THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 384 11: Theoretical Statistics

Lecture 4: Sub-gaussian and sub-exponential random variables

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

www.cs.cmu.edu/∼psarkar/teaching

Sub-Gaussian random variables

Theorem

For $X_1, \ldots, X_n$ independent sub-gaussian random variables with sub-gaussian parameters σ_i and $E[X_i] = \mu_i$, for $\forall t > 0$,

$$P\left(\sum_i (X_i - \mu_i) \geq t\right) \leq e^{-\frac{t^2}{2\sum_i \sigma_i^2}}$$

- If $X_i \in [a, b]$, $E[X_i] = 0$, using Hoeffding's lemma we have: $\sigma_i^2 = (b-a)^2/4$.
- So, the above theorem immediately gives the original Hoeffding inequality back.

$$P\left(\sum_i X_i \geq t\right) \leq e^{-\frac{2t^2}{n(b-a)^2}}$$

Sub-exponential random variables

Definition

X is sub-exponential with parameters (ν, b) if, $\forall |\lambda| < 1/b$,

$$\log M_{X-\mu}(\lambda) \le \frac{\lambda^2 \nu^2}{2}$$

Examples:

- Sub-Gaussian X with parameter σ is sub-exponential with parameters (σ, b) $\forall b > 0$.
- How about the converse?

Sub-exponential but not sub-gaussian

Example

Let $Z \sim N(0,1)$ and consider the random variable $X = Z^2$. For $\lambda < 1/2$, we have:

- The MGF is only defined for $\lambda < 1/2$. So this is a sub-exponential random variable with parameter $(2,4)$, but not a sub-gaussian random variable.
- We use $\log(1+x) \geq \frac{x}{2}\frac{2+x}{1+x}$ for $-1 \leq x \leq 0$.

Sub-exponential but not sub-gaussian

Example

Let $Z \sim N(0,1)$ and consider the random variable $X = Z^2$. For $\lambda < 1/2$, we have:

$$
\begin{aligned}
E[e^{\lambda(X-1)}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(z^2-1)} e^{-z^2/2} dz \\
&= e^{-\lambda} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2(1-2\lambda)/2} dz \\
&= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \\
&\leq e^{2\lambda^2} \qquad \forall |\lambda| < 1/4
\end{aligned}
$$

- The MGF is only defined for $\lambda < 1/2$. So this is a sub-exponential random variable with parameter $(2, 4)$, but not a sub-gaussian random variable.
- We use $\log(1+x) \geq \frac{x}{2} \frac{2+x}{1+x}$ for $-1 \leq x \leq 0$.

Concentration

Theorem

Let X be a sub-exponential random variable with parameters (ν, b). Then,

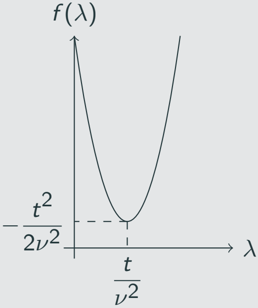
$$P(X \geq \mu + t) \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \textit{if } 0 \leq t \leq \frac{\nu^2}{b} \\ e^{-\frac{t}{2b}} & \textit{if } t \geq \frac{\nu^2}{b} \end{cases}$$

- For small t this is sub-gaussian in nature, whereas for large t the exponent decays linearly with t.

Proof

Proof.

$$
\begin{aligned}
P(X \geq t) &\leq \inf_{\lambda \geq 0} e^{-\lambda t} E[e^{\lambda X}] \\
&\leq \inf_{\lambda \geq 0} \exp\left(\underbrace{-\lambda t + \lambda^2 \nu^2 / 2}_{f(\lambda)}\right) \qquad \text{When } 0 \leq \lambda < 1/b
\end{aligned}
$$



- If $\frac{t}{\nu^2} \leq \frac{1}{b}$,

  $\inf_{\lambda \geq 0} f(\lambda) = f(t/\nu^2) = -\frac{t^2}{2\nu^2}$

- If $\frac{t}{\nu^2} > \frac{1}{b}$, then $f(\lambda)$ is minimized at the boundary $\lambda' = 1/b$.

  $f(\lambda') = -t/b + \nu^2/2b^2 \leq -\frac{t}{2b}$

$\square$

A moment condition

- Its typically difficult to check if a random variable is subexponential.

A moment condition

- Its typically difficult to check if a random variable is subexponential.
- We can also characterize a random variable by how quickly its moments grow.

Definition

A random variable with mean μ and variance σ^2 satisfies the Bernstein condition with parameter $b > 0$, if $E[(X-\mu)^k] \leq \frac{1}{2}k!\sigma^2 b^{k-2}$ for $k \geq 2$.

A moment condition

- Its typically difficult to check if a random variable is subexponential.
- We can also characterize a random variable by how quickly its moments grow.

