HW02

Who I had discussions with:

Went to physical OHs and got help from various Students. Also got help/helped Madeline (Don't know her last name) Also Looked in the CS189 discord for help usually when stuck. Same thing for the EDStem posts on the homework.

"I certify that all solutions are entirely my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."

X

1)

1) $1\{a \geq b\} = \begin{cases} 1 & \text{if } a \geq b \\ 0 & \text{if } a < b \end{cases}$ $\leftarrow$ Indicator function

$$P(A) = E[1\{A\}]$$

$$A = X \geq t$$

$$P(X \geq t) = E[1\{X \geq t\}]$$

$$1\{X \geq t\} \leq \frac{X}{t}$$

$1\{a \geq b\} \leq \frac{a}{b}$ because if $a = b$ then RHS $= 1$ which is the highest value the indicator funct. can output.

$$P(X \geq t) \leq E\left[\frac{X}{t}\right]$$

Note: $E\left[\frac{X}{t}\right] = \frac{1}{t} E[X]$ because $\frac{1}{t}$ is a constant

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

2)

$$X = (|\hat{\mu} - \mu|)^2$$

$$(P(|\hat{\mu} - \mu| \geq t))^2 \leq \frac{E[(|\hat{\mu} - \mu|)^2]}{(t)^2}$$

Taken from previous part $P(X \geq t) \leq E[X]/t$ but $x = |\hat{\mu} - \mu|$

$$(P(|\hat{\mu} - \mu| \geq t))^2 \leq \frac{1}{n(t)^2}$$

$$\boxed{P(|\hat{\mu} - \mu| \geq t) \leq \frac{1}{\sqrt{n}\, t}}$$

2) 1) $x^T \Sigma x \geq 0$ $\longleftarrow$ A definition of PSD (1)

$$x^T E[(z-\mu)(z-\mu)^T]x \geq 0$$

$$E[\underbrace{x^T (z-\mu)}\underbrace{(z-\mu)^T x}] \geq 0$$

The two underlined Products are identical we will call ω

$$E[\omega \cdot \omega] \geq 0$$

$$E[\omega^2] \geq 0$$

Something squared is always greater than 0, so covariance matrix satisfies a PSD equation

2) i) $0.3 \cdot 0.4 = 0.12$ $\boxed{12\%}$

ii) $0.4 \cdot 0.7 = 0.28 = \boxed{28\%}$

iii) $0.28 \cdot 0.7 = 0.2016 = \boxed{20\%}$

iv) $0.6 = \boxed{60\%}$

3) $$\int_0^{1/\sqrt{3}} 4 \frac{2}{\pi(1+x^2)} dx + \int_{1/\sqrt{3}}^{1} 3 \frac{2}{\pi(1+x^2)} dx + \int_1^{\sqrt{3}} 2 \frac{2}{\pi(1+x^2)} dx$$

$$\frac{8}{\pi}[\arctan(x)]_0^{1/\sqrt{3}} + \frac{6}{\pi}[\arctan(x)]_{1/\sqrt{3}}^{1} + \frac{4}{\pi}[\arctan(x)]_1^{\sqrt{3}}$$

$$\frac{4}{3} + \frac{1}{2} + \frac{1}{3} = \boxed{\frac{13}{6}}$$

3) 1) a) $\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}$

$$\begin{bmatrix} 0 & I_n \\ AB & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ B & 1 \end{bmatrix}$$

$$\begin{bmatrix} B & I_n \\ 0 & AB \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & AB \end{bmatrix}$$

$$\begin{bmatrix} B & I_n \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix}$$

$$\begin{bmatrix} B & A I_n \\ 0 & A \end{bmatrix} \cdot \begin{bmatrix} 0 & A \cdot I_n \cdot 1 \\ 0 & A \end{bmatrix}$$

$$\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$$

b)

<u>Upper Bound:</u>

rank of $\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$ must be the same as the rank of $\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}$ which this rank is n (Because rank $I_n = n$) + rank(AB) and rank of first matrix is rank(A) + rank(B), we have: $n + \text{rank}(AB) = \text{rank}(A) + \text{rank}(B)$

$$\boxed{\text{rank}(AB) \geq \text{rank}(A) + \text{rank}(B) - n}$$

Lower Band:

Since the final matrix dimension is $m \times p$, this means the rowspace of AB is in the subspace of A and the column the subspace of B. This means the rank is dependent on A and B, so that means it cannot exceed the lower one of the two, leaving us with
$\text{rank}(AB) \leq \min(\text{rank } A, \text{rank } B)$

c) $\det(M) != 0$ Therefore M must be full rank with a row space dimension of r and column space dimension of r.

