Total: 95/100 (A)

ISYE 8803: Mathematical Data Science HW1 Solutions

Cameron Khanpour

February 14, 2025

Quick Links: Problem 1 Problem 2 Problem 3 Problem 4 Problem 5 Problem 6 Problem 7

Problem 1 (MGF vs Moment Bounds).

a) Show that if $X > 0$ a.s., then for any $u > 0$,

$$\inf_{\lambda>0} M_X(\lambda)e^{-\lambda u} \geq \inf_{k\in\mathbb{Z}_+} \mathbb{E}[X^k]u^{-k}$$

Proof.

$$M_X(\lambda) := \mathbb{E}[e^{\lambda X}] = \mathbb{E}[\sum_{n=0}^{\infty} \frac{(\lambda X)^n}{n!}] = \sum_{n=0}^{\infty} \frac{\lambda^n \mathbb{E}[X^n]}{n!}$$

$$\implies M_X(\lambda)e^{-\lambda u} = \boxed{\sum_{n=0}^{\infty} \frac{\lambda^n \mathbb{E}[X^n]}{n!} \Big/ \sum_{n=0}^{\infty} \frac{\lambda^n \cancel{X^n} u^n}{n!}}$$

typo (no worries)

\left[... \right] to scale the brackets

It is clear that for n = 0, 1, 2, . . . ,

$$\inf_{k\in\mathbb{Z}_+} \mathbb{E}[X^k]u^{-k} \leq \boxed{\frac{\lambda^n \mathbb{E}[X^n]}{n!} \Big/ \frac{\lambda^n \cancel{X^n} u^n}{n!}}$$

(✓) ← this inequality is wrong though

where not all ratios are identical, and n can be indexed differently.

$$\implies M_X(\lambda)e^{-\lambda u} \geq \inf_{k\in\mathbb{Z}_+} \mathbb{E}[X^k]u^{-k},\ \forall\lambda \geq 0$$

Lemma $\forall a_k, b_k > 0$

$$\frac{\sum_k a_k}{\sum_k b_k} \geq \min_j \frac{a_j}{b_j}. \quad \square$$

Proof: let $r_j = \frac{a_j}{b_j}$, and $j^* = \arg\min r_j$, then

$$r_{j^*} \sum_k b_k \leq \sum_k r_k b_k = \sum_k a_k. \quad \square$$

Therefore, since the above holds $\forall\lambda \geq 0$,

+/-

$$\inf_{\lambda>0} M_X(\lambda)e^{-\lambda u} \geq \inf_{k\in\mathbb{Z}_+} \mathbb{E}[X^k]u^{-k}$$

b) Show that if X is symmetric, then for any $u > 0$,

$$\inf_{\lambda>0} M_X(\lambda)e^{-\lambda u} \geq \frac{1}{2} \inf_{k\in\mathbb{Z}_+} \mathbb{E}[X^{2k}]u^{-2k}$$

Proof. Since X is symmetric,

$$M_X(\lambda) = \mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[\cosh(\lambda X)\right] \geq \frac{1}{2}\mathbb{E}\left[e^{\lambda|X|}\right].$$

Thus, for any $\lambda > 0$ and $u > 0$,

$$M_X(\lambda)e^{-\lambda u} \geq \frac{1}{2}\mathbb{E}\left[e^{\lambda|X|}\right]e^{-\lambda u}.$$

From part a), with $|X| > 0$ a.s. we have

$$\inf_{\lambda>0} \mathbb{E}\left[e^{\lambda|X|}\right]e^{-\lambda u} \geq \inf_{k\in\mathbb{Z}_+} \mathbb{E}\left[|X|^k\right] u^{-k}.$$

Since X is symmetric, $\mathbb{E}[|X|^{2k}] = \mathbb{E}[X^{2k}]$, so that

$$\inf_{\lambda>0} M_X(\lambda)e^{-\lambda u} \geq \frac{1}{2} \inf_{k\in\mathbb{Z}_+} \mathbb{E}[X^{2k}]\, u^{-2k}.$$

+

□

Problem 2 (Convexity of CGF).

Show that $K_X := \log \mathbb{E}[e^{tX}]$ is convex. Use Young's inequality: for $a, b \in \mathbb{R}^d$ and $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$,

$$|a^\top b| \leq \|a\|_p \|b\|_q$$

You can assume that X has a discrete distribution.

Proof. $f(t)$ is convex if $f(\theta t_1 + (1-\theta)t_2) \leq \theta f(t_1) + (1-\theta) f(t_2)$, for all $t_1, t_2 \in \mathbb{R}$ and $\theta \in [0, 1]$

It suffices to show that $M_X(t)$ is log-convex:

(Equivalently, ...) $$M_X(\theta t_1 + (1-\theta)t_2) \leq M_X(t_1)^\theta M_X(t_2)^{1-\theta}$$

since taking log on both sides results in

$$\log M_X(\theta t_1 + (1-\theta)t_2) \leq \theta \log M_X(t_1) + (1-\theta) \log M_X(t_2)$$

Let $t = \theta t_1 + (1-\theta)t_2$. Then by discrete distribution assumption,

$$M_X(t) = \sum_i p_i e^{(\theta t_1 + (1-\theta)t_2)x_i} = \sum_i p_i (e^{t_1 x_i})^\theta (e^{t_2 x_i})^{1-\theta}$$

Let $p = \frac{1}{\theta}$ and $q = \frac{1}{1-\theta}$, then by Young's inequality,

$$M_X(t) \leq \left(\sum_i p_i e^{t_1 x_i}\right)^\theta \left(\sum_i p_i e^{t_2 x_i}\right)^{1-\theta}$$

Thus, $M_X(t) \leq M_X(t_1)^\theta M_X(t_2)^{1-\theta}$ is log-convex. □

+

[Cameron, consider breaking it down for "normies" :)
E.g.: "taking $p = \frac{1}{\theta}$ and $q = \frac{1}{1-\theta}$ we get,
since $\frac{1}{p} + \frac{1}{q} = 1$, that ..."]

