$,* $\mathbb{P}(|\sum_k D_k| \geq t) \leq \begin{cases} 2\exp\left\{\frac{-t^2}{2\sum \nu_k^2}\right\}, & t \leq \frac{\sum_k \nu_k^2}{\max \alpha_k} \\ 2\exp\left\{\frac{-t}{2\max \alpha_k}\right\}, & t > \frac{\sum_k \nu_k^2}{\max \alpha_k} \end{cases}$

We will prove part 1 of these results below.

Proof: Fix λ. $\mathbb{E}[e^{\lambda \sum_{k=1}^n D_k}] = \mathbb{E}[\mathbb{E}[e^{\lambda \sum_{k=1}^n D_k}|D_1, ..., D_n]] = \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}\, \mathbb{E}[e^{\lambda D_n}|D_1, ..., D_{n-1}]]$. We know that $\mathbb{E}[e^{\lambda D_n}|D_1, ..., D_{n-1}] \leq e^{\frac{\lambda \nu_n^2}{2}}$ if $|\lambda| ¡ 1\frac{}{\alpha_n}$. Thus, $\mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}\, \mathbb{E}[e^{\lambda D_n}|D_1, ..., D_{n-1}]] \leq e^{\lambda \nu_n^2} 2\, \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}]$. We can repeat this process of using the law of total expectation to eventually arrive at $e^{\lambda \nu_n^2} 2\, \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}] \leq e^{\lambda \sum_{k=1}^n \frac{\nu_k^2}{2}}$ where $|\lambda| < \frac{1}{\max_k \alpha_k}$. ■

Corollary (Azuma-Hoeffding Inequality): If $a_k \leq D_k \leq b_k$ a.e. for all k, then $\mathbb{P}(|\sum_{k=1}^n D_k| \geq t) \leq 2\exp\left\{\frac{-2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right\}$ because $D_k \in \text{SG}((\frac{b_k - a_k}{2}^2))$ where D_k is conditioned on $D_1, ..., D_{k-1}$.

6.2.2 Bounded Difference Property

Now that we took this brief detour of Martingales, let's go back to bounding functions of random variables. We know $X_1, ..., X_n$ are independent and $Z = f(X_1, ..., X_n)$, and we want to show concentration for $Z-$. We need a condition that tells us if we change one coordinate at a time, the difference is still controlled. A regularity condition on f that leads to an application of the Azuma-Hoeffding inequality is that of having bounded differences.

Definition: $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the **bounded difference property (BDP)** if $\exists\ L_1, ..., L_n \in \mathbb{R}_+$ such that $\forall\ (x_1, ..., x_n)$ in domain of f, for any coordinate, k:

$$\sup_{x_1, ..., x_n} |f(x_1, x_2, ..., x_k, ..., x_n) - f(x_1, x_2, ..., x_k', ..., x_n)| \leq L_k$$

where x_k' represents that the kth coordinate was changed from its original value. This implies that the value of the function f doesn't depend much on the value of any one coordinate.

Also, if $\|f\|_\infty \leq b$, then $L_k \leq 2b$ for all k. $\|f\| = \sup_{x_1,...,x_n} |f(x_1, ..., x_n)|$

We will now present the final theorem, which gives a tail bound for functions that satisfy the bounded difference property.

Theorem 6.3 *Let $X_1, ..., X_n$ be independent and $f : \mathbb{R}^n \to \mathbb{R}$ satisfy the BDP. Let $Z = f(X_1, ..., X_n)$. Then*

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq 2 \exp\left\{\frac{-2t^2}{\sum_{k=1}^n L_k^2}\right\}$$

References

[SB12] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, 2012, Ch. 1.