MATH466/MATH766
Math of machine learning
04/07-04/09 Lecture 21-22 Concentration Inequality

References:

- Ch2 of High-Dimensional Probability, An Introduction with Applications in Data Science, by Roman Vershynin
-

Todays contents:

- Markov inequality
- Chebyshev inequality
- Hoeffding inequality
- Bernstein inequality

Important concepts:

- asymptotic vs non-asymptotic

Recommend reading:

-

Motivation example

1. Toss a fair coin n times.

$X_i = \begin{cases} 1 & \text{head} \\ 0 & \text{tail} \end{cases}$ $\quad S_n := \sum_{i=1}^{n} X_i, \quad \mathbb{E}\left[\frac{1}{n} S_n\right] = \frac{1}{2}$

CLT : $\frac{2}{\sqrt{n}}\left(S_n - \frac{n}{2}\right) \to N(0,1) \quad n \to +\infty$

Want : with high probability, say at least $1-\varepsilon$

$\frac{1}{n} S_n$ is an α-accurate estimation of $\frac{1}{2}$

Mathematically, want to ask :

To have $\Pr\left[\left|S_n - \frac{n}{2}\right| > \alpha\right] \le \varepsilon$

at least how many tosses we need ?

2. ERM

$f^* := \min_{f \in \mathcal{F}} R(f) := \int l(f(x), y)\, dP(x,y)$

$f_n^* := \min_{f \in \mathcal{F}} R_n(f) := \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i) \quad (x_i, y_i) \sim_{iid} P(x,y)$

when $n \to +\infty$, does $f_n^* \to f^*$? how fast is the convergence ?

Concentration Inequality (Non-asymptotic)

The goal is to control large deviation of independent sum of r.v.'s

Setting :

Let $X_1, X_2, \cdots, X_n$ be independent r.v.'s (may not be identical)

$S_n := \sum_{i=1}^{n} X_i$

- Two basic inequalities

Prop 1. (Markov inequality). Let X be a random variable such that $\mathbb{E}|X| < \infty$. Then for any $t > 0$,

$$\Pr[|X| > t] \le \frac{\mathbb{E}|X|}{t}$$

Pf. For any $t > 0$,

$$\begin{aligned}
\mathbb{E}|X| &= \mathbb{E}\left[|X| 1_{\{|X|>t\}} + |X| 1_{\{|X|\le t\}}\right] \\
&= \mathbb{E}\left[|X| 1_{\{|X|>t\}}\right] + \mathbb{E}\left[|X| 1_{\{|X|\le t\}}\right] \quad \text{(linearity of expectation)} \\
&\ge \mathbb{E}\left[|X| 1_{\{|X|>t\}}\right] \\
&\ge t\,\mathbb{E}\left[1_{\{|X|>t\}}\right] = t \Pr[|X|>t]
\end{aligned}$$

#

Prop 2. (Chebyshev's inequality). Let X be a random variable such that $\mathbb{E}X^2 < \infty$. Then for any $t > 0$,

$$\Pr[|X - \mathbb{E}X| > t] \le \frac{\mathbb{E}X^2}{t^2}$$

Pf. For any $t > 0$,

$$\Pr[|X - \mathbb{E}X| > t] = \Pr[|X - \mathbb{E}X|^2 > t^2].$$

Since $\mathbb{E}|X - \mathbb{E}X|^2 = \mathbb{E}\left[X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2\right]$

$$= \mathbb{E}X^2 - (\mathbb{E}X)^2 \le \mathbb{E}X^2 < +\infty.$$

We can apply Markov inequality to $|X - \mathbb{E}X|^2$ and get $\Pr[|X - \mathbb{E}X| > t] \le \frac{\mathbb{E}|X - \mathbb{E}X|^2}{t^2}$

$$\le \frac{\mathbb{E}X^2}{t^2}.$$

#.