Problem 3 (Gaussian Tails).

Mills Ratio

Let $\phi(\cdot)$ be the p.d.f of $\mathcal{N}(0,1)$, i.e. $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. For any $u \geq 0$, let $\Phi(u) := \int_{t \geq u} \phi(t)dt$ be the c.d.f

a) Prove the following bounds for all $u \geq 0$

$$\left(\frac{1}{u} - \frac{1}{u^3}\right)\phi(u) \leq \Phi(u) \leq \frac{1}{u}\phi(u)$$

$d\phi(t) = -t\phi(t)$

Proof. To obtain the upper bound, integration by parts is necessary.

$$\Phi(u) = \int_u^\infty \phi(t)dt = \int_u^\infty \frac{1}{t}(t\phi(t))dt \qquad \begin{array}{c|c} \textbf{D} & \textbf{I} \\ \hline 1/t & t\phi(t) \\ -1/t^2 & -\phi(t) \end{array}$$

$$\implies -\frac{1}{t}\phi(t)\Big|_u^\infty - \underbrace{\int_u^\infty \frac{1}{t^2}\phi(t)dt}_{\geq\ 0,\ \forall\ u\geq 0} \leq \cancel{\underbrace{\lim_{t\to\infty}\left(-\frac{1}{t}\phi(t)\right)}_{=\ 0}} - \left(-\frac{1}{u}\phi(u)\right) = \frac{1}{u}\phi(u)$$

$$\implies \Phi(u) \leq \frac{1}{u}\phi(u)$$

you don't need those, do you?

To obtain the lower bound, another iteration of integration by parts on the remaining integral is necessary.

$$\int_u^\infty -\frac{1}{t^2}\phi(t)dt = \int_u^\infty -\frac{1}{t^3}(t\phi(t))dt \qquad \begin{array}{c|c} \textbf{D} & \textbf{I} \\ \hline -1/t^3 & t\phi(t) \\ 3/t^4 & -\phi(t) \end{array}$$

$$\implies \frac{1}{t^3}\phi(t)\Big|_u^\infty + \underbrace{\int_u^\infty \frac{3}{t^4}\phi(t)dt}_{\geq\ 0,\ \forall\ u\geq 0} \geq \cancel{\underbrace{\lim_{t\to\infty}\left(\frac{1}{t^3}\phi(t)\right)}_{=\ 0}} - \left(\frac{1}{u^3}\phi(u)\right) = -\frac{1}{u^3}\phi(u)$$

$$\implies \Phi(u) \geq \left(\frac{1}{u} - \frac{1}{u^3}\right)\phi(u)$$

Combining the upper and lower bound gives the final bound as follows

$$\left(\frac{1}{u} - \frac{1}{u^3}\right)\phi(u) \leq \Phi(u) \leq \frac{1}{u}\phi(u)$$

$\square$

b) Now using this trick, prove a new sharper upper bound from the previous lower bound:

$$\Phi(u) \leq \left(\frac{1}{u} - \frac{1}{u^3} + \frac{3}{u^5}\right)\phi(u)$$

Proof.

$$\int_u^\infty \frac{3}{t^4}\phi(t)dt = \int_u^\infty \frac{3}{t^5}(t\phi(t))dt \qquad \begin{array}{c|c} \textbf{D} & \textbf{I} \\ \hline 3/t^5 & t\phi(t) \\ -15/t^6 & -\phi(t) \end{array}$$

$$\implies -\frac{3}{t^5}\phi(t)\Big|_u^\infty - \underbrace{\int_u^\infty \frac{15}{t^6}\phi(t)dt}_{\geq\ 0,\ \forall\ u\geq 0} \leq \underbrace{\lim_{t\to\infty}\left(-\frac{3}{t^5}\phi(t)\right)}_{=\ 0} - \left(-\frac{3}{u^5}\phi(u)\right) = \frac{3}{u^5}\phi(u)$$

$$\implies \Phi(u) \leq \left(\frac{1}{u} - \frac{1}{u^3} + \frac{3}{u^5}\right)\phi(u)$$

$\square$

c) It is clear from above that $\Phi(u)$ can continually be approximated with higher powers as you repeat the integration by parts trick to arrive to the Mills ratio as shown in Lecture 2 Theorem 2.1.

Power series for c.d.f

Show that

$$\frac{1}{2} - \Phi(u) = \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k u^{2k+1}}{2^k k!(2k+1)}$$

Proof. Note that since $\phi(u)$ is the p.d.f of $\mathcal{N}(0,1)$, the following is true:

$$\Phi(0) = \int_0^\infty \phi(t)dt = \frac{1}{2}$$

$$\implies \Phi(u) := \int_u^\infty \phi(t)dt = \frac{1}{2} - \int_0^u \phi(t)dt$$

$$\implies \frac{1}{2} - \Phi(u) = \int_0^u \phi(t)dt = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{t^2}{2}} dt$$

Let $t = ux$, such that $dt = udx$. Then,

$$\frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{t^2}{2}} dt = \frac{u}{\sqrt{2\pi}} \int_0^1 e^{-\frac{u^2}{2}x^2} dx$$

By Taylor expansion of e^x centered at 0,

$$\frac{u}{\sqrt{2\pi}} \int_0^1 e^{-\frac{u^2}{2}x^2} dx = \frac{u}{\sqrt{2\pi}} \int_0^1 \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \frac{u^{2k}}{2^k} x^{2k} dx$$

