CS 189: Homework 2

Vinay Maruri

February 11, 2019

I did not work with anybody on this homework.
"I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."
Signature: Vinay Maruri

1 Identities with Expectation

1. Let X be a random variable with pdf $f(x) = \lambda e^{-\lambda x}$ for x > 0 (and zero everywhere else.) Use induction on k to show that $E[X^k] = \frac{k!}{\lambda^k}$.

Proof. Assumption: k $\geq$ 0
Base Case: Let k = 0.
E$[X^0] = \int_0^\infty X^0(\lambda e^{-\lambda x})dx$
E$[X^0] = \int_0^\infty \lambda e^{-\lambda x}dx$
E$[X^0] = \lambda \int_0^\infty e^{-\lambda x}dx$
E$[X^0] = \lambda\ [\frac{-1}{\lambda} * e^{-\lambda x}]_0^\infty$
E$[X^0] = [e^{-\lambda x}]_0^\infty$
E$[X^0] = 1$
or equivalently:
E$[X^0] = \frac{0!}{\lambda^0} = 1$

Induction Hypothesis:
For $0 \leq k \leq n$, $E[X^k] = \frac{k!}{\lambda^k}$.

Induction Step: Let k = n+1
$E[X^{n+1}] = \int_0^\infty X^{n+1}(\lambda e^{-\lambda x})dx$
We proceed by integration by parts.
u = X^{n+1}, dv = $\lambda e^{-\lambda x}$
du = $(n+1)X^n$, v = $-e^{-\lambda x}$
$E[X^{n+1}] = (X^{n+1} - e^{-\lambda x})_0^\infty$ - $\int_0^\infty (n+1)X^n(-e^{-\lambda x})dx$
$E[X^{n+1}] = 0$ + (n+1)$\int_0^\infty X^n(e^{-\lambda x})dx$
We notice the integral term is equivalent to $\frac{1}{\lambda}$ $E[X^n]$

Applying the induction hypothesis:
$E[X^{n+1}] = \frac{n+1}{\lambda} * \frac{n!}{\lambda^n}$
$E[X^{n+1}] = \frac{(n+1)!}{\lambda^{n+1}}$
Thus, the statement is proven by induction. $\square$

2. Assume that X is a non-negative real-valued random variable. Prove the following identity:
E[X] = $\int_0^\infty P(X \geq t)dt$

Proof. I assume first that X has density $f(x)$ and a CDF $F(x)$.
Since X is continuous, we use the continuous version of the expectation of a random variable:
E[X] = $\int_0^\infty Xf(x)dx$
E[X] =

$$\int_0^\infty \int_0^x f(x)dtdx$$

By Fubini's Theorem,
E[X] =

$$\int_0^\infty \int_t^\infty f(x)dxdt$$

Thus, E[X] = $\int_0^\infty P(X \geq t)dt$

$\square$

3. Again assume X $\geq$ 0, but now additionally let $E[X^2] < \infty$. Prove the following:
$P(X > 0) \geq \frac{(E[X])^2}{E[X^2]}$.

Proof. By assumption, P(X $\geq$ 0) = 1.
We can subdivide this probability into 2 groups:
P(X $\geq$ 0) = P(X = 0) + P(X > 0)
For probability event A, $P(A) = E[1\{A\}]$, where $1\{.\}$ is the indicator function.
$E[X] = E[X * 1\{X = 0\}] + E[X * 1\{X > 0\}]$
However, notice that when $1\{X = 0\}$ takes on a value of 1, X is 0. Thus:
$E[X] = E[X * 1\{X > 0\}]$
Now, I invoke the Cauchy-Schwarz inequality.
$E[X * 1\{X > 0\}] \leq \sqrt{< X, X >< 1\{X > 0\}, 1\{X > 0\} >}$

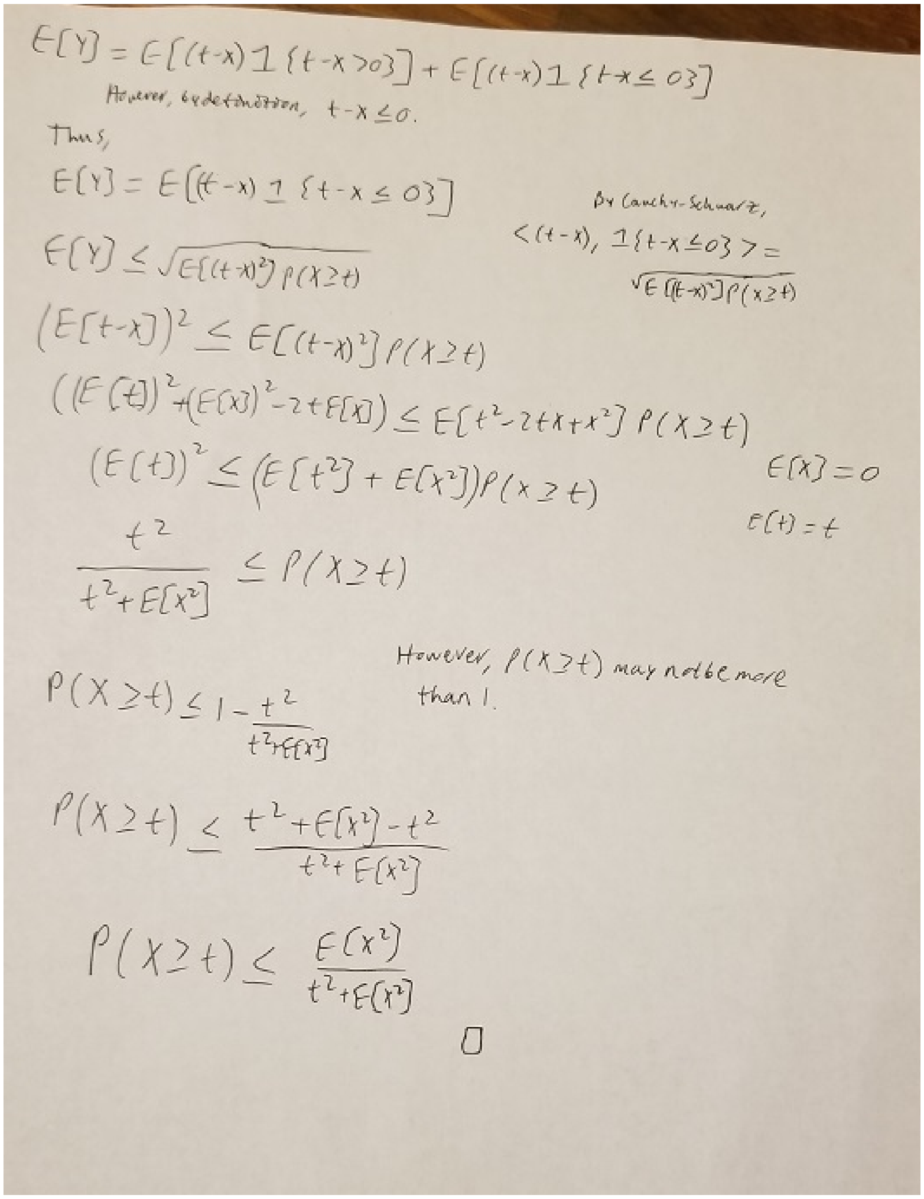
$E[X * 1\{X > 0\}] \leq \sqrt{E[X^2]P(X > 0)}$
Thus,
$E[X] \leq \sqrt{E[X^2]P(X > 0)}$
$(E[X])^2 \leq E[X^2]\ P(X > 0)$
$\frac{(E[X])^2}{E[X^2]} \leq P(X > 0)$

$\square$

4. Now assume $E[X^2] \leq \infty$, and additionally assume that $E[X] = 0$ (X no longer has to be non-negative). Prove the following inequality:
$P(X \geq t) \leq \frac{E[X^2]}{E[X^2]+t^2}$, for any t $\geq$ 0

Proof. Let Y be a random variable such that:
Y = t - X
The rest of the proof is included on the next page.

$\square$



$$E[Y] = E[(t-x)\mathbb{1}\{t-x>0\}] + E[(t-x)\mathbb{1}\{t-x\le 0\}]$$

However, by definition, $t-x \le 0$.

Thus,

$$E[Y] = E[(t-x)\,\mathbb{1}\{t-x\le 0\}]$$

By Cauchy-Schwarz,

$$\langle (t-x), \mathbb{1}\{t-x\le 0\}\rangle = \sqrt{E[(t-x)^2]P(x\ge t)}$$

$$E[Y] \le \sqrt{E[(t-x)^2]P(X\ge t)}$$

$$(E[t-x])^2 \le E[(t-x)^2]P(X\ge t)$$

$$((E[t])^2 + (E[X])^2 - 2tE[X]) \le E[t^2 - 2tX + X^2]P(X\ge t)$$

$$(E[t])^2 \le (E[t^2] + E[X^2])P(X\ge t)$$

$E[X] = 0$

$E(t) = t$

$$\frac{t^2}{t^2 + E[X^2]} \le P(X\ge t)$$

However, $P(X\ge t)$ may not be more than 1.

$$P(X\ge t) \le 1 - \frac{t^2}{t^2+E[X^2]}$$

$$P(X\ge t) \le \frac{t^2 + E[X^2] - t^2}{t^2 + E[X^2]}$$

$$P(X\ge t) \le \frac{E(X^2)}{t^2+E(X^2)}$$

$\square$

Figure 1: Completion of Proof for question 1.4

2 Properties of Gaussians

1. Prove that $E[e^{\lambda x}] = e^{\sigma^2 \lambda^2 / 2}$ where $\lambda \in R$ is a fixed constant and X $\sim N(0, \sigma^2)$.
Solution is on the next page.

Let X have pdf $f_X(X)$ as follows:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$$

Let $M_X(\lambda)$ be a moment generating function for X.

$$M_X(\lambda) = \int_{-\infty}^{\infty} e^{\lambda x} f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} e^{\lambda x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx$$

$$= \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(-2\lambda x + \left(\frac{x}{\sigma}\right)^2\right)} \frac{1}{\sigma\sqrt{2\pi}}\, dx$$

$$= e^{\frac{1}{2}\sigma^2\lambda^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(-2\lambda x + \left(\frac{x}{\sigma}\right)^2 + \sigma^2\lambda^2\right)} \frac{1}{\sigma\sqrt{2\pi}}\, dx$$

(we complete the square.)