CLT suggests exponential decay

The key is to compute the exponential moments.

$$\mathbb{E}\, e^{tS_n} = \prod_{i=1}^{n} \mathbb{E}\, e^{tX_i}. \qquad \text{(by independence)}$$

Theorem 3. (Hoeffding's inequality). Suppose $X_1, \cdots, X_n$ are independent random variables which are bounded, $a_i \le X_i \le b_i$.

Let $S_n := \sum_{i=1}^{n} X_i$, then for any $\alpha > 0$,

$$\Pr[|S_n - \mathbb{E}S_n| > \alpha] \le 2\exp\left(-\frac{2\alpha^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right).$$

Pf. Suppose $\mathbb{E}X_i = 0$. Otherwise redefine $X_i - \mathbb{E}X_i$ as X_i, then $(b_i - a_i)$ is the same. Thus it suffices to prove the case where $\mathbb{E}X_i = 0$.

Then $\mathbb{E}S_n = \mathbb{E}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathbb{E}X_i = 0$.

We first consider the upper tail:

Want $\Pr[S_n > \alpha] \le \exp\left(-\frac{2\alpha^2}{\sum_i (b_i-a_i)^2}\right) \quad \forall \alpha > 0$.

Standard framework

For any $t > 0$,

$$\Pr[S_n > \alpha] = \Pr[e^{tS_n} > e^{t\alpha}]$$

$$\le e^{-t\alpha}\, \mathbb{E}\, e^{tS_n} \qquad \text{(by Markov ineqn.)}$$

$$= e^{-t\alpha} \prod_{i=1}^{n} \mathbb{E}\, e^{tX_i} \qquad (*) \quad \text{(by independence)}$$

Lemma (Hoeffding's Lemma). Suppose X is bounded r.v. $a \le X \le b$, then for any $t > 0$,

$$\mathbb{E}\, e^{t(X-\mathbb{E}X)} \le \exp\left(\frac{t^2}{8}(b-a)^2\right).$$

By the Lemma, $\mathbb{E}\, e^{tX_i} \le \exp\left(\frac{t^2}{8}(b_i-a_i)^2\right)$

Then $\Pr[S_n > \alpha] \le e^{-t\alpha} \prod_{i=1}^{n} \mathbb{E}\, e^{tX_i}$ (by (*))

$\le \exp\left(-t\alpha + \frac{t^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2\right)$ (by Hoeffding's Lemma)

$=: e^{f(t)}$

Want to bound $e^{f(t)}$

$f(t) := -t\alpha + \frac{t^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2$ quadratic in t

$f'(t) = 0 \Rightarrow 4\alpha = t\sum_{i=1}^{n}(b_i-a_i)^2 \Rightarrow t^* = \frac{4\alpha}{\sum_{i=1}^{n}(b_i-a_i)^2}$

$f(t)$ achieves min at $f(t^*) = \frac{-2\alpha^2}{\sum_{i=1}^{n}(b_i-a_i)^2}$

Thus $\Pr[S_n > \alpha] \le \exp(f(t^*)) = \exp\left(\frac{-2\alpha^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$.

The lower tail can be proved similarly. (Ex) #.

Remark:

$\Pr\left[\frac{1}{n}S_n > \alpha\right] \le \exp\left(\frac{-2(n\alpha)^2}{n(b-a)^2}\right) = \exp(-Cn\alpha^2)$

Lemma (Hoeffding's Lemma). Suppose X is bounded r.v. $a \le X \le b$, then for any $t > 0$,

$$\mathbb{E}\, e^{t(X-\mathbb{E}X)} \le \exp\left(\frac{t^2}{8}(b-a)^2\right).$$

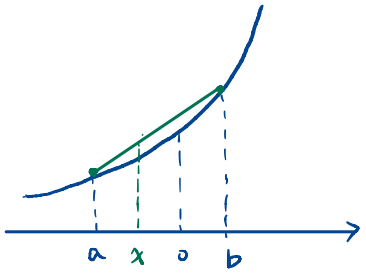
Pf. • By that redefining $X - \mathbb{E}X$ to be X does not change $L = (b-a)$ where $X \in [a,b]$, it suffices to show the case when $\mathbb{E}X = 0$.