Since this sum converges absolutely, apply Fubini's Theorem and collect terms,

$$\frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k u^{2k+1}}{2^k k!} \int_0^1 x^{2k} dx = \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k u^{2k+1}}{2^k k!} \frac{x^{2k+1}}{2k+1}\Big|_{x=0}^{x=1} = \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k u^{2k+1}}{2^k k!(2k+1)}$$

$$\implies \frac{1}{2} - \Phi(u) = \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k u^{2k+1}}{2^k k!(2k+1)}$$

+ Good work! □

Problem 4 (Paley-Zygmund and Friends).

a) Prove the Paley-Zygmund inequality:

If X is a non-negative random variable with $\mathbb{E}[X^2] < \infty$, then for any $t \in [0,1]$ one has

$$\mathbb{P}(X \geq (1-t)\mathbb{E}[X]) \geq t^2 \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}$$

Proof. Let $a = (1-t)\mathbb{E}[X]$,

$$\mathbb{E}[X] = \int_0^\infty x f_X(x)dx = \int_0^a x f_X(x)dx + \int_a^\infty x f_X(x)dx$$

Since $x \leq a = (1-t)\mathbb{E}[X]$ over the interval $[0,a]$,

$$\int_0^a x f_X(x)dx \leq (1-t)\mathbb{E}[X] \underbrace{\int_0^a f_X(x)dx}_{\leq 1} \leq (1-t)\mathbb{E}[X]$$

$$\implies \mathbb{E}[x] \leq (1-t)\mathbb{E}[X] + \int_a^\infty x f_X(x)dx \implies t\mathbb{E}[X] \leq \int_a^\infty x f_X(x)dx$$

$$\implies t^2(\mathbb{E}[X])^2 \leq \left(\int_a^\infty x f_X(x)dx\right)^2$$

By Cauchy-Schwarz,

$$\left(\int_a^\infty x f_X(x)dx\right)^2 = \left(\int_a^\infty \left(x\sqrt{f_X(x)}\right)\left(\sqrt{f_X(x)}\right)dx\right)^2 \leq \Bigg(\underbrace{\int_a^\infty x^2 f_X(x)dx}_{E[X^2]}\Bigg)\Bigg(\underbrace{\int_a^\infty f_X(x)dx}_{\mathbb{P}(X \geq (1-t)\mathbb{E}[X])}\Bigg)$$

$$\implies t^2(\mathbb{E}[X])^2 \leq E[X^2]\mathbb{P}(X \geq (1-t)\mathbb{E}[X])$$

$$\implies \mathbb{P}(X \geq (1-t)\mathbb{E}[X]) \geq t^2 \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}$$

+

□

b) Now strengthen Paley-Zygmund inequality to Cantelli's inequality:

$$\mathbb{P}(X \geq (1-t)\mathbb{E}[X]) \geq t^2 \frac{(\mathbb{E}[X])^2}{t^2(\mathbb{E}[X])^2 + \mathrm{Var}[X]}$$

Give an example where this inequality is sharp.

Proof. Let $a = (1-t)\mathbb{E}[X]$.

$$\mathbb{E}\left[(X-a)^2\right] = \int_0^a (x-a)^2 f_X(x)\,dx + \int_a^\infty (x-a)^2 f_X(x)\,dx.$$

For $x \geq \cancel{a}$ μ, $(x-a)^2 \geq t^2(\mathbb{E}[X])^2$, Wrong. (Works if you take $a=\mu$)

$$\int_a^\infty (x-a)^2 f_X(x)\,dx \geq t^2(\mathbb{E}[X])^2 \int_a^\infty f_X(x)\,dx = t^2(\mathbb{E}[X])^2\,\mathbb{P}(X \geq a).$$

$$\boxed{\implies \mathbb{E}\left[(X-a)^2\right] \geq t^2(\mathbb{E}[X])^2\,\mathbb{P}(X \geq a).} \text{ - Wrong}$$

Working on the inside expression,

$$X - a = (X - \mathbb{E}[X]) + t\,\mathbb{E}[X].$$

$$\implies (X-a)^2 = (X-\mathbb{E}[X])^2 + 2t\,\mathbb{E}[X](X-\mathbb{E}[X]) + t^2(\mathbb{E}[X])^2.$$

If $x \geq \mu$

$\Rightarrow x - a = x - \mu + t\mu \geq t\mu \qquad (x-a)^2 \geq t^2\mu^2.$

$\mathbb{E}[(X-\mu)^2]$ $(a-\mu)^2$

Proof

Since $\mathbb{E}[X - \mathbb{E}[X]] = 0$,

$$\mathbb{E}\big[(X-a)^2\big] = \mathrm{Var}(X) + t^2(\mathbb{E}[X])^2. \quad +$$

$$\implies \mathrm{Var}(X) + t^2(\mathbb{E}[X])^2 \geq t^2(\mathbb{E}[X])^2\, \mathbb{P}(X \geq a).$$

$$\implies \mathbb{P}\Big(X \geq (1-t)\mathbb{E}[X]\Big) \geq \frac{t^2(\mathbb{E}[X])^2}{t^2(\mathbb{E}[X])^2 + \mathrm{Var}(X)}.$$

Pay attention ...

□

Example (Sharpness): Assume $\mathbb{E}[X] = 1$ and define the following discrete random variable

$$X = \begin{cases} 1-t, & \text{with probability } p, \\ 1+\dfrac{t}{p}, & \text{with probability } 1-p \end{cases}$$

By definition of discrete RV,

$$\mathbb{E}[X] = p(1-t) + (1-p)\left(1+\frac{t}{p}\right) = 1$$

$$\implies 1 - pt + \frac{t(1-p)}{p} = 1 \implies p^2 + p - 1 = 0$$

$$\implies p = \varphi^{-1},\ 1-p = 1-\varphi^{-1}$$

Where φ is the golden ratio. The variance for a Bernoulli random variable is as follows,

$$\mathrm{Var}(X) = p(1-p)\left(t\left(\frac{1}{p}+1\right)\right)^2 = t^2p(1-p)\left(\frac{1+p}{p}\right)^2 = t^2p(1+p)^2$$

Wrong ...

