CS 189 HW 2

Question 1) Honor Code

1 Honor Code

1. **List all collaborators. If you worked alone, then you must explicitly state so.**
2. **Declare and sign the following statement**:

   "I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

   Signature : _____

   While discussions are encouraged, *everything* in your solution must be your (and only your) creation. Furthermore, all external material (i.e., *anything* outside lectures and assigned readings, including figures and pictures) should be cited properly. We wish to remind you that the consequences of academic misconduct are *particularly severe*!

Question 2 : Probability Potpourri

1) (+) Semidefinite : $n \times n$ symmetric matrix where all of its eigenvalues are non-negative (greater or equal to 0)

↳ $z^T A z \geq 0$; any vector x

$\Sigma_{ij} = \text{cov}(z_i, z_j)$

↳ det semidefinite!

$z^T \cdot \Sigma \cdot z \quad ; \quad \underbrace{z \in \mathbb{R}^n}_{z \text{ in real numbers}}$

↓

$z^T \cdot E[(z-N)(z-N)^T] \cdot z \; 0$

↳ $E[\underbrace{z^T \cdot (z-N)}\underbrace{(z-N)^T \cdot z}]$; Linearity of Expectation

↳ $E[((z-N) \cdot z)^2] \geq 0$

* Since the values are being squared, the expectation is that all values will at least be ≥ 0.

2) X: Archer hits target wind $P(X) = \frac{4}{10}$

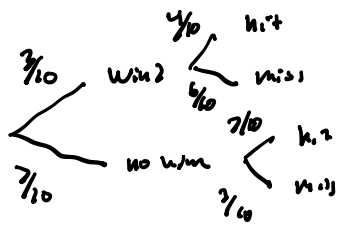
Y: Archer hits target not wind $P(Y) = \frac{7}{10}$

Z: Gust of wind $P(Z) = \frac{3}{10}$

i) $P(Z \cap X) = \frac{3}{10} \cdot \frac{4}{10} = \frac{12}{100}$

ii) $P(\text{1st shot hit}) =$

↳ $= \frac{3}{10}\left(\frac{4}{10}\right) + \frac{7}{10}\left(\frac{7}{10}\right)$



$P(\text{1st shot win}) = \frac{61}{100}$

iii) P(Hits exactly 1 time in 2 trials)

Binomial Distribution: k successes in n trials

$= \binom{2}{1} P(\text{Hit})^1 \, P(\text{Miss})^1$

$= 2 \cdot \left[\frac{3}{10}\left(\frac{4}{10}\right) + \frac{7}{10}\left(\frac{7}{10}\right) \right] \cdot \left[\frac{3}{10}\left(\frac{6}{10}\right) + \frac{7}{10}\left(\frac{3}{10}\right) \right]$

$= 2 \cdot \frac{61}{100} \cdot \frac{39}{100}$

$= 0.4758$

↳ P(Hits exactly 1 time in 2 trials) = 0.4758

iv) $P(\text{no wind} \mid \text{miss}) = \dfrac{P(\text{miss} \mid \text{no wind}) \, P(\text{no wind})}{P(\text{miss})}$

$P(\text{no wind} \mid \text{miss}) = \dfrac{\frac{3}{10} \cdot \frac{7}{10}}{1 - \frac{61}{100}} = \frac{21}{100} \times \frac{100}{39}$

$P(\text{no wind} \mid \text{miss}) = \frac{7}{13}$

3) $E[X] = \int x \cdot f(x) \, dx \quad ; x \geq 0$

Caveat: Our area is restricted by radii $\frac{1}{\sqrt{3}}, 1, \sqrt{2}$ ft

Split expression into

$E[x] = 4 \cdot P\left(x \leq \frac{1}{\sqrt{3}}\right) + 3 \cdot P\left(\frac{1}{\sqrt{3}} < x \leq 1\right) + 2 \cdot P\left(1 < x \leq \sqrt{3}\right)$

$$P\left(x \le \frac{1}{\sqrt{3}}\right) = \int_0^{1/\sqrt{3}} f(x)\,dx \quad ; f(x) = \frac{2}{\pi(1+x^2)}$$

$$= \frac{2}{\pi}\left[\arctan(x)\right]_0^{1/\sqrt{3}} \quad ; \frac{d}{dx}\arctan = \frac{1}{1+x^2}$$