Definition

A random variable with mean μ and variance σ^2 satisfies the Bernstein condition with parameter $b > 0$, if $E[(X-\mu)^k] \leq \frac{1}{2} k! \sigma^2 b^{k-2}$ for $k \geq 2$.

A moment condition

- Its typically difficult to check if a random variable is subexponential.
- We can also characterize a random variable by how quickly its moments grow.

Definition

A random variable with mean μ and variance σ^2 satisfies the Bernstein condition with parameter $b > 0$, if $E[(X-\mu)^k] \leq \frac{1}{2}k!\sigma^2 b^{k-2}$ for $k \geq 2$.

- A bounded random variable with $|X - \mu| \leq b$ satisfies the above.

Bernstein's condition and the sub-exponential property

Theorem

If X ($E[X] = \mu, var(X) = \sigma^2$) satisfies the Bernstein condition with parameter $b > 0$, then X is sub-exponential with $(\sqrt{2}\sigma, 2b)$.

Proof.

$$
\begin{aligned}
E[e^{\lambda(X-\mu)}] &= \sum_{k=0}^{\infty} \frac{\lambda^k E[(X-\mu)^k]}{k!} \\
&= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{|\lambda|^k \sigma^2 b^{k-2}}{2} \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2}\left(1 + \sum_{k=1}^{\infty} (|\lambda| b)^k\right) \\
&= 1 + \frac{\lambda^2\sigma^2}{2(1-|\lambda| b)} \qquad \text{For } |\lambda| < 1/b \\
&\leq e^{\frac{\lambda^2\sigma^2}{2(1-|\lambda|b)}} \leq e^{\lambda^2\sigma^2} = e^{\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}} \qquad \text{For } |\lambda| < 1/2b
\end{aligned}
$$

Bernstein's inequality

Theorem

If X with mean μ and variance σ^2 satisfies the Bernstein condition with parameter $b > 0$, then

$$P(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2(\sigma^2 + bt)}} \tag{1}$$

- Why not use Hoeffding?

Bernstein's inequality

Theorem

If X with mean μ and variance σ^2 satisfies the Bernstein condition with parameter $b > 0$, then

$$P(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2(\sigma^2 + bt)}} \tag{1}$$

- Why not use Hoeffding?
- For small t, Bernstein gives us a subgaussian tail with parameter σ

Bernstein's inequality

Theorem

If X with mean μ and variance σ^2 satisfies the Bernstein condition with parameter $b > 0$, then

$$P(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{2(\sigma^2 + bt)}} \tag{1}$$

- Why not use Hoeffding?
- For small t, Bernstein gives us a subgaussian tail with parameter σ
- In contrast, Hoeffding always gives us a subgaussian tail with parameter $b \geq \sigma$.

Bernstein's inequality

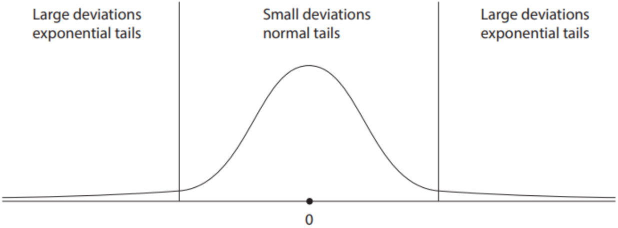


Figure 2.3 Bernstein's inequality for a sum of sub-exponential random variables gives a mixture of two tails: sub-gaussian for small deviations and sub-exponential for large deviations.

Figure 1: Taken from the High dimensional prob. book by R. Vershynin.

Bernstein's inequality

Proof.

$$
\begin{aligned}
P(X-\mu \geq t) &\leq \inf_{\lambda \in [0,1/b)} e^{-\lambda t} M_{X-\mu}(\lambda) \\
&= \inf_{\lambda \in [0,1/b)} e^{-\lambda t + \frac{\lambda^2 \sigma^2/2}{1-b\lambda}} \\
&\leq e^{-\frac{t^2}{2(bt+\sigma^2)}} \qquad \text{Setting } \lambda = \frac{t}{bt+\sigma^2} \in [0, 1/b)
\end{aligned}
$$

□

sub-exponential property

- The sub-exponential property is preserved under summation of independent random variables.
- Consider $X_k, k = 1, \dots, n$ independent sub-exponential (ν_k, b_k) random variables with $E[X_k] = \mu_k$.
-

$$
\begin{aligned}
E\left[e^{\lambda \sum_k (X_k - \mu_k)}\right] &= \prod_{i=1}^{n} E\left[e^{\lambda (X_i - \mu_i)}\right] \\
&\leq \prod_{i=1}^{n} e^{\frac{\lambda^2 \nu_k^2}{2}} \qquad \text{For } |\lambda| \leq 1/\max_i b_i
\end{aligned}
$$