d) Nullspace of A^TA is defined as $A^TA\,x = 0$. If the rank (A^TA) is less than rank(A), then that means there are less Lin. Ind. columns in A^TA than A. Because of rank nullity theorem this means if rank is smaller than nullspace must be bigger which implies there must be a non-trivial solution in there. If there is a non-trivial solution to $A^TAx = 0$ then that means it must be in the column space of A and nullspace of A^T which are orthogonal, which means they cannot be shared together This causes a contradiction,

e) AR^n is the matrix sized to fit the R^n space. Specifically for the rowspace sin R^n is the $n \times 1$ matrix.
A^TAR^n is the same as mentioned before but instead it applies to the row and column space

2) $A \to B$:

$$Ax = \lambda x$$

$$x^T A x \geq 0$$

$$x^T \lambda x \geq 0 \quad \leftarrow \lambda \text{ is just a number}$$

$$\lambda x^T x \geq 0 \quad \leftarrow x^T x \geq 0 \text{ due to squared nature}$$

So the only way the above inequality is true is if $\lambda \geq 0$

$B \to C$:

$$A = \lambda D \lambda^T$$

- λ contains eigenvectors
- D diagonal matrix with eigenvalues
- Diagonalized the matrix

$$D = D^{1/2} D^{1/2}$$ Aka singular values

$$A = \lambda D^{1/2} D^{1/2} \lambda^T$$

$$A = \lambda D^{1/2} (\lambda D^{1/2})^T$$

$$U = \lambda D^{1/2}$$

$$\boxed{A = UU^T}$$

$C \Rightarrow A$:

$$x^T A x$$

$$x^T (U U^T) x$$

$$x^T U \, U^T x$$

$$x^T U (x^T U)^T$$

$$\omega = x^T U$$

$$\omega \, \omega^T \geq 0$$

Therefore inequality holds true.

3) a)

$$\langle A, xy^T \rangle$$

$$\text{trace}(A^T xy^T) \qquad \text{order does not matter}$$

$$\text{trace}(x A y^T)$$

$$\text{trace}(x^T A y)$$

Note on dimensions: $1 \times \cancel{m} \quad \cancel{m} \times n \quad n \times 1$

$$1 \times \cancel{A} \quad n \times 1 = 1 \times 1$$

So $x^T A y$ is a scalar

The trace of a scalar is itself.

b) For PSD we know $x^T A x \geq 0$
$x^T B x \geq 0$

therefore since A, B are both are positive, then the trace as a result would also be positive.

c) Create a diagonal matrix with the diagonal entries are

$$\langle A, B \rangle = \text{trace}(A B)$$

$x^T A B x \geq 0$ scaling by a constant $\lambda_{max}(A)$ we will be able to ensure holds true

$$\boxed{\underbrace{\|A\|_F \|B\|_F}_{\langle A, B \rangle} \leq \sqrt{n}\, \lambda_{max}(A) \|B\|_F}$$

4) $$r(AA^T, u) = \frac{u^T (AA^T) u}{u^T u} \qquad \frac{u^T (A^T A) u}{u^T u} = r(A^T A, u)$$

$$\frac{u^T (AA^T) u}{u^T u} = \frac{u^T (A^T A) u}{u^T u}$$

As a result of this, the singular value of A as shown in SVD decomposition $U \Sigma V^T$ the σ would be represented as $\boxed{\max_{\substack{u \in \mathbb{R}^m,\ v \in \mathbb{R}^n \\ \|u\|=1\ \|v\|=1}} u^T A v}$

4) 1)

$$\begin{bmatrix} 2A_{11}\cos(A_{11}^2 + e^{A_{11}+A_{22}})e^{A_{11}+A_{22}} + x^T A y & \cos(A_{11}^2\, e^{A_{11}+A_{22}}) + x^T A y \\ \cos(A_{11}^2\, e^{A_{11}+A_{22}}) + x^T A y & \cos(A_{11}^2 + e^{A_{11}+A_{22}})e^{A_{11}+A_{22}} + x^T A y \end{bmatrix}$$