$$P\left(\frac{1}{\sqrt{3}} < x \le 1\right) = \int_{1/\sqrt{3}}^{1} f(x)\,dx$$

$$= \frac{2}{\pi}\left[\arctan(x)\right]_{1/\sqrt{3}}^{1} \quad ; \frac{d}{dx}\arctan = \frac{1}{1+x^2}$$

$$P\left(1 < x \le \sqrt{3}\right) = \int_1^{\sqrt{3}} f(x)\,dx \quad ; f(x) = \frac{2}{\pi(1+x^2)}$$

$$= \frac{2}{\pi}\left[\arctan(x)\right]_1^{\sqrt{3}}$$

$$E[x] = 4 \cdot P\left(x \le \frac{1}{\sqrt{3}}\right) + 3 \cdot P\left(\frac{1}{\sqrt{3}} < x \le 1\right) + 2 \cdot P\left(1 < x \le \sqrt{3}\right)$$

$$= \frac{2}{\pi}\left[4\left(\arctan\left(\frac{1}{\sqrt{3}}\right) - \arctan(0)\right) + 3\left(\arctan(1) - \arctan\left(\frac{1}{\sqrt{3}}\right)\right) + 2\left(\arctan(\sqrt{3}) - \arctan(1)\right)\right]$$

$$= 13/6$$

$\therefore$ $E[x] = 13/6$

4)

$$\text{entropy}(x) = H(X) = -\sum_{x \in Z} p(x) \ln(p(x))$$

i) $X \sim$ Bernoulli Distribution (p) $\begin{cases} P(X=1) = p \\ P(X=0) = 1-p \end{cases}$

$$H(X) = -\left[p(x) \ln(p(x)) + (1-p(x)) \ln(1-p(x)) \right]$$

$$H'(x) = -\left[\ln(p(x)) \cdot 1 + p(x) \cdot \frac{1}{p(x)} - \ln(1-p(x)) - \frac{1-p(x)}{1-p(x)} \right]$$

$$= -\left[\ln(p(x)) + 1 - \ln(1-p(x)) - 1 \right]$$

$$= -\ln(p(x)) + \ln(1-p(x))$$

$$H''(x) = -\frac{1}{p(x)} + \frac{-1}{1-p(x)}$$

$$= -\left(\frac{1}{p(x)} + \frac{1}{1-p(x)} \right)$$

Since $p(x) \geq 0$, this value is always positive

$H''(x) \leq 0$

$H(x)$ concave in p if $-H(x)$ convex function of p

ii) Sample space Σ w/ n states:

Discrete Dist

Rand var X

→ Over all possible PMF's, entropy $H(X)$ maxed by uniform dist.

$$H(X) = -\sum_{i=1}^{n} p_i \ln(p_i) \quad ; \text{ all } p \text{ equal } \quad p_i = \frac{1}{n} \text{ for all } i$$

Lagrange Multipliers:

Strategy finding local maxima/minima subject to equation constraints

(constraint $\sum_{i=1}^{n} p_i = 1$

↳ $$L = -\sum_{i=1}^{n} p_i \ln(p_i) + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right)$$

$$\frac{\partial L}{\partial p_i} = -(\ln(p_i) + 1) + \lambda = 0$$

↳ $\ln(p_i) = \lambda - 1 \rightarrow p_i = e^{\lambda - 1}$

Plug into prob function

↳ $\sum_{i=1}^{n} e^{\lambda - 1} = p_i = 1$; identical terms

$$n e^{\lambda - 1} = 1$$

$e^{\lambda - 1} = p_i = \frac{1}{n}$; for all i

↳ $H(X) = -\sum_{i=1}^{n} \frac{1}{n} \ln\left(\frac{1}{n}\right) = n\left(\frac{1}{n}\right) \ln\left(\frac{1}{n}\right)$

$= \ln(n)$

↳ $H(X) = \ln(n)$ max entropy

Question 3: Linear Algebra Review

1) a)

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

3 elementary Row ops:

1) interchange two rows / columns
2) Multiplication of row/column by number nonzero
3) Multiplication of row/column by nonzero number and add result to other row/column

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

Left multiply 1st row by A and add to 2nd row

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

right multiply 1st column by B and subtract from 2nd column

Swap column 1 with column 2

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

b) Find Upper/Lower Bounds Rank (AB)

