Chapter 3: Concentration Ineq.

Obj:

↳ Suppose we have $X_1, \dots, X_n$ are indp. RVs.

↳ In many cases, it will be useful to find bounds on some func. of $X_1, \dots, X_n$.
That is, we want to bound

(1) $P\{f(x_1, \dots, x_n) \geq t\}$

for some $t > 0$.

↳ For the first part of this unit, we'll focus on func. f of the form

$$f : (x_1, x_2, \dots, x_n) \longmapsto \frac{1}{n}\sum_{i=1}^{n} x_i$$

or

$$f : (x_1, \dots, x_n) \longmapsto \left| \frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}\left[\frac{1}{n}\sum x_i\right] \right|$$

↳ We could consider (1) under asymp.

Eg: if $X_1, \dots, X_n$ iid and $\mu = \mathbb{E}[X]$ and $\sigma = \sqrt{Var(X)}$, then the CLT says that

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$P\left\{\bar{X}_n \geq \mu + \frac{\sigma t}{\sqrt{n}}\right\} \xrightarrow{n\to\infty} 1 - \Phi(t)$$

where Φ is the CDF of a standard normal random variable.

But: We usually finite sample n, so asymp. guarantees won't provide an exact bound on (2) above.

Question: What can we say in finite samples?

For example, what can we say about

$$P\{\bar{X}_n \geq \mu + t\}$$

for (possibly large) $t > 0$?

During this unit, we'll obtain bounds on (1) using the following 3 approaches.

(1) **Bounds on moments:** $\begin{cases} \text{Markov} \\ \text{Chebyshev} \end{cases}$

(2) Bounds on moment generating function (MGF): Chernoff, Hoeffding, Sub-Gaussian, sub-Exponential, Bernstein Ineq.

(3) Martingale arguments $\begin{cases} \text{Azuma-Hoeffding} \\ \text{Bounded difference} \end{cases}$

Bounds on Moments

Thm. (Markov)

If $X \geq 0$, $\mathbb{E}[X] < \infty$, and $t > 0$, then

$$\boxed{P\{X \geq t\} \leq \frac{\mathbb{E}[X]}{t}}$$

Pf: For any $t > 0$, we have the following

$$P\{X \geq t\} = \int_t^\infty dP(X) \qquad \left(1 \leq \frac{x}{t} \text{ when } x \geq t\right)$$

$$\leq \int_t^\infty \frac{x}{t}\, dP(X)$$

$$\leq \int_0^\infty \frac{x}{t}\, dP(X)$$

$$= \frac{\mathbb{E}[X]}{t}$$

Note: All of concentration inequalities use Markov's inequality as a building block!

Thm. Suppose $\mathbb{E}|X|<\infty$. $h:[0,\infty)\longrightarrow[0,\infty)$ is a non-decreasing func. for which $h(t)>0$ for all $t>0$, and $\mathbb{E}[h(|X-\mathbb{E}[X]|)]<\infty$. For all $t>0$,

$$P\{|X-\mathbb{E}[X]|\geq t\}\leq\frac{\mathbb{E}[h(|X-\mathbb{E}X|)]}{h(t)}$$

Proof: Because h is non-decreasing,

say constant func.

$$\{|X-\mathbb{E}X|\geq t\}\subseteq\{h(|X-\mathbb{E}X|)\geq h(t)\}$$

Hence,

$$P\{|X-\mathbb{E}X|\geq t\}\leq P\{h(|X-\mathbb{E}X|)\geq h(t)\}$$

$$\overset{\text{(Markov)}}{\leq}\frac{\mathbb{E}[h(|X-\mathbb{E}X|)]}{h(t)}$$

Example: when $h(t)=t^2$, the previous result

$$P\{|X-\mathbb{E}X|\geq t\}\leq\frac{\mathbb{E}[|X-\mathbb{E}(X)|^2]}{t^2}=\frac{\mathrm{Var}(X)}{t^2}$$

Chebyshev

Jan 25

Remark: Suppose $l<K$ and both $\mathbb{E}[|X-\mathbb{E}X|^l]$ and $\mathbb{E}[|X-\mathbb{E}X|^K]$ are finite.