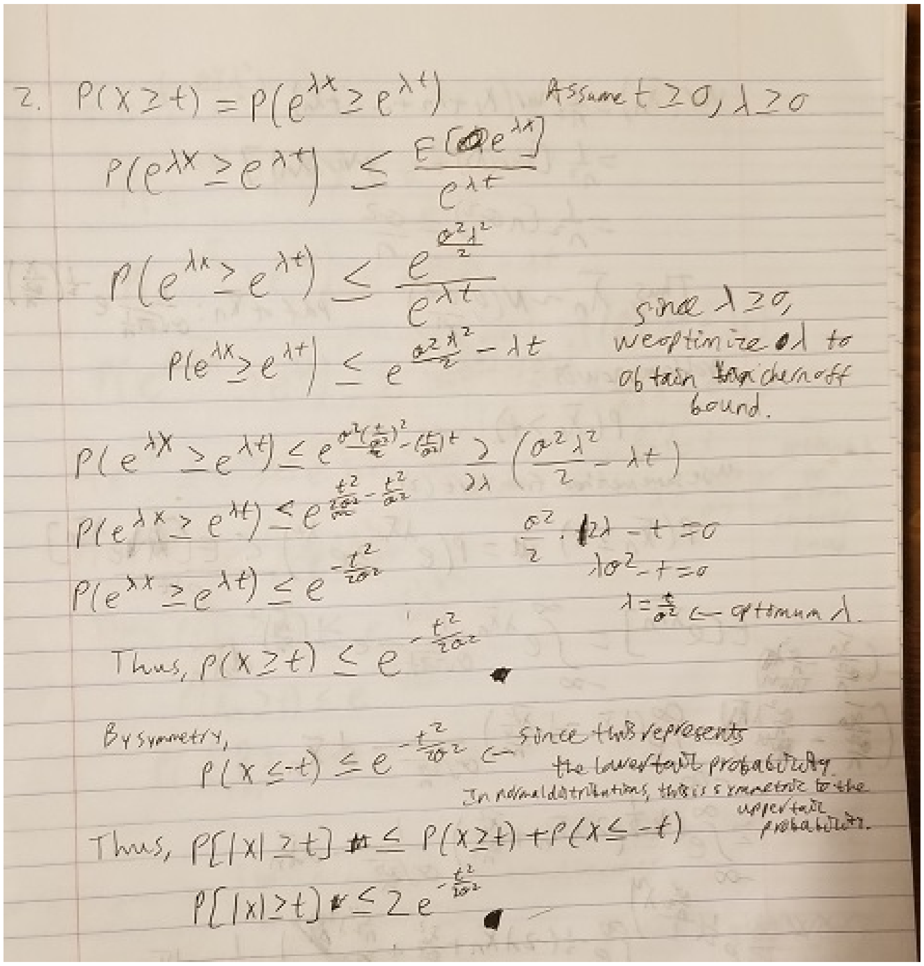
$$= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{1}{\sigma}x - \sigma\lambda\right)^2} \frac{1}{\sigma\sqrt{2\pi}}\, dx$$

(however, this is the pdf of $N(\sigma^2\lambda, \sigma^2)$)

$$= e^{\frac{\sigma^2\lambda^2}{2}} [1] = e^{\frac{\sigma^2\lambda^2}{2}}$$

$$E[e^{\lambda x}] = M_X(\lambda) = e^{\frac{\sigma^2\lambda^2}{2}} \quad \square$$

Figure 2: Proof for question 2.1



2. $P(X \geq t) = P(e^{\lambda X} \geq e^{\lambda t})$ Assume $t \geq 0, \lambda \geq 0$

$P(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{E[e^{\lambda X}]}{e^{\lambda t}}$

$P(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{e^{\frac{\sigma^2\lambda^2}{2}}}{e^{\lambda t}}$

$P(e^{\lambda X} \geq e^{\lambda t}) \leq e^{\frac{\sigma^2\lambda^2}{2} - \lambda t}$

since $\lambda \geq 0$, we optimize λ to obtain chernoff bound.

$\frac{\partial}{\partial \lambda}\left(\frac{\sigma^2\lambda^2}{2} - \lambda t\right)$

$\frac{\sigma^2}{2}\cdot 2\lambda - t = 0$

$\lambda\sigma^2 - t = 0$

$\lambda = \frac{t}{\sigma^2}$ ← optimum λ.

$P(e^{\lambda X} \geq e^{\lambda t}) \leq e^{\frac{\sigma^2 (\frac{t}{\sigma^2})^2}{2} - (\frac{t}{\sigma^2})t}$

$P(e^{\lambda X} \geq e^{\lambda t}) \leq e^{\frac{t^2}{2\sigma^2} - \frac{t^2}{\sigma^2}}$

$P(e^{\lambda X} \geq e^{\lambda t}) \leq e^{-\frac{t^2}{2\sigma^2}}$

Thus, $P(X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$

By symmetry,

$P(X \leq -t) \leq e^{-\frac{t^2}{2\sigma^2}}$ ← Since this represents the lower tail probability. In normal distributions, this is symmetric to the upper tail probability.

Thus, $P[|X| \geq t] \leq P(X \geq t) + P(X \leq -t)$

$P[|X| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$

Figure 3: Proof for question 2.2

2. Prove that $P(X \geq t) \leq exp(-t^2/2\sigma^2)$, and conclude that $P(|x| \geq t) \leq 2$ $exp(-t^2/2\sigma^2)$

Solution is above.

3. Let $X_1, \cdots, X_n \sim N(0, \sigma^2)$ be iid. Can you prove a similar concentration result for the average of n Gaussians: $P(\frac{1}{n}\sum_{i=1}^n X_i \geq t)$?
Solution is on next 2 pages.

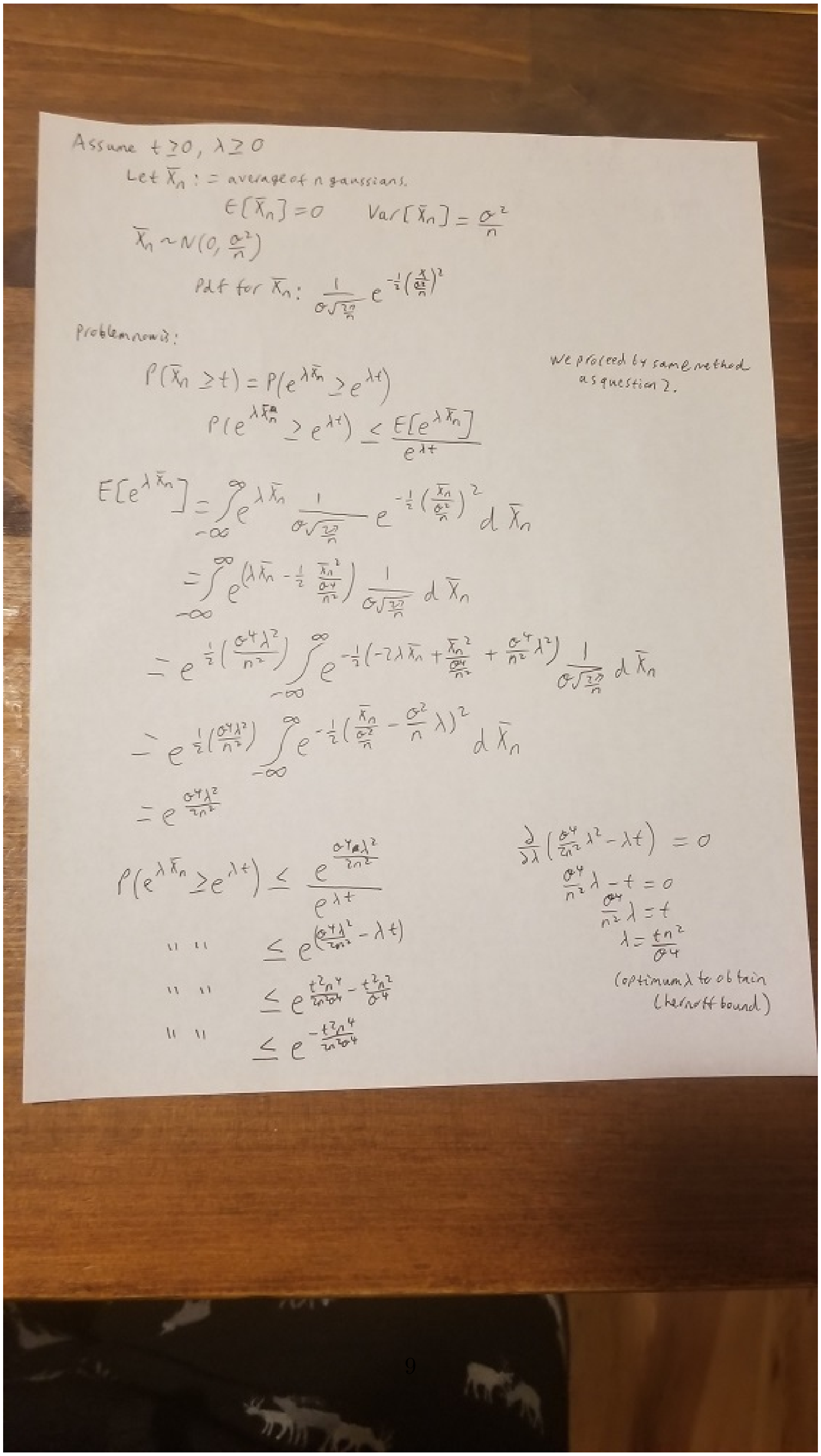


Figure 4: Proof for question 2.3

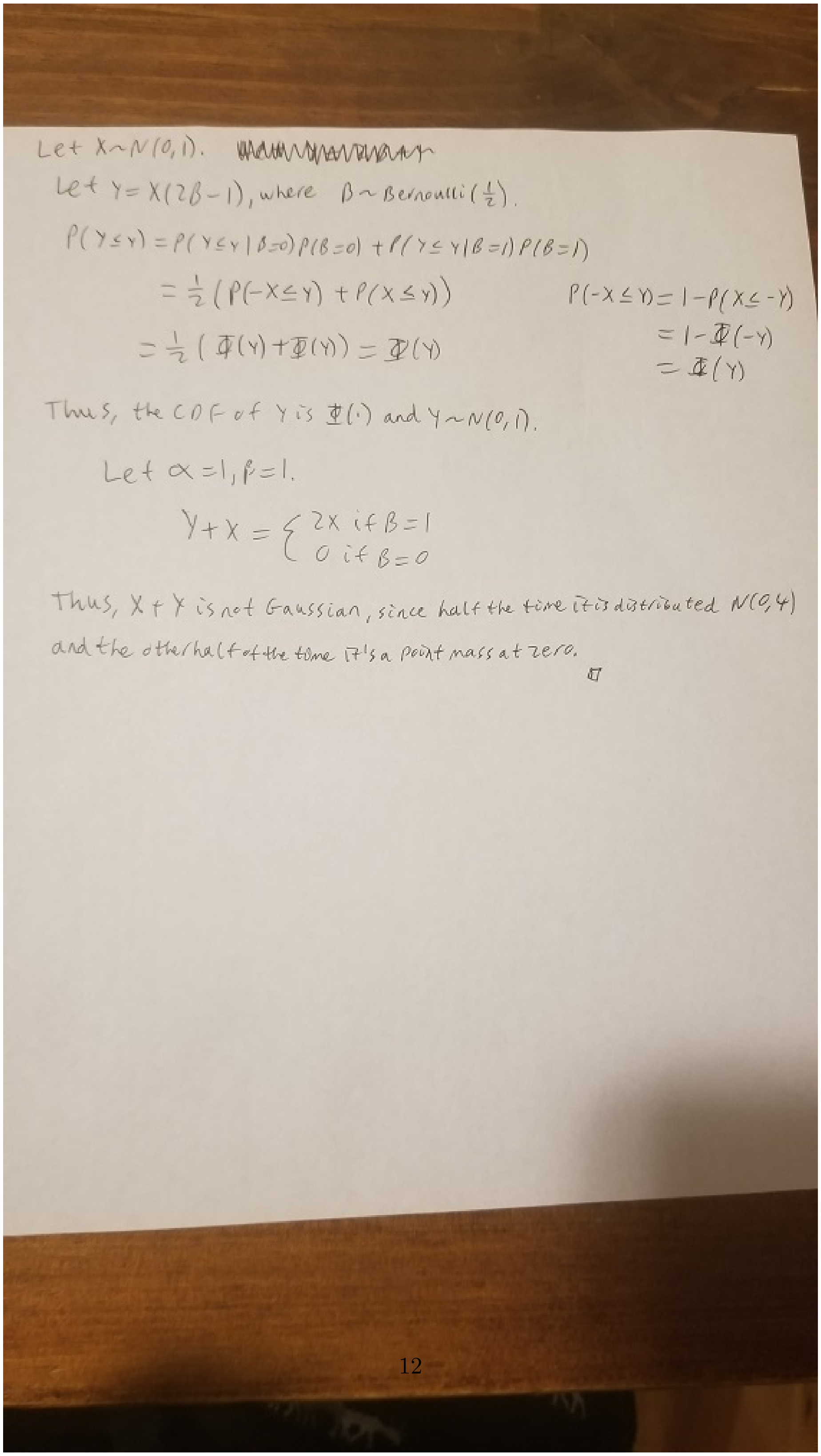
Thus, $P(\bar{X}_n \geq t) \leq e^{-\frac{t^2 n^2}{2a^4}}$

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq e^{-\frac{t^2 n^2}{2a^4}}$$

$\square$

Figure 5: Proof continued for question 2.3

4. Give an example of two Gaussian random variables X and Y, such that there exists no linear combination $\alpha X + \beta Y$, for some $\alpha, \beta \in R$, which is not Gaussian.
Solution is on the next page.