2) a) $\sigma_{max}(A) = \sqrt{\lambda_{max}(A^T A)}$

b) $\max\limits_{1 \le i \le m} \sum\limits_{j=1}^{n} |a_{ij}|$

3) a) $\sum\limits_{i=1}^{n} \frac{\partial \alpha}{\partial B_i} y_i \ln B_i$

$$\boxed{\sum_{i=1}^{n} \frac{y_i}{B_i}}$$

b) $\boxed{\frac{\partial y_i}{\partial p_j} = A}$

c) $K \times \cancel{m}$ $\quad \cancel{m} \times n$ $\quad K \times n$

$$\boxed{\frac{d\vec{z}}{dy} \, \frac{dy}{dx}}$$

d) $\boxed{\nabla_x Y^T z + Y^T \nabla_x z}$

4) a) $\boxed{\sum_{i=1}^{m} \sum_{j=1}^{n} (X W^T - Y)^2}$ ← with summation

$\boxed{(X W^T - Y)^2}$ ← without summation

b) $\boxed{\langle RSS(\omega), \Delta W \rangle}$

c) $\mu'_{\Delta\theta}(\theta) = \Delta\theta \cdot \mu(\theta)$

$= \boxed{\sum_{j=1}^{L} \omega_{>j} \Delta w_j \, w_{<j}}$

$\mu(\theta)$ term is irrelevant due to summation

d) The chain rule for scalar functions is as follows: $\frac{d}{dx} f(g(x))|_{x=x_0} = \frac{d}{dx} f(y)|_{y=g(x_0)}$

$$\cdot \frac{d}{dx} g(x)|_{x=x_0}$$

applying this onto RSS and μ, we get

$$\boxed{\frac{d}{d\theta} \|XW^T - Y\|_F^2 \cdot \frac{d}{d\theta} \mu(\theta)}$$

e) $J(\theta) = RSS(\mu(\theta))$

$$= \|XW^T - Y\|_F^2$$

$$\langle XW^T - Y \rangle^2 \quad (1)$$

$$\mu(\theta) = \sum_{j=1}^{L} W_{>j} \Delta W_j W_{<j} \quad (2)$$

$$W = \mu(W_L, W_{L-1}, \dots W_2, \dots) \quad (3)$$

Putting these all together 1, 2, 3 gives us.

$$\boxed{\begin{array}{l} (2(\mu(\theta))X^T - Y^T)XW^T_{<L} \cdots \\ 2W^T_{>j}(\mu(\theta)X^T - Y^T)X \\ W^T_{<j} \cdots \quad 2W^T(\mu(\theta)X^T - Y^T) \\ X) \end{array}}$$

5) 1) $E[e^{\lambda X}] = e^{\sigma^2 \lambda^2 / 2}$

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

$M_X(t) = E[e^{tX}]$

PDF = Since Normal Gaussian, then

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

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

2) $P(X \geq t) \leq \frac{E[X]}{t}$

$X = e^{\lambda X}$ $\quad t = e^{\lambda t}$

$$\frac{E[e^{\lambda X}]}{e^{\lambda t}}$$

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

$$\lambda = \frac{t}{\sigma^2}$$

$$P(X \geq t) \leq \frac{\exp\left(\frac{\sigma^2\left(\frac{t^2}{\sigma^4}\right)}{2}\right)}{\exp\left(\frac{t}{\sigma^2}t\right)}$$

$$\frac{\exp\left(\frac{t^2}{2\sigma^2}\right)}{\exp\left(\frac{t^2}{\sigma^2}\right)}$$

$$\frac{t^2}{2\sigma^2} - \frac{2t^2}{2\sigma^2} = \frac{-t^2}{2\sigma^2}$$

$$\boxed{e^{\frac{-t^2}{2\sigma^2}}}$$

3) As $n \to \infty$ the concentration inequality will go closer and closer to the average

4) Yes U_x and V_x are both independant because of X being i.i.d. However, if X becomes independant, but not identically distributed, then U_x and V_x become dependant as now the values in X are dependant on some i, when before, they were not, it was just $N(0,1)$.