For large t, the bound based on K

will be much tighter than the bound based on l.

This is somewhat intuitive: we should expect the bound based on k will improve upon the one based on l in some sense given the requiring $\mathbb{E}[|X-\mathbb{E}X|^k]<\infty$ is stronger than requiring $\mathbb{E}[|X-\mathbb{E}X|^l]<\infty$.

Why?

$\frac{1}{t^k}$ decays to zero much more quickly than $\frac{1}{t^l}$ when $t\to\infty$.

__Question__: Can we get even sharper tail inequalities than the one above if we assume even more about the distr. of X?

__Ans__: Yes! Provided X has a moment generating function*.

__Remark*__: The $\boxed{\text{bound based on } \mathbb{E}[|X-\mathbb{E}X|^k]}$ implies
$$P\{|X-\mathbb{E}X|\geq t\}\leq \inf_{k\in\mathbb{N}} \frac{\mathbb{E}[|X-\mathbb{E}X|^k]}{t^k}$$
It turns out that this inequality is sharper than the one we're about to derive based on the MGF.

__But__: the bound above is hard to work with in practice.

Thm. (Chernoff): Suppose that X has a MGF in a neighbourhood of zero, meaning that there exists $b>0$ s.t. $\mathbb{E}[\exp\{\lambda X\}] < \infty$ for all $|\lambda| \le b$. Then, for $t>0$ and $\lambda \in (0, b]$,

$$P\{X - \mathbb{E}X \geq t\} = P\{e^{\lambda(X-\mathbb{E}X)} \geq e^{\lambda t}\}$$
$$\leq \frac{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]}{e^{\lambda t}}$$
$$=: \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}},$$

where $\mu = \mathbb{E}[X]$.

Hence,

$$P\{X - \mathbb{E}X \geq t\} \leq \inf_{\lambda > 0} \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}}$$

Equivalently,

$$\log P\{X - \mathbb{E}X \geq t\} \leq -\sup_{\lambda}\{\lambda t - \underbrace{\log M_{X-\mu}(\lambda)}_{\text{"Cumulant generating function"}}\}$$

Example: Gaussian X.

Suppose $X \sim N(\mu, \sigma^2)$. In this case,

$$M_{X-\mu}(\lambda) = \mathbb{E}[\exp\{(X-\mu)\lambda\}]$$
$$= \frac{1}{\sqrt{2\pi}\,\sigma} \int \exp\left\{\lambda(x-\mu) - \frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

$(z = x-\mu)$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \int \exp\left\{\lambda z - \frac{z^2}{2\sigma^2}\right\} dz$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\} \int \exp\left\{-\frac{(z/\sigma - \lambda\sigma)^2}{2}\right\} dz$$

$(y = z/\sigma)$

$$= \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\} \int \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y-\lambda\sigma)^2}{2}\right\}}_{N(\lambda\sigma,\,1)} dy$$

$$= \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\}$$

Plugging the form of $M_{X-\mu}(\lambda)$ into $\boxed{\text{Chernoff's bound}}$, we find that