Since the lower value of X is $1-t$, the event $\{X \geq 1-t\}$ occurs when $X = 1+\frac{t}{p}$. Therefore,

$$\mathbb{P}(X \geq 1-t) = 1-p$$

With Cantelli's inequality and $\mathbb{E}[X] = 1$,

$$\mathbb{P}\Big(X \geq (1-t)\mathbb{E}[X]\Big) \geq \frac{t^2(\mathbb{E}[X])^2}{t^2(\mathbb{E}[X])^2 + \mathrm{Var}(X)} \implies \mathbb{P}\Big(X \geq (1-t)\Big) = \frac{t^2}{t^2 + \mathrm{Var}(X)}$$

Substituting the formulas from above,

$$\implies 1-p \geq \frac{t^2}{t^2 + t^2p(1+p)^2} \implies 1-p \geq \frac{1}{1+p(1+p)^2}$$

For this inequality to be sharp, we need to set the equations equal to each other,

$$1-p = \frac{1}{1+p(1+p)^2} \implies p^2+p-1 = 0$$

$$\implies p = \varphi^{-1},\ 1-p = 1-\varphi^{-1}$$

+ Great Job

Which is the same probability values derived from the first moment assumption.

Therefore Cantelli's inequality is sharp for the following discrete random variable:

$$X = \begin{cases} 1-t, -\varepsilon & \text{with probability } \varphi^{-1}, \\ 1+\varphi t, & \text{with probability } 1-\varphi^{-1} \end{cases}$$

where φ is the golden ratio.

$\mathrm{Var}(X) = \varphi^{-1}t^2 + (1-\varphi^{-1})\varphi^2 t^2 = t^2\varphi^{-1}(1+\varphi^3-\varphi^2) = t^2(\varphi^2-1) = t^2\varphi$

$\boxed{\varphi^2 - \varphi - 1 = 0}$ ok

$\mu = \varphi^{-1}(1-t) + (1-\varphi^{-1})(1+\varphi t) = 1 - \varphi^{-1}t + \varphi t - t = 1 - \varphi^{-1}t\underbrace{(1-\varphi^2+\varphi)}_{0} = 1$

$\boxed{\mathbb{P}\{X \geq a\} = 1-\varphi^{-1}}$ $\mathrm{RHS} = 1 - \frac{\sigma^2}{t^2+\sigma^2} = 1 - \frac{\cancel{t^2}\varphi}{\cancel{t^2}(1+\varphi)} = \frac{1}{1+\varphi}$

c) Now prove the generalized Paley-Zygmund inequality assuming $\mathbb{E}[|X|^p] < \infty$, for some $p > 1$,

$$\mathbb{P}(X \geq (1-t)\mathbb{E}[X]) \geq \left(t^p \frac{(\mathbb{E}[X])^p}{\mathbb{E}[|X|^p]}\right)^{\frac{1}{p-1}}$$

Proof. Following the proof for Paley-Zygmund in part a), let $a = (1-t)\mathbb{E}[X]$:

$$\mathbb{E}[X] = \int_0^\infty x f_X(x)dx = \int_0^a x f_X(x)dx + \int_a^\infty x f_X(x)dx$$

Since $x \leq a = (1-t)\mathbb{E}[X]$ over the interval $[0, a]$,

$$\int_0^a x f_X(x)dx \leq (1-t)\mathbb{E}[X] \underbrace{\int_0^a f_X(x)dx}_{\leq 1} \leq (1-t)\mathbb{E}[X]$$

$$\implies \mathbb{E}[x] \leq (1-t)\mathbb{E}[X] + \int_a^\infty x f_X(x)dx \implies t\mathbb{E}[X] \leq \int_a^\infty x f_X(x)dx$$

$$\implies t^p(\mathbb{E}[X])^p \leq \left(\int_a^\infty x f_X(x)dx\right)^p$$

Let $\frac{1}{p} + \frac{1}{q} = 1$, then by Holder's inequality,

$$\left(\int_a^\infty x f_X(x)dx\right) \leq \left(\underbrace{\int_a^\infty x^p f_X(x)dx}_{\leq \mathbb{E}[|X|^p]}\right)^{\frac{1}{p}} \left(\underbrace{\int_a^\infty f_X(x)dx}_{\mathbb{P}(X \geq (1-t)\mathbb{E}[X])}\right)^{\frac{1}{q}}$$

$$\implies \int_a^\infty x f_X(x)dx \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} \Big(\mathbb{P}(X \geq (1-t)\mathbb{E}[X]\Big)^{\frac{1}{q}}$$

$$\implies (t\mathbb{E}[X])^p \leq \mathbb{E}[|X|^p]\Big(\mathbb{P}(X \geq (1-t)\mathbb{E}[X]\Big)^{\frac{p}{q}}$$

Since $\frac{1}{p} + \frac{1}{q} = 1$, $\frac{p}{q} = p - 1$

$$\implies t^p(\mathbb{E}[X])^p \leq \mathbb{E}[|X|^p]\Big(\mathbb{P}(X \geq (1-t)\mathbb{E}[X]\Big)^{p-1}$$

Rearranging gives the final expression,

Yup

+

$$\implies \mathbb{P}(X \geq (1-t)\mathbb{E}[X]) \geq \left(t^p \frac{(\mathbb{E}[X])^p}{\mathbb{E}[|X|^p]}\right)^{\frac{1}{p-1}}$$

