

Total: 75/100

ISYE 8803 Homework 1

Samuel Talkington

February 7, 2025

1 MGF method vs. moment bounds

It is natural to compare the best bound on the tails obtained via MGF and by bounding the moments. As it turns out, the moment bounds are sharper, even if we only use the integer moments.

Definition 1.1 (Moment-generating function (MGF)). For a random variable X , the MGF of X , $M_X : \mathbb{R} \rightarrow \mathbb{R}_+$ is given as

$$M_X(\lambda) := \mathbb{E} [e^{\lambda X}].$$

1.1 Part a

Proposition 1 (Markov's Inequality). Let $X > 0$ almost surely, and let $t > 0$. Then,

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Lemma 1. Let $X > 0$ almost surely. 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. For any $u > 0$, we have that

$$\begin{aligned} \inf_{\lambda > 0} M_X(\lambda) e^{-\lambda u} &:= \inf_{\lambda > 0} \mathbb{E}[e^{\lambda X}] e^{-\lambda u} \\ &= \inf_{\lambda > 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda u}} \\ &\stackrel{\square}{\geq} \inf_{\lambda > 0} \Pr(e^{\lambda X} \geq e^{\lambda u}) \\ &\stackrel{(i)}{=} \inf_{\lambda > 0} \Pr(\lambda X \geq \lambda u) \\ &= \Pr(X \geq u) \end{aligned}$$

where in step (i) we applied $\log(\cdot)$ to both sides of the event. □

1.2 Part b

This is what
I call
a thorough
person :)

⊖

2 Convexity of the cumulant-generating function

Definition 2.1 (Cumulant-generating function (CGF)). The cumulant-generating function (CGF) of a random variable X $K_X : \mathbb{R} \rightarrow \mathbb{R}_+$ is given as

$$K_X(\lambda) := \log(\mathbb{E}[e^{\lambda X}]) = \log(M_X(\lambda)).$$

Proposition 2 (Convexity of the CGF). Let X be a discrete random variable distributed over $\{x_i\}_{i=1}^{\infty}$, and let $p_i := \Pr(X = x_i)$. Then, the CGF K_X is convex over the extended real line $[-\infty, \infty]$.

Proof using second derivative. First, note that the first and second derivatives of the MGF are

$$M'_X(\lambda) = \mathbb{E}[Xe^{\lambda X}], \quad M''_X(\lambda) = \mathbb{E}[X^2 e^{\lambda X}].$$

Recall that a twice-differentiable function $f : \text{dom}(f) \rightarrow \mathbb{R}$ is convex if and only if

$$f''(x) \geq 0 \quad \forall x \in \text{dom}(f).$$

Observe that K_X is differentiable, where

$$K'_X(\lambda) := \frac{\partial}{\partial \lambda} K_X(\lambda) = \frac{M'_X(\lambda)}{M_X(\lambda)} = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]},$$

and

$$\begin{aligned} K''_X(\lambda) &= \frac{\partial^2}{\partial \lambda^2} K_X(\lambda) = \frac{M_X(\lambda)M''_X(\lambda) - (M'_X(\lambda))^2}{M_X(\lambda)^2} \\ &= \frac{\mathbb{E}[e^{\lambda X}] \mathbb{E}[X^2 e^{\lambda X}] - (\mathbb{E}[Xe^{\lambda X}])^2}{(\mathbb{E}[e^{\lambda X}])^2} \\ &\geq \frac{\mathbb{E}[e^{\lambda X}] \mathbb{E}[X^2 e^{\lambda X}] - \mathbb{E}[X^2 e^{2\lambda X}]}{(\mathbb{E}[e^{\lambda X}])^2} \end{aligned}$$

$e^t \geq 0 \forall t \in \mathbb{R}$

≥ 0
by Cauchy-Schwarz (check)
- you were very close!

Now, recall that for a discrete probability mass function, we have

$$M_X(\lambda) := \sum_{i=0}^{\infty} e^{\lambda x_i} p_i,$$

moreover,

$$M_X^{(n)}(\lambda) :=$$

□

Proof using first derivative. Recall that a differentiable function $f : \text{dom}(f) \rightarrow \mathbb{R}$ is convex if and only if

$$f(x_1) \geq f(x_2) + f'(x_2)(x_1 - x_2) \quad \forall x_1, x_2 \in \text{dom}(f).$$

Furthermore, for all $\lambda_1, \lambda_2 \in \mathbb{R}$, we have

$$K_X(\lambda_1) - K_X(\lambda_2) = \log \mathbb{E}[e^{\lambda_1 X}] - \log \mathbb{E}[e^{\lambda_2 X}]$$

$$= \log \left(\frac{\mathbb{E}[e^{\lambda_1 X}]}{\mathbb{E}[e^{\lambda_2 X}]} \right)$$

$$= \log \left(\mathbb{E}[e^{(\lambda_1 - \lambda_2) X}] \right)$$

$$\stackrel{(1)}{\geq} \mathbb{E}[(\lambda_1 - \lambda_2) X]$$

this is wrong!

where step (1) is the reverse Jensen's inequality.