• Suppose $(b-a) > 0$, otherwise X is a constant and the claim holds.

• Furthermore, $a < 0 < b$. because if $a = 0$, then $\mathbb{E}X = 0$ means that $X \equiv 0$, then claim holds. Similarly $b = 0$ is a trivial case.

For any $x \in [a,b]$,

$$x = \frac{b-x}{b-a}\, a + \frac{x-a}{b-a}\, b$$



Thus, for each realization of random variable X,

$$e^{tx} \le \frac{b-x}{b-a}\, e^{ta} + \frac{x-a}{b-a}\, e^{tb}.$$

• Taking expectation, by $\mathbb{E}X = 0$

$$\mathbb{E}\, e^{tX} \le \underbrace{\frac{b}{b-a}\, e^{ta} - \frac{a}{b-a}\, e^{tb}}$$

write this as $e^{[\ ]}$ and relax to simplify

$$= e^{ta}\left(\underbrace{\frac{b}{b-a}}_{\text{cst}} - \underbrace{\frac{a}{b-a}}_{\text{cst}}\, e^{t(b-a)}\right) \qquad \theta := -\frac{a}{b-a}$$

$$= e^{ta}\left(1 - \theta + \theta\, e^{t(b-a)}\right)$$

$$= \exp\left(\underbrace{ta + \ln\left(1 - \theta + \theta\, e^{t(b-a)}\right)}\right)$$

simplify this

$$= \exp\left(t(b-a)\frac{a}{b-a} + \ln\left(1-\theta+\theta e^{t(b-a)}\right)\right)$$

$$= \exp\left(-\theta u + \ln\left(1-\theta+\theta e^{u}\right)\right)$$

Compare to Lemma statement

suffice to show $-\theta u + \ln(1-\theta+\theta e^u) \le \frac{1}{8}u^2$

Let $\varphi(u) := -\theta u + \log(1-\theta+\theta e^u)$

$\varphi(0) = 0$

$\varphi'(u) = -\theta + \frac{\theta e^u}{1-\theta+\theta e^u}$, $\varphi'(0) = 0$

$\varphi''(u) = \frac{-(\theta-1)\theta e^u}{(1-\theta+\theta e^u)^2} = \frac{\theta e^u}{1-\theta+\theta e^u} \cdot \frac{1-\theta}{1-\theta+\theta e^u}$

$\le \frac{1}{4}$ $\quad \forall u \ge 0$ $\quad$ by $\sqrt{xy} \le \frac{x+y}{2}$ $x, y \ge 0$

Thus $\forall u > 0$,

$\varphi(u) = \varphi(0) + u\varphi'(0) + \frac{1}{2}\varphi''(v)u^2$ for some $v > 0$

$\le \frac{1}{8}u^2$

Hoeffding inequality only use the boundedness of the random variables.

Think about $X_i \sim_{i.i.d}$ Bernoulli distribution with $Pr(X_i=1)=p$.

The tail of $\frac{1}{n}\sum_{i=1}^{n} X_i$ when $p=0.1$ and when $p=0.9$ can have different decay rate, but this is not reflected in Hoeffding inequality.

- Chernoff bound: Let $X_1, \cdots, X_n$ be i.i.d Bernoulli random variables with mean p.

Then for any $\lambda > 0$,

$$Pr\left[\sum_i X_i < np - \lambda\right] \leq \exp\left(-\frac{\lambda^2}{2np}\right) \quad \text{(HW)}$$

$$Pr\left[\sum_i X_i > np + \lambda\right] \leq \exp\left(-\frac{\lambda^2}{2(np+\lambda/3)}\right)$$

Similarly, the standard deviation σ_i of independent random variables X_i can affect the decay rate of $\frac{1}{n}\sum_i X_i$.

But Hoeffding inequality does not reflect this.

The following Bernstein's inequality shows how σ affects the tail behavior.

