IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: P Rigollet
Scribes: Adam Block & Patrik Gerber

Lecture 3
Feb. 11, 2020

Goals: In the previous lecture, we introduced the notion of subGaussian random variables and explored some of their basic properties, including their connection to an Orlicz norm and Hoeffding's inequality. In this lecture, we apply analogous methods to consider a wider class of random variables: the subExponential distributions. We then introduce another Orlicz norm and prove basic properties about the subExponential random variables and their connection to the norm. We conclude by proving Bernstein's inequality, the analogue of Hoeffding's inequality in this more general regime.

1. SUBEXPONENTIAL RANDOM VARIABLES

Last lecture, we discussed subGaussian random variables and some of their properties. This is a nice, large class of random variables including, for example, the set of random variables with compact support (as seen by Hoeffding's lemma). Unfortunately, some random variables are not subGaussian. For instance, if we consider $Z \sim \mathcal{N}(0,1)$ then we might consider Z^2. Then we can directly compute the moment generating function. Indeed, set $s < \frac{1}{2}$ and we see

$$M(s) = \mathbb{E}\left[e^{sZ^2}\right] = \frac{1}{\sqrt{2\pi}} \int e^{sx^2 - \frac{1}{2}x^2} \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{x^2}{\frac{2}{1-2s}}\right) \mathrm{d}x \tag{1.1}$$

$$= \frac{1}{\sqrt{1-2s}} \frac{1}{\sqrt{\frac{2\pi}{1-2s}}} \int \exp\left(-\frac{x^2}{\frac{2}{1-2s}}\right) \mathrm{d}x = \frac{1}{\sqrt{1-2s}} \tag{1.2}$$

where the last equality follows from the fact that we are integrating the density of a centred Gaussian with variance $\frac{1}{1-2s}$. We note further that $M(s)$ is only finite for $s < \frac{1}{2}$ so we clearly cannot have $M(s) \leq e^{cs^2}$ for all s. In this lecture we consider a more general class of random variables, with heavier tails.

Definition (subExponential Random Variables): A random variable X is subExponential with parameter $\lambda > 0$ if $\mathbb{E}[X] = 0$ and for all $|s| \leq \frac{1}{\lambda}$, we have

$$M(s) = \mathbb{E}\left[e^{sX}\right] \leq e^{s^2\lambda^2} \tag{1.3}$$

We abbreviate this by saying that $X \sim \mathsf{subE}(\lambda)$.

We note as a side remark that this definition is a weakening of that of a subGaussian random variable, where the inequality of moment generating functions must hold for the entire real line. Much as in the case of subGaussian random variables, we have equivalent definitions providing tail bounds and moment estimates for subExponential random variables:

Proposition: Let X be a centred random variable. Then, the following are equivalent:

(i) There exists a constant $c_1 > 0$ such that for $|s| \leq \frac{1}{c_1}$, we have $\mathbb{E}\left[e^{sX}\right] \leq e^{s^2 c_1^2}$.

(ii) There is a constant $c_2 > 0$ such that $\mathbb{P}(|X| > t) \leq 2\, e^{-c_2 t}$.

(iii) There is a constant $c_3 > 0$ such that $\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq c_3 p$ for all $p \in \mathbb{N}$

Proof. We prove this result in much the same way that we proved the analogous proposition in the case of subGaussian random variables.

(i) implies (ii): We apply a Chernoff bound. Indeed, for $t > 0$, we have for $0 < s < \frac{1}{c_1}$

$$\mathbb{P}(X > t) \leq \mathbb{E}\left[e^{sX}\right] e^{-st} \leq e^{c_1^2 s^2 - st} \tag{1.4}$$

Now, we minimize $c_1^2 s^2 - st$ for $s \in \left(0, \frac{1}{c_1}\right)$. We know that this is a parabola and so it either attains its minimum at its vertex or at the right end point of this interval, depending on whether the abscissa of its vertex falls in this interval. Elementary calculus tells us that the vertex is at $s = \frac{t}{2c_1^2}$. Thus if $t > 2c_1$ then we have

$$\mathbb{P}(X > t) \leq e^{1 - \frac{t}{c_1}} \leq e^{-\frac{t}{2c_1}}$$

Using the same argument for $-X$ together with a union bound, we get that for $t > 2c_1$, we have

$$\mathbb{P}(|X| > t) \leq 2e^{-\frac{t}{2c_1}}$$

To deal with the case where $t \leq 2c_1$, observe that

$$\mathbb{P}(|X| > t) \leq 1 \leq \frac{2}{\sqrt{e}} \leq 2e^{-\frac{t}{4c_1}}\,.$$

where in the last inequality, we used $t \leq 2c_1$.