Prove $\text{rank}(A) + \text{rank}(B) - n \leq \text{rank}(AB)$

def Rank Matrix: dimension of vector space generated by its columns

maximal number of Linearly independent columns of A

initial matrix rank = final matrix rank

because elementary row operations don't impact matrix rank

$$\text{rank}\left(\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}\right)$$

prove rank A + rank B − n ≤ rank (AB)

$$n + \text{rank}(AB) = \text{rank}\left(\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}\right) \geq \text{rank } A + \text{rank } B$$

↳ Since AB results from A · B, col space (AB) ≤ intersection col space A col space B

Thus:

rank (AB) ≤ min (rank A, rank B)

c) rank (A) = n det (M) ≠ 0

↳ col space (A) spanned n linearly independent columns implying that the dimensions of col space are at least n

↳ row space (A) consists of n independent row vectors implying that its dimension is also at least n

* Since rank of matrix is number of linearly independent rows / col, this confirms standard facts

d) Relationship between col space A & null space (A^T)

Establish that rank (A^TA) = rank (A)

↳ rank (A^TA) ≤ min (rank (A^T), rank (A))

Since rank (A^T) = rank (A)

↳ rank (A^TA) ≤ rank (A)

Fundamental theorem Linear algebra:

Null space (A^T) ⊥ column space (A)

Suppose rank (A^TA) < rank (A)

↳ Implies: $A^TAv = 0$; v non zero vector

↳ $Av = 0$ → contradicts Fundamental theorem Linear algebra

Thus: rank (A^TA) = rank (A)

c) $A\mathbb{R}^n$ and $A^TA\mathbb{R}^n$

↳ $\mathbb{R}^n$ = input space of A

Applying A to all $\mathbb{R}^n$ results in columns space of A

$A\mathbb{R}^n$ = Col(A)

output of A spans entire range of A

↳ transformation of entire span by A^TA

$A^TA\mathbb{R}^n$ = Row space (A)

A^TA symmetric matrix that projects vectors onto the row space of A preserves dimensions

2) $A \in \mathbb{R}^{n \times n}$ symmetric matrix

a) $\forall\ x \in \mathbb{R}^n,\ x^TAx \geq 0$

$v^TAv = \lambda v^Tv = \lambda \|v\|^2$; λ eigenvalue of A corresponding eigenvector v

Thus fulfills the conditions

b) From part a → $\lambda \|v\|^2 \geq 0$

Thus $\lambda \geq 0$

c) Eigen decomposition of A: $A = V \Lambda V^T$; Λ diagonal matrix entries of eigenvalues of A

↳ $U = V\sqrt{\Lambda}$

↳ $A = UU^T$

3) $Q_A(x) = x^T A x$; $A = I$ and $x \in \mathbb{R}^2$

Introduction knowledge:

- $A = I$; $x^T A x = 1$ represents circle or sphere higher dim
- A diagonal w/ diff vals, eq represents an ellipse
- A multi 1, line or flat degenerate shape

a) $A = 2I$ $x \in \mathbb{R}^3$, $Q_A(x) = 32$

↳ $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ $\xrightarrow{x^T A x}$ $2(x_1^2 + x_2^2 + x_3^2) = 32$

* Sphere at origin with radius $\sqrt{16} = 4$

b) $A = aa^T$; $a \in \mathbb{R}^3$ nonzero vector, $x \in \mathbb{R}^3$, $Q_A(x) = 25$

↳ $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}$

↳ $x^T A x = (x^T a)(a^T x) = (a^T x)^2$

↳ $(a^T x)^2 = 25$

$a^T x = \pm 5$

* 2 parallel planes in $\mathbb{R}^3$ perpendicular to a normal is / unit away from the origin

c) $A = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}$ $x \in \mathbb{R}^2$ $Q_A(x) = 24$

↳ $\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 5x_1 + x_2 \\ 1x_1 + 5x_2 \end{bmatrix} = 5x_1^2 + x_1x_2 + x_1x_2 + 5x_2^2$