Theorem 4. (Bernstein's inequality). Let $X_1, \cdots, X_n$ be independent random variables, $\mathbb{E}X_i = 0$, $|X_i| \le L$ for all i, $L > 0$ is a constant and $\sum_{i=1}^n \mathbb{E}X_i^2 = n\sigma^2$. Then for any $\alpha > 0$,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n X_i < -\alpha\right], \Pr\left[\frac{1}{n}\sum_{i=1}^n X_i > \alpha\right] \le \exp\left(-\frac{n\alpha^2}{2\sigma^2 + \frac{2}{3}\alpha L}\right).$$

Pf. Let $t = n\alpha > 0$, for any $s > 0$, we have

$$\begin{aligned}
\Pr\left[\sum_{i=1}^n X_i > t\right] &= \Pr\left[e^{s\sum_i X_i} > e^{st}\right] && \text{by monotonicity of } e^{sx} \\
&\le e^{-st}\,\mathbb{E}e^{s\sum_i X_i} && \text{by Markov ineqn} \\
&= e^{-st}\prod_{i=1}^n \mathbb{E}e^{sX_i} && (1)\ \text{by independence}
\end{aligned}$$

Denote $\mathbb{E}X_i^2 = \sigma_i^2$

$$\begin{aligned}
\mathbb{E}e^{sX_i} &= \mathbb{E}\left[1 + sX_i + \sum_{m=2}^{\infty}\frac{s^m X_i^m}{m!}\right] && \text{by Taylor expansion} \\
&= 1 + 0 + \sum_{m=2}^{\infty}\frac{s^m}{m!}\mathbb{E}X_i^2 X_i^{m-2} && \text{by linearity of } \mathbb{E} \\
&\le 1 + \sum_{m=2}^{\infty}\frac{s^m}{m!}L^{m-2}\sigma_i^2 && \text{by } |X_i| \le L \text{ linearity of } \mathbb{E}.
\end{aligned}$$

$$= 1 + \frac{\sigma_i^2}{L^2} \sum_{m=2}^{\infty} \frac{(sL)^m}{m!}$$

$$= 1 + \frac{\sigma_i^2}{L^2} \left(e^{sL} - 1 - sL \right) \qquad \text{by Taylor expansion}$$

$$\le \exp\left(\frac{\sigma_i^2}{L^2} \left(e^{sL} - 1 - sL \right) \right) \qquad \text{by } 1 + x \le e^x \ (\forall x).$$

Putting back to (1), we have

$$\Pr\left[\sum_{i=1}^{n} X_i > t \right] \le e^{-st} \prod_{i=1}^{n} \exp\left(\frac{\sigma_i^2}{L^2} \left(e^{sL} - 1 - sL \right) \right)$$

$$= \exp\left(-st + \sum_{i=1}^{n} \frac{\sigma_i^2}{L^2} \left(e^{sL} - 1 - sL \right) \right)$$

$$= \exp\left(-st + \frac{n\sigma^2}{L^2} \left(e^{sL} - 1 - sL \right) \right) \qquad (2) \ \forall s > 0$$

Next we want to find the minimum value of

$$g(s) := -st + \frac{n\sigma^2}{L^2} \left(e^{sL} - 1 - sL \right), \quad s > 0.$$

$$g'(s) = \frac{n\sigma^2}{L^2} \left(L e^{sL} - L \right) - t = \frac{n\sigma^2}{L} \left(e^{sL} - 1 \right) - t \qquad \nearrow \text{ w.r.t. } s$$