The above two displays yield that for every $t \geq 0$, it holds

$$\mathbb{P}(|X| > t) \leq 2e^{-c_2 t}\,, \qquad c_2 = \frac{1}{4c_1}\,.$$

(ii) implies (iii): We apply Fubini's theorem, in the same way that we did for the subGaussian case. Indeed, we note

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty \mathbb{P}\left(|X|^p > u\right) \mathrm{d}u = \int_0^\infty p t^{p-1} \mathbb{P}(|X| > t) \mathrm{d}t \leq 2 \int_0^\infty p t^{p-1} e^{-c_2 t} \mathrm{d}t \tag{1.5}$$

where we substituted $u = t^p$ and took advantage of the fact that $x \mapsto x^p$ is increasing on the positive half line, followed by the tail bound. Now, let $r = c_2 t$ to get that

$$\mathbb{E}[|X|^p] \leq \frac{2p}{c_2^p} \int_0^\infty r^{p-1} e^{-r} \mathrm{d}r = \frac{2p}{c_2^p} \Gamma(p) \tag{1.6}$$

Taking p^{th} roots and recalling that $\Gamma(p) \leq p^p$ gives

$$||X||_p \leq \frac{(2p)^{\frac{1}{p}}}{c_2} p \tag{1.7}$$

Finally, note that $(2p)^{\frac{1}{p}}$ converges as $p \to \infty$ so we may take a supremum to get that $||X||_p \leq c_3 p$ as desired.

(iii) implies (i): Let $s > 0$ and recall the definition of the exponential

$$e^{sX} = \sum_{p=0}^{\infty} \frac{s^p X^p}{p!}. \tag{1.8}$$

By the hypothesis we have

$$\mathbb{E}\ e^{sX} \leq \mathbb{E}e^{s|X|} = \sum_{p=0}^{\infty} s^p \frac{\mathbb{E}|X|^p}{p!} \leq \sum_{p=0}^{\infty} \frac{(c_3 sp)^p}{p!}. \tag{1.9}$$

We can easily find the radius of convergence of the above power series (for example using the ratio-test), which yields that $\mathbb{E}e^{s|X|} < \infty$ provided that $s < \frac{1}{ec_3}$. Thus by the Dominated Convergence Theorem, we have for $s < \frac{1}{ec_3}$ that

$$\mathbb{E}\left[e^{sX}\right] = \sum_{p=0}^{\infty} s^p \frac{\mathbb{E}X^p}{p!} \tag{1.10}$$

$$= 1 + \sum_{p=2}^{\infty} s^p \frac{\mathbb{E}X^p}{p!} \tag{1.11}$$

where we used that $\mathbb{E}X = 0$. Recall that Stirling's approximation says that $p! \geq \left(\frac{p}{e}\right)^p$. Using this and the hypothesis, we get

$$\mathbb{E}\left[e^{sX}\right] \leq 1 + \sum_{p=2}^{\infty} s^p \frac{(c_3 p)^p}{(p/e)^p} \tag{1.12}$$

$$= 1 + \sum_{p=2}^{\infty} s^p (ec_3)^p \tag{1.13}$$

$$= 1 + \frac{(ec_3 s)^2}{1 - ec_3 s}. \tag{1.14}$$

Let us further restrict s to be $s < \frac{1}{2ec_3}$. Using the trivial inequality $1 + x \leq e^x$ for all $x \geq 0$ we obtain

$$1 + \frac{(ec_3 s)^2}{1 - ec_3 s} \leq 1 + 2(ec_3 s)^2 \tag{1.15}$$

$$\leq e^{2(ec_3 s)^2} \tag{1.16}$$

so that $X \sim \mathsf{subE}(2ec_3)$ as required. □

Important Remark: Notice that we proved something slightly stronger: we proved that there exists a universal constant $c > 0$ such that

1. if (i) holds with c_1 then (ii) holds with $\frac{1}{c_2} \leq cc_1$
2. if (ii) holds with c_2 then (iii) holds with $c_3 \leq \frac{c}{c_2}$

3. if (iii) holds with c_3 then (i) holds with $c_1 \leq c\, c_3$.