□

Problem 5 (Tail bound for χ^2_d)

Let $X \sim \chi^2_{2d}$, that is $X = \|Z\|^2 = Z_1^2 + \cdots + Z_{2d}^2$ where $Z \sim \mathcal{N}(0, I_d)$. Define $M_{2d}(\cdot)$ as the MGF of $X \sim \chi^2_{2d}$,

$$M_{2d}(t) := \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}$$

in particular, $M_2(t) = \mathbb{E}[e^{t(Z_1^2+Z_2^2)}]$. Our ultimate goal here is to prove that, with probability $\geq 1 - \delta$,

$$X - 2d \leq \sqrt{Cd \log\left(\frac{1}{\delta}\right)} + c \log\left(\frac{1}{\delta}\right)$$

for some numerical constants $C, c > 0$.

a) Derive the explicit form of $M_2(t)$:

$$M_2(t) \begin{cases} \frac{1}{1-2t}, & t < \frac{1}{2} \\ +\infty, & t \geq \frac{1}{2} \end{cases}$$

Proof.

$$M_2(t) = \mathbb{E}[e^{t(Z_1^2+Z_2^2)}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t(z_1^2+z_2^2)} \frac{1}{2\pi} e^{-(z_1^2+z_2^2)/2} dz_1 dz_2$$

Transform integral to polar coordinates $(z_1, z_2) \mapsto (r, \theta)$ with $r = \sqrt{z_1^2 + z_2^2}$

$$\frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} r e^{r^2(t-1/2)} dr d\theta = \int_0^{\infty} r e^{-r^2(-t+1/2)} dr$$

Let $u = r^2 \implies du = 2rdr$

$$\frac{1}{2} \int_0^{\infty} e^{-u(-t+1/2)} du = \frac{1}{2(t-1/2)} e^{-u(-t+1/2)} \Big|_{u=0}^{u \to \infty} = \frac{1}{1-2t}, \quad t < \frac{1}{2}$$

$$\therefore \quad M_2(t) \begin{cases} \frac{1}{1-2t}, & t < \frac{1}{2} \\ +\infty, & t \geq \frac{1}{2} \end{cases}$$

(And BTW, its also clear that $M_1(t) = \frac{1}{\sqrt{1-2t}}$, isn't it? □

Given this, it is clear that as you add more squares of standard gaussians, i.e. chi-squared with higher degrees of freedom, that its moment generating function follows (See ProofWiki):

$$M_{2d}(t) = \frac{1}{(1-2t)^d}, \quad t < \frac{1}{2}$$

+

b) Using Chernoff's method, bound the tail function $\mathbb{P}(X > x)$, for any $x > 2d$ as

$$\mathbb{P}(X > x) \leq \inf_{t < \frac{1}{2}} \frac{e^{-tx}}{(1-2t)^d} = \exp\left(d \log\left(\frac{x}{2d}\right) - \frac{x-2d}{2}\right)$$

Proof. Since $u \mapsto \log(u) \in \mathbb{R}_+$ is monotonically increasing,

$$\inf_{t<\frac{1}{2}} \frac{e^{-tx}}{(1-2t)^d} = \exp\left(\inf_{t<\frac{1}{2}} \log\left(\frac{e^{-tx}}{(1-2t)^d}\right)\right) = \exp\left(\inf_{t<\frac{1}{2}} \underbrace{\Big(-tx - d\log(1-2t)\Big)}_{g(t)}\right)$$

Optimizing $g(t)$ yields the following,

$$g'(t) = 0 \implies -x + \frac{2d}{1-2t} = 0 \implies t^\star := t = (1/2)\left(1 - \frac{2d}{x}\right)$$

Plugging in $g(t^\star)$ and doing simple algebraic manipulations clearly lead to the final expression:

$$\mathbb{P}(X > x) \leq \exp\left(d \log\left(\frac{x}{2d}\right) - \frac{x-2d}{2}\right)$$

+

□

c) **Bonus.** Derive subexponential concentration for chi-squared distribution.

(i) Show that

$$\mathbb{P}(X - 2d > z) \leq \begin{cases} \exp\left(-\dfrac{z^2}{16d}\right) & \text{for } \ 0 \leq z \leq 2d \\ \exp\left(-\dfrac{z}{8}\right) & \text{for } \ z > 2d \end{cases}$$

Proof. Let $z = x - 2d$, so that $x = 2d + z$ and $z \geq 0$. From part b) we have

$$\mathbb{P}(X > x) = \mathbb{P}(X - 2d > z) \leq \exp\Big(d \log\Big(\frac{x}{2d}\Big) - \frac{x - 2d}{2}\Big) = \exp\Big(d \log\Big(1 + \frac{z}{2d}\Big) - \frac{z}{2}\Big)$$

For when $0 \leq z \leq 2d$:
Let $u = \frac{z}{2d}$, so that $0 \leq u \leq 1$,

$$\implies d \log\Big(1 + \frac{z}{2d}\Big) - \frac{z}{2} = d \log(1 + u) - d\,u$$

Since for $0 \leq u \leq 1$ we have

$$\log(1 + u) \leq u - \frac{u^2}{4},$$

$$\implies d \log(1 + u) - d\,u \leq -\frac{d\,u^2}{4} = -\frac{z^2}{16d}$$

$$\implies \mathbb{P}(X - 2d > z) \leq \exp\Big(-\frac{z^2}{16d}\Big) \quad \text{for } 0 \leq z \leq 2d$$

For when $z > 2d$:
Let $u = \frac{z}{2d}$, so that $u > 1$

$$\implies d \log\Big(1 + \frac{z}{2d}\Big) - \frac{z}{2} = d \log(1 + u) - d\,u$$