$$\log P\{X - \mathbb{E}X \geq t\}$$

$$\leq -\sup_{\lambda>0} \{\lambda t - \log M_{X-\mu}(t)\}$$

$$= -\sup_{\lambda>0} \left\{\lambda t - \frac{\lambda^2\sigma^2}{2}\right\}$$

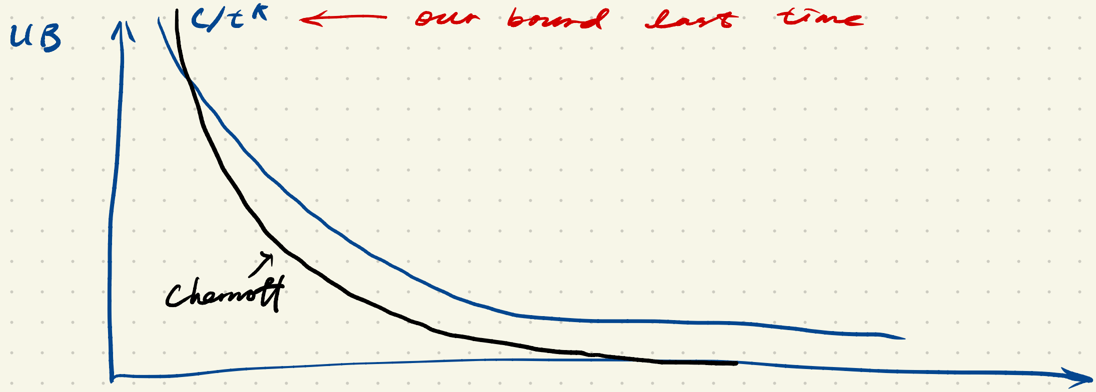
Solving for the λ^* that maximizes the above, we find that

$$\lambda^* = t/\sigma^2.$$

Hence,

$$\log P\{X - \mathbb{E}X \geq t\} \leq -\frac{t^2}{2\sigma^2}$$

$$\Rightarrow \quad P\{X - \mathbb{E}X \geq t\} \leq \exp\left\{-\frac{t^2}{2\sigma^2}\right\}.$$



Sub-Gaussian random variables

Def: A R.V. X is called a sub-Gaussian (sub-G) with params σ^2 if

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2 \sigma^2}{2} \quad \text{for all } \lambda \in \mathbb{R}.$$

Note: A normal $\mathcal{N}(\mu, \sigma^2)$ R.V. is sub-G w/ param σ^2.

By Chernoff, a sub-G R.V. X w/ σ^2 satisfies

$$P\{X - \mathbb{E}X \geq t\} \leq \exp\left\{-\frac{t^2}{2\sigma^2}\right\}$$

This is the same bound we derived for $X \sim \mathcal{N}(\mu, \sigma^2)$!
Tail decays faster than Gaussian.

Alternative Characterization of sub-G R.V. (HDS, Chap 2)

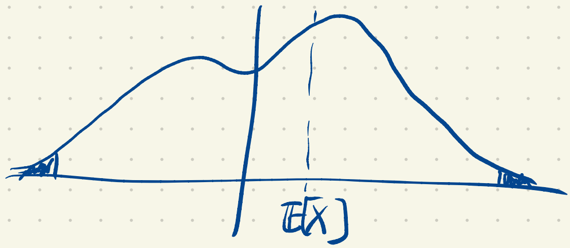
A R.V. X is sub-G iff there exists $c > 0$, and $s > 0$ s.t. $\forall t > 0$,

$$P\{|X - \mathbb{E}X| \geq t\} \leq c\, P\{|sZ| \geq t\}, \text{ where } Z \sim N(0,1)$$

s is not necessary σ^2.
explicit expression exists

Counterexample: An Exp(1) R.V. is not sub-G. Why?

$$P\{X \geq t\} = \exp(-t)$$



Thm. (Hoeffding): If X is a R.V. w/ support $[a,b]$, $(|a|, |b| < \infty)$, then X is sub-G w/ $\sigma^2 = \frac{(b-a)^2}{4}$.

Pf: WLOG, suppose $\mathbb{E}X = 0$.

Let $f : \lambda \mapsto \log M_{X-\mu}(\lambda)$

Noting that $f'(0) = \frac{\mathbb{E}[X]}{\mathbb{E}[\exp\{t \cdot 0\}]} = \mathbb{E}[X] = 0$

we have that

$$(1) \qquad f(\lambda) = \int_0^\lambda f'(r)\,dr \qquad = f'(r) - f'(0)$$

$$= \int_0^\lambda \int_0^r f''(s)\,ds\,dr$$

Hence, to bound f, it's enough to get a pointwise upper bound on f''.