Let $X \sim N(0,1)$.

Let $Y = X(2B-1)$, where $B \sim \text{Bernoulli}(\frac{1}{2})$.

$$P(Y \le y) = P(Y \le y \mid B=0)P(B=0) + P(Y \le y \mid B=1)P(B=1)$$
$$= \tfrac{1}{2}\left(P(-X \le y) + P(X \le y)\right)$$
$$= \tfrac{1}{2}\left(\Phi(y) + \Phi(y)\right) = \Phi(y)$$

$$P(-X \le y) = 1 - P(X \le -y)$$
$$= 1 - \Phi(-y)$$
$$= \Phi(y)$$

Thus, the CDF of Y is $\Phi(\cdot)$ and $Y \sim N(0,1)$.

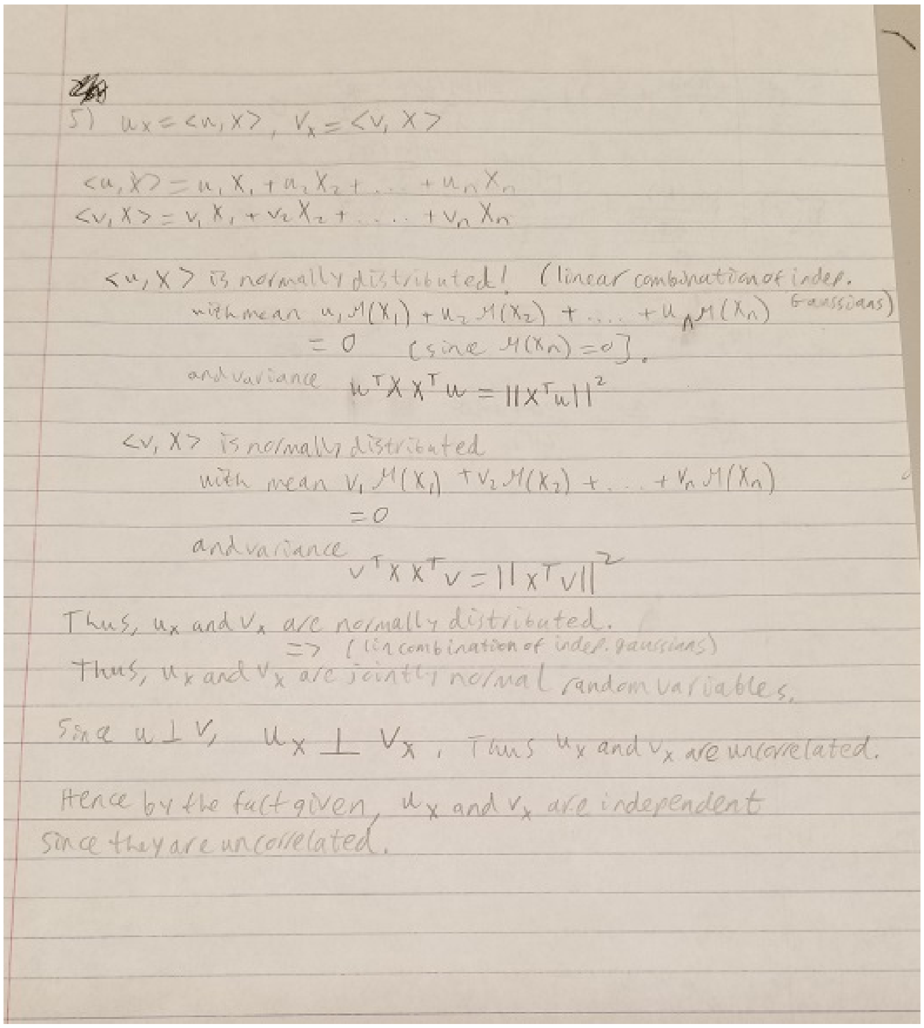
Let $\alpha = 1, \beta = 1$.

$$Y + X = \begin{cases} 2X & \text{if } B = 1 \\ 0 & \text{if } B = 0 \end{cases}$$

Thus, $X + Y$ is not Gaussian, since half the time it is distributed $N(0,4)$ and the other half of the time it's a point mass at zero. $\square$

Figure 6: Proof for question 2.4

5. Take two orthogonal vectors $u, v \in R^n$, u $\perp$ v, and let $X = (X_1, \cdots, X_n)$ be a vector of n iid standard Gaussians, $X_i \sim N(0, 1)$, $\forall i \in$ [n]. Let u$_x$ =$\langle u, X\rangle$ and v$_x$ =$\langle v, X\rangle$. Are u_x and v_x independent?
Solution is on the next page.



5) $u_x = \langle u, X \rangle$, $v_x = \langle v, X \rangle$

$\langle u, X \rangle = u_1 X_1 + u_2 X_2 + \dots + u_n X_n$

$\langle v, X \rangle = v_1 X_1 + v_2 X_2 + \dots + v_n X_n$

$\langle u, X \rangle$ is normally distributed! (linear combination of indep. Gaussians)

with mean $u_1 \mu(X_1) + u_2 \mu(X_2) + \dots + u_n \mu(X_n)$

$= 0$ [since $\mu(X_n) = 0$].

and variance $u^T X X^T u = \|X^T u\|^2$

$\langle v, X \rangle$ is normally distributed

with mean $v_1 \mu(X_1) + v_2 \mu(X_2) + \dots + v_n \mu(X_n)$

$= 0$

and variance $v^T X X^T v = \|X^T v\|^2$

Thus, u_x and v_x are normally distributed.

$\Rightarrow$ (lin combination of indep. gaussians)

Thus, u_x and v_x are jointly normal random variables.

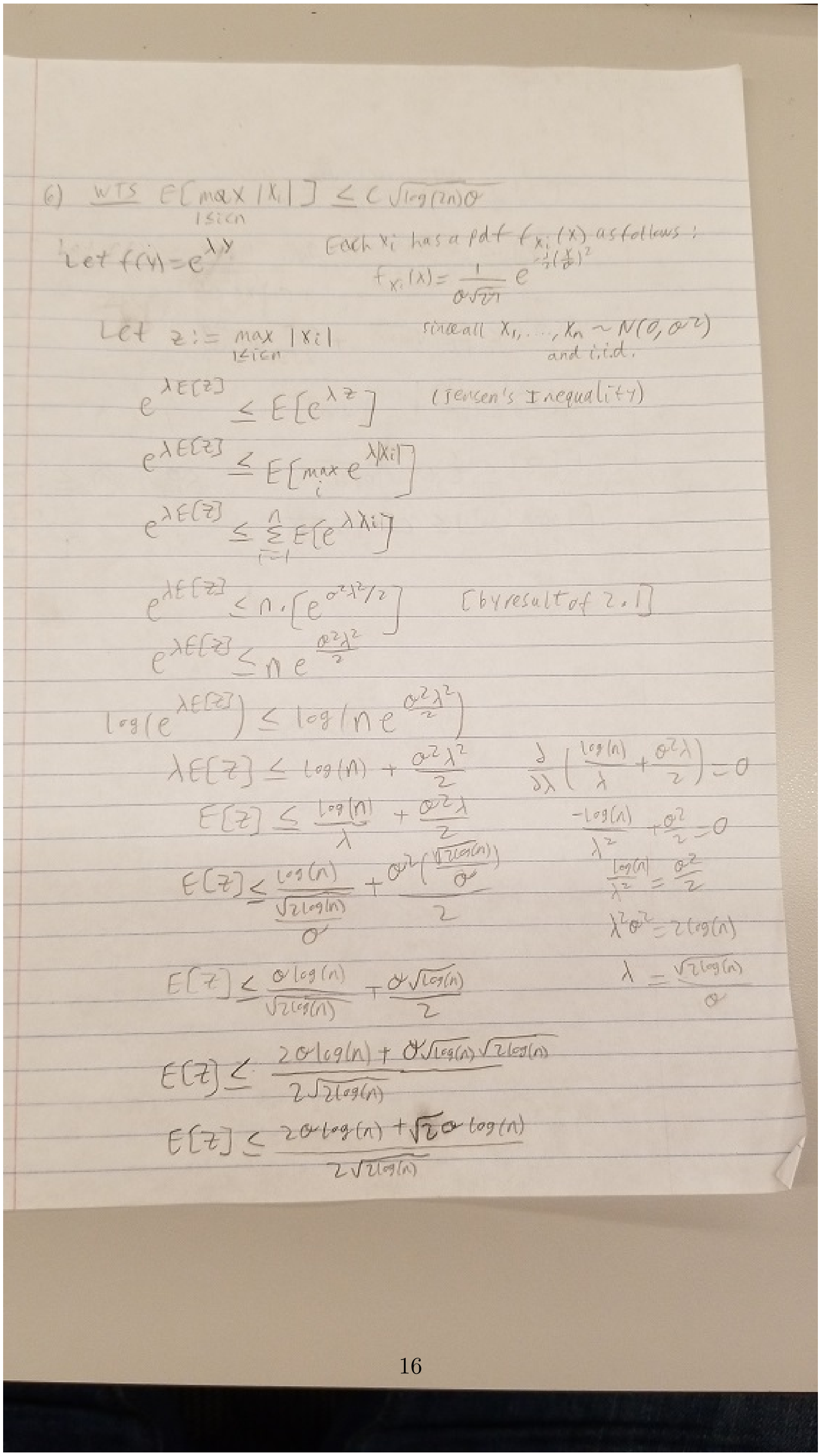
Since $u \perp v$, $u_x \perp v_x$. Thus u_x and v_x are uncorrelated.

Hence by the fact given, u_x and v_x are independent since they are uncorrelated.

Figure 7: Proof for question 2.5

6. Prove that $E[max_{1\leq i\leq n}|X_i|] \leq \mathrm{C}\sqrt{log(2n)}\sigma$, where $X_1, \cdots, X_n \sim N(0, \sigma^2)$ are iid.

Solution is on the next 2 pages.



6) WTS $E[\max_{1\le i\le n}|X_i|] \le C\sqrt{\log(2n)}\sigma$

Let $f(y)=e^{\lambda y}$

Each X_i has a pdf $f_{X_i}(x)$ as follows:

$f_{X_i}(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x}{\sigma})^2}$

Let $z := \max_{1\le i\le n}|X_i|$

since all $X_1,\dots,X_n \sim N(0,\sigma^2)$ and i.i.d.