$5x_1^2 + 2x_1x_2 + 5x_2^2 = 24$

* tilted ellipse centered at origin

d) $5x_1^2 + 2x_1x_2 + 5x_2^2 = 0$

* Only solution is the origin or $x_1, x_2 = 0$

No geometric shape, simply a point on the origin

4) Frobenius inner product 2 matrices $A, B \in \mathbb{R}^{m \times n}$:

$$\langle A, B \rangle = \text{trace}(A^T B) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

Frobenius norm of matrix:

$$\|A\|_F = \sqrt{\langle A, A \rangle} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}}$$

a) $x^T A y = \langle A, xy^T \rangle \quad \forall x \in \mathbb{R}^m, y \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}$

↳ $\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} \cdot x_i \cdot y_j$

b) Using 2A:

$A = UU^T$ and $B = VV^T$

$$\text{trace}(AB) = \text{trace}(UU^T VV^T)$$

$$= \text{trace}(U^T V (UV^T))$$

$$= \text{trace}((U^T V)^2) \geq 0$$

Since trace $M \geq 0$ since it's sum of all eigenvalues

or simply the trace (AB) is essentially the trace of a PSD matrix

c) PSD matrix using $\lambda_{max}(A)$

↳ $\lambda_{max}(A) I_n - A \succeq 0$

proved in 4b) trace $(C(\lambda_{max}(A) I_n - A)B) \geq 0$

↳ trace $(AB) \leq \lambda_{max}(A)$ trace (I_n, B)

↳ trace $(I_n, B) \leq \sqrt{n} \|B\|_F$

Question 4: Matrix / Vector Calculus

1) $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A$ Gradient of $\sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) + x^T A y$

↳ To find the gradient, calculate the derivative:

$$\frac{d}{dA_{11}} \sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) + x^T A y$$

$$\frac{d}{dA_{11}} \sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) = \cos\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) \cdot \left(2A_{11} + e^{A_{11}+A_{22}}\right)$$

$$\frac{d}{dA_{22}} \sin\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) = \cos\left(e^{A_{11}+A_{22}}\right) \cdot \left(e^{A_{11}+A_{22}}\right)$$

$$\downarrow$$

$$\begin{bmatrix} \cos\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) \cdot \left(2A_{11} + e^{A_{11}+A_{22}}\right) & 0 \\ 0 & \cos\left(e^{A_{11}+A_{22}}\right) \cdot \left(e^{A_{11}+A_{22}}\right) \end{bmatrix}$$

$$\frac{d}{dA} x^T A y = \frac{d}{dA} \langle A, x^T y \rangle$$

$$= \frac{d}{dA} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} \cdot x_i^T \cdot y_j$$

$$= \frac{d}{dA} x_1 (A_{11} y_1 + A_{12} y_2) + x_2 (A_{21} y_1 + A_{22} y_2)$$

$$\frac{d}{dA_{11}} (x^T A y) = x_1 y_1 \qquad \frac{d}{dA_{12}} (x^T A y) = x_1 y_2$$

$$\frac{d}{dA_{21}} (x^T A y) = x_2 y_1 \qquad \frac{d}{dA_{22}} (x^T A y) = x_2 y_2$$

$$\nabla_A f(A) =$$

$$\begin{bmatrix} \cos\left(A_{11}^2 + e^{A_{11}+A_{22}}\right) \cdot \left(2A_{11} + e^{A_{11}+A_{22}}\right) + x_1 y_1 & x_1 y_2 \\ x_2 y_1 & \cos\left(e^{A_{11}+A_{22}}\right) \cdot \left(e^{A_{11}+A_{22}}\right) + x_2 y_2 \end{bmatrix}$$

2) a) $\alpha = \sum_{i=1}^{n} \gamma_i \ln B_i \quad ; \quad \gamma, B \in \mathbb{R}^n$

$$\frac{d\alpha}{dB_i} = \frac{\gamma_i}{B_i}$$

b) $\gamma = Ap + b \quad ; \quad b \in \mathbb{R}^n \quad p \in \mathbb{R}^m \quad A \in \mathbb{R}^{n \times m}$

$$\frac{d\gamma_i}{dp_j} = A_{ij}$$

c) $$\left(\frac{dz}{dx}\right)_{ij} = \frac{dz_i}{dx_j} = \sum_{l=1}^{m} \frac{dz_i}{dy_l} \frac{dy_l}{dx_j}$$

$$\text{i.e.} \quad \frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

$\frac{dz}{dx}$ matrix product of other 2 Jacobians

d) $$\frac{d\, y^T z}{dx_i} = \sum_{j=1}^{m} \frac{dy_j}{dx_i} z_j + \frac{dz_j}{dx_i} y_j$$

$$\nabla_x y^T z = \frac{dy}{dx}^T z + \frac{dz}{dx}^T y$$

3) Prove $f(x) - f(x^*) \leq \frac{\rho}{2}$

$$f(x) = f(x^*) + \nabla f(x^*)^T (x - x^*) + \frac{1}{2} (x - x^*)^T H(x') (x - x^*)$$

