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

Lecture 4: January 29

Lecturer: Alessandro Rinaldo *Scribes: Tudor Manole*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

4.1 Comparing Hoeffding and Chernoff Bounds

Last time we proved Hoeffding's Inequality, which states that for any independent random variables $X_1, \dots, X_n$ with $X_i \in \text{SG}(\sigma_i)$ and $\mu_i = \mathbb{E}(X_i)$, $i = 1, \dots, n$, we have

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2 n^2}{2\sum_{i=1}^{n}\sigma_i^2}\right), \quad \forall t > 0.$$

In particular, if $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$, for some $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, we have

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

Example 1. Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, for some $0 < p < 1$. Set $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. We know $(X_i - p) \in \text{SG}(1/4)$ by results from the last class, for all $i = 1, \dots, n$. Therefore, by Hoeffding's inequality,

$$\mathbb{P}\left\{|\bar{X}_n - p| \geq t\right\} \leq 2\exp\left(-2nt^2\right), \quad \forall t > 0.$$

To invert this bound, set $\delta = 2\exp(-2nt^2) \in (0, 1)$. Solving for t, we see that with probability at least $1 - \delta$,

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

One often sets $\delta = \frac{1}{n^c}$ for some $c > 0$. For example, we have that with probability at least $1 - \frac{1}{n}$,

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

Alternatively, one has

$$\bar{X}_n - p = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Hoeffding's inequality is not, however, the sharpest concentration inequality in general. Note that the above calculations would similarly hold for any bounded random variable. Therefore, we could hope that we would obtain a tighter inequality by using more information about the X_i than merely their boundedness. It indeed turns out that Chernoff's Inequality yields an improvement on Hoeffding's inequality whenever p is small. Recall that if $X_1, \ldots, X_n$ are random variables supported in $[0,1]$, such that $\mathbb{E}(X_i) = p_i$, then Chernoff's Inequality yields

$$\mathbb{P}\left\{\sum_{i=1}^n X_i \geq (1+\epsilon)\sum_{i=1}^n p_i\right\} \leq \begin{cases} \exp\left\{-\frac{\epsilon^2 \sum_{i=1}^n p_i}{3}\right\}, & \epsilon \in (0,1] \\ \exp\left\{-\frac{\epsilon^2 \sum_{i=1}^n p_i}{2+\epsilon}\right\}, & \epsilon > 1 \end{cases}, \tag{4.1}$$

and,

$$\mathbb{P}\left\{\sum_{i=1}^n X_i \leq (1-\epsilon)\sum_{i=1}^n p_i\right\} \leq \exp\left\{-\frac{\epsilon^2 \sum_{i=1}^n p_i}{2}\right\}, \quad \forall \epsilon \in (0,1). \tag{4.2}$$

See also [HR90] for more on these inequalities.

Example 2 (Example 1 Continued)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$. Hoeffding's inequality gave

$$p - \bar{X}_n \leq \sqrt{\frac{1}{2n}\log(1/\delta)}, \tag{4.3}$$

with probability at least $1-\delta$. On the other hand, (4.1) yields

$$\mathbb{P}\left\{p - \bar{X}_n \geq \epsilon p\right\} \leq \exp\left(-\frac{np\epsilon^2}{2}\right), \quad \forall \epsilon \in (0,1).$$

Therefore, provided $p \geq \frac{2}{n}\log(1/\delta)$,

$$p - \bar{X}_n \leq \sqrt{\frac{2p}{n}\log(1/\delta)}, \tag{4.4}$$

with probability at least $1-\delta$. Clearly, when p is fixed, Chernoff's bound (4.4) is nearly the same as Hoeffding's bound (4.3), as far as rates are concerned. On the other hand, if we let $p \equiv p_n$ so that $p_n \to 0$, then (4.4) provides a significant improvement upon (4.3). This happens becase the variance of a Bernoulli random variable is upper bounded by p, so if $p_n \to 0$, its the variance is shrinking as the sample size grows. Chernoff's inequality incorporates this information, thus yielding a tighter bound.