$e^{\lambda E[z]} \le E[e^{\lambda z}]$ (Jensen's Inequality)

$e^{\lambda E[z]} \le E[\max_i e^{\lambda|X_i|}]$

$e^{\lambda E[z]} \le \sum_{i=1}^{n} E[e^{\lambda X_i}]$

$e^{\lambda E[z]} \le n\cdot[e^{\sigma^2\lambda^2/2}]$ [by result of 2.1]

$e^{\lambda E[z]} \le n e^{\frac{\sigma^2\lambda^2}{2}}$

$\log(e^{\lambda E[z]}) \le \log(n e^{\frac{\sigma^2\lambda^2}{2}})$

$\lambda E[z] \le \log(n) + \frac{\sigma^2\lambda^2}{2}$

$\frac{\partial}{\partial\lambda}\left(\frac{\log(n)}{\lambda}+\frac{\sigma^2\lambda}{2}\right)=0$

$E[z] \le \frac{\log(n)}{\lambda}+\frac{\sigma^2\lambda}{2}$

$\frac{-\log(n)}{\lambda^2}+\frac{\sigma^2}{2}=0$

$\frac{\log(n)}{\lambda^2}=\frac{\sigma^2}{2}$

$\lambda^2\sigma^2=2\log(n)$

$\lambda=\frac{\sqrt{2\log(n)}}{\sigma}$

$E[z] \le \frac{\log(n)}{\frac{\sqrt{2\log(n)}}{\sigma}} + \frac{\sigma^2\left(\frac{\sqrt{2\log(n)}}{\sigma}\right)}{2}$

$E[z] \le \frac{\sigma\log(n)}{\sqrt{2\log(n)}} + \frac{\sigma\sqrt{\log(n)}}{2}$

$E[z] \le \frac{2\sigma\log(n)+\sigma\sqrt{\log(n)}\sqrt{2\log(n)}}{2\sqrt{2\log(n)}}$

$E[z] \le \frac{2\sigma\log(n)+\sqrt{2}\sigma\log(n)}{2\sqrt{2\log(n)}}$

Figure 8: Proof for question 2.6

$$E[Z] \leq \frac{\sigma(2\log(n) + \sqrt{2}\log(n))}{2\sqrt{2\log(n)}}$$

$$E[Z] \leq \frac{\sigma(2\log(n))}{2\sqrt{2\log(n)}} + \frac{\sigma\sqrt{2}\log(n)}{2\sqrt{2\log(n)}}$$

$$E[Z] \leq \frac{2\sigma}{2\sqrt{2}}\sqrt{\log(n)} + \frac{\sigma\sqrt{2}}{2\sqrt{2}}\sqrt{\log(n)}$$

$$E[Z] \leq \left(\frac{\sigma}{\sqrt{2}} + \frac{\sigma}{2}\right)\sqrt{\log(n)}$$

$$E[Z] \leq c\,\sigma\sqrt{\log(n)}$$

$c\sigma\sqrt{\log(n)} \leq c\sigma\sqrt{\log(2n)}$ for $n \geq 1$.

$$E\left[\max_{1\leq i\leq n} |X_i|\right] \leq c\,\sigma\sqrt{\log(2n)}$$

Figure 9: Proof for question 2.6 continued

3 Linear Algebra Review

Question 1:
Solution is on the next page.

1\. 3. Let $A \in \mathbb{R}^{n \times n}$ be symmetric. Prove equivalence.

1) For all $x \in \mathbb{R}^n$, $x^T A x \geq 0$
2) All eigenvalues of A are non-negative.
3) There exists matrix $U \in R^{n \times n}$, such that $A = UU^T$

(1) => (2)

Let λ be an eigenvalue of A.

Then, there exist eigenvector $v \in V$ such that $Av = \lambda v$.

So, $v^T A v \geq 0$ and $\lambda v^T v \geq 0$

since $v^T v$ is positive for all v, $\lambda \geq 0$.

(2) => (3)

Since A is symmetric, it has a spectral decomposition.

$$A = \sum_i \lambda_i x_i x_i^T \qquad y_i = \sqrt{\lambda_i}\, x_i \quad (\text{since } \lambda_i \geq 0)$$

$$A = \sum_i y_i y_i^T$$

Let B be a matrix whose columns are y_i.

Then $A = BB^T$. Thus, there exist matrix $U \in R^{n \times n}$ such that $A = UU^T$.

(3) => (1)

$\exists$ matrix $U \in R^{n \times n}$ such that $A = UU^T$.

$$x^T (UU^T) x = x^T U x U^T = (xU)^T (xU) = \|Ux\|_2$$

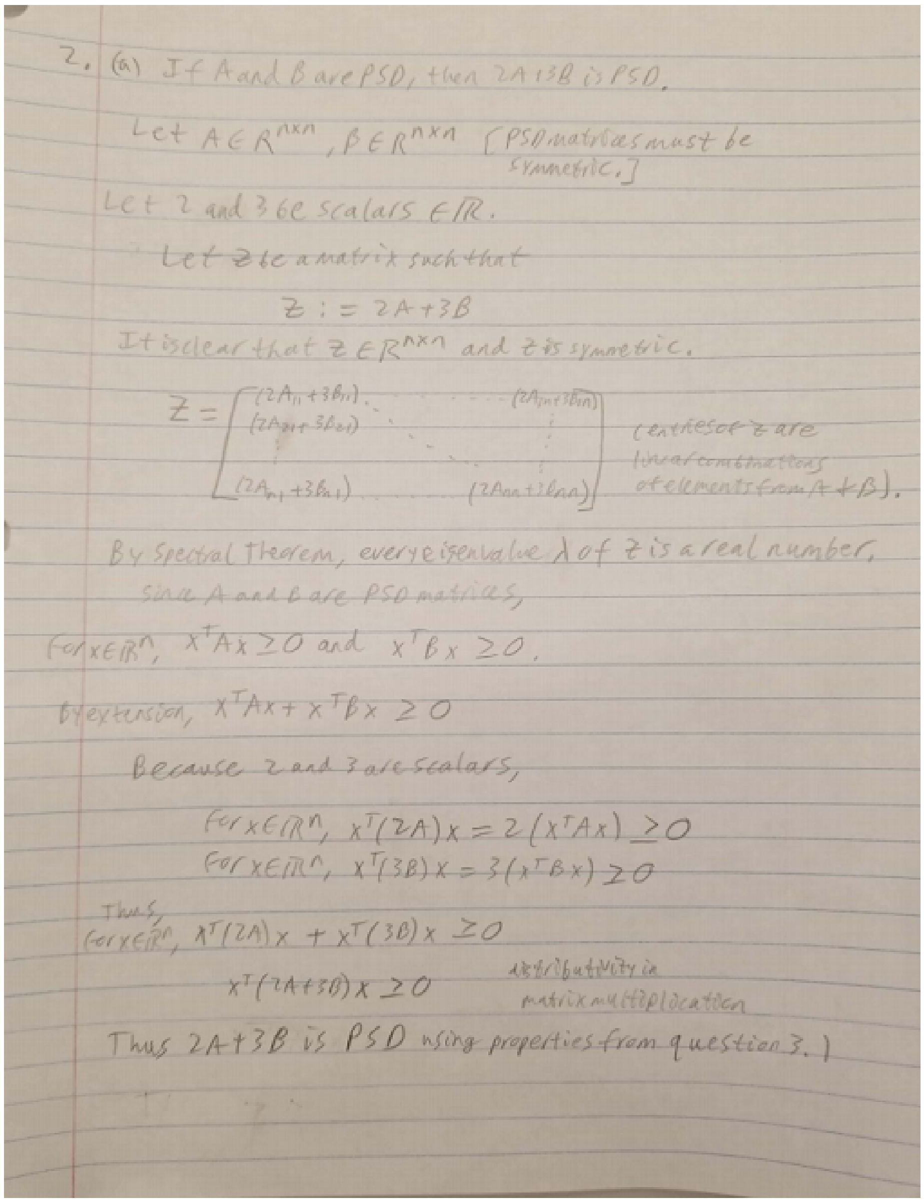
← but this is the Euclidean norm!

by definition $\|Ux\|_2 \geq 0$,

thus, $x^T (UU^T) x = x^T A x \geq 0$. □

Figure 10: Proof for question 3.1

Question 2(a):
Solution is on the next page.



2. (a) If A and B are PSD, then $2A+3B$ is PSD.

Let $A \in R^{n\times n}$, $B \in R^{n\times n}$ [PSD matrices must be symmetric.]

Let 2 and 3 be scalars $\in \mathbb{R}$.

~~Let Z be a matrix such that~~

$$Z := 2A+3B$$

It is clear that $Z \in R^{n\times n}$ and Z is symmetric.

$$Z = \begin{bmatrix} (2A_{11}+3B_{11}) & \cdots & (2A_{1n}+3B_{1n}) \\ (2A_{21}+3B_{21}) & \ddots & \vdots \\ \vdots & & \vdots \\ (2A_{n1}+3B_{n1}) & \cdots & (2A_{nn}+3B_{nn}) \end{bmatrix}$$

(entries of Z are linear combinations of elements from $A+B$).

By Spectral Theorem, every eigenvalue λ of Z is a real number.

~~Since A and B are PSD matrices,~~

For $x \in \mathbb{R}^n$, $x^TAx \geq 0$ and $x^TBx \geq 0$.

By extension, $x^TAx + x^TBx \geq 0$

Because 2 and 3 are scalars,

For $x \in \mathbb{R}^n$, $x^T(2A)x = 2(x^TAx) \geq 0$

For $x \in \mathbb{R}^n$, $x^T(3B)x = 3(x^TBx) \geq 0$

Thus,

For $x \in \mathbb{R}^n$, $x^T(2A)x + x^T(3B)x \geq 0$

$x^T(2A+3B)x \geq 0$ distributivity in matrix multiplication

Thus $2A+3B$ is PSD using properties from question 3.1

Figure 11: Proof for question 3.2(a)

Question 2(b):

Solution is on the next page.

(b) If A is PSD, all diagonal entries of A are non-negative, $A_{ii} \geq 0, \forall i \in [n]$.

Let matrix E be defined as:

$$E = \begin{bmatrix} e_1 & e_2 & e_3 & \ldots & e_n \end{bmatrix}$$

$E \in \mathbb{R}^{n \times n}$ where e_n is a column vector of all zeros except in the nth place where there is a 1.

Since A is PSD,

for all $x \in \mathbb{R}^n$, $x^T A x \geq 0$.