Taylor series, for some x' between x^* and x

x^* global min $\rightarrow \nabla f(x^*) = 0$!

$$f(x) - f(x^*) = \frac{1}{2} (x - x^*)^T H(x') (x - x^*)$$

$H(x')$ positive semidefinite so max eigenvalue is 1:

$$\lambda_{max}(H(x')) \leq 1$$

Quel fun fact :

$$(x - x^*)^T H(x') (x - x^*) \leq \lambda_{max}(H(x')) \| x - x^* \|^2$$

$$\text{i.e.} \quad f(x) - f(x^*) \leq \frac{1}{2} \| x - x^* \|^2$$

x belongs in spherical region

$$\mathcal{X} = \{x \mid \|x - x^*\|^2 \leq r\}$$

↳ $f(x_2) - f(x^*) \leq \frac{1}{2} D$

thus:

$$f(x) - f(x^*) \leq \frac{D}{2} \qquad \forall x \in \mathcal{X}$$

Question 6: Properties of Normal Dist

1) Prove: $E[e^{\lambda x}] = e^{\sigma^2 \lambda^2 / 2}$; Hint MGF

↳ $$E[e^{\lambda x}] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\lambda x} e^{-x^2/2\sigma^2} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\lambda x \sigma} e^{-z^2/2} dz$$

$$= e^{\sigma^2\lambda^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-\lambda\sigma)^2/2} dz$$

$$= e^{\sigma^2\lambda^2/2}$$

2)

Gaussian RV are independent if not correlated:

$$E[u_x v_x] = E\left(\left[\sum_{i=1}^{n} u_i x_i\right]\left[\sum_{i=1}^{n} v_i x_i\right]\right)$$

$$= \sum_{i=1}^{n} u_i v_i E[x_i^2]$$

↳ $\langle u, v \rangle = 0$

$$= 0$$

* U_x and V_x are independent. However, if RV are not iid, this can cause problems with the correlation value. We are only able to derive cov = 0 due to iid given. Thus, w/o this fact, we cannot say that U_x and V_x are independent

Question 7: Multivariate Normal Distribution

1) $E[X] = \mu$

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

$$E[X] = \int_{\mathbb{R}^d} x \cdot f(x)\, dx$$

Change of variables

let z: $\Sigma^{-1/2}(x-\mu)$; standardizes x so $Z \sim N(0, I_d)$

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

↳ $f_z(z) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2} z^T z}$

↳ $E[X] = \int_{\mathbb{R}^d} (\mu + \Sigma^{1/2}) f_z(z)\, dz$

$$= \mu + \Sigma^{1/2} \int_{\mathbb{R}^d} z\, f_z(z)\, dz$$

$= 0$ w/ symmetry of standard multivariate normal dist. $E[Z] = 0$

$$E[X] = \mu$$

2) $\text{Var}(X) = \Sigma$

$\text{Var}(X) = \text{cov}(X, X) = E[(X-\mu)(X-\mu)^T] = E[XX^T] - \mu\mu^T$

$X - \mu = \Sigma^{1/2} Z$; Change of variable : $X - \mu = \Sigma^{1/2} Z$

$\text{Var}(X) = E[(\Sigma^{1/2} Z)(\Sigma^{1/2} Z)^T]$

$= E[(\Sigma^{1/2} Z Z^T \Sigma^{1/2})]$; $Z \sim N(0, I_d)$

$\hookrightarrow E[ZZ^T] = I_d$

$= \Sigma^{1/2} E[ZZ^T] \Sigma^{1/2}$

$= \Sigma^{1/2} I_d \Sigma^{1/2}$

$= \Sigma$

$\hookrightarrow \text{Var}(X) = \Sigma$

Question 8: Gradient Descent

optimization problem: $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T A x - b^T x$; $A \in \mathbb{R}^{n \times n}$ PSD w/ $0 < \lambda_{min}(A) < \lambda_{max}(A) < 1$