See [SN06], for an application of Bernoulli random variables with parameter $p_n \to 0$.

4.2 Equivalent Definitions of Sub-Gaussian Random Variables

Sub-Gaussianity can equivalently be characterized using Orlicz norms, as will be explored in the second assignment. It turns out that Sub-Gaussian random variables are also uniquely characterized by their moments, which we describe in the following proposition.

Proposition 1. *Let $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ be the Gamma function. If $X \in SG(\sigma^2)$, then*

$$\mathbb{E}[|X|^p] \leq p2^{p/2}\sigma^p\Gamma(p/2), \quad \forall p > 0.$$

In particular, there exists $C > 0$ not depending on p such that

$$(\mathbb{E}[|X|^p])^{\frac{1}{p}} \leq C\sigma\sqrt{p}.$$

Proof. We have,

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \int_0^\infty \mathbb{P}(|X|^p \geq u)du \\
&= \int_0^\infty \mathbb{P}(|X| \geq u^{\frac{1}{p}})du \\
&\leq 2\int_0^\infty \exp\left\{-\frac{u^2}{2\sigma^2}\right\}du \\
&= \mathbb{E}[|X|^p] \leq (2\sigma^2)^{\frac{p}{2}}p\int_0^\infty e^t t^{\frac{p}{2}-1}dt \qquad \left(\text{where } t = \frac{u^{\frac{2}{p}}}{2\sigma^2}\right) \\
&= (2\sigma^2)^{\frac{p}{2}}p\Gamma\left(\frac{p}{2}\right).
\end{aligned}$$

□

Remark. For example, if $X \sim \mathcal{N}(0, \sigma^2)$, we have

$$\mathbb{E}[|X|^p] = \frac{\sigma^p 2^{\frac{p}{2}}\Gamma(\frac{p+1}{2})}{\sqrt{\pi}}.$$

4.3 Sub-Exponential Random Variables.

In this section, we consider a broader class of distributions than the Sub-Gaussian family, call the Sub-Exponential family. We will see that interesting tail bounds can still be derived for random variables belonging to this collection. One motivation for its definition is that Sub-Gaussian random variables are not closed under taking squares, in the sense that $X \in \mathrm{SG}(\sigma^2)$ does not imply X^2 is Sub-Gaussian. For example, the square of a standard Gaussian is a Chi-Squared random variable, which cannot be Sub-Gaussian since its moment generating function is not defined on the entire real line.

Example 3. Let $X \sim \text{Laplace}(b)$ for $b > 0$. Then it can be shown that

$$\mathbb{P}(|X| \geq t) \leq \exp(-tb), \quad \forall t > 0.$$

This is a different tail behaviour than what we are used to for Sub-Gaussian random variables, and indeed,

we note that $X \notin \mathrm{SG}(\sigma^2)$ since its moment generating function is only defined on a subset of the real line:

$$\mathbb{E}[e^{\lambda X}] = \frac{1}{1-\lambda^2} b^2, \quad \forall |\lambda| < \frac{1}{b}.$$

Definition 1 (Sub-Exponential Random Variable)**.** We say that a random variable X is Sub-Exponential with parameters $\nu, \alpha > 0$, and we write $X \in \mathrm{SE}(\nu^2, \alpha)$, if

$$\mathbb{E}\left[e^{\lambda(X - \mathbb{E}(X))}\right] \leq e^{\frac{\lambda^2 \nu^2}{2}}, \quad \forall |\lambda| < \frac{1}{\alpha}.$$

Remark. An immediate consequence of the definition is that $\mathrm{SG}(\sigma^2) \subseteq \mathrm{SE}(\sigma^2, 0)$. Thus, all Sub-Gaussian random variables are also Sub-Exponential.