Thus, for all e_n in E, $e_n^T A e_n \geq 0$. [since $e_n \in \mathbb{R}^n$]

Now assume towards contradiction that diagonal entries of A were negative. ($A_{ii} < 0, \forall i \in [n]$)

Then, $e_i^T A e_i = 0 \cdot A_{11} \cdot 0 + 1 \cdot A_{12} \cdot 0 + \ldots + 1 \cdot A_{ii} \cdot 1 + \ldots + 0 \cdot A_{nn} \cdot 0$

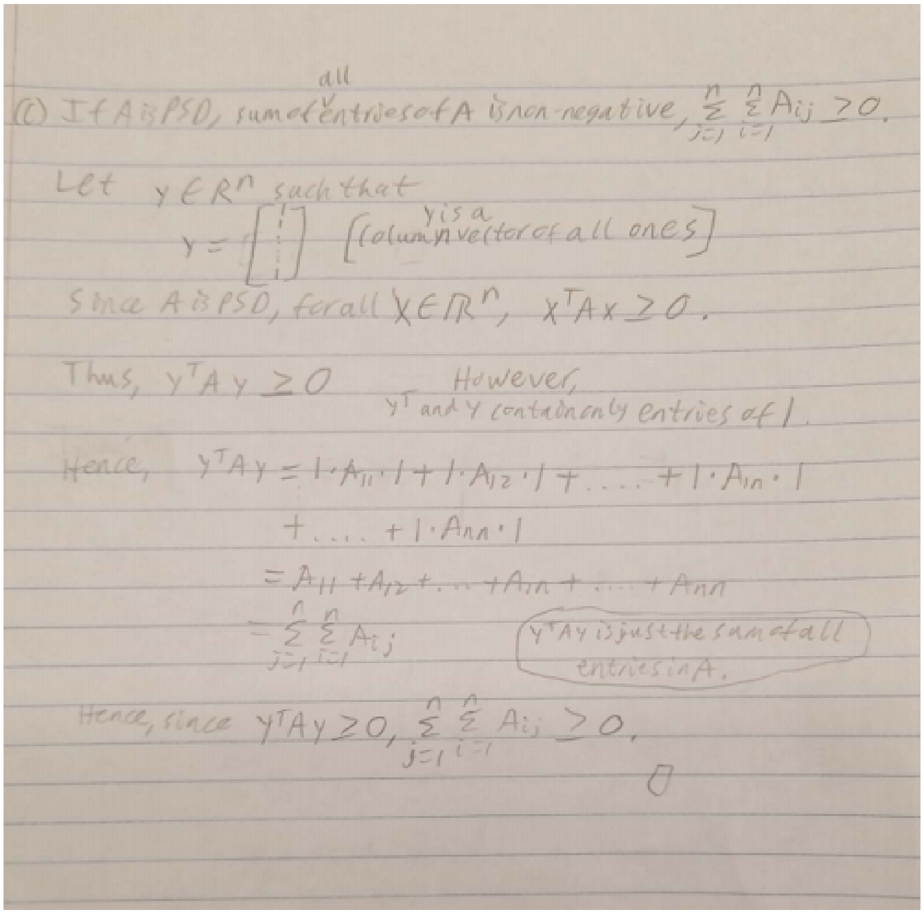
$e_i^T A e_i = 1 \cdot A_{ii} \cdot 1 = A_{ii}$

Then, $e_i^T A e_i < 0$ contradicting the fact that $e_i^T A e_i$ must be ≥ 0 since A is PSD.

Thus, all diagonal entries of A must be non-negative.

□

Figure 12: Proof for question 3.2(b)



(c) If A is PSD, sum of all entries of A is non-negative, $\sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} \geq 0$.

Let $y \in \mathbb{R}^n$ such that

$$y = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$ [y is a column vector of all ones]

Since A is PSD, for all $x \in \mathbb{R}^n$, $x^T A x \geq 0$.

Thus, $y^T A y \geq 0$. However, y^T and y contain only entries of 1.

Hence, $y^T A y = 1 \cdot A_{11} \cdot 1 + 1 \cdot A_{12} \cdot 1 + \dots + 1 \cdot A_{1n} \cdot 1 + \dots + 1 \cdot A_{nn} \cdot 1$

$= A_{11} + A_{12} + \dots + A_{1n} + \dots + A_{nn}$

$= \sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij}$

$y^T A y$ is just the sum of all entries in A.

Hence, since $y^T A y \geq 0$, $\sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} \geq 0$. □

Figure 13: Proof for question 3.2(c)

Question 2(c):
Solution is above.

Question 2(d):

Solution is on the next page.

(d) If A and B are PSD, then $Tr(AB) \geq 0$.

As proven in part (b) of this question, for a PSD matrix, all diagonal entries are non-negative.
This means $Tr(A) \geq 0$ and $Tr(B) \geq 0$.

Let $A \in R^{n \times n}$ be PSD and symmetric.
Let $B \in R^{n \times n}$ be PSD and symmetric.

$AB \in R^{n \times n}$ as result of matrix multiplication.

Let $B^{\frac{1}{2}}$ be PSD square root of B.

$$AB = B^{-\frac{1}{2}}(B^{\frac{1}{2}} A B^{\frac{1}{2}}) B^{\frac{1}{2}}$$

now AB is similar to $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ and $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ is congruent to A.

Thus, $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ is PSD with non-negative eigenvalues.

Since AB is similar to $B^{\frac{1}{2}} A B^{\frac{1}{2}}$, AB has the same non-negative eigenvalues.

Since AB is square matrix, $tr(AB) =$ sum of eigenvalues, which is non-negative since no eigenvalue is negative.

Thus, $tr(AB) \geq 0$ $\square$

Figure 14: Proof for question 3.2(d)

Question 2(e):

Solution is on the next 2 pages.

(e) If A and B are PSD, then $tr(AB) = 0$ if and only if $AB = 0$.

($\Leftarrow$) Assume $AB = 0$. AB is the all zero matrix.
clearly, $tr(AB) = 0$ since all diagonal entries are zero.

($\Rightarrow$) Assume $tr(AB) = 0$.
Let $B^{\frac{1}{2}}$ be the PSD square root of B.
Let $A^{\frac{1}{2}}$ be the PSD square root of A.

$tr(AB) = tr(BA)$

Proof: $tr(AB) = \sum_{i=1}^{n} (AB)_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} B_{ji} = \sum_{i=1}^{n} \sum_{j=1}^{n} B_{ji} A_{ij}$

$= \sum_{j=1}^{n} (BA)_{jj} = tr(BA)$

$tr(AB) = tr(B^{\frac{1}{2}} A B^{\frac{1}{2}})$

$B^{\frac{1}{2}} A B^{\frac{1}{2}} = (A^{\frac{1}{2}} B^{\frac{1}{2}})^T \cdot (A^{\frac{1}{2}} B^{\frac{1}{2}})$

thus, it's PSD

since $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ is PSD, all its eigenvalues are ≥ 0.
$B^{\frac{1}{2}} A B^{\frac{1}{2}}$ is diagonalizable as well since it's PSD.

Let P be matrix with eigenvectors for $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ as its columns.

$P^{-1} (B^{\frac{1}{2}} A B^{\frac{1}{2}}) P = D$

where D is a matrix with eigenvalues of $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ on diagonal and zero everywhere else.

Then, since $P^{-1} P = I_n$ [I_n is identity matrix]

$B^{\frac{1}{2}} A B^{\frac{1}{2}} = D$

Figure 15: Proof for question 3.2(e)

Now, if $\text{trace}(AB) = 0$,

then $tr(B^{\frac{1}{2}}AB^{\frac{1}{2}}) = 0$ and $tr(D) = 0$

since $B^{\frac{1}{2}}AB^{\frac{1}{2}}$ only has non-negative eigenvalues,

$D = 0$ [0 is zero matrix],

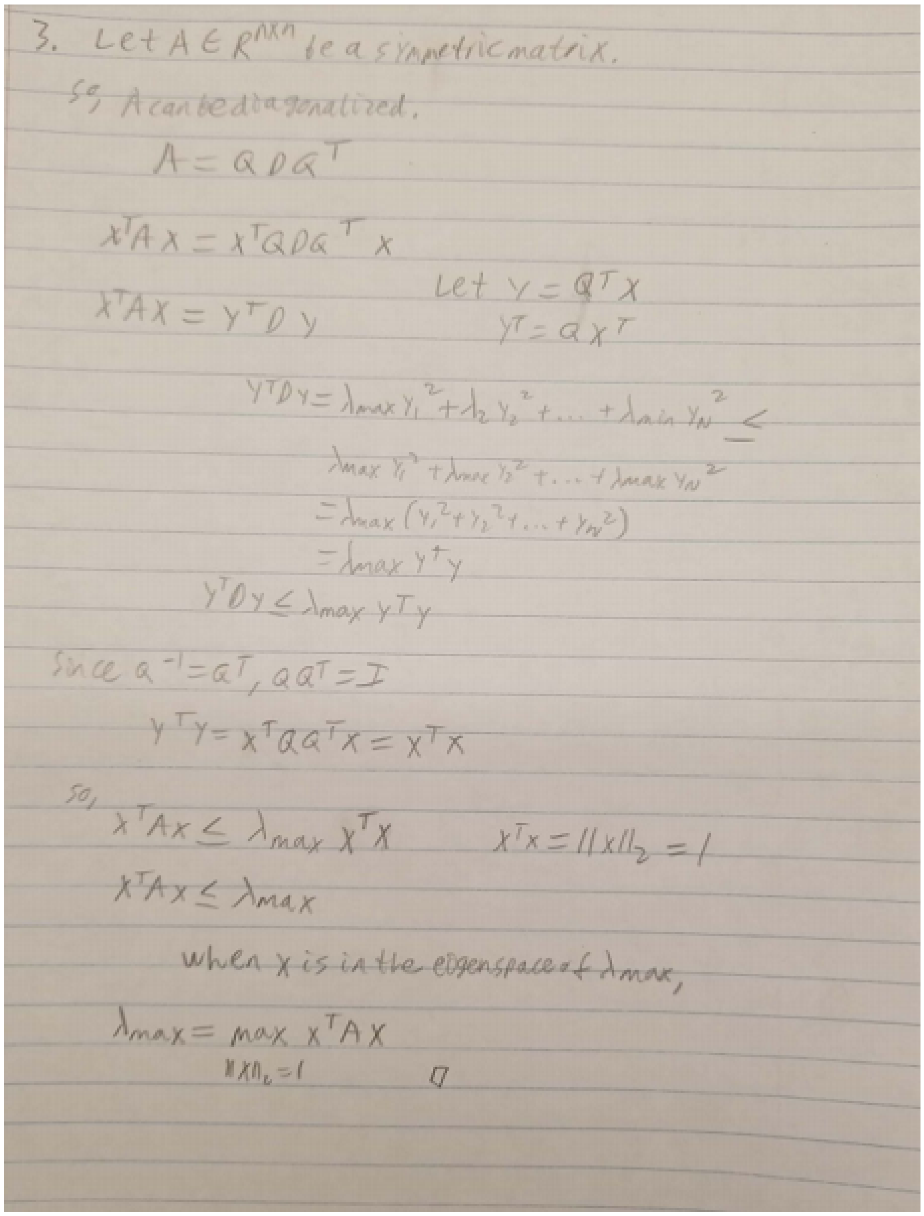
Thus, $B^{\frac{1}{2}}AB^{\frac{1}{2}} = 0$

and $(A^{\frac{1}{2}}B^{\frac{1}{2}}) \cdot (A^{\frac{1}{2}}B^{\frac{1}{2}}) = B^{\frac{1}{2}}AB^{\frac{1}{2}} = AB$

Thus, $AB = 0$ $\square$

Figure 16: Proof for question 3.2(e) continued.

Question 3:
Solution is on the next page.