Let $g'(s^*) = 0$ we have $s^* = \frac{1}{L} \log\left(1 + \frac{tL}{n\sigma^2} \right) = \frac{1}{L} \log(1 + t')$, where $t' := \frac{tL}{n\sigma^2}$

$$e^{s^* L} = 1 + \frac{tL}{n\sigma^2} = 1 + t'$$

$$g(s^*) = -\frac{t}{L} \log(1 + t') + \frac{n\sigma^2}{L^2} \left(t' - \log(1 + t') \right)$$

$$= -\frac{n\sigma^2}{L^2} \left(t' \log(1 + t') - t' + \log(1 + t') \right)$$

$$= -\frac{n\sigma^2}{L^2} \left((1 + t') \log(1 + t') - t' \right) \qquad (3)$$

Want the upper-bound of $g(s^*)$

i.e. want the lower-bound of $(1+x)\log(1+x) - x$

Let $h(x) := (1+x)\log(1+x) - x$

Then $h(x) \geq \frac{x^2}{2 + \frac{2x}{3}}$ (appendix)

and (3) continuous as

$$g(s^*) \leq -\frac{n\sigma^2}{L^2} \frac{x^2}{2 + \frac{2x}{3}}$$

Plugging the definition of $x = \frac{tL}{n\sigma^2}$

$$= -\frac{n\sigma^2}{L^2} \frac{tL/n\sigma^2}{2n\sigma^2/tL + 2/3} = -\frac{t^2}{2n\sigma^2 + \frac{2}{3}tL}$$

Back to (2) $\Pr\left[\sum_{i=1}^{n} X_i > t\right] \leq e^{g(s^*)} \leq \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}tL}\right)$.

Since $t = n\alpha$, this proves

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i > \alpha\right] \leq \exp\left(-\frac{n\alpha^2}{2\sigma^2 + \frac{2}{3}\alpha L}\right)$$

The lower-tail $\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i < \alpha\right]$ is proved similarly.

Appendix How to observe $h(x) = (1+x)\log(1+x) - x \geq \dfrac{x^2}{2 + \frac{2x}{3}}$

$x \to 0$ $\quad h(x) = (1+x)\left(x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots\right) - x = \frac{x^2}{2} - \frac{x^3}{6} + \cdots$

dominated by x^2

$x \to +\infty$ $\quad h(x)$ growth slower than x^2, faster than x

Assume $\tilde{h}(x) = h(x) - \dfrac{x^2}{a+bx}$

Want to find a, b s.t. $\tilde{h}(x) \geq 0$ for any $x \geq 0$

want the ineqn to be tight

① need $a > 0$, $b \geq 0$,

② with ①, $\tilde{h}(0) = 0$

$$\tilde{h}'(x) = \log(1+x) - \frac{2x}{a+bx} + \frac{bx^2}{(a+bx)^2} \qquad h'(0) = 0$$

To have $\tilde{h}(x) \geq 0$, at least near 0, need $\tilde{h}''(x) \geq 0$

$$\tilde{h}''(x) = \frac{1}{1+x} - \frac{2}{a+bx} + \frac{2bx}{(a+bx)^2} + \frac{2bx}{(a+bx)^2} - \frac{2b^2x^2}{(a+bx)^3}$$

$$= \frac{1}{1+x} - \frac{2}{a+bx}\left(1 - \frac{bx}{a+bx}\right)^2$$

$$= \frac{1}{1+x} - \frac{1}{(a+bx)^3/2a^2} = \frac{1}{1+x} - \frac{1}{a/2 + \frac{3}{2}bx + \frac{3}{2a}b^2x^2 + \frac{1}{2a}b^3x^3}$$

To have $\tilde{h}''(x) \geq 0$ near $x = 0$

$$a = 2, \quad b = \frac{2}{3}$$

Interpretation of Bernstein bound

$$\Pr[Y > t] \leq \exp\left(-\frac{t^2}{C^2}\right) \quad \text{for any } t > 0$$

is called a **sub-Gaussian tail** at the magnitude of C

the decay of a sub-Gaussian tail is as fast as the tail of a Gaussian distribution

$$\Pr[Y > t] \leq \exp\left(-\frac{t}{C}\right) \quad \text{for any } t > 0$$

is called a **sub-exponential tail** at the magnitude of C

the decay of a sub-exponential tail is as fast as the tail of an exponential distribution