Example 4. Let $Z \sim \mathcal{N}(0,1)$, and $X = Z^2 \sim \chi^2_{(1)}$, $\mathbb{E}(X) = 1$. Let $\lambda < \frac{1}{2}$. Then,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(X-1)}\right] &= \frac{1}{\sqrt{2\pi}} e^{-\lambda} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}(1-2\lambda)} dz. \\
&= e^{-\lambda} \frac{1}{\sqrt{1-2\lambda}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \qquad (y = z\sqrt{1-2\lambda}) \\
&= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \\
&\leq \exp\left\{\frac{\lambda^2}{1-2\lambda}\right\} \qquad (\star) \\
&\leq \exp\left\{\frac{4\lambda^2}{2}\right\}.
\end{aligned}$$

Thus, $X \in \mathrm{SE}(4,4)$. Note that $(\star)$ follows from the following elementary inequality

$$-\log(1-u) - u \leq \frac{u^2}{2(1-u)}, \quad \forall u \in (0,1),$$

with $u = 2\lambda$. Also, it is possible to show that

$$\mathbb{P}\left(X - 1 > 2t + 2\sqrt{t}\right) \leq e^{-t}, \quad \forall t > 0.$$

Note that there is an additional term $2t$ here compared with the usual bounds we had for Sub-Gaussian random variables. This accounts for the possibly thicker tails of X.

Properties of SE(ν^2, α).

(P1) Squares and products of centered sub-Gaussians are Sub-Exponential:

$$X \in \mathrm{SG}(\sigma^2) \Longrightarrow X^2 \in \mathrm{SE}(256\sigma^4, 16\sigma^2).$$

(P2) Suppose X is a random variable with $\mathrm{Var}[X] = \sigma^2$ and $|X - \mathbb{E}(X)| \leq b$ almost everywhere, for some

$b > 0$. Then, $X \in \mathrm{SE}(2\sigma^2, 2b)$. Unlike Sub-Gaussian bounded random variables, the variance of X appears in the Sub-Exponential parameters.

Proof. Let $|\lambda| < \frac{1}{2b}$. Then,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(X-\mathbb{E}(X))}\right] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{\lambda^n \mathbb{E}[(X-\mathbb{E}(X))^k]}{k!} \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2} \\
&= 1 + \frac{\lambda^2\sigma^2}{2} \sum_{k=0}^{\infty} [|\lambda| b]^k] \\
&= 1 + \frac{\lambda^2\sigma^2}{2} \frac{1}{1-|\lambda| b} && \text{(Since } |\lambda| < \frac{1}{b}\text{)} \\
&\leq \exp\left\{\frac{\lambda^2\sigma^2}{1-|\lambda| b}\right\} \\
&\leq \exp\left\{\lambda^2\sigma^2\right\}.
\end{aligned}$$

□

Tail Bounds for Sub-Exponential Random Variables. We are now in a position to derive a tail bound for Sub-Exponential random variables.

Theorem 1. *Let $X \in SE(\nu^2, \alpha)$, and $t > 0$. Then,*

$$\mathbb{P}\left\{X - \mathbb{E}(X) \geq t\right\} \leq \begin{cases} \exp\left\{-\frac{t^2}{2\nu^2}\right\}, & t \leq \frac{\nu^2}{\alpha} \\ \exp\left\{-\frac{t}{2\alpha}\right\}, & t > \frac{\nu^2}{\alpha} \end{cases}.$$

Equivalently,

$$\mathbb{P}\{X - \mathbb{E}(X) \geq t\} \leq \exp\left\{-\frac{1}{2}\min\left(\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right)\right\}.$$

References

[SN06] C. Scott and R. Nowak, "Minimax-optimal classification with dyadic decision trees," *IEEE transactions on information theory*, 2006, 52, 1335–1353.

[HR90] T. Hagerup and C. Rub, "A guided tour of Chernoff bounds," *Information processing letters*, 1990, 33, 305–308.