- So $\sum_k (X_k - \mu_k)$ is sub-exponential with parameters $(\sqrt{n}\nu_*, b_*)$.

$$
b_* = \max_k b_k, \text{ and } \nu_*^2 = \sum_i \nu_i^2 / n \tag{2}
$$

Concentration of sub-exponential mean

- Plugging into our previous tail bound we have:

$$P\left(\bar{X}_n - \mu \geq t\right) \leq \begin{cases} e^{-\frac{nt^2}{2\nu_*^2}} & \text{for } 0 \leq t \leq \frac{\nu_*^2}{b_*} \\ e^{-\frac{nt}{2b_*}} & \text{for } t > \frac{\nu_*^2}{b_*} \end{cases}$$

Application: the wonders of Johnson-Lindenstrauss embedding

- Given m data points $u_i, i = 1 : m$ in $\mathbb{R}^d$, one wants to compute low dimensional projections $F(u_i)$, $F : \mathbb{R}^d \to \mathbb{R}^n$ with $n << d$.
- The goal is to preserve distances, so that distance-based algorithms can work "almost as well" on the low dimensional space.

Application: the wonders of Johnson-Lindenstrauss embedding

- Given m data points $u_i, i = 1 : m$ in $\mathbb{R}^d$, one wants to compute low dimensional projections $F(u_i)$, $F : \mathbb{R}^d \to \mathbb{R}^n$ with $n << d$.
- The goal is to preserve distances, so that distance-based algorithms can work "almost as well" on the low dimensional space.
- We define "almost as well" by:

$$\|u_i - u_j\|^2(1-\epsilon) \leq \|F(u_i) - F(u_j)\|^2 \leq \|u_i - u_j\|^2(1+\epsilon) \qquad (3)$$

Application: the wonders of Johnson-Lindenstrauss embedding

- Given m data points $u_i, i = 1 : m$ in $\mathbb{R}^d$, one wants to compute low dimensional projections $F(u_i)$, $F : \mathbb{R}^d \to \mathbb{R}^n$ with $n << d$.
- The goal is to preserve distances, so that distance-based algorithms can work “almost as well” on the low dimensional space.
- We define “almost as well” by:

$$\|u_i - u_j\|^2(1-\epsilon) \leq \|F(u_i) - F(u_j)\|^2 \leq \|u_i - u_j\|^2(1+\epsilon) \tag{3}$$

- Construct a random matrix $X \in \mathbb{R}^{n\times d}$ with $X_{ij} \sim N(0,1)$.

Application: the wonders of Johnson-Lindenstrauss embedding

- Given m data points $u_i, i = 1 : m$ in $\mathbb{R}^d$, one wants to compute low dimensional projections $F(u_i)$, $F : \mathbb{R}^d \to \mathbb{R}^n$ with $n << d$.
- The goal is to preserve distances, so that distance-based algorithms can work "almost as well" on the low dimensional space.
- We define "almost as well" by:

$$\|u_i - u_j\|^2(1 - \epsilon) \leq \|F(u_i) - F(u_j)\|^2 \leq \|u_i - u_j\|^2(1 + \epsilon) \qquad (3)$$

- Construct a random matrix $X \in \mathbb{R}^{n \times d}$ with $X_{ij} \sim N(0, 1)$.
- Define $F(u)$ as $Xu/\sqrt{n}$

Theorem

As long as $m > 2$, and $u_i \neq u_j, \forall i \neq j$ and $n = \Omega(\log(m/\delta)/\epsilon^2)$, Equation (3) is satisfied with probability at least $1 - \delta$.

-

We can do this easily with our tools

Proof.

- $u' = u/\|u\|$. We will assume that $u \neq 0$.
- Let $Y := \frac{\|F(u)\|^2}{\|u\|^2} = \sum_i (Xu')_i^2$.
- But $Y_i := (Xu')_i = \sum_j X_{ij} u'_j \sim N(0,1)$
- Note that Y_i^2 is sub-exponential with parameters $(2,4)$. So by the summation property, Y is sub-exponential $(2\sqrt{n}, 4)$.
- So $P\left(|\frac{Y}{n} - 1| \geq t\right) \leq 2e^{-\frac{nt^2}{8}}$ for $t \in (0,1)$.
- $P\left(\left|\frac{\|F(u_i - u_j)\|^2}{\|u_i - u_j\|^2} - 1\right| \geq \epsilon \text{ For some } u_i \neq u_j\right) \leq 2\binom{m}{2} e^{-\frac{n\epsilon^2}{8}}$
- If $m \geq 2$ and $n > \frac{16}{\epsilon^2}\log(m/\delta)$, the above probability can be made as small as δ.

$\square$