3. Let $A \in R^{n \times n}$ be a symmetric matrix.

so, A can be diagonalized.

$$A = QDQ^T$$

$$x^TAx = x^TQDQ^T x$$

$$x^TAx = y^TDy \qquad \text{Let } y = Q^Tx, \quad y^T = Qx^T$$

$$y^TDy = \lambda_{max} y_1^2 + \lambda_2 y_2^2 + \dots + \lambda_{min} y_N^2 \le$$

$$\lambda_{max} y_1^2 + \lambda_{max} y_2^2 + \dots + \lambda_{max} y_N^2$$

$$= \lambda_{max}(y_1^2 + y_2^2 + \dots + y_N^2)$$

$$= \lambda_{max} y^T y$$

$$y^TDy \le \lambda_{max} y^T y$$

Since $Q^{-1} = Q^T$, $QQ^T = I$

$$y^Ty = x^TQQ^Tx = x^Tx$$

so,

$$x^TAx \le \lambda_{max} x^Tx \qquad x^Tx = ||x||_2 = 1$$

$$x^TAx \le \lambda_{max}$$

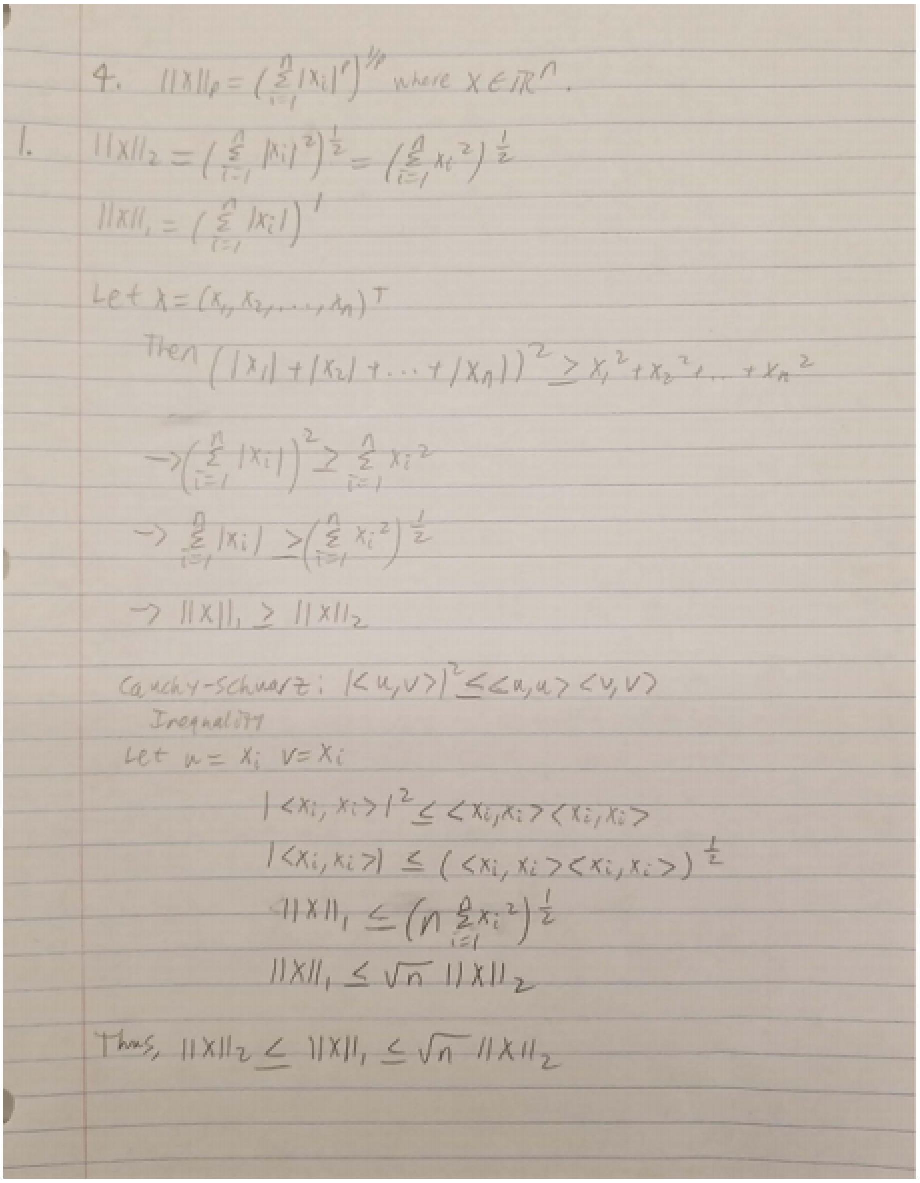
when x is in the eigenspace of λ_{max},

$$\lambda_{max} = \max_{||x||_2 = 1} x^TAx \qquad \square$$

Figure 17: Proof for question 3.3

4 Gradients and Norms

Question 1:
Solution is on the next page.



4. $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ where $x \in \mathbb{R}^n$.

1. $\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$

$\|x\|_1 = \left(\sum_{i=1}^{n} |x_i|\right)^1$

Let $x = (x_1, x_2, \dots, x_n)^T$

Then $\left(|x_1| + |x_2| + \dots + |x_n|\right)^2 \geq x_1^2 + x_2^2 + \dots + x_n^2$

$\rightarrow \left(\sum_{i=1}^{n} |x_i|\right)^2 \geq \sum_{i=1}^{n} x_i^2$

$\rightarrow \sum_{i=1}^{n} |x_i| \geq \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$

$\rightarrow \|x\|_1 \geq \|x\|_2$

Cauchy-Schwarz Inequality: $|\langle u,v\rangle|^2 \leq \langle u,u\rangle \langle v,v\rangle$

Let $u = x_i$ $v = x_i$

$|\langle x_i, x_i\rangle|^2 \leq \langle x_i, x_i\rangle \langle x_i, x_i\rangle$

$|\langle x_i, x_i\rangle| \leq \left(\langle x_i, x_i\rangle \langle x_i, x_i\rangle\right)^{\frac{1}{2}}$

$\|x\|_1 \leq \left(n \sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$

$\|x\|_1 \leq \sqrt{n}\, \|x\|_2$

Thus, $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\, \|x\|_2$

Figure 18: Proof for question 4.1

Question 2(a) - 2(c):
Solutions are on the next page.

2.

(a) $\alpha = \sum_{i=1}^{n} y_i \ln(\beta_i) \qquad y, \beta \in \mathbb{R}^n$

$$\frac{\partial \alpha}{\partial \beta_i} = \sum_{i=1}^{n} y_i \cdot \frac{1}{\beta_i} = \sum_{i=1}^{n} \frac{y_i}{\beta_i}$$

$\square$

(b) Let $\beta = \sinh(\gamma)$ for $\gamma \in \mathbb{R}^n$.

$$\beta = \frac{e^{\gamma} - e^{-\gamma}}{2}$$

$$\frac{\partial \beta_i}{\partial \gamma_j} = \frac{\partial}{\partial \gamma_j}\left(\frac{e^{\gamma_j}}{2}\right) - \frac{\partial}{\partial \gamma_j}\left(\frac{e^{-\gamma_j}}{2}\right)$$

$$\frac{\partial \beta_i}{\partial \gamma_j} = \frac{e^{\gamma_j}}{2} + \frac{e^{-\gamma_j}}{2}$$

$$\frac{\partial \beta_i}{\partial \gamma_j} = \frac{e^{\gamma_j} + e^{-\gamma_j}}{2} = \cosh(\gamma_j)$$

$$\frac{\partial \beta_i}{\partial \gamma_j} = \cosh(\gamma)$$

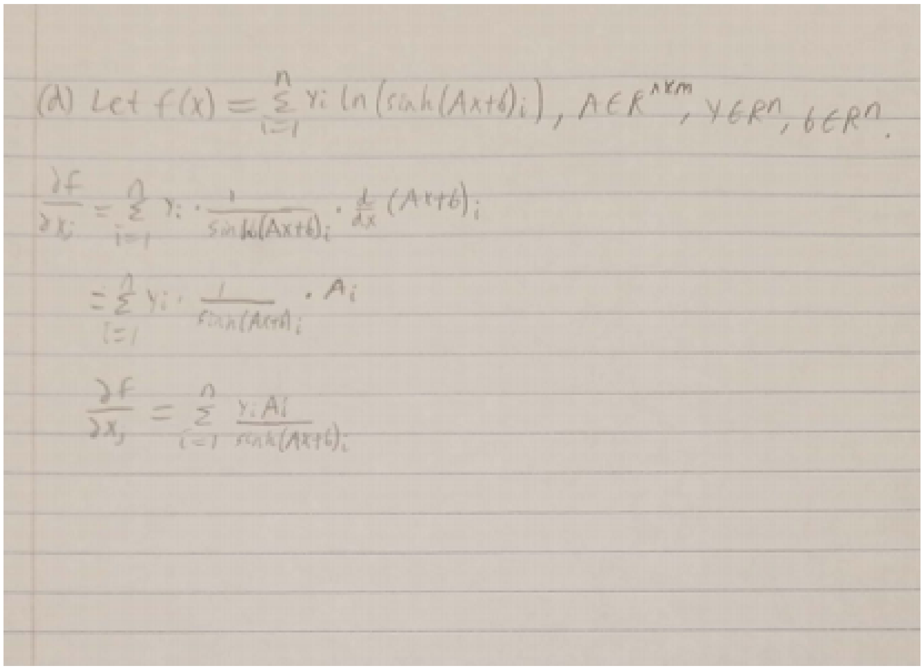
$\square$

(c) Let $\gamma = A\rho + b$ for $b \in \mathbb{R}^n, \rho \in \mathbb{R}^m, A \in \mathbb{R}^{n \times m}$

$\frac{\partial \gamma_i}{\partial \rho_j} = (A_1 \; A_2 \; \ldots \; A_n)^T$ where A_n represents the dot product of row n of A with entry n of ρ.

result is a column vector with dimensions n rows, 1 column.

Figure 19: Proofs for questions 4.2(a-c)



(d) Let $f(x) = \sum_{i=1}^{n} y_i \ln(\sinh(Ax+b)_i)$, $A \in R^{n \times m}$, $y \in R^n$, $b \in R^n$.

$$\frac{\partial f}{\partial x_i} = \sum_{i=1}^{n} y_i \cdot \frac{1}{\sinh(Ax+b)_i} \cdot \frac{d}{dx}(Ax+b)_i$$

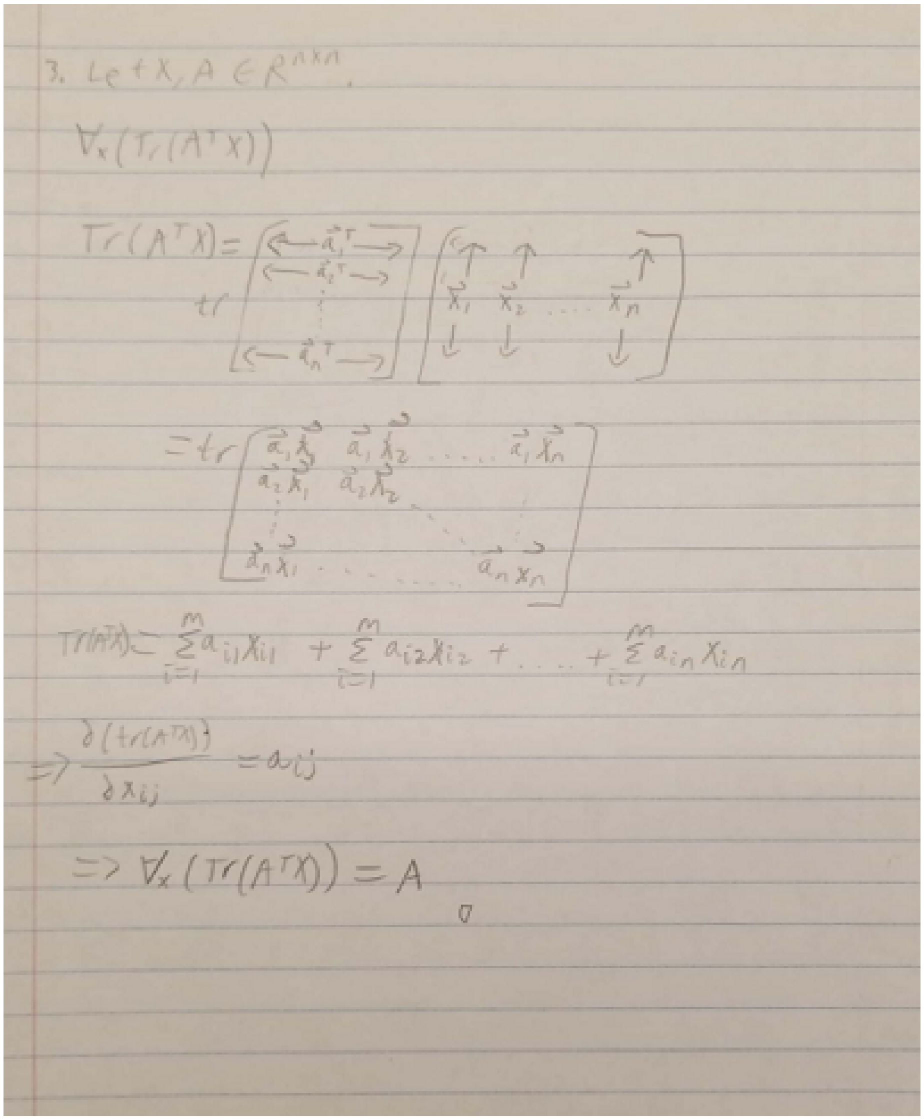
$$= \sum_{i=1}^{n} y_i \cdot \frac{1}{\sinh(Ax+b)_i} \cdot A_i$$

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^{n} \frac{y_i A_i}{\sinh(Ax+b)_i}$$

Figure 20: Proof for question 4.2(d)

Question 2(d):
Solution is above.