6) 1)

$$\int_{R^d} x f(x; \mu, \Sigma_1)\, dx$$

$$z = \Sigma_1^{-1/2}(x-\mu)$$

$$\Sigma_1^{1/2} z = x - \mu$$

$$\Sigma_1^{1/2} z + \mu = x$$

$$\int_{R^d} (\Sigma_1^{1/2} z + \mu) f(\Sigma_1^{1/2} z + \mu; \mu, \Sigma_1)$$

$$f(\Sigma_1^{1/2} z + \mu; \mu, \Sigma_1) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_1|}} \exp\left(-\frac{1}{2}(\Sigma_1^{1/2} z + \cancel{\mu - \mu})^T \Sigma_1^{-1} (\Sigma_1^{1/2} z + \cancel{\mu - \mu}\right)$$

$$\frac{1}{\sqrt{2(\pi)^d |\Sigma_1|}} \cdot \exp\left[-\frac{1}{2}(\Sigma_1^{1/2} z)^T \Sigma_1^{-1} (\Sigma_1^{1/2} z)\right]$$

$$\text{"} \cdot \exp\left[-\frac{1}{2}\left(z^T \underbrace{(\Sigma_1^{1/2})^{T} \Sigma_1^{-1} \Sigma_1^{1/2}}_{I} z\right)\right]$$

$$\text{"} \cdot \exp\left[-\frac{1}{2}(z^T z)\right]$$

$$\int_{R_d} (\quad) dz$$

$$\underbrace{\int_{R_d} \Sigma_1^{1/2} z (\quad) dz}_{\Sigma_1^{1/2} \cdot 0} + \underbrace{\int_{R_d} \mu (\quad) dz}_{\mu \cdot 1}$$

$$\mu$$

6) 2) $Z = \Sigma_1^{1/2}(X-\mu)$

$Var(Z) = E[ZZ^T] - E[Z]E[Z]^T$

For $E[ZZ^T]$:

$$E[ZZ^T] = E[\Sigma_1^{1/2}(X-\mu)(\Sigma_1^{1/2}(X-\mu))^T]$$

$$\Sigma_1^{1/2} E[(X-\mu)(X-\mu)\Sigma_1^{1/2}]$$

$$\Sigma_1^{1/2} E[(X-\mu)(X-\mu)]\Sigma_1^{1/2}$$

$$\Sigma_1^{1/2}\Sigma_1\Sigma_1^{1/2} = \Sigma_1$$

<u>For $E[Z]E[Z^T]$:</u>

$$E[Z]E[Z^T]$$

$$E[\Sigma_1^{1/2}(X-\mu)] \cdot E[(X-\mu)\Sigma_1^{1/2}]$$

$$\Sigma_1^{1/2} E[X-\mu] \cdot E[X-\mu]\Sigma_1^{1/2}$$

$$\Sigma_1^{1/2} E[X]-\mu \cdot E[X]-\mu \cdot \Sigma_1^{1/2}$$

$$\Sigma_1^{1/2}(\mu-\mu) \cdot \mu-\mu \cdot \Sigma_1^{1/2} = 0$$

Therefore $\Sigma_1 - 0 = \boxed{\Sigma_1' = Var(x)}$

7) i) Gradient = $Ax-b$

Local min @ Gradient $=0$

$$Ax-b=0$$
$$+b \quad +b$$

$$\cancel{A^{-1}}Ax = A^{-1}b$$
$$\boxed{x^* = A^{-1}b} \leftarrow \text{optimizer}$$

2) Next pt. = curr. pt. − Step · gradient

$$\boxed{x^{k+1} = x^k - 1\cdot(Ax^k-b)}$$

3) $$x^k - (Ax^k-b) = x^{k+1}$$
$$x^k - Ax^k + b$$
$$x^k(I-A)+b$$
$$x^{k+1} - x^* = x^k - (Ax^k+b) - x^*$$
$$x^{k+1} - (A^{-1}b) = x^k - (Ax^k+b) - (A^{-1}b)$$
$$x^{k+1} - x^* = x^k - (Ax^k + Ax^*) - (\cancel{A^{-1}}Ax^*)$$
$$x^k - x^* - (Ax^k + x^*)$$

$$x^{k} - Ax^{k} - x^{*} - x^{*}$$

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

This is expressed as
k+1 and k, but
you could just
do -1 on all
the k exponents
to get the formula
in the question

$Av = \lambda v$

4) $\|Ax\|^2 = (Ax)^T Ax = x^T A^T A x$

$= x^T A^2 x$?

~~only~~
if symmetric

Rayleigh quotient

$x^T A x \leq \lambda_{max}(A) \cdot \|x\|_2$

$= x^T A^2 x$

Yes PSD
means
symmetric

$$\sqrt{x^T A^2 x} \leq \sqrt{\lambda_{max}(A^2) \cdot \|x_2\|^2}$$

$$\boxed{\|Ax\|^2 \leq \lambda_{max}(A) \cdot \|x_2\|}$$

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

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

6) k should be $\boxed{\dfrac{\|x^k - x^*\|_2}{\rho \epsilon}}$