In other words, all constants are within constant factor of each other.

Just as in the case of subGaussian random variables, we have another condition, equivalent to the above, in terms of the Orlicz norm. We first need to introduce the relevant Orlicz norm, however.

Definition (Orlicz ψ_1-norm): The Orlicz ψ_1-norm of a random variable X is

$$\|X\|_{\psi_1} = \inf\left\{t > 0 \,:\, \mathbb{E}\left[e^{\frac{|X|}{t}}\right] \leq 2\right\}. \tag{1.17}$$

Just like for subGaussian random variables, subExponentiality can be characterized using an Orlicz norm.

Proposition: There exist universal constants $0 < c_1, c_2 < \infty$ such that for any centered random variable X,

$$X \sim \mathsf{subE}(1) \quad \Longrightarrow \quad \|X\|_{\psi_1} \leq c_1. \tag{1.18}$$

and

$$\|X\|_{\psi_1} \leq 1 \quad \Longrightarrow \quad X \sim \mathsf{subE}(c_2). \tag{1.19}$$

Proof. Suppose that $X \sim \mathsf{subE}(1)$. In this case we have that for all $|s| < 1$,

$$\mathbb{E}\left[e^{sX}\right] \leq e^{s^2} \tag{1.20}$$

Suppose that $s \sim \gamma\rho$ where $\gamma \in (0,1)$ to be specified later and $\rho \sim \mathsf{Rad}\left(\frac{1}{2}\right)$. By Fubini's theorem, we have that

$$\mathbb{E}_s\left[\mathbb{E}_x\left[e^{sX}\right]\right] = \mathbb{E}_x\left[\mathbb{E}_s\left[e^{sX}\right]\right] = \mathbb{E}_x\left[\frac{e^{\gamma x} + e^{-\gamma x}}{2}\right] = \mathbb{E}[\cosh(\gamma X)] \tag{1.21}$$

We prove in the lemma below that $\cosh(\gamma x) \geq \frac{2}{3} e^{\frac{|x|}{2}}$. Thus we get that

$$\mathbb{E}[\cosh(\gamma X)] \geq \frac{2}{3}\mathbb{E}\left[e^{\frac{\gamma|X|}{2}}\right] \tag{1.22}$$

Rearranging, we have

$$\mathbb{E}\left[e^{\frac{\gamma|X|}{2}}\right] \leq \frac{3}{2}\mathbb{E}_s\left[e^{s^2}\right] = \frac{3}{2}e^{\gamma^2} \tag{1.23}$$

because $s^2 = \gamma^2$ because $|\rho| = 1$. Let $\gamma = \sqrt{\log\frac{4}{3}}$ and note that as $1 < \frac{4}{3} < e$ we have that $0 < \gamma < 1$. Then we see that

$$\mathbb{E}\left[e^{\frac{\gamma|X|}{2}}\right] \leq 2 \tag{1.24}$$

Thus by definition of the Orlicz norm, $\|X\|_{\psi_1} \leq \frac{2}{\gamma} = c_1$ as desired.

Suppose that $\|X\|_{\psi_1} \leq 1$. Then by Markov's inequality,

$$\mathbb{P}(|X| > t) = \mathbb{P}\left(e^{\frac{|X|}{\|X\|_{\psi_1}}} > e^{\frac{t}{\|X\|_{\psi_1}}}\right) \leq \mathbb{E}\left[e^{\frac{|X|}{\|X\|_{\psi_1}}}\right] e^{-\frac{t}{\|X\|_{\psi_1}}} \leq 2e^{-\frac{t}{\|X\|_{\psi_1}}} \leq 2e^{-t}. \quad (1.25)$$

By Remark 1 we know that there exists a universal constant c_2 such that the above implies that $X \sim \mathsf{subE}(c_2)$.