□

3 Gaussian tails

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

3.1 Part 1: Mills ratio

3.1.1 Part 1.a

Lemma 2. For all $u \geq 0$ the following holds:

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

Proof of the upper bound. This is easily shown by integration by parts. To begin, observe that

$$\begin{aligned} \Phi(u) &:= \int_{t \geq u} \phi(t) dt := \frac{1}{\sqrt{2\pi}} \int_{t=u}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt \\ &\stackrel{(1)}{=} \frac{1}{\sqrt{2\pi}} \int_{t=u}^{\infty} \frac{1}{t} t \exp\left(-\frac{t^2}{2}\right) dt \\ &:= \int_{t=u}^{\infty} \frac{1}{t} \cdot (t\phi(t)) dt, \end{aligned}$$

where in step (1) we multiplied the integrand by t and $1/t$.

Let $u(t) := \frac{1}{t}$, and $\frac{dv}{dt} := t\phi(t)$. Then, for some constant C :

$$v = \int dv = \int t\phi(t) dt = \int t e^{-\frac{t^2}{2}} dt = -e^{-\frac{t^2}{2}} + C := -\phi(t) + C.$$

Additionally, we have the infinitesimal relation

$$\frac{du}{dt} = -\frac{1}{t^2} \implies du = -\frac{1}{t^2} dt,$$

So, applying integration by parts, we simply obtain the identities

$$\Phi(u) = -\frac{1}{t} \phi(t) \Big|_{t=u}^{\infty} - \int_{t=u}^{\infty} \frac{1}{t^2} \phi(t) dt = \left(\lim_{b \rightarrow \infty} -\frac{1}{b} \phi(b) + \frac{1}{u} \phi(u) \right) - \int_{t=u}^{\infty} \frac{1}{t^2} \phi(t) dt = \frac{1}{u} \phi(u) - \int_{t=u}^{\infty} \frac{1}{t^2} \phi(t) dt.$$

Now, to obtain the upper bound, observe that

$$\Phi(u) = \frac{1}{u} \phi(u) - \underbrace{\int_{t=u}^{\infty} \frac{1}{t^2} \phi(t) dt}_{\geq 0} \stackrel{(1)}{\leq} \frac{1}{u} \phi(u) \tag{2}$$

where inequality (1) is by applying the assumption that $u \geq 0$, and dropping the integral. □

Proof of the lower bound. Obtaining the lower bound again amounts to integration by parts. Considering the integral that we dropped in (2), we continue with integration by parts to obtain

$$\Phi(u) = \int_{t=u}^{\infty} -\frac{1}{t^2} \phi(t) dt = \int_{t=u}^{\infty} -\frac{1}{t^3} \cdot (t\phi(t)) dt$$

=



□


3.1.2 Part 1.b.



3.1.3 Part 1.c.



3.2 Part 2: Power series for the CDF

Proposition 3. *Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the* 

4 Payley-Zygmund and friends

Below we introduce a counterpart of Markov's inequality. A nonnegative random variable cannot be much *smaller* than its expectation.

4.1 Part i: Paley-Zygmund inequality

Lemma 3 (Paley-Zygmund Inequality). *Let X be a non-negative random variable with finite second moment; $E[X^2] < \infty$. Then, for any $t \in [0, 1]$,*

$$\Pr(X \geq (1-t)E[X]) \geq t^2 \frac{(E[X])^2}{E[X^2]}. \quad (3)$$

Proof. As X is nonnegative by assumption, by definition, we can bound the expectation of X as

$$\begin{aligned} E[X] &:= \int_{x=0}^{\infty} x f_X(x) dx \\ &\stackrel{(1)}{=} \underbrace{\int_{x=0}^k x f_X(x) dx}_{:=T_1} + \int_{x=k}^{\infty} x f_X(x) dx, \end{aligned}$$

where step (1) is valid for any finite $k \in \mathbb{R}_{++}$. Now, set $k := (1-t)E[X]$ for any $t \in [0, 1]$. Hence, we have $E[X] \geq k \geq 0$. And consequently, as $x \in [0, k]$ over the interval, the first component of the integral summation is:

$$\begin{aligned} T_1 &= \int_{x=0}^k x f_X(x) dx \\ &\leq (1-t)E[X] \underbrace{\int_{x=0}^k f_X(x) dx}_{\leq 1} \quad (\text{typo}) \\ &\leq (1-t)E[X]. \end{aligned}$$

Combining the above with the previous inequality, we have

$$E[X] \leq T_1 + T_2 \leq (1-t)E[X] + \int_{x=k}^{\infty} x f_X(x) dx.$$

step (2) is because $\int_0^t x f_X(x) dx \geq 0$ for the nonnegative realizations $x \in \mathbb{R}_+$, and step (3) is because $x \geq k$ in the integral.