Want to show that

$$d \log(1 + u) - d\,u \leq -\frac{z}{8}$$

Since $z = 2d\,u$, this is equivalent to

$$d \log(1 + u) - d\,u \leq -\frac{d\,u}{4} \quad \iff \quad \log(1 + u) \leq \frac{3u}{4}$$

Define

$$h(u) = \frac{3u}{4} - \log(1 + u)$$

Then,

$$h'(u) = \frac{3}{4} - \frac{1}{1 + u}$$

For $u \geq 1$,

$$h'(u) \geq \frac{3}{4} - \frac{1}{2} = \frac{1}{4} > 0,$$

so that $h(u)$ is increasing on $[1, \infty)$. At $u = 1$,

$$h(1) = \frac{3}{4} - \log 2 \geq 0$$

Thus, $h(u) \geq 0$ for all $u \geq 1$, i.e.,

$$\log(1 + u) \leq \frac{3u}{4} \quad \text{for } u \geq 1$$

$$\implies \mathbb{P}(X - 2d > z) \leq \exp\Big(-\frac{z}{8}\Big) \quad \text{for } z > 2d$$

Having shown both cases, the final expression is as follows:

$$\mathbb{P}(X - 2d > z) \leq \begin{cases} \exp\Big(-\dfrac{z^2}{16d}\Big) & 0 \leq z \leq 2d \\ \exp\Big(-\dfrac{z}{8}\Big) & z > 2d \end{cases}$$

$\square$

(ii) Reformulating the last bound to

$$\mathbb{P}(X - 2d > z) \leq \exp\left(- \min\left\{ \frac{z^2}{16d}, \frac{z}{8} \right\}\right)$$

and letting $\mathbb{P}(X - 2d > z) \leq \delta$, "invert" the last inequality to obtain the inequality we wanted to prove at beginning, with $C = 16$ and $c = 8$. Hint: $\max\{a, b\} \leq a + b$ for $a, b \geq 0$.

Proof. From part (i) we have

$$\mathbb{P}(X - 2d > z) \leq \exp\left(- \min\left\{ \frac{z^2}{16d}, \frac{z}{8} \right\}\right).$$

Inverting this inequality,

$$\exp\left(- \min\left\{ \frac{z^2}{16d}, \frac{z}{8} \right\}\right) \leq \delta$$

With $\delta \in (0, 1)$ take log of both sides,

$$- \min\left\{ \frac{z^2}{16d}, \frac{z}{8} \right\} \leq \ln \delta$$

$$\implies \min\left\{ \frac{z^2}{16d}, \frac{z}{8} \right\} \geq = \log \frac{1}{\delta}$$

Thus,

$$\frac{z^2}{16d} \geq \ln \frac{1}{\delta} \quad \text{and} \quad \frac{z}{8} \geq \ln \frac{1}{\delta}$$

$$\implies z \geq \sqrt{16d \ln \frac{1}{\delta}} \quad \text{and} \quad z \geq 8 \ln \frac{1}{\delta}$$

$$\implies z \geq \max\left\{ \sqrt{16d \ln \frac{1}{\delta}}, 8 \ln \frac{1}{\delta} \right\}$$

Using the hint that for any $a, b \geq 0$, $\max\{a, b\} \leq a + b$,

$$z \leq \sqrt{16d \ln \frac{1}{\delta}} + 8 \ln \frac{1}{\delta}$$

Therefore, with probability $\geq 1 - \delta$,

$$X - 2d \leq \sqrt{16d \ln \frac{1}{\delta}} + 8 \ln \frac{1}{\delta}$$

$+$

Welcome to the $\log(1/\delta)$ world !

□

Problem 6 (Stein's Paradox)

Consider the problem of estimating the mean μ in the multivariate Gaussian location family:

$$P_\mu = \mathcal{N}(\mu, I_d), \quad \mu \in \mathbb{R}^d,$$

where I_d is the $d \times d$ identity matrix, from a single observation $X \sim P_\mu$. Note that here, X itself is the maximum likelihood estimator (MLE) for μ. Defining for any estimator $\hat{\mu} = \hat{\mu}(X)$ of μ the variance

$$\mathrm{Var}_\mu[\hat{\mu}] := \mathbb{E}_\mu \|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2$$

and the quadratic risk

$$\mathrm{Risk}_\mu[\hat{\mu}] := \mathbb{E}_\mu \|\hat{\mu} - \mu\|^2,$$

where $\|x\| := \left(\sum_i x_i^2\right)^{1/2}$ is the Euclidean norm of $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we see that for any $\mu \in \mathbb{R}^d$,

$$\mathrm{Risk}_\mu[X] = \mathrm{Var}_\mu[X] = d.$$

Intuitively, one can suspect that no better estimator of X can be found: really, what can be done with only a single observation of the mean? Yet, this turns out to be false: one may improve over the MLE uniformly on the family (3) when $d > 2$. This celebrated result was established by James and Stein in 1976, and our goal is to reproduce it. But first, let us establish the terminology.

Definition 1. An estimator $\hat{\mu}$ is **dominated** by some other estimator $\hat{\mu}'$ if $\mathrm{Risk}_\mu[\hat{\mu}'] \le \mathrm{Risk}_\mu[\hat{\mu}]$ for any μ, and there exists a parameter value $\overline{\mu}$ such that $\mathrm{Risk}_{\overline{\mu}}[\hat{\mu}'] < \mathrm{Risk}_{\overline{\mu}}[\hat{\mu}]$.

Definition 2. An estimator $\hat{\mu}$ is called **admissible** if it is not dominated by any other estimator. Otherwise, it is called **inadmissible**.