Asymptotically ($t \to +\infty$), sub-Gaussian tail decays faster

If Y_1 and Y_2 both have sub-Gaussian (sub-exponential tail) but at the magnitude C_1 and C_2 ($C_1 < C_2$)

then the one with smaller magnitude decays faster

Comparison

Hoeffding $\Pr\left[\frac{1}{n} S_n > \alpha\right] \leq \exp\left(\frac{-2(n\alpha)^2}{n(b-a)^2}\right) \overset{\substack{a=L \\ b=-L}}{=} \exp\left(-\frac{n\alpha^2}{2L}\right)$

Bernstein $\Pr\left[\frac{1}{n} S_n > \alpha\right] \leq \exp\left(-\frac{n\alpha^2}{2\sigma^2 + \frac{2}{3}\alpha L}\right)$

Bernstein: $\exp\left(-\frac{n\alpha^2}{2\sigma^2+\frac{2}{3}\alpha L}\right)$

- α is the deviation threshold
- σ^2 is the variance of the random variable.
- L is the boundedness of the random variable.

$$\exp\left(-\frac{n\alpha^2}{\frac{2}{3}L\left(\frac{3\sigma^2}{L}+\alpha\right)}\right) \leq \begin{cases} \exp\left(-\frac{\alpha^2}{4\sigma^2/n}\right), & \text{near tail} \\ & \text{"}\alpha < \frac{3\sigma^2}{L}\text{"} \\ \exp\left(-\frac{\alpha}{\frac{4}{3}L/n}\right), & \text{far tail} \\ & \text{"}\alpha > \frac{3\sigma^2}{L}\text{"} \end{cases}$$

sub-Gaussian decay of α at the magnitude order of $\frac{\sigma}{\sqrt{n}}$

sub-exponential decay of α at the magnitude order of $\frac{L}{n}$

(1) near tail: Bernstein — sub-Gaussian — order of $\frac{\sigma}{\sqrt{n}}$

Hoeffding — order of $\frac{L}{\sqrt{n}}$

Bernstein takes advantage when $\sigma \ll L$

(2) target a deviation α at the proper scale proportional to σ^2 : take advantage of the sub-Gaussian decay.

(3) near tail: Bernstein sub-Gaussian. order of $\frac{\sigma}{\sqrt{n}}$

Central Limit Theorem: $X_1, \cdots, X_n$ i.i.d, $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma^2$

$n \to +\infty$, $\frac{1}{n}\sum_{i=1}^{n} X_i \to N(0, \frac{\sigma^2}{n})$ ~ sub-Gaussian of $\frac{\sigma}{\sqrt{n}}$.

More on concentration inequality (Optional)

$X_1, \dots, X_n$ independent, $S_n = \sum_{i=1}^n X_i$

Hoeffding ineqn shows that if $X_1, \dots, X_n$ are bounded then S_n has sub-Gaussian tail

Q: To ensure S_n has sub-Gaussian tail, what assumption is necessary / sufficient for X_i ?

Sub-Gaussian Random Variables

- def. (i) For all $t \geq 0$ $\Pr(|X| \geq t) \leq 2\exp(-t^2/K_1^2)$
- prop. (ii) $\|X\|_{L^p} := (\mathbb{E}[|X|^p])^{1/p} \leq K_2\sqrt{p}$ for all $p \geq 1$

  (iii) $\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(\lambda^2 K_3^2)$ for all λ s.t. $|\lambda| \leq \frac{1}{K_3}$

  (iv) $\mathbb{E}[\exp(X^2/K_4^2)] \leq 2$ for some K_4

  (v) if $\mathbb{E}X = 0$, then $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 K_5^2)$ for all $\lambda \in \mathbb{R}$

  We skip the proof

- examples of sub-Gaussian random variables

  (1) (Gaussian) $X \sim N(0, \sigma^2)$ $\|X\|_{\psi_2} \leq C\sigma$

  (2) (Sym Bern) $\Pr[X=1] = \Pr[X=-1] = \frac{1}{2}$ $\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}$

  (3) (Bounded) $\|X\|_\infty < \infty$ $\|X\|_{\psi_2} \leq \|X\|_\infty / \sqrt{\ln 2}$