□

Before proving Bernstein's inequality, we wish to compare the norm introduced last lecture ($\|\cdot\|_{\psi_2}$) with that introduced this lecture ($\|\cdot\|_{\psi_2}$). We have the following proposition:

Proposition: Let X and Y be random variables. Then

(i) $\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$

(ii) $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$

Proof. **(i):** By definition of ψ_2, we have

$$\mathbb{E}\left[e^{\frac{X^2}{\|X\|_{\psi_2}}}\right] \leq 2 \quad (1.26)$$

and thus by the definition of the ψ_1 norm, we have $\|X\|_{\psi_2}^2 \geq \left\|X^2\right\|_{\psi_1}$. Moreover, for all $t > 0$ such that

$$\mathbb{E}\left[e^{\frac{X^2}{t^2}}\right] \leq 2 \quad (1.27)$$

we have $\|X\|_{\psi_2}^2 \leq t^2$ by definition. In particular, this holds for $t^2 = \left\|X^2\right\|_{\psi_1}$ and so $\|X\|_{\psi_2}^2 \leq \left\|X^2\right\|_{\psi_1}$, proving the other side. Thus they are equal.

(ii): Without loss of generality, by homogeneity of norms, we may assume that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. Recall that because $(X-Y)^2 \geq 0$, we have $|XY| \leq \frac{X^2}{2} + \frac{Y^2}{2}$. Thus we have by Cauchy-Schwarz,

$$\mathbb{E}\left[e^{|XY|}\right] \leq \mathbb{E}\left[e^{\frac{X^2}{2}} e^{\frac{Y^2}{2}}\right] \leq \sqrt{\mathbb{E}\left[e^{X^2}\right] \mathbb{E}\left[e^{Y^2}\right]} \leq 2 \quad (1.28)$$

because we have $\mathbb{E}\left[e^{X^2}\right] \leq 2$ and similarly for Y by the fact that their Orlicz norms are both one. But this implies that $\|XY\|_{\psi_1} \leq 1$ by definition. □

We note in passing that the second part of the above proposition together with the antecedent result and its analogy from last lecture together imply that if we centre the product of two subGaussian random variables, then we have a subExponential random variable. Thus, the example that we saw above, where $Z \sim \mathcal{N}(0,1)$ is subGaussian, tells us that $Z^2 - 1$ is subExponential. Indeed, we have

$$\left\|Z^2 - 1\right\|_{\psi_1} \leq \left\|Z^2\right\|_{\psi_1} + \|1\|_{\psi_1} \leq \|Z\|_{\psi_2}^2 + \|1\|_{\psi_1} < \infty \quad (1.29)$$

and so the result above implies that $Z^2 - 1$ is subExponential.

2. BERNSTEIN'S INEQUALITY

In this section we develop a tail bound for sums of independent subexponential random variables similar to Hoeffding's inequality.

Theorem (Bernstein's inequality): Let $X_1, ..., X_n$ be independent subExponential random variables and let

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \|X\|_{\psi_1}^2 \qquad \text{and} \qquad \sigma_{\max} = \max_{1\le i\le n} \|X\|_{\psi_1}. \tag{2.30}$$

Then for every $t \geq 0$ the following inequality holds

$$\mathbb{P}(|\overline{X}_n| > t) \leq 2\exp\left(-Cn\left[\frac{t^2}{\overline{\sigma}^2} \wedge \frac{t}{\sigma_{\max}}\right]\right) \tag{2.31}$$

for some positive constant C.

Proof. Since X_i is subExponential, there exists a universal constant $c > 0$ such that

$$X_i \sim \mathsf{subE}(c\,\|X_i\|_{\psi_1}) \tag{2.32}$$

holds for each i. In particular, for any $0 < s < \frac{1}{c\sigma_{\max}}$ we have

$$\mathbb{E}\left[e^{sX_i}\right] \leq e^{s^2c^2\|X_i\|_{\psi_1}^2}. \tag{2.33}$$