More generally, the second moment of X can be bounded as

$$\begin{aligned} E[X^2] &:= \int_{x=0}^{\infty} x^2 f_X(x) dx \\ &= \int_{x=0}^k x^2 f_X(x) dx + \int_{x=k}^{\infty} x^2 f_X(x) dx \\ &\geq \int_{x=k}^{\infty} x^2 f_X(x) dx \\ &\geq k^2 \Pr(X \geq k). \end{aligned}$$

Now, suppose that $k := (1 - t) E[X]$ for some $t \in [0, 1]$; consequently, we have $E[X] \geq k \geq 0$. Then, from the first inequality chain, we have

$$E[X] \geq (1 - t) E[X] \Pr(X \geq (1 - t) E[X]);$$

equivalently, we can divide $E[X] > 0$ from both sides, yielding,

$$1 \geq (1 - t) \Pr(X \geq (1 - t) E[X]) \iff \Pr(X \geq (1 - t) E[X]) \geq \frac{\Pr(X \geq (1 - t) E[X])}{t}.$$

Next, note that $k^2 = 1 - 2t + t^2$, and from the second inequality chain,

$$E[X^2] \geq (1 - 2t + t^2) \Pr(X \geq 1 - 2t + t^2)$$



4.2 Part ii: Cantelli's inequality

Lemma 4 (Cantelli's Inequality). Under the same conditions as Lemma [3](#) we have that

$$\Pr(X \geq (1 - t) E[X]) \geq t^2 \frac{(E[X])^2}{t^2 (E[X])^2 + \text{var}(X)}.$$



4.3 Part iii: Generalized Paley-Zygmund inequality

Theorem 1 (Generalized Paley-Zygmund Inequality). Assume that $E[|X|^p] < \infty$ for some $p > 1$. Then, [\(3\)](#) can be generalized to

$$\Pr(X \geq (1 - t) E[X]) \geq \left(t^p \frac{(E[X])^p}{E[|X|^p]} \right)^{\frac{1}{p-1}}. \quad (4)$$

Note that when $p > 2$, [\(4\)](#) gives an improvement over [\(3\)](#) for small t , which is important in applications where X is the sample average of i.i.d. random variables Y_1, \dots, Y_n .

Proof. We apply the same technique as part i, but raise to the power $p > 2$.

More precisely?

5 Tail bound for χ_d^2

5.1 Part a

Proposition 4. Let $Z \sim \mathcal{N}(0, \mathbf{I}_d)$ be a d dimensional standard Gaussian, and let $X \sim \chi_d^2$ be a Chi-Squared random variable with d degrees of freedom i.e. $X := Z^\top Z$. Let $M_2(t)$ be the moment-generating function of X when $d = 2$. Then,

$$M_2(t) := \mathbb{E} \left[e^{t(Z_1^2 + Z_2^2)} \right] = \begin{cases} \frac{1}{1-2t}, & t < \frac{1}{2}, \\ +\infty & t > \frac{1}{2}, \\ \text{undefined} & t = \frac{1}{2}. \end{cases} \quad (5)$$

Proof. The probability density function (PDF) of a multivariate Gaussian $\mathbf{Z} = [Z_1, \dots, Z_n]^\top \in \mathbb{R}^n$ with covariance $\Sigma \succeq 0$, $\Sigma \in \mathbb{R}^{n \times n}$, is given as follows:

$$f_Z(z_1, \dots, z_n) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp \left(-\frac{1}{2} (\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top \Sigma^{-1} (\mathbf{Z} - \mathbb{E}[\mathbf{Z}]) \right).$$

In the special case where $n = 2$:

$$f_Z(z_1, z_2) = \frac{1}{2\pi} \exp \left(-\frac{1}{2} (z_1^2 + z_2^2) \right).$$

Then, for any $t \in \mathbb{R}$, by the law of the unconscious statistician (LOTUS), we have

$$\begin{aligned} M_2(t) &= \mathbb{E} \left[e^{t(Z_1^2 + Z_2^2)} \right] \\ &= \int_{z_1=-\infty}^{\infty} \int_{z_2=-\infty}^{\infty} f_Z(z_1, z_2) e^{t(z_1^2 + z_2^2)} dz_1 dz_2 \\ &= \frac{1}{2\pi} \int_{z_1=-\infty}^{\infty} \int_{z_2=-\infty}^{\infty} e^{(z_1^2 + z_2^2)(t - \frac{1}{2})} dz_1 dz_2. \end{aligned}$$

Plugging the above integral into Mathematica yields exactly the desired result, 5.