1) Objective is convex $\rightsquigarrow$ optimizer stationary point of objective:

$$Ax - b = 0 \quad ; \text{ A invertible}$$

↳ optimizer: $x^* = A^{-1} b$

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

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

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

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

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

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

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

↳ $\|Ax\|_2 = x^T A^2 x \leq \lambda_{max}(A)^2$

If $x \neq 0$, then:

$$\| A (x / \|x\|_2) \|_2^2 \leq (\lambda_{max}(A))^2$$

which proves $\|Ax\|_2 \leq \lambda_{max}(A) \|x\|_2$ after scaling manipulation

5) $\| x^{(k)} - x^* \|_2 \leq \rho \| x^{k-1} - x^* \|_2$

since $\lambda_{max}(A) < 1$; $I - A > 0$:

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

$\rho = \lambda_{max}(I-A) = 1 - \lambda_{min}(A)$:

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

6) recursion proceeds:

$$\| x^k - x^* \|_2 \leq \rho^k \| x^0 - x^* \|_2$$

conditioning on $\rho^k \| x^0 - x^* \|_2 \leq \varepsilon$

$$\hookrightarrow \log(\rho^k) \cdot \log(\| x^k - x^* \|_2) \leq \log(\varepsilon)$$

$$k \log(\rho) \cdot \log(\| x^k - x^* \|_2) \leq \log(\varepsilon)$$

$$k \log(\rho) \leq \log\left(\frac{\| x^k - x^* \|_2}{\varepsilon} \right)$$

$$k \leq \frac{1}{\log(\rho)} \log\left(\frac{\| x^k - x^* \|_2}{\varepsilon} \right)$$

Question 5: Linear Neural Networks

Apply multivariate chain rule to simple Neural Net → Linear Neural Network

- Only vectorized Linear regression / Decision func

- $n \times d$ design matrix X
  - Each row of X training point
    - $\hookrightarrow$ n training points w/ d features

- $n \times k$ matrix Y
  - Each row Y set of k labels for each X train point

Goal: Learn $k \times d$ matrix W of weights such that:

$$Y \approx XW^T$$

Find W minimizes cost func:

$$RSS(W) = \|XW^T - Y\|_F^2$$

Instead of optimizing W over $k \times d$ space, write W we seek as product of multiple matrices

Linear Neural Net:

$$W = \nu(W_L, W_{L-1} \dots W_1) = W_L W_{L-1} \dots W_1$$

- ν: matrix multiplication map
- W_1: real matrix $d_1 \times d_0$, matrix

weight vector: $\theta = (W_L \dots W_1) \in \mathbb{R}^d$

$$\nu(\theta) = \nu(W_L \dots W_1)$$

- θ: column vector

Goal: Find θ that min RSS

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

Using Gradient Descent

1) $G = \nabla_W RSS(W)$ $k \times 2$ matrix

↳ $RSS(W) = \sum_{i=1}^{n} \sum_{j=1}^{k} \left((xw^T)_{ij} - y_{ij} \right)^2$; $\|A\|_F^2 = \sum_{ij} A_{ij}^2$

$(xw)^T_{ij} = \sum_{m=1}^{2} x_{im} w_{jm}$

$RSS(W) = \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\sum_{m=1}^{2} x_{im} w_{jm} - y_{ij} \right)^2$

$G_{pq} = \frac{\partial RSS(W)}{\partial w_{pq}}$

↳ $\frac{\partial}{\partial w_{pq}} RSS(W) = 2 \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\sum_{m=1}^{d} x_{im} w_{jm} - y_{ij} \right) \cdot \frac{\partial}{\partial w_{pq}} \sum_{m=1}^{d} x_{im} w_{jm}$; $j=p$ $m=q$ for w_{jm}

↳ $\frac{\partial}{\partial w_{pq}} \sum_{m=1}^{2} x_{im} w_{jm} = x_{iq} \delta_{jp}$