Once again we can use Markov's inequality and our control of the moment-generating function to derive the tail-bounds. For any $0 < s < \frac{1}{c\sigma_{\max}}$ we have

$$\begin{aligned}
\mathbb{P}(\overline{X}_n > t) &= \mathbb{P}\left(e^{s\sum_{i=1}^{n} X_i} > e^{nst}\right) && (2.34)\\
&\leq \mathbb{E}\left[e^{s\sum_{i=1}^{n} X_i}\right] e^{-stn} && (2.35)\\
&= e^{-stn}\prod_{i=1}^{n}\mathbb{E}\left[e^{sX_i}\right] && (2.36)\\
&\leq e^{-stn}\prod_{i=1}^{n} e^{c^2s^2\|X_i\|_{\psi_1}^2} && (2.37)\\
&= \exp\left(c^2s^2n\overline{\sigma}^2 - nst\right). && (2.38)
\end{aligned}$$

As usual, the next step is to minimize the RHS. Looking at the exponent, differentiating with respect to s and setting equal to 0 we get the minimizer

$$s^* = \frac{t}{2c^2\overline{\sigma}^2}. \tag{2.39}$$

Now, it might be that s^* falls outside the interval $(0, (c\sigma_{\max})^{-1})$ in which case the best possible bound is given by plugging in $s = (c\sigma_{\max})^{-1}$. In the latter case, the RHS becomes

$$\exp\left(\frac{n\overline{\sigma}^2}{\sigma_{\max}^2} - \frac{nt}{c\sigma_{\max}}\right) \leq \exp\left(-\frac{nt}{2c\sigma_{\max}}\right), \tag{2.40}$$

where we substituted $\frac{t}{2c} \geq \frac{\overline{\sigma}^2}{\sigma_{\max}}$. Summarising, we have

$$\mathbb{P}(\overline{X}_n > t) \leq \begin{cases} \exp\left(-\frac{nt^2}{4c^2\overline{\sigma}^2}\right) & \text{if } t \leq \frac{2c\overline{\sigma}^2}{\sigma_{\max}} \\ \exp\left(-\frac{nt}{2c\sigma_{\max}}\right) & \text{otherwise.} \end{cases} \tag{2.41}$$

This can conveniently be written as the expression

$$\mathbb{P}\left(\overline{X}_n > t\right) \leq \exp\left(-n\left[\frac{t^2}{4c^2\overline{\sigma}^2} \wedge \frac{t}{2c\sigma_{\max}}\right]\right). \tag{2.42}$$

Taking $C = \frac{1}{2c} \wedge \frac{1}{4c^2}$ together with a union bound yields (2.31). □

Let us compare the above result to what we've seen from the Central Limit Theorem (CLT). We get

$$\mathbb{P}\left(\sqrt{n}\overline{X}_n > t\right) \leq \exp\left(-\left[\frac{t^2}{4c^2\overline{\sigma}^2} \wedge \frac{\sqrt{n}t}{2c\sigma_{\max}}\right]\right). \tag{2.43}$$

We see that there is a window of width $2c\sqrt{n}$ where the rescaled average has subGaussian tails in line with the CLT while outside that growing window we only get a subExponential tail bound.

Summary:

- **subExponential random variables:** TFAE
  1. $\exists\, c_1 > 0$ such that $\mathbb{E}\left[e^{sX}\right] \leq e^{s^2c_1^2}$ for all $|s| < \frac{1}{c_1}$.
  2. $\exists\, c_2 > 0$ such that $\mathbb{P}(|X| > t) \leq 2e^{-c_2 t}$ for all $t \geq 0$.
  3. $\exists\, c_3 > 0$ such that $\|X\|_p \leq c_3 p$ for all $p \in bN$.
- **Orlicz ψ_1-norm**: There exist constants c_1, c_2 such that for any centered random variable X
  1. $X \sim \mathsf{subE}(1) \quad \Longrightarrow \quad \|X\|_{\psi_1} \leq c_1$
  2. $\|X\|_{\psi_1} \leq 1 \quad \Longrightarrow \quad X \sim (c_2)$.
- **Bernstein's Inequality**: For $X_1, ..., X_n$ independent subExponential random variables

$$\mathbb{P}\left(|\overline{X}_n| > t\right) \leq 2\exp\left(-Cn\left[\frac{t^2}{\overline{\sigma}^2} \wedge \frac{t}{\sigma_{\max}}\right]\right) \tag{2.44}$$

  for all $t \geq 0$ where

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \|X\|_{\psi_1}^2 \qquad \text{and} \qquad \sigma_{\max} = \max_{1 \leq i \leq n} \|X\|_{\psi_1}. \tag{2.45}$$