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

Lecture 3: January 24

Lecturer: Alessandro Rinaldo *Scribes: Nil-Jana Akpinar*

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.

3.1 Basic concentration inequalities

3.1.1 Motivation

We know that Gaussian random variables concentrate around their mean, i. e. for $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, it holds

$$P\left(\left|\bar{X}_n - \mu\right| \geq t\right) \leq \exp\left(\frac{-nt^2}{2\sigma^2}\right)$$

for every $t \geq 0$. Thus, the probability that the sample average $\bar{X}_n$ is far away from the mean μ decays rapidly. We want to replicate this type of behavior for other random variables in a manner that allows us to (1) obtain finite samples guarantees (i. e. for every n), and (2) circumvent the need for too many distributional assumptions on $X_1, \ldots, X_n$.

Goal: Given some $X \sim P$ with mean μ, we want to derive an upper bound on $P(|X - \mu| \geq t)$ which holds for all $t \geq 0$.

3.1.2 Markov inequality

We make a first attempt at bounding the above probability in terms of moments of X based on Markov's inequality.

Theorem 3.1 (Markov inequality) *Let X be a random variable and $\mathbb{E}[X] = \mu$. Then,*

$$P\left(|X - \mu| \geq t\right) \leq \min_{q \in \mathbb{N}} \frac{\mathbb{E}[|X - \mu|^q]}{t^q}.$$

This procedure often yields an analytically sharp bound. However, it requires us to compute the centered moments of X which is often infeasible or computationally expensive.

3.1.3 Chernoff bound

For a second approach to bounding of the above probability, we draw on the moment generating function of the centered version of X, i. e. $\psi_X(\lambda) := \log(\mathbb{E}[e^{\lambda(X-\mu)}])$, which is well-defined for all $\lambda \in (-b, b)$ for some

$0 \leq b \leq \infty$. Assuming a $0 < \lambda \leq b$, we get with Markov's inequality that

$$\begin{aligned} P(X - \mu \geq t) &= P\left(e^{X-\mu} \geq e^t\right) \\ &= P\left(e^{\lambda(X-\mu)} \geq e^{\lambda t}\right) \\ &\leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}} \\ &= \exp\left(\psi_X(\lambda) - \lambda t\right), \end{aligned}$$

which results in the following bound.

Theorem 3.2 (Chernoff bound) *Let X be a random variable and $\mathbb{E}[X] = \mu$. Then,*

$$P(X - \mu \geq t) \leq \exp\left(-\psi_X^*(t)\right),$$

where $\psi_X^*(t) := \sup_{\lambda \in (0,b)}(\lambda t - \psi_X(\lambda))$.

In some sense, deriving a Chernoff bound does not require less knowledge about a distribution than a Markov-based bound since we need the moment generating function of $X - \mu$. In fact, we have to assume the existence of infinity many moments. A main advantage is that these moments do not have to be painstakingly calculated, and in turn, Chernoff bounds are usually the way to go when having enough knowledge about the distribution although they are not as sharp as the Markov-based bounds.

Example 3.3 (Chernoff bound for Gaussian) *Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}\left[e^{\lambda X}\right] = e^{\mu\lambda + \sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. So, we have*

$$\sup_{\lambda > 0}\left(\lambda t - \log\left(\mathbb{E}\left[e^{\lambda(X-\mu)}\right]\right)\right) = \sup_{\lambda > 0}\left(\lambda t - \frac{\lambda^2\sigma^2}{2}\right) = \frac{t^2}{2\sigma^2},$$

which yields the bound

$$P(X - \mu \geq t) \leq e^{\frac{-t^2}{2\sigma^2}}$$

for all $t > 0$.

Theorem 3.4 (Two-sided Chernoff bound) *Let X be a random variable and $\mathbb{E}[X] = \mu$. Then,*

$$P(|X - \mu| \geq t) \leq 2\exp\left(-\psi_X^*(t)\right),$$

where $\psi_X^*(t) := \sup_{\lambda \in (-b,b)}(\lambda t - \psi_X(\lambda))$.

3.1.4 Sub-Gaussian random variables

In order to be able to derive Chernoff bounds, we need (a bound for) $\psi_X(\lambda)$ which is not always easily attainable. A sufficient condition in this setting is that the random variable is sub-Gaussian, i. e. its tails decay faster than the tails of some Gaussian. An extensive overview over sub-Gaussian random variables can be found in [BK00].

Definition 3.5 (Sub-Gaussian) *A random variable X is sub-Gaussian with parameter σ if*

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

for all $\lambda \in \mathbb{R}$. In that case, we write $X \in SG(\sigma^2)$.

A first simple observation is given by $X \in \mathrm{SG}(\sigma^2)$ iff $-X \in \mathrm{SG}(\sigma^2)$.

Now, if $X \in \mathrm{SG}(\sigma^2)$, then the mgf of X can be bounded by the Gaussian mgf which yields the same Chernoff bound as in Example 3.3, i. e.

$$P\left(|X-\mu| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2\sigma^2}\right).$$

Proposition 3.6 *We observe several properties of sub-Gaussian random variables.*

(1) Let $X \in SG(\sigma^2)$, then $\mathbb{V}[X] \leq \sigma^2$ with $\mathbb{V}[X] = \sigma^2$ if X is Gaussian.

(2) If there are $a, b \in \mathbb{R}$ such that $a \leq X - \mu \leq b$ almost everywhere, then $X \in SG\left(\left(\frac{b-a}{2}\right)^2\right)$.

(3) Let $X \in SG(\sigma^2)$ and $Y \in SG(\tau^2)$, then

(i) $X\alpha \in SG(\sigma^2\alpha^2)$ for all $\alpha \in \mathbb{R}$ with $\alpha \neq 0$,

(ii) $X + Y \in SG((\tau+\sigma)^2)$, and

(iii) if $X \perp Y$, $X + Y \in SG(\tau^2 + \sigma^2)$.