As statisticians, ideally, we would like to compare two estimators over the whole family at once, without specifying a value of μ. Two admissible estimators cannot be compared this way, but at the very least we can rule out any inadmissible estimator, as for it there exists a uniformly better one. You will show that the MLE is inadmissible when $d \ge 3$, by constructing a dominating estimator.

a) Consider **shrinkage estimators** $\hat{\mu} = sX$ with $s \in \mathbb{R}$, and compute their risks for any s. Show that one can restrict attention to $s \in [0, 1]$ (hence "shrinkage") by finding a dominating estimator for $\hat{\mu}$ with $s < 0$ or $s > 1$.

Proof.

$$\mathrm{Risk}_\mu[\hat{\mu}] = \mathbb{E}_\mu[\|sX - \mu\|^2] = \underbrace{\mathbb{E}_\mu[\|sX\|^2]}_{(a)} - 2s \underbrace{\mathbb{E}_\mu[X^\top \mu]}_{(b)} + \mathbb{E}_\mu[\|\mu\|^2] \tag{1}$$

For $s < 0$, let a new estimator be $s'X$ where $s' := -s$. This new estimator dominates (1) because for $s < 0$, the (b) term becomes positive, but for the new s' shrinkage estimator, that term stays negative. This means that the risk for that new estimator is less than or equal to the original shrinkage estimator, and for $\overline{\mu} = 1$, it is clear this is new risk strictly less than (1).

For $s > 1$, let a new estimator be $s'X = X$, where $s' = 1$. This estimator is dominating to (1) since the (a) term is quadratic in s such that for $s > 1$, that term is larger than the (b) term. Trivially for $\overline{\mu} = 0$, the new estimator risk is strictly less than (1). □

b) Show that, for given μ, the best value of s—i.e., the one minimizing the risk—is given by

$$s^* = \frac{\|\mu\|^2}{d + \|\mu\|^2} = 1 - \frac{d}{d + \|\mu\|^2}.$$

Proof. Since $X \sim \mathcal{N}(\mu, I_d)$, $\mathbb{E}[\|X - \mu\|^2] = d$:

$$\mathbb{E}[\|sX - \mu\|^2] = s^2 \mathbb{E}[\|X - \mu\|^2] + (1 - s)^2 \|\mu\|^2 = s^2 d + (1 - s)^2 \|\mu\|^2$$

$$\frac{\partial}{\partial s}\left(s^2 d + (1-s)^2\|\mu\|^2\right) = 0 \implies \frac{\partial}{\partial s}\left(s^2\left(d+\|\mu\|^2\right) - 2s\|\mu\|^2 + \|\mu\|^2\right) = 0$$

$$\implies 2s\left(d+\|\mu\|^2\right) - 2\|\mu\|^2 = 0 \implies s^\star := s = \frac{\|\mu\|^2}{d+\|\mu\|^2} = 1 - \frac{d}{d+\|\mu\|^2}$$

$\square$

c) Unfortunately, $\hat{\mu}^* = s^* X$ is not a proper estimator. (*Why?*) Instead of it, one may consider

$$\left(1 - \frac{d}{\|X\|^2}\right) X,$$

which is an actual estimator. Can you explain the heuristic motivation behind this estimator?

This optimized shrinkage estimator is not a proper estimator because it uses μ, which is what you are trying to estimate in the first place; in other words this estimator is circular. The heuristic motivation behind the new estimator comes from the fact that $d + \|\mu\|^2 = \mathbb{E}[\|X\|^2] \approx \|X\|^2$, which is the actual data we can observe.

I'm not aware of the term but ok. ;)

d) Assuming that $d \geq 2$, derive the **James-Stein estimator**

$$\hat{\mu}^{\mathrm{JS}} = \left(1 - \frac{d-2}{\|X\|^2}\right) X$$

by minimizing over $\delta \in \mathbb{R}$ the risk of the estimator

$$\hat{\mu}^\delta = \left(1 - \frac{\delta}{\|X\|^2}\right) X$$

for a fixed μ. In order to show that $R(\delta) = \mathrm{Risk}_\mu[\hat{\mu}^\delta]$ is minimized at $d-2$, use Stein's lemma:
Lemma 1. Let $X \sim \mathcal{N}(\mu, I)$ and $g(x)$ be a function on $\mathbb{R}^d$ differentiable almost everywhere, and such that $\mathbb{E}_\mu\left[\left|\frac{\partial}{\partial x_i} g(X)\right|\right] < \infty$ and $\mathbb{E}_\mu\|(X_i - \mu_i) g(X)\| < \infty$ for any $i \in [d] := \{1, 2, \dots, d\}$. Then

$$\mathbb{E}_\mu[(X_i - \mu_i) g(X)] = \mathbb{E}_\mu\left[\frac{\partial}{\partial x_i} g(X)\right], \quad i \in [d].$$

When applying Stein's lemma to the right function $g(X)$, please do check the absolute integrability conditions in its premise, and explain why the argument does not work for $d = 1$.
Finally, verify that $R(\delta)$ is strictly convex when $d \geq 3$ (thus $\hat{\mu}^{\mathrm{JS}}$ indeed dominates the MLE).