To show this analytically, transform to polar coordinates by setting $r := \sqrt{z_1^2 + z_2^2}$, and $\theta := \arctan \left(\frac{z_2}{z_1} \right) \in$

$[0, 2\pi]$. Then,

$$\begin{aligned}
M_2(t) &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} r e^{r^2(t-\frac{1}{2})} dr d\theta \\
&\stackrel{(1)}{=} \int_{r=0}^{\infty} r e^{r^2(t-\frac{1}{2})} dr \\
&\stackrel{(2)}{=} \frac{1}{2} \int_{r=0}^{\infty} e^{u(t-\frac{1}{2})} du \\
&= \frac{1}{2(t-\frac{1}{2})} \cdot \lim_{c \rightarrow \infty} \left(e^{u(t-\frac{1}{2})} \Big|_{u=0}^c \right) \\
&= \frac{1}{2t-1} \lim_{c \rightarrow \infty} \left(e^{-c(\frac{1}{2}-t)} - 1 \right) \\
&= \begin{cases} \frac{1}{(1-2t)} & t < \frac{1}{2} \\ +\infty & t > \frac{1}{2} \\ \text{undefined} & t = \frac{1}{2}, \end{cases}
\end{aligned}$$

where in step (1) we integrated over the circumference of a circle and cancelled out the $1/2\pi$ factor, and in step (2) we applied a simple u -substitution with $u := r^2$. This completes the proof. + \square

Corollary 1.1. In general, let $Z \sim \mathcal{N}(0, \mathbf{I}_{2d})$ where $d \geq 2$. Then

$$M_d(t) = \frac{1}{(1-2t)^{d/2}}, \quad t < \frac{1}{2}. \quad (6)$$

Proof. First, note that $\det(\text{cov}(Z)) = 1$. Then, for any $t \in \mathbb{R}$, we have

$$\begin{aligned}
M_d(t) &= \mathbb{E} \left[e^{tZ^\top Z} \right] \\
&\stackrel{(1)}{=} \int_{z_1=-\infty}^{\infty} \cdots \int_{z_d=-\infty}^{\infty} f_Z(z_1, \dots, z_d) e^{tz^\top z} dz_1 \cdots dz_d \\
&= \frac{1}{\sqrt{(2\pi)^d}} \int_{z_1=-\infty}^{\infty} \cdots \int_{z_d=-\infty}^{\infty} e^{(\sum_{i=1}^d z_i^2)(t-\frac{1}{2})} dz_1 \cdots dz_d \\
&= \frac{1}{(2\pi)^{d/2}} \prod_{i=1}^d \left(\int_{z_i=-\infty}^{\infty} e^{z_i^2(t-\frac{1}{2})} dz_i \right) \\
&= \boxed{6}.
\end{aligned}$$

where step (1) is again by the LOTUS. This completes the proof. + \square

5.2 Part b

Proposition 5. Let $Z \sim \mathcal{N}(0, \mathbf{I}_2)$ and let $X := Z^\top Z = Z_1^2 + Z_2^2$, that is, $Z \sim \chi_2^2$. Then, for any $x > 2d$, we have the tail bound

$$\Pr(X > x) \leq \exp \left(d \log \left(\frac{x}{2d} \right) - \frac{x - 2d}{2} \right). \quad (7)$$

Proof. Applying the Cramer-Chernoff technique,

$$\begin{aligned}
\Pr(X > x) &= \Pr(e^{tX} > e^{tx}) \\
&\stackrel{(1)}{\leq} \frac{\mathbb{E}[e^{tX}]}{e^{tx}} := M_2(t)e^{-tx}, \\
&\leq \inf_{t < 1/2} \frac{e^{-tx}}{(1-2t)^d} \\
&\stackrel{(2)}{=} \exp\left(\log\left(\inf_{t < 1/2} \frac{e^{-tx}}{(1-2t)^d}\right)\right) \\
&\stackrel{(3)}{=} \exp\left(\inf_{0 < t < 1/2} \log\left(\frac{e^{-tx}}{(1-2t)^d}\right)\right) \\
&\stackrel{(4)}{=} \exp\left(\inf_{0 < t < 1/2} -tx - d \log(1-2t)\right). \quad +
\end{aligned}$$

where step (1) is Markov's inequality, as X is nonnegative. In step (2), we simply apply \exp and \log . Step (3) is legal because $\log(\cdot)$ is monotonically increasing over the positive reals, so $\log(a) \leq \log(b)$ for any real numbers a, b such that $-\infty < a \leq b < \infty$. In step (4), we used $\log(a/b) = \log(a) - \log(b)$.