Note that

$$f'(\lambda) = \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]},$$

$$f''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \right)^2$$

Note that f'' is the variance of a R.V. Z_λ with density

$$x \longmapsto \frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda x}]} p(x)$$

Hence, $f''(\lambda) = \mathrm{Var}(Z_\lambda)$

$$= \mathrm{Var}\left(Z_\lambda - \frac{a+b}{2}\right)$$

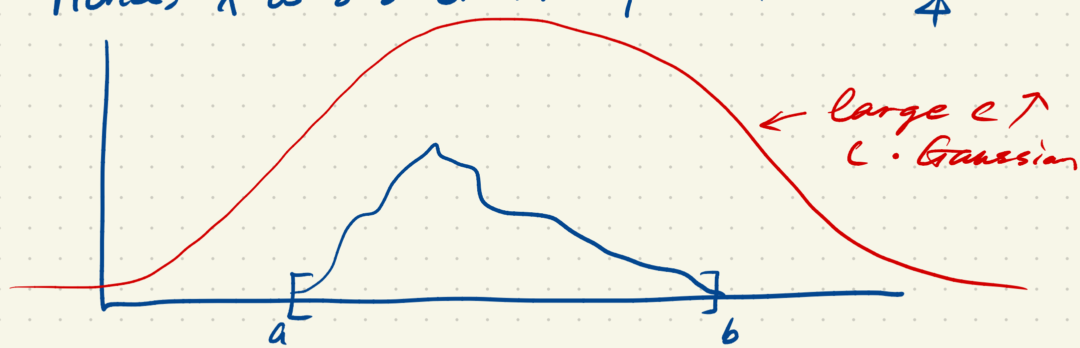
$$\leq \mathbb{E}\left[\left(Z_\lambda - \frac{a+b}{2}\right)^2\right]$$

$$\leq \frac{(b-a)^2}{4}$$

Plugging in (1),

$$\log M_{X-\mu}(\lambda) = f(\lambda) \leq \frac{(b-a)^2}{4} \int_0^\lambda \int_0^r ds\,dr = \frac{(b-a)^2}{4} \frac{\lambda^2}{2}$$

Hence, X is sub-G with param. $\sigma^2 = \frac{(b-a)^2}{4}$



Implication of Thm. above

By Chernoff, for any X taking values in $[a,b]$,

$$P\{X - \mathbb{E}X \geq t\} \leq \exp\left\{-\frac{2t^2}{(b-a)^2}\right\}$$

This inequality is known as <u>Hoeffding Inequality</u>.

More general form,

Suppose that $X_1, X_2 \cdots X_n$ are indp and each has support $[a,b]$

$$P\{\bar{X}_n - \mathbb{E}\bar{X}_n \geq t\} \leq \exp\left\{-\frac{2nt^2}{(b-a)^2}\right\}$$

See a proof when $n=2$.

To do this, we'll show if X_1 and X_2 are indp. with sub-G σ_1^2, σ_2^2, then $X_1 + X_2$ is sub-G with param. $\sigma_1^2 + \sigma_2^2$.

To see this,

$$M_{X_1+X_2}(\lambda) = \mathbb{E}[\exp\{\lambda(X_1+X_2)\}] \quad \text{def. of MGF}$$

$$= \mathbb{E}[\exp\{\lambda X_1\}]\,\mathbb{E}[\exp\{\lambda X_2\}] \quad \text{indp.}$$

$$= M_{X_1}(\lambda)\, M_{X_2}(\lambda)$$

Similarly, $M_{X_1+X_2-\mu_1-\mu_2}(\lambda) = M_{X_1-\mu_1}(\lambda)\, M_{X_2-\mu_2}(\lambda)$

By sub-G of X_1, X_2,

$$\log M_{X_1+X_2-\mu_1-\mu_2}(\lambda) = \sum_{i=1}^{2} \log M_{X_i-\mu_i}(\lambda)$$

$$\leq \sum_{i=1}^{2} \frac{\lambda^2 \sigma_i^2}{2}$$

$$= (\sigma_1^2 + \sigma_2^2) \lambda^2/2$$