Proof. Consider the estimator

$$\hat{\mu}^\delta = \left(1 - \frac{\delta}{\|X\|^2}\right) X, \quad X \sim \mathcal{N}(\mu, I_d).$$

Its risk is

$$R(\delta) = \mathbb{E}_\mu\left[\|\hat{\mu}^\delta - \mu\|^2\right].$$

Write

$$\hat{\mu}^\delta - \mu = \left(1 - \frac{\delta}{\|X\|^2}\right) X - \mu = (X - \mu) - \frac{\delta}{\|X\|^2} X.$$

Then,

$$\|\hat{\mu}^\delta - \mu\|^2 = \|X - \mu\|^2 - 2\frac{\delta}{\|X\|^2}(X-\mu)^\top X + \frac{\delta^2}{\|X\|^4}\|X\|^2$$

$$= \|X - \mu\|^2 - 2\frac{\delta}{\|X\|^2}(X-\mu)^\top X + \frac{\delta^2}{\|X\|^2}.$$

Taking expectation and noting that $\mathbb{E}_\mu\|X - \mu\|^2 = d$, we get

$$R(\delta) = d - 2\delta\, \mathbb{E}_\mu\left[\frac{(X-\mu)^\top X}{\|X\|^2}\right] + \delta^2\, \mathbb{E}_\mu\left[\frac{1}{\|X\|^2}\right].$$

Define

$$A = \mathbb{E}_\mu\left[\frac{(X-\mu)^\top X}{\|X\|^2}\right], \quad B = \mathbb{E}_\mu\left[\frac{1}{\|X\|^2}\right].$$

Thus,

$$R(\delta) = d - 2\delta A + \delta^2 B.$$

For each coordinate i, set

$$g_i(X) = \frac{X_i}{\|X\|^2}.$$

Then by Stein's lemma,

$$\mathbb{E}_\mu\left[(X_i-\mu_i)g_i(X)\right] = \mathbb{E}_\mu\left[\frac{\partial}{\partial x_i}g_i(X)\right].$$

Since

$$\frac{\partial}{\partial x_i}\left(\frac{x_i}{\|x\|^2}\right) = \frac{\|x\|^2 - 2x_i^2}{\|x\|^4}, \qquad +$$

we have

$$\mathbb{E}_\mu\left[\frac{(X_i-\mu_i)X_i}{\|X\|^2}\right] = \mathbb{E}_\mu\left[\frac{\|X\|^2 - 2X_i^2}{\|X\|^4}\right]. \qquad +$$

Summing over $i = 1, \ldots, d$:

$$\begin{aligned} A &= \sum_{i=1}^d \mathbb{E}_\mu\left[\frac{\|X\|^2 - 2X_i^2}{\|X\|^4}\right] = \mathbb{E}_\mu\left[\frac{d\|X\|^2 - 2\sum_{i=1}^d X_i^2}{\|X\|^4}\right] \\ &= \mathbb{E}_\mu\left[\frac{d\|X\|^2 - 2\|X\|^2}{\|X\|^4}\right] = \mathbb{E}_\mu\left[\frac{d-2}{\|X\|^2}\right] = (d-2)B. \qquad + \end{aligned}$$

The use of Stein's lemma does not work for $d = 1$ because that means $g(x) = 1/x$, which makes this not have a definite integral from 0 to infinity to be less than infinity.

Yup. (And for $d=2$ it does work but makes no difference.)

Substitute $A = (d-2)B$ into the risk:

$$R(\delta) = d - 2\delta(d-2)B + \delta^2 B = d + B\left(\delta^2 - 2(d-2)\delta\right).$$

Minimize the quadratic $f(\delta) = \delta^2 - 2(d-2)\delta$. Its derivative is

$$f'(\delta) = 2\delta - 2(d-2) = 0 \quad \Longrightarrow \quad \delta = d-2.$$

Thus, the minimizer is $\delta^* = d-2$, and the corresponding estimator is

$$\hat{\mu}^{\mathrm{JS}} = \left(1 - \frac{d-2}{\|X\|^2}\right)X.$$

Since $B > 0$, $R(\delta)$ is strictly convex in δ (and for $d \geq 3$ we have $d - 2 > 0$). Hence, the James–Stein estimator strictly dominates the MLE when $d \geq 3$. $\square$

Problem 7 (Planar Venn Diagrams)

Prove that one cannot draw a planar Venn diagram for $n \geq 5$ sets by shifting a circle.
Use **Euler's formula**: any planar graph with V vertices, E edges, and F faces satisfies

$$V - E + F = 2$$

Proof. For the $n-1$ circles, assume Euler's formula holds

$$V_{n-1} - E_{n-1} + F_{n-1} = 2$$

For any graph to realize all intersections and be a valid Venn Diagram, the vertices must equal:

$$\boxed{V_n = V_{n-1} + 2(n-1)}$$ +

Each intersection splits the new circle into at most $2(n-1)$ edges and each existing circle gains 2 edges:

$$\boxed{E_n \leq E_{n-1} + 4(n-1)}$$ +

A Venn diagram for n sets must have exactly 2^n:

$$F_n = 2^n$$

This new Venn diagram must satisfy Euler's formula,

$$V_n - E_n + F_n = 2 \implies [V_{n-1} + 2(n-1)] - [E_{n-1} + 4(n-1)] + F_n \leq 2$$

$$\implies \underbrace{V_{n-1} - E_{n-1} + F_{n-1}}_{=\ 2 \text{ by assumption}} - 2(n-1) + F_n \leq F_{n-1} + 2$$

$$\implies F_n \leq F_{n-1} + 2(n-1)$$

Using proof by induction to show $F_n \leq F_{n-1} + 2(n-1) = n^2 - n + 2$,
Base Case: For $n = 1$,

$$F_1 = 2 = 1^2 - 1 + 2.$$

Inductive Step: Assume that for some $k \geq 1$,

$$F_k \leq k^2 - k + 2.$$

Then

$$F_{k+1} \leq F_k + 2k \leq \left(k^2 - k + 2\right) + 2k = k^2 + k + 2 = (k+1)^2 - (k+1) + 2.$$

Thus,

$$F_n \leq n^2 - n + 2 \quad \text{for all } n \geq 1.$$ +

Since $F_n = 2^n$, a valid Venn diagram occurs only when

$$2^n \leq n^2 - n + 2$$ (n=4: 16-4+2=14)

And since for $n \geq 4$ that inequality does not hold, there is no way to make a Venn diagram from shifting 4 or more circles. ☐

+

Well done!