Proof: *(1) It holds by assumption that $\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ for all $\lambda \in \mathbb{R}$, and hence,*

$$1 + \lambda \underbrace{\mathbb{E}[X-\mu]}_{=0} + \lambda^2 \frac{\mathbb{E}[(X-\mu)^2]}{2} + o(\lambda^2) \leq 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2).$$

We divide both sides of this inequality by λ^2 (and assume $\lambda \neq 0$), and let $\lambda \to 0$.

(2) WLOG, let $\mu = 0$. We show that $\log\left(\mathbb{E}\left[e^{\lambda X}\right]\right) \leq \frac{(b-a)^2\lambda^2}{8}$ for all $\lambda \in \mathbb{R}$. First, notice that $\mathbb{V}[X] \leq \left(\frac{b-a}{2}\right)^2$. For any $\lambda \in \mathbb{R}$, let X_λ be a RV with distribution that has density of the form

$$x \mapsto e^{\lambda X} e^{-\psi_X(\lambda)} f_X(x)$$

if $a \leq x \leq b$. Then, $\mathbb{V}[X_\lambda] = \psi_X''(\lambda) \leq \left(\frac{b-a}{2}\right)^2$. Since $\psi_\lambda(0) = \psi_\lambda'(0) = 0$, we have with the fundamental theorem of calculus that

$$\psi_X(\lambda) = \int_0^\lambda \psi_X'(u)du = \int_0^\lambda \int_0^\mu \psi_X''(w)dw \leq \int_0^\lambda \int_0^\mu \lambda^2 \frac{(b-a)^2}{4} dwdu = \lambda\frac{(b-a)^2}{8}.$$

(3) We prove (ii) and (iii) and assume that $\mathbb{E}[X] = \mathbb{E}[Y]$. If $X \perp Y$, the proof is immediate. If not, it holds for every $\lambda \in \mathbb{R}$ that $\mathbb{E}\left[e^{\lambda(X+Y)}\right] = \mathbb{E}\left[e^{\lambda X}e^{\lambda Y}\right]$. We use Hölder's inequality (see below) and obtain

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X+Y)}\right] = \mathbb{E}\left[e^{\lambda X}e^{\lambda Y}\right] &\leq \left(\mathbb{E}\left[e^{\lambda p X}\right]\right)^{1/p}\left(\mathbb{E}\left[e^{\lambda q Y}\right]\right)^{1/q} \overset{SG}{\leq} \exp\left(\frac{\lambda^2 p^2\sigma^2}{2}\frac{1}{p} + \frac{\lambda^2 q^2\tau^2}{2}\frac{1}{q}\right) \\ &= \exp\left(\frac{\lambda^2}{2}\left(p\sigma^2 + q\tau^2\right)\right) = \exp\left(\frac{\lambda^2}{2}(\sigma+\tau)^2\right),\end{aligned}$$

where we set $p = \tau/\sigma + 1$ in the last step. ■

Detour:
If $p, q > 0$ with $\frac{1}{p} + \frac{1}{q} = 1$, it holds that

$$\mathbb{E}\left[|X_1 X_2|\right] \leq \left(\mathbb{E}\left[|X_1|^p\right]\right)^{1/p} \left(\mathbb{E}\left[|X_2|^q\right]\right)^{1/q}$$

which is called **Hölder inequality**. The special case with $p = q = 2$ is referred to as **Cauchy-Schwartz inequality**. The Cauchy-Schwartz inequality can, for example, be used to show that

$$\left| \frac{\mathbf{Cov}[X_1, X_2]}{\sqrt{\mathbb{V}[X_1]}\sqrt{\mathbb{V}[X_2]}} \right| \leq 1.$$

Remark 3.7 *Let $X_1, \ldots, X_m$ be centered $SG(\sigma^2)$ random variables. Then we see with a union bound that*

$$P\left(\max_i |X_i| \geq t\right) = P\left(\cup_{i=1}^m \{|X_i| \geq t\}\right) \leq \sum_{i=1}^m P(|X_i| \geq t) \leq m e^{-t^2/(2\sigma^2)} = \exp\left(\frac{-t^2}{2\sigma^2} + \log(m)\right).$$

3.1.5 Hoeffding inequality

Theorem 3.8 (Hoeffding inequality) *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in SG(\sigma_i^2)$ for all i. Then,*

$$P\left(\left|\sum_{i=1}^n \frac{X_i - \mathbb{E}[X_i]}{n}\right| \geq t\right) \leq 2\exp\left(\frac{-n^2 t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Usually, we have $\sigma_i^2 = \sigma^2$ for all i. In this case, it holds that

$$2\exp\left(\frac{-n^2 t^2}{2\sum_{i=1}^n \sigma_i^2}\right) = 2\exp\left(\frac{-nt^2}{2\sigma^2}\right).$$

Example 3.9 (Hoeffding for Bernoulli RV) *Let $X_1, \ldots, X_n$ be independent RV with $X_i \sim Bernoulli(p_i)$ for some $p_i \in (0, 1)$. Then, $X_i \in SG(1/4)$ and thus,*

$$P\left(\left|\bar{X}_n - \bar{p}_n\right| \geq t\right) \leq 2\exp\left(-2nt^2\right)$$

where $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{p}_n := \frac{1}{n}\sum_{i=1}^n p_i$. Thus, we have that

$$\left|\bar{X}_n - \bar{p}_n\right| \leq \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}$$

with probability at least $1 - \delta$.

References

[BK00] V.V. Buldygin and Yu.V. Kozachenko, "Metric Characterization of Random Variables and Random Processes", 2000.