To optimize over the final bound in step (4), note that the objective function of the $\inf(\cdot)$ is convex in t . A necessary condition for optimality of a candidate infimizer $0 \leq t_* \leq \frac{1}{2}$ is thus

$$-x + \frac{2d}{1-2t_*} = 0 \implies x - 2xt_* = 2d \implies t_* = \frac{1}{2} - \frac{d}{x}.$$



Plugging in the above boxed expression for t_* into the objection function of the previous inequality chain completes the proof. \square

5.3 Part c

5.3.1 Part c.1

There is a class of random variables that will be of use to us.

Definition 5.1. A random variable $X \in \mathbb{R}$ with mean $\mu = \mathbb{E}[X]$ is *sub-exponential* if there are non-negative parameters (ν, b) such that the inequality $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2 \nu^2 / 2}$ holds for all $|\lambda| \leq \frac{1}{b}$.

Theorem 2 (sub-Exponential Concentration (Wainwright, Proposition 2.9)). *Let $X \in \mathbb{R}$ be a sub-Exponential random variable with parameters (ν, b) . Then, we have the following concentration inequality:*

$$\Pr(X - \mu \geq t) \leq \begin{cases} \exp\left(-\frac{t^2}{2\nu^2}\right) & 0 \leq t \leq \frac{\nu^2}{b}, \\ \exp\left(-\frac{t}{2b}\right) & t > \frac{\nu^2}{b}. \end{cases}$$

Lemma 5 (Sub-exponentiality of Chi-squared distribution (Wainwright, Example 2.8)). *Let $Z \sim \mathcal{N}(0, 1)$ and consider the random variable $X = Z^2 \sim \chi^2(1)$. Then X is sub-exponential with parameters $(2, 4)$.*

Proof. We have

$$\mathbb{E} \left[e^{\lambda(X-1)} \right] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2} \quad \forall |\lambda| \leq 1/4.$$

□

Lemma 6 (Sums of sub-exponentials are sub-exponential). *Let X_1, \dots, X_n be a sequence of sub-exponential random variables with parameters $\{(\nu_k, b_k)\}_{k=1}^n$. Then, the sum $\sum_{k=1}^n (X_k - \mathbb{E} X_k)$ is sub-Exponential with parameters*

$$b_* := \max_{k=1, \dots, n} b_k, \quad \nu_* := \sqrt{\sum_{k=1}^n \nu_k^2}.$$

Corollary 2.1. *Let $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{2d})$, and let $X = Z^\top Z$ as before. Then we have*

$$\Pr(X - 2d \geq t) \leq \begin{cases} \exp\left(-\frac{t^2}{16d}\right) & 0 \leq t \leq 2d \\ \exp\left(-\frac{t}{8}\right) & t > 2d. \end{cases}$$

Proof. By Lemma 5 and Lemma 6 we have that the random variable X is sub-exponential with parameters

$$b_X = \max_{k \in [2d]} b_k = 4, \quad \nu_X := \sqrt{\sum_{k=1}^{2d} \nu_k^2} = \sqrt{\sum_{k=1}^{2d} 2^2} = \sqrt{8d}.$$

Then, by Theorem 2 we have

$$\begin{aligned} \Pr(X - \mathbb{E} X \geq t) &= \Pr(X - 2d \geq t) \\ &= \begin{cases} \exp\left(-\frac{t^2}{2\nu_X^2}\right) & 0 \leq t \leq \frac{\nu_X^2}{b_X} \\ \exp\left(-\frac{t}{2b_X}\right) & t > \frac{\nu_X^2}{b_X} \end{cases} \\ &\leq \begin{cases} \exp\left(-\frac{t^2}{16d}\right) & 0 \leq t \leq 2d \\ \exp\left(-\frac{t}{8}\right) & t > 2d. \end{cases} \end{aligned}$$

Ⓣ.

This completes the proof.

5.3.2 Part c.2

FINISH LATER

Lol, I was not aware that my constants are the same as Woodward's! □

6 Stein's paradox

Let $\mu \in \mathbb{R}^d$ be the mean of the multivariate Gaussian location family

$$\Pr_{\mu} := \mathcal{N}(\mu, \mathbf{I}_d).$$

Let $x \sim \Pr_{\mu}$ be a single observation—this is the *maximum likelihood estimator* (MLE) for μ . Let $\hat{\mu} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be an estimator for μ as a function of the observation $\hat{\mu} := \hat{\mu}(x)$. Define the variance of the estimator as

$$\text{var}_{\mu}(\hat{\mu}) := \mathbb{E} \left[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|_2^2 \right],$$

and the *quadratic risk*

$$\text{risk}_{\mu}(\hat{\mu}) := \mathbb{E} \left[\|\hat{\mu} - \mu\|_2^2 \right].$$

We can see that

$$\text{risk}_{\mu}(x) = \text{var}_{\mu}(x) = d.$$

Definition 6.1 (Dominance of an estimator). An estimator $\hat{\mu}$ is said to be *dominated* by some other estimator $\hat{\mu}'$ if:

1. $\text{risk}(\hat{\mu}') \leq \text{risk}(\hat{\mu})$ for any $\mu \in \mathbb{R}^d$, and
2. There exists a parameter value $\bar{\mu}$ such that $\text{risk}_{\bar{\mu}}(\hat{\mu}') < \text{risk}_{\bar{\mu}}(\hat{\mu})$

Definition 6.2 (Admissibility of an estimator). An estimator $\hat{\mu}$ is called *admissible* if it is not dominated by any other estimator. Otherwise, it is called inadmissible.