- **sub-Gaussian norm** if X is a sub-Gaussian random variable

  $\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$

  is a norm on the space of sub-Gaussian random variables

From the definition and property

$$\Pr[|X| \geq t] \leq 2\exp(-ct^2/\|X\|_{\psi_2}^2) \quad \text{for any } t \geq 0$$

$$\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p} \quad \text{for all } p \geq 1$$

$$\text{if } \mathbb{E}X = 0, \text{ then } \mathbb{E}[\exp(\lambda X)] \leq \exp(C\lambda^2\|X\|_{\psi_2}^2) \text{ for all } \lambda \in \mathbb{R}$$

- Sum of independent sub-Gaussian

  if $X_1, \cdots, X_n$ independent, mean zero, sub-Gaussian

  then $S_n = \sum_i X_i$ is sub-Gaussian and

  $$\|S_n\|_{\psi_2}^2 \leq C\sum_{i=1}^n \|X_i\|_{\psi_2}^2$$

- Centering

  If X is sub-Gaussian, then $X - \mathbb{E}X$ is sub-Gaussian

  and $\|X - \mathbb{E}X\|_{\psi_2} \leq C\|X\|_{\psi_2}$

- General Hoeffding's inequality

  $X_1, \cdots, X_n$ independent, mean zero, sub-Gaussian

  Then for any $t \geq 0$

  $$\Pr\left[\left|\sum_{i=1}^n X_i\right| \geq t\right] \leq 2\exp\left(-\frac{ct^2}{\sum_i \|X_i\|_{\psi_2}^2}\right)$$

Sub-Exponential random variables

Motivation

Consider $\underline{X} \sim N(\underline{0}, I_d)$. $\|\underline{X}\|_2^2 = \sum_i X_i^2$

Want to study the concentration property of $\|\underline{X}\|_2^2$

But X_i^2 is not sub-Gaussian

$$\Pr[X_i^2 > t] = \Pr[|X_i| > \sqrt{t}] \sim \exp(-t/2)$$

Sub-Exponential Random Variables

- def. (i) For all $t \geq 0$ $\Pr(|X| \geq t) \leq 2\exp(-t/K_1)$
- prop. (ii) $\|X\|_{L^p} := (\mathbb{E}[|X|^p])^{1/p} \leq K_2 p$ for all $p \geq 1$

  (iii) $\mathbb{E}[\exp(\lambda|X|)] \leq \exp(\lambda K_3)$ for all λ s.t. $0 \leq \lambda \leq \frac{1}{K_3}$

  (iv) $\mathbb{E}[\exp(|X|/K_4)] \leq 2$ for some K_4

  (v) if $\mathbb{E}X = 0$, then $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 K_5^2)$ for all λ s.t. $|\lambda| \leq \frac{1}{K_5}$

  We skip the proof

- **sub-exponential norm** if X is a sub-exponential random variable

  $\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \leq 2\}$

  is a norm on the space of sub-exponential random variables
- Sub-Gaussian is sub-exponential

Sub-Gaussian square is sub-exponential

X sub-Gau $\Leftrightarrow$ X^2 sub-exp. Moreover $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$

product of sub-Gaussian is sub-exponential.

X, Y sub-Gau $\Rightarrow XY$ sub-exp Moreover $\|XY\|_{\psi_1} \le \|X\|_{\psi_2} \|Y\|_{\psi_2}$

- **Bernstein Inequality** $X_1, \dots, X_n$ independent, mean zero, sub-exp

$S_n = \sum_i X_i$ Then for any $t \ge 0$

$$\Pr[\,|S_n| \ge t\,] \le 2\exp\left[-c \min\left\{ \frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2},\ \frac{t}{\max_i \|X_i\|_{\psi_1}} \right\}\right]$$

a mixture of sub-Gau and sub-exp dist.

From CLT, we expect a sub-Gau tail

But the heavy tail of sub-exp prevents S_n to have sub-Gau tail everywhere, so we expect sub-exp tail as well

It is produced by the X_i with maximum sub-exp norm.