Question 3:
Solution is on the next page.



3. Let $X, A \in R^{n \times n}$.

$$\nabla_X \left(Tr(A^T X) \right)$$

$$Tr(A^T X) = tr \begin{bmatrix} \leftarrow \vec{a}_1^T \rightarrow \\ \leftarrow \vec{a}_2^T \rightarrow \\ \vdots \\ \leftarrow \vec{a}_n^T \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

$$= tr \begin{bmatrix} \vec{a}_1 \vec{x}_1 & \vec{a}_1 \vec{x}_2 & \cdots & \vec{a}_1 \vec{x}_n \\ \vec{a}_2 \vec{x}_1 & \vec{a}_2 \vec{x}_2 & & \vdots \\ \vdots & & \ddots & \\ \vec{a}_n \vec{x}_1 & \cdots & \cdots & \vec{a}_n \vec{x}_n \end{bmatrix}$$

$$Tr(A^T X) = \sum_{i=1}^{m} a_{i1} x_{i1} + \sum_{i=1}^{m} a_{i2} x_{i2} + \cdots + \sum_{i=1}^{m} a_{in} x_{in}$$

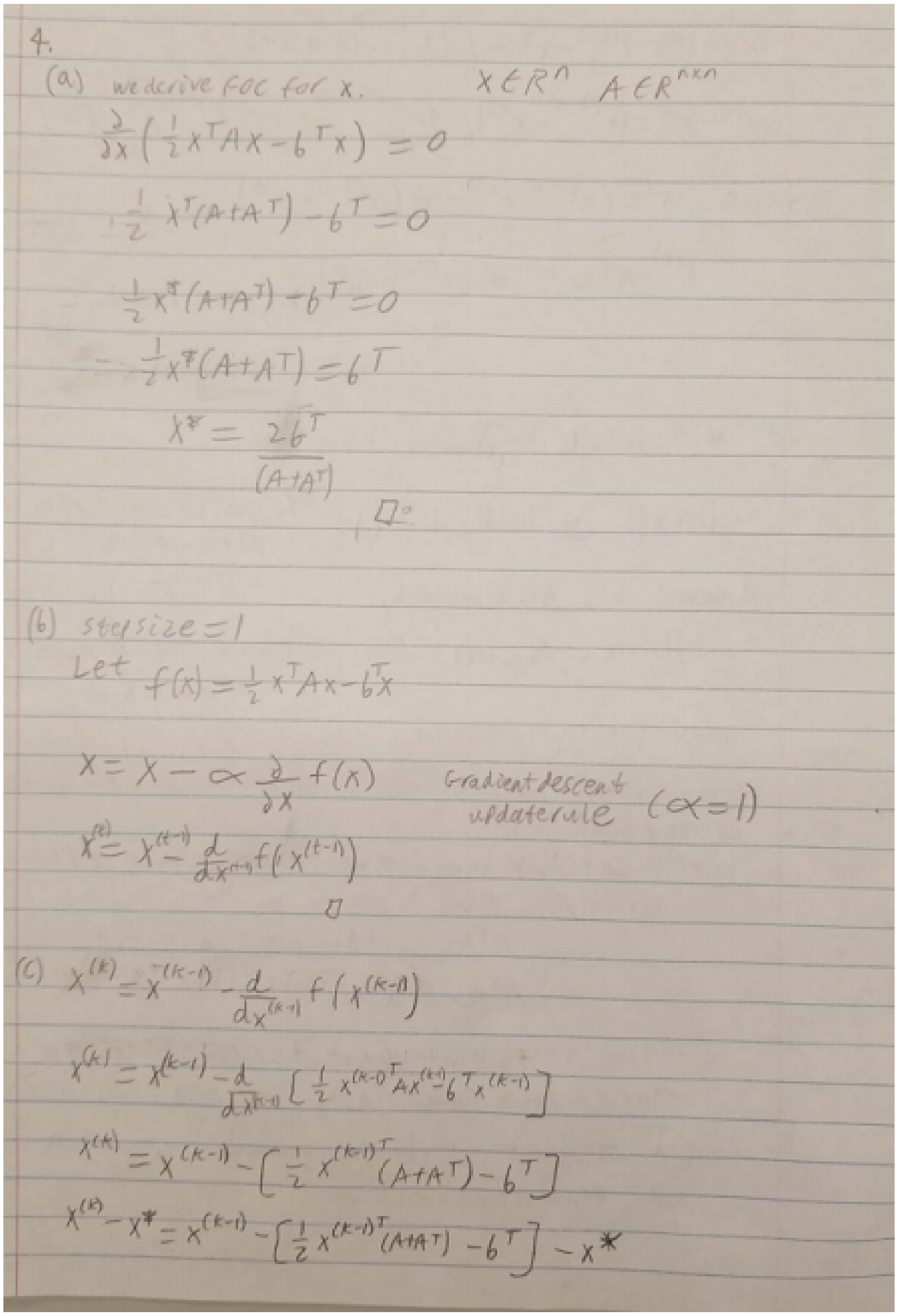
$$\Rightarrow \frac{\partial(tr(A^T X))}{\partial x_{ij}} = a_{ij}$$

$$\Rightarrow \nabla_X \left(Tr(A^T X) \right) = A \quad \square$$

Figure 21: Proof for question 4.3

Question 4(a-e):

Solutions are on the next 3 pages.



4.

(a) we derive FOC for x. $x \in R^n \quad A \in R^{n \times n}$

$$\frac{\partial}{\partial x}\left(\frac{1}{2}x^T A x - b^T x\right) = 0$$

$$\frac{1}{2}x^T(A+A^T) - b^T = 0$$

$$\frac{1}{2}x^{*}(A+A^T) - b^T = 0$$

$$\frac{1}{2}x^{*}(A+A^T) = b^T$$

$$x^{*} = \frac{2b^T}{(A+A^T)}$$

$\square$

(b) stepsize $= 1$

Let $f(x) = \frac{1}{2}x^T A x - b^T x$

$$x = x - \alpha \frac{\partial}{\partial x} f(x)$$ Gradient descent update rule $(\alpha = 1)$

$$x^{(t)} = x^{(t-1)} - \frac{d}{dx^{(t-1)}} f(x^{(t-1)})$$

$\square$

(c) $$x^{(k)} = x^{(k-1)} - \frac{d}{dx^{(k-1)}} f(x^{(k-1)})$$

$$x^{(k)} = x^{(k-1)} - \frac{d}{dx^{(k-1)}}\left[\frac{1}{2}x^{(k-1)T} A x^{(k-1)} - b^T x^{(k-1)}\right]$$

$$x^{(k)} = x^{(k-1)} - \left[\frac{1}{2}x^{(k-1)T}(A+A^T) - b^T\right]$$

$$x^{(k)} - x^{*} = x^{(k-1)} - \left[\frac{1}{2}x^{(k-1)T}(A+A^T) - b^T\right] - x^{*}$$

Figure 22: Proofs for questions 4.4(a-e)

$$x^{(k)} - x^* = x^{(k-1)} - x^* - \left[\tfrac{1}{2}x^{(k-1)T}(A+A^T) - b^T\right]$$

$$x^{(k)} - x^* = \left(x^{(k-1)} - x^*\right) - \left[\tfrac{1}{2}x^{(k-1)T}A + \tfrac{1}{2}x^{(k-1)T}A^T - b^T\right]$$

plug in x^*

$$x^{(k)} - x^* = \left(x^{(k-1)} - x^*\right) - A\left(x^{(k-1)} - x^*\right)$$

$$x^{(k)} - x^* = (I-A)\left(x^{(k-1)} - x^*\right)$$

□

(d) WTS $\|Ax\|_2 \leq \lambda_{max(A)} \|x\|_2$

$\|Ax\|_2 \leq \|A\|_2 \, \|x\|_2$ (holds for induced norms, which we assume since in 3.3, $\|x\|_2 = 1$).

However, since A is a matrix,

$\|A\|_2 = \sigma_{max}(A)$ $\quad$ $\sigma_{max}(A)$ is largest singular value of matrix A.

However, A is a PSD matrix.

Lemma For PSD matrices, the singular values are the same as the eigenvalues.

proof: Let λ, v be an eigenvalue, eigenvector pair for PSD matrix A. $Av = \lambda v$.

$$A^TAv = A^T\lambda v \qquad A^T = A \text{ (for PSD)}$$

$$A^TAv = \lambda^2 v$$

Thus, λ^2 is an eigenvalue for A^TA, which is the square of a singular value for matrix A. Since A is PSD, $\lambda \geq 0$ hence $\sqrt{\lambda^2} = \lambda$. Thus, the singular value for matrix A is equal to the eigenvalue.

Figure 23: Proofs for questions 4.4(a-e)

Thus, for PSD matrix A,

$$\|A\|_2 = \lambda_{max(A)}$$

$\lambda_{max(A)}$ is largest eigenvalue of A.

Hence, $\|Ax\|_2 \leq \lambda_{max(A)} \|x\|_2$.

$\square$

(e) From part (d),

$$\|Ax\|_2 \leq \lambda_{max(A)} \|x\|_2$$

From part (c),

$$x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$$

$$\|x^{(k)} - x^*\|_2 \leq \|(I-A)\|_2 \, \|x^{(k-1)} - x^*\|_2$$

$$\|x^{(k)} - x^*\|_2 \leq \sigma_{max}(I-A) \, \|x^{(k-1)} - x^*\|_2$$

Since $I-A$ is a matrix, $\|(I-A)\|_2 = \sigma_{max}(I-A)$

I is a PSD matrix since all eigenvalues of I are non-negative.

$I-A$ is a PSD matrix since PSD-ness is preserved under matrix addition

Using lemma I proved in part (d),

$$\sigma_{max}(I-A) = \lambda_{max}(I-A)$$

Let $\rho = \lambda_{max}(I-A)$

condition that $0 < \rho < 1$ is preserved since

$$0 \leq \lambda_{max(I-A)} \leq 1$$

Thus, $\|x^{(k)} - x^*\|_2 \leq \rho \, \|x^{(k-1)} - x^*\|_2, \quad 0 < \rho < 1$

$\square$

Figure 24: Proofs for questions 4.4(a-e)

Question 4(f):
Solution is on the next page.