6.1 Part a

6.1.1 Part a my solution

Lemma 7. Let $\hat{\mu} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be the parameterized shrinkage estimator for μ , where $\hat{\mu}(s) := s\mathbf{x}$, for any given hyperparameter $s \in \mathbb{R}$. Then, for any $s \notin [0, 1]$ there exists an alternative estimator $\hat{\mu}' \neq \hat{\mu}(s)$ that dominates $\hat{\mu}(s)$. That is, $\hat{\mu}(s)$ is admissible only if $s \in [0, 1]$.

Proof. First, suppose that $s < 0$. Then, for $\psi := -s$, we have that

$$\begin{aligned} \text{risk}_{\mu}(\hat{\mu}(s)) &= \text{risk}_{\mu}(s\mathbf{x}) \\ &:= \mathbb{E} \left[s^2 \|\mathbf{x}\|_2^2 - 2s \langle \mathbf{x}, \mu \rangle + \|\mu\|_2^2 \right] \\ &\geq \mathbb{E} \left[\psi^2 \|\mathbf{x}\|_2^2 - 2\psi \langle \mathbf{x}, \mu \rangle + \|\mu\|_2^2 \right] \\ &:= \text{risk}_{\mu}(\psi\mathbf{x}) = \text{risk}_{\mu}(\hat{\mu}(\psi)), \end{aligned}$$

where the inequality is due to the fact that $-2s \langle \mathbf{x}, \mu \rangle \geq -2\psi \langle \mathbf{x}, \mu \rangle = 2s \langle \mathbf{x}, \mu \rangle$. To conclude, we can set $\bar{\mu} = 1$, and we see that $\hat{\mu}(s)$ is dominated by $\hat{\mu}(\psi)$. Next, suppose that $s > 1$. Choose $\psi := 1$. The claim trivially follows. **[FINISH]**. □

6.1.2 Part A guessing per time

Let $X \sim \mathcal{N}(\mu, I_d)$ and consider the shrinkage estimator

$$\hat{\mu}(s) = sX, \quad s \in \mathbb{R}.$$

Write $X = \mu + Z$ with $Z \sim \mathcal{N}(0, I_d)$. Then

$$\hat{\mu}(s) - \mu = s(\mu + Z) - \mu = (s - 1)\mu + sZ.$$

Hence, the quadratic risk is

$$\begin{aligned} \text{risk}_\mu(s) &= \mathbb{E}[\|\hat{\mu}(s) - \mu\|^2] \\ &= \mathbb{E}[\|(s - 1)\mu + sZ\|^2] \\ &= (s - 1)^2 \|\mu\|^2 + s^2 \mathbb{E}[\|Z\|^2] \\ &= (s - 1)^2 \|\mu\|^2 + s^2 d. \end{aligned}$$

We now show that if $s \notin [0, 1]$ then there exists an estimator dominating $\hat{\mu}(s)$.

Case 1: $s < 0$.

Consider the estimator with $s' = 0$, i.e., $\hat{\mu}(0) = \mathbf{0}$. Its risk is

$$\text{risk}_\mu(0) = \|\mu\|^2.$$

For any $s < 0$, note that $(s - 1)^2 > 1$ and $s^2 > 0$, so

$$\text{risk}_\mu(s) = (s - 1)^2 \|\mu\|^2 + s^2 d > \|\mu\|^2 = \text{risk}_\mu(0),$$

with strict inequality (in particular, when $\mu = \mathbf{0}$ we have $\text{risk}_0(s) = s^2 d > 0 = \text{risk}_0(0)$). Hence, $\hat{\mu}(0)$ dominates $\hat{\mu}(s)$ for $s < 0$.

Case 2: $s > 1$.

Now, consider the estimator with $s' = 1$, namely $\hat{\mu}(1) = X$. Its risk is

$$\text{risk}_\mu(1) = (1 - 1)^2 \|\mu\|^2 + 1^2 d = d.$$

For any $s > 1$, we have $s^2 > 1$, so

$$\text{risk}_\mu(s) = (s - 1)^2 \|\mu\|^2 + s^2 d \geq s^2 d > d = \text{risk}_\mu(1),$$

again with strict inequality (e.g. when $\mu = \mathbf{0}$, $\text{risk}_0(s) = s^2 d > d$). Thus, $\hat{\mu}(1)$ dominates $\hat{\mu}(s)$ for $s > 1$.

