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 CS159 discord for help usually cuban stock- 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 students solutions. I have given credit to all external sources I consulted."

X _____

1) 1) $1\{a \geq b\} = \begin{cases} 1 & \text{if } a \geq b \\ 0 & \text{if } a < b \end{cases}$ ← 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: $E[\frac{X}{t}] = \frac{1}{t} E[X]$ 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) $x^T \Sigma x \geq 0$ ← A definition of PSD (1)

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

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

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) $0.3 \cdot 0.4 = 0.12$ 12%

ii) $0.4 \cdot 0.7 = 0.28$ = 28%

iii) $0.28 \cdot 0.7 = 0.2016$ = 20%

iv) 0.6 = 60%

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

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

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

3) is a)

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

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

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

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

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

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

b)

Upper Bound:

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 Bound:

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's, 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) $\det(M) \neq 0$ 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 $\text{rank}(A^TA)$ is less than $\text{rank}(A)$, 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 then 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 spaces. Specifically for the rowspace sin R^n is the $n \times 1$ matrix.
$A^TA\,R^n$ is the same as mentioned before but instead it applies to the row and column space

2) $A \to B$:

$Ax = \lambda x$

$x^T Ax \geq 0$

$x^T \lambda x \geq 0$ ← λ is just a number

$\lambda x^T x \geq 0$ ← $x^T x \geq 0$ 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→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) $\langle A, xy^T \rangle$

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

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

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

Note on dimensions: $1 \times n \quad n \times n \quad n \times 1$

$1 \times n \quad n \times 1 = 1 \times 1$

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

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

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

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

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

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

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 $\max_{v \in \mathbb{R}^n, \text{mean} \|v\|_2 = 1} v^T A v$

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

b) $\max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$

3) a) $\sum_{i=1}^{n} \frac{\partial \alpha}{\partial B_i} \gamma_i \ln B_i$

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

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

c) $K \times m \quad m \times n \quad K \times n$

$$\boxed{\frac{d_2}{d\gamma} \frac{d\gamma}{dx}}$$

d) $\nabla_x Y^T z + Y^T \nabla_y z$

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

$(X W^T - Y)^2$ ← without summation

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

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

$= \sum_{j=1}^{L} \omega_{>j} \Delta W_{ij} W_{<j}$

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

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

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

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

$$W = \mu(W_L, W_{L-1}, \dots W_{2,1}) \quad (3)$$

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

$$\boxed{(2(\mu(\theta))X^T - Y^T)XW^T_{<L} \cdots 2W^T_{2j}(\mu(\theta)X^T - Y^T)X W^T_{<j} \cdots 2W^T(\mu(\theta)X^T - Y^T)X)}$$

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

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

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

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

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

$$P(X \geq t) \leq \frac{\exp\left(\frac{\sigma^2\left(\frac{t^2}{\sigma^4}\right)}{2}\right)}{\exp\left(\frac{1}{\sigma^2}t\right)}$$

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

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

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

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)

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

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

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

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

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

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

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

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

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

$$\int_{R_d} (\;) \, dz$$

$$\int_{R_d} \Sigma^{1/2} z (\;) dz + \int_{R_d} \mu (\;) dz$$

$\Sigma^{1/2} \cdot 0$ $\qquad$ $\mu \cdot 1$

μ

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

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

For $E[ZZ^T]$:

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

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

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

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

For $E[Z]\,E[Z^T]$:

$$E[Z]\,E[Z^T]$$

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

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

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

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

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

7) 1) Gradient = $Ax - b$

Local min @ Gradient = 0

$Ax - b = 0$

$+b \quad +b$

$A^{-1}Ax = A^{-1}b$

$\boxed{x^* = A^{-1}b}$ ← optimizer

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

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

3) $x^k - (Ax^k - b) = x^{k+1}$

$x^k - Ax^k + b$

$x^k(I - A) + b$

$x^{k+1} - x^* = x^k - (Ax^k + b) - x^*$

$x^{k+1} - (A^{-1}b) = x^k - (Ax^k + b) - (A^{-1}b)$

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

$x^k - x^* - (Ax^k + x^*)$

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

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

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

$= x^T A^2 x$? only if symmetric

Rayleigh quotient

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

$= x^T A^2 x$

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

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

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

References:

- https://math.stackexchange.com/questions/326784/prove-that-the-rank-of-a-block-diagonal-matrix-equals-the-sum-of-the-ranks-of-th
- https://math.stackexchange.com/questions/349738/prove-operatornamerankata-operatornameranka-for-any-a-in-m-m-times-n
- https://math.stackexchange.com/questions/456025/what-is-the-relationship-between-the-null-space-and-the-column-space
- https://en.wikipedia.org/wiki/Rank_(linear_algebra)
- https://math.stackexchange.com/questions/29072/how-is-the-column-space-of-a-matrix-a-orthogonal-to-its-nullspace
- https://en.wikipedia.org/wiki/Frobenius_inner_product
- https://math.stackexchange.com/questions/3516928/why-does-xtx-x2
- https://en.wikipedia.org/wiki/Diagonalizable_matrix
- https://ocw.mit.edu/courses/res-18-011-algebra-i-student-notes-fall-2021/mit18_701f21_lect12.pdf
- https://www.khanacademy.org/math/linear-algebra/alternate-bases/orthonormal-basis/v/lin-alg-orthogonal-matrices-preserve-angles-and-lengths
- https://math.stackexchange.com/questions/2958603/if-a-pdpt-does-p-have-to-be-orthogonal
- https://math.libretexts.org/Bookshelves/Linear_Algebra/Fundamentals_of_Matrix_Algebra_(Hartman)/03%3A_Operations_on_Matrices/3.01%3A_The_Matrix_Transpose
- https://en.wikipedia.org/wiki/Trace_(linear_algebra)
- https://en.wikipedia.org/wiki/Definite_matrix
- https://en.wikipedia.org/wiki/Cauchy%E2%80%93Schwarz_inequality
- https://en.wikipedia.org/wiki/Singular_value_decomposition
- https://math.stackexchange.com/questions/213653/show-that-norm-of-matrix-a-is-given-by-the-square-root-of-the-largest-eigenval
- https://en.wikipedia.org/wiki/Singular_value
- https://en.wikipedia.org/wiki/Markov%27s_inequality

- https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf
- https://en.wikipedia.org/wiki/Concentration_inequality
- https://en.wikipedia.org/wiki/Estimator
- https://en.wikipedia.org/wiki/Jensen%27s_inequality
- https://math.stackexchange.com/questions/3148049/reverse-of-jensens-inequality
- https://www.geeksforgeeks.org/covariance-matrix/
- https://en.wikipedia.org/wiki/Covariance
- https://math.stackexchange.com/questions/1524424/prove-that-the-average-of-iid-gaussian-random-variables-is-gaussian
- https://en.wikipedia.org/wiki/Rank_(linear_algebra)
- https://byjus.com/jee/rank-of-a-matrix-and-special-matrices/
- https://www2.math.upenn.edu/~moose/240S2013/slides7-22.pdf
- https://math.stackexchange.com/questions/64350/what-is-the-difference-between-the-rowspace-and-the-columnspace-in-linear-algebr
- https://en.wikipedia.org/wiki/Row_and_column_spaces
- https://stattrek.com/matrix-algebra/elementary-operations
- https://www.khanacademy.org/math/linear-algebra/vectors-and-spaces/null-column-space/v/introduction-to-the-null-space-of-a-matrix
- https://en.wikipedia.org/wiki/Rank%E2%80%93nullity_theorem
- https://www.algebrapracticeproblems.com/how-to-diagonalize-a-matrix-diagonalizable-diagonalization/
- https://en.wikipedia.org/wiki/Moment-generating_function#:~:text=1%20For%20a%20discrete%20probability%20mass%20function%2CM%20X,%28t%29%3Dint%20_%20%7B-infty%20%7D%5E%20%7Binfty%20%7De%5E...%20More%20
- https://ocw.mit.edu/courses/18-600-probability-and-random-variables-fall-2019/d68c0ee3be77495da2aeaf921f7fc081_MIT18_600F19_lec26.pdf
- https://www.scribbr.com/methodology/independent-and-dependent-variables/

- https://brilliant.org/wiki/expected-value/
- https://en.wikipedia.org/wiki/Matrix_norm
- https://math.stackexchange.com/questions/289989/first-and-second-derivative-of-a-summation
- https://en.wikipedia.org/wiki/Partial_derivative
- https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant
- https://en.wikipedia.org/wiki/Gradient
- https://en.wikipedia.org/wiki/Artificial_neural_network#:~:text=The%20simplest%20kind%20of%20feedforward%20neural%20network%20%28FNN%29,and%20the%20inputs%20is%20calculated%20at%20each%20node.
- https://www.mathworks.com/help/deeplearning/ug/linear-neural-networks.html
- https://en.wikipedia.org/wiki/Design_matrix
- https://builtin.com/machine-learning/cost-function
- https://en.wikipedia.org/wiki/Linear_least_squares
- https://en.wikipedia.org/wiki/Matrix_norm#Frobenius_norm
- https://math.stackexchange.com/questions/1898839/relation-between-frobenius-norm-and-trace
- https://en.wikipedia.org/wiki/Projection_(mathematics)
- https://math.iit.edu/~fass/477577_Chapter_10.pdf
- https://math.libretexts.org/Bookshelves/Calculus/Calculus_(OpenStax)/14%3A_Differentiation_of_Functions_of_Several_Variables/14.05%3A_The_Chain_Rule_for_Multivariable_Functions