(f) $\|x^{(k)} - x^*\|_2 \le \rho \|x^{(k-1)} - x^*\|_2$ from (e)

Let $x^{(k-1)} = x^0$

$$x^{(k)} - x^* = (I - A)(x^0 - x^*) \quad \text{(part(e))}$$

from part (e), this is

$$\|x^{(k)} - x^*\|_2 \le \rho \|x^0 - x^*\|_2$$

now, $\|x^{(k)} - x^*\|_2 \le \epsilon$ is target.

$$\rho \|x^0 - x^*\|_2 = \epsilon$$

$$\|x^0 - x^*\|_2 = \frac{\epsilon}{\rho}$$

$$\left(\sum_{i=1}^{k} |x^0 - x^*|_2^2\right)^{\frac{1}{2}} = \frac{\epsilon}{\rho}$$

$$k \|x_0 - x^*\|_2 = \frac{\epsilon}{\rho}$$

$$\boxed{k = \frac{\epsilon}{\rho} \|x_0 - x^*\|_2}$$

□

Figure 25: Proof for question 4.4

Question 5:
Solution is on the next page.

5. Assuming X has full column rank, compute $\theta^* = \underset{\theta}{\operatorname{argmin}}\, L(\theta)$ in terms of X and y.

$$L(\theta) = ||Y - X\theta||_2^2$$

$$\underset{\theta}{\operatorname{argmin}}\, ||Y - X\theta||_2^2$$

$$\frac{\partial}{\partial\theta}\left((Y - X\theta)^T(Y - X\theta)\right) = 0$$

$$\frac{\partial}{\partial\theta}\left((Y^T - X^T\theta^T)(Y - X\theta)\right) = 0$$

$$\frac{\partial}{\partial\theta}\left(y^Ty - y^TX\theta - X^T\theta^Ty + X^TX\theta^T\theta\right) = 0$$

$$\frac{\partial}{\partial\theta}\left(y^Ty - 2X^T\theta^Ty + X^TX\theta^T\theta\right) = 0$$

$$-2X^Ty + 2X^TX\theta = 0$$

$$2X^TX\theta = 2X^Ty$$

$$(X^TX)\theta = X^Ty$$

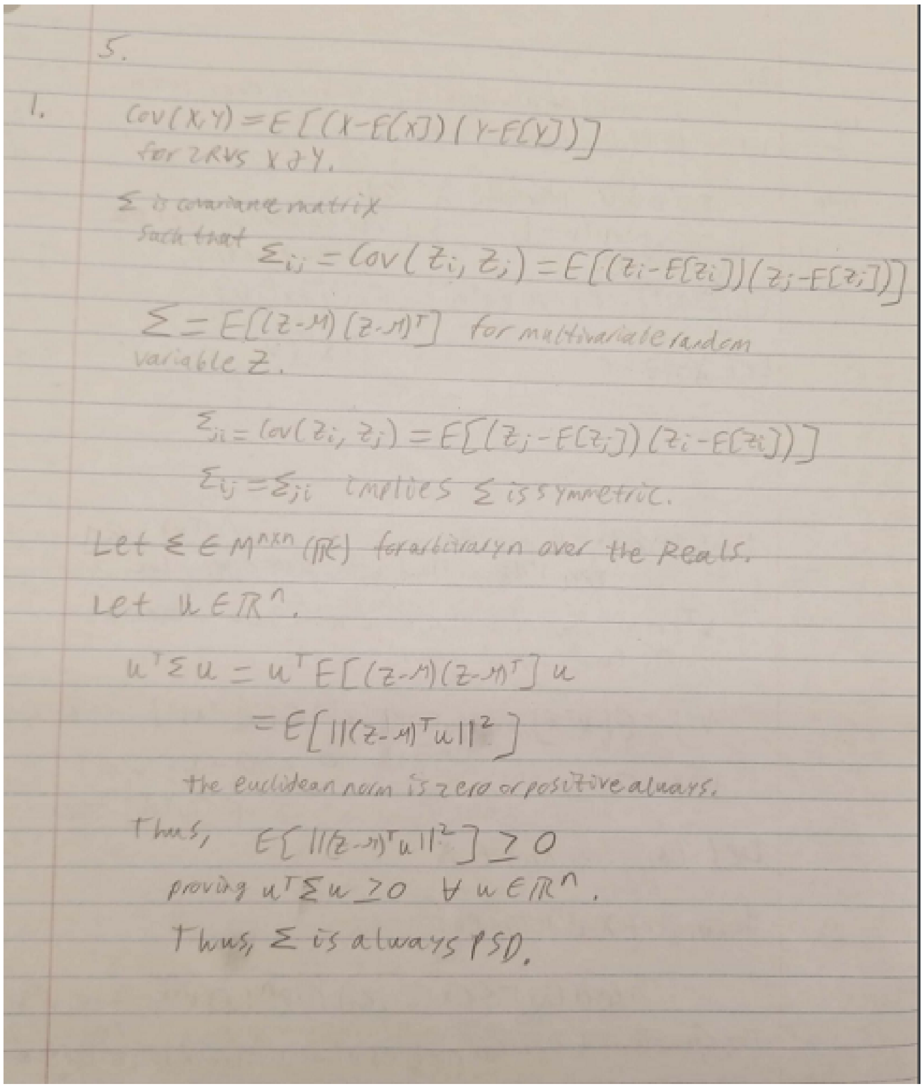
$$\boxed{\theta^* = (X^TX)^{-1}(X^Ty)}$$

Figure 26: Proof for question 4.5

5 Covariance Practice

Question 1:

Solution is on the next page.



5.

1. $Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$
for 2 RVs X & Y.

~~Σ is covariance matrix~~
~~such that~~ $\Sigma_{ij} = Cov(z_i, z_j) = E[(z_i - E[z_i])(z_j - E[z_j])]$

$\Sigma = E[(z-\mu)(z-\mu)^T]$ for multivariate random variable z.

$\Sigma_{ji} = Cov(z_i, z_j) = E[(z_j - E[z_j])(z_i - E[z_i])]$

$\Sigma_{ij} = \Sigma_{ji}$ implies Σ is symmetric.

~~Let $\Sigma \in M^{n \times n}(\mathbb{R})$ for arbitrary n over the Reals.~~

Let $u \in \mathbb{R}^n$.

$u^T \Sigma u = u^T E[(z-\mu)(z-\mu)^T] u$
$= E[\|(z-\mu)^T u\|^2]$

the euclidean norm is zero or positive always.

Thus, $E[\|(z-\mu)^T u\|^2] \geq 0$

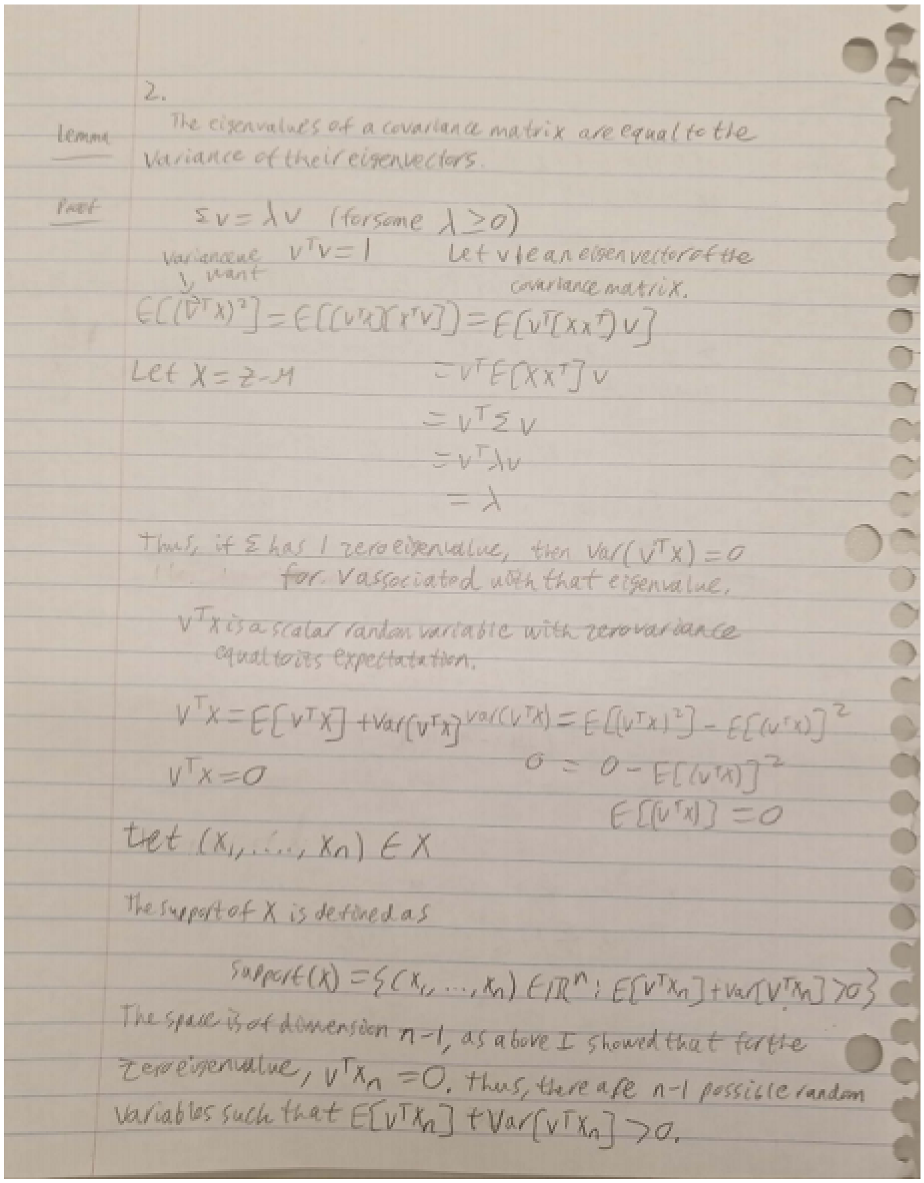
proving $u^T \Sigma u \geq 0 \quad \forall u \in \mathbb{R}^n$.

Thus, Σ is always PSD.

Figure 27: Proof for question 5.1

Question 2:

Solution is on the next 2 pages.



2.

lemma The eigenvalues of a covariance matrix are equal to the variance of their eigenvectors.

Proof $\Sigma v = \lambda v$ (for some $\lambda \geq 0$)

variance we want $v^T v = 1$ Let v be an eigenvector of the covariance matrix.

$E[(v^T x)^2] = E[(v^T x)(x^T v)] = E[v^T (x x^T) v]$

Let $x = z - \mu$

$= v^T E[x x^T] v$

$= v^T \Sigma v$

$= v^T \lambda v$

$= \lambda$

Thus, if Σ has 1 zero eigenvalue, then $Var(v^T x) = 0$ for v associated with that eigenvalue.

$v^T x$ is a scalar random variable with zero variance equal to its expectation.

$v^T x = E[v^T x] + Var(v^T x)$ $\quad Var(v^T x) = E[(v^T x)^2] - E[(v^T x)]^2$

$v^T x = 0$ $\quad 0 = 0 - E[(v^T x)]^2$

$E[(v^T x)] = 0$

Let $(x_1, \ldots, x_n) \in X$

The support of X is defined as

$$support(X) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : E[v^T x_n] + Var[v^T x_n] > 0\}$$

The space is of dimension $n-1$, as above I showed that for the zero eigenvalue, $v^T x_n = 0$. thus, there are $n-1$ possible random variables such that $E[v^T x_n] + Var[v^T x_n] > 0$.

Figure 28: Proof for question 5.2

To construct a new $\tilde{X}$ so that no information is lost from the original distribution, but that the covariance matrix has no zero eigenvalues, you can add in a noise term generated by a fixed probability distribution. That way, the information about the original distribution is preserved (since it is shifted by a constant that we know), and the covariance matrix would have no zero eigenvalues by construction.

If Σ has $m \leq n$ zero eigenvalues, $\tilde{X}$ would resemble a univariate or bivariate random variable depending on the value of m. As m increases, $\tilde{X}$ would behave more like a bivariate or univariate random variable.

Figure 29: Proof for question 5.2 continued.