Since any estimator $\hat{\mu}(s)$ with $s < 0$ or $s > 1$ is dominated by either $\hat{\mu}(0)$ or $\hat{\mu}(1)$, it suffices to restrict attention to shrinkage estimators with $s \in [0, 1]$.

6.2 Part b

Lemma 8. Given $\mu \in \mathbb{R}^d$, the s_\star that minimizes the risk is given as

$$s_\star := \frac{\|\mu\|_2^2}{d + \|\mu\|_2^2} = 1 - \frac{d}{d + \|\mu\|_2^2}$$

Proof. Write $X = \mu + Z$ with $Z \sim \mathcal{N}(0, I_d)$. It is well known that

$$\mathbb{E} \|X - \mu\|_2^2 = \mathbb{E} \|Z\|_2^2 = d.$$

Moreover, we can write

$$\hat{\mu}(s) - \mu = s(\mu + Z) - \mu = (s - 1)\mu + sZ.$$

Hence, the quadratic risk is

$$\begin{aligned} \text{risk}_\mu(s) &= \mathbb{E} \left[\|\hat{\mu}(s) - \mu\|^2 \right] \\ &= \mathbb{E} \left[\|(s - 1)\mu + sZ\|^2 \right] \\ &= (s - 1)^2 \|\mu\|^2 + s^2 \mathbb{E} \left[\|Z\|^2 \right] \\ &= (s - 1)^2 \|\mu\|^2 + s^2 d. \end{aligned}$$

To find the s_* that minimizes the above, a necessary condition is

$$0 = \frac{\partial}{\partial s} \text{risk}_\mu(s_*) = 2s_*d + 2\|\mu\|_2^2(s_* - 1) \implies s_* = \frac{\|\mu\|_2^2}{d + \|\mu\|_2^2}.$$



This completes the proof. □

6.3 Part c

Remark. This value of s_* that minimizes risk is absurd, and X_{s_*} is an absurd estimator because s_* depends on the true parameter μ , which we wish to estimate.

Note that

$$\mathbb{E} \|X\|_2^2 = \mathbb{E} \|Z + \mu\|_2^2 = \mathbb{E} \left[\|Z\|_2^2 + 2\langle Z, \mu \rangle + \|\mu\|_2^2 \right] = d + \|\mu\|_2^2.$$



Hence, ***the proposed heuristic is a plugin estimator.***

6.4 Part d

Lemma 9 (Stein's Lemma). *Let $X \sim \mathcal{N}(\mu, I_d)$ and let $g : \mathbb{R}^d \rightarrow \mathbb{R}^n$ be differentiable almost everywhere, with*

$$\mathbb{E} \left[\left| \frac{\partial}{\partial x_i} g(X) \right| \right] < \infty \quad \text{and} \quad \mathbb{E} [| (X_i - \mu_i) g(X) |] < \infty$$

for any $i = 1, \dots, d$. Then

$$\mathbb{E} [(X_i - \mu_i) g(X)] = \mathbb{E} \left[\frac{\partial}{\partial x_i} g(X) \right], \quad i = 1, \dots, d.$$

Theorem 3. *Given a $\mu \in \mathbb{R}^d$ and $\delta \in \mathbb{R}$, define the parameterized estimator*

$$\hat{\mu}_\delta := \left(1 - \frac{\delta}{\|X\|_2^2} \right) X,$$

and define the minimum-risk parameter δ_* as

$$\delta_* := \arg \min_{\delta \in \mathbb{R}} \text{risk}_\mu(\hat{\mu}_\delta).$$

This estimator, known as the James-Stein estimator, $\hat{\mu}^{\text{JS}}$, can be written as

$$\hat{\mu}^{\text{JS}} := \hat{\mu}_{\delta_*} = \left(1 - \frac{d-2}{\|X\|_2^2}\right) X.$$

Proof. First, note that

$$\hat{\mu}_\delta - \mu = \left(1 - \frac{\delta}{\|X\|_2^2}\right) X - \mu = X - \mu - \frac{\delta X}{\|X\|_2^2}$$

We have that the minimum value of the risk is given as

$$\begin{aligned} \min_{\delta \in \mathbb{R}} \text{risk}_\mu(\hat{\mu}_\delta) &:= \min_{\delta \in \mathbb{R}} \mathbb{E} \left[\|\hat{\mu}_\delta - \mu\|_2^2 \right] \\ &= \min_{\delta \in \mathbb{R}} \mathbb{E} \left[\left\| X - \mu - \frac{\delta X}{\|X\|_2^2} \right\|_2^2 \right] \\ &= \min_{\delta \in \mathbb{R}} \mathbb{E} \left[\left(X - \mu - \frac{\delta X}{\|X\|_2^2} \right)^\top \left(X - \mu - \frac{\delta X}{\|X\|_2^2} \right) \right] \\ &= \min_{\delta \in \mathbb{R}} \mathbb{E} \left[\|X - \mu\|_2^2 - \frac{2\delta}{\|X\|_2^2} \langle X - \mu, X \rangle + \frac{\delta^2}{\|X\|_2^2} \right] \\ &= d + \min_{\delta \in \mathbb{R}} \mathbb{E} \left[\frac{\delta^2}{\|X\|_2^2} - \frac{2\delta \langle X - \mu, X \rangle}{\|X\|_2^2} \right]. \end{aligned}$$

In the final equality, we applied the fact that $\mathbb{E} \|X - \mu\|_2^2 = d$ for any $X \sim \mathcal{N}(\mu, I_d)$. To minimize the above expression, we must differentiate the objective function. Note that the risk of a parameter δ is

$$R(\delta) := \text{risk}_\mu(\hat{\mu}_\delta) = \mathbb{E} \left[\|\hat{\mu}_\delta - \mu\|_2^2 \right] = d + \delta^2 \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right] - 2\delta \mathbb{E} \left[\frac{\langle X - \mu, X \rangle}{\|X\|_2^2} \right].$$

Then, define $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ as

$$g(X) = \frac{X}{\|X\|_2^2}.$$

We see that

$$\frac{\partial}{\partial x_i} g_i(X) = \frac{1}{\|X\|_2^2} - \frac{2X_i^2}{\|X\|_2^4};$$

so, by Stein's Lemma, **[CHECK INTEGRAL CONDITIONS????]**

not to worry -

$$\mathbb{E} [(X_i - \mu_i) g_i(X)] = \mathbb{E} \left[\frac{1}{\|X\|_2^2} - \frac{2X_i^2}{\|X\|_2^4} \right], \quad i = 1, \dots, d.$$

(+)

Then, we see that the parameter risk can be given as

$$\begin{aligned}
R(\delta) &= d + \delta^2 \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right] - 2\delta \sum_{i=1}^d \mathbb{E} \left[(X_i - \mu_i) \frac{\mu_i}{\|X\|_2^2} \right] \\
&= d + \delta^2 \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right] - 2\delta \sum_{i=1}^d \mathbb{E} \left[\frac{1}{\|X\|_2^2} - \frac{2X_i^2}{\|X\|_2^4} \right] \\
&= d + \delta^2 \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right] - 2\delta \mathbb{E} \left[\frac{d}{\|X\|_2^2} - 2 \frac{\|X\|_2^2}{\|X\|_2^4} \right] \\
&= d + \delta^2 \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right] - 2\delta \left(\frac{d}{\|X\|_2^2} - 2 \frac{d}{\|X\|_2^4} \right) \\
&= d + \delta^2 \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right] - 2\delta (d - 2) \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right].
\end{aligned}$$

This is a quadratic in δ . To find the minimum, we differentiate with respect to δ and find the stationary point:

$$\frac{d}{d\delta} R(\delta) = -2(d - 2) \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right] + 2\delta \mathbb{E} \left[\frac{1}{\|X\|_2^2} \right] = 0 \implies \delta_\star = d - 2.$$

This completes the argument.



7 Planar Venn diagrams

Definition 7.1 (Congruent Venn Diagram). Given a *base* subspace $\mathcal{A} \subset \mathbb{R}^d$ and a set of n locations $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^d$, define the shifted sets:

$$\mathcal{A}_j := \{\mathbf{a} + \mathbf{a}_j, \quad \mathbf{a} \in \mathcal{A}\}, \quad \forall j = 1, \dots, n.$$

Then, a *congruent Venn diagram* in \mathbb{R}^d for n sets is any such instance of the aforementioned object such that for any subset of indices $\mathcal{I} \subseteq \{1, \dots, n\}$, we have that

$$\mathcal{A}_{\mathcal{I}} := \bigcap_{j \in \mathcal{I}} \mathcal{A}_j \neq \emptyset.$$

Theorem 4. One cannot draw a planar ($d = 2$) Venn diagram for $n \geq 5$ sets by shifting a circle.

Proof. Fix an $n \in \mathbb{N}$ and consider the diagram at the $(n-1)$ -th set (circle), defined by $\mathcal{V}_{n-1}, \mathcal{E}_{n-1}, \mathcal{F}_{n-1}$.

By Def. 7.1 we must have a non-empty intersection of all index subsets $\mathcal{I} \subseteq \{1, 2, \dots, n\}$; hence,

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

each intersection splits the new circle into at most $2(n-1)$ edges.

Since $F_n = 2^n$,

$$F_n = 2^n \leq n^2 - n + 2$$

+
Second condition?

(+) / -

□

References