↳ $G_{pq} = 2 \sum_{i=1}^{n} (xw^T - y)_{ip} x_{ip}$ Summation form

$R = xw^T - y$

$G = 2R^T x$

↳ $G = 2(xw^T - y)^T x$

$G = 2(wx^T x - y^T x)$ Gradient form

2) Directional Derivative of $RSS(W)$

$RSS'_{\Delta w}(W) = \lim_{\varepsilon \to 0} \frac{RSS(W + \varepsilon \Delta W) - RSS}{\varepsilon}$

First order approximation

$RSS(W + \varepsilon \Delta W) \approx RSS(W) + \varepsilon \nabla_W RSS(W) \cdot \Delta W + O(\varepsilon^2)$

$RSS'_{\Delta w}(W) = \nabla_W RSS(W) \cdot \Delta W$; $\nabla_W RSS(W) = 2(xw^T - y)^T x$

↳ $RSS'_{\Delta w}(W) = 2(xw^T - y)^T x \cdot \Delta W$

$$\langle A, B \rangle_F = \sum_{i,j} A_{ij} B_{ij}$$

$$= Tr(A^T B)$$

$$RSS'_{\Delta w}(w) = 2\, Tr\left((xw^T - y)^T x \Delta w \right)$$

$$= 2\, Tr\left((xw^T - y)^T x \Delta w \right)$$

$$\langle A, B \rangle_F = Tr(A^T B)$$

$$\hookrightarrow \boxed{RSS'_{\Delta w}(w) = 2 \langle xw^T - y, x \Delta w^T \rangle_F}$$

3)

$$w = p(w_L, w_{L-1}, \dots, w_1)$$

$$= w_L \dots w_1$$

$$N'_{\Delta\theta}(\theta) \quad ; \quad \theta = (w_L \dots w_1) \quad \Delta\theta = (\Delta w_L, \Delta w_{L-1}, \dots, \Delta w_1)$$

$$N'_{\Delta\theta}(\theta) = \sum_{j=1}^{L} w_j^{>} \Delta w_j w_j^{<} \quad ; \quad w_j^{>} = w_L w_{L-1} \dots w_{j+1}$$

$$w_j^{<} = w_{j-1} w_{j-2} \dots w_1$$

$$w_L^{>} = I_{d_L} \text{ (identity matrix of size } d_L \times d_L)$$

$$w_1^{<} = I_{d_0} \text{ (identity matrix of size } d_0 \times d_0)$$

$$RSS'_{\Delta w}(w) = \lim_{\varepsilon \to 0} \frac{RSS(w + \varepsilon \Delta w) - RSS}{\varepsilon} \quad ; \quad N(\theta) = w_L w_{L-1} \dots w_1$$

$$N'_{\Delta\theta}(\theta) = \sum_{j=1}^{L} w_L \dots w_{j+1} \Delta w_j w_{j-1} w_{j-2} \dots w_1$$

$$w_j^{>} = w_L \dots w_{j+1} \qquad w_j^{<} = w_{j-1} \dots w_1$$

$$\hookrightarrow \boxed{N'_{\Delta\theta}(\theta) = \sum_{j=1}^{L} w_j^{>} \Delta w_j w_j^{<}}$$

4.1 Multivariate Chain Rule $J'_{\Delta\theta}(\theta)$

↳ $J'_{\Delta\theta}(\theta) = R(w)'_{\Delta w}(w) \cdot w'_{\Delta\theta}(\theta)$

Multivariate chain rule

$$\frac{\partial}{\partial x} f(g(x)) = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$$

↳ Q2: $R(J)'_{\Delta w}(w) = 2\langle xw^T - y, x\Delta w^T \rangle_F$

↳ Q3: $w'_{\Delta\theta}(\theta) = \sum_{j=1}^{L} w_j^{>} \Delta w_j w_j^{<}$

↳ Chain Rule: $J'_{\Delta\theta}(\theta) = R(J)'_{w'_{\Delta\theta}(\theta)}(w)$

$$J'_{\Delta\theta}(\theta) = 2\langle xw^T - y, x\left(\sum_{j=1}^{L} w_j^{>} \Delta w_j w_j^{<}\right)^T \rangle_F$$

Linearity of inner product!

$$J'_{\Delta\theta}(\theta) = 2\sum_{j=1}^{L} \langle xw^T - y, x(w_j^{<})^T \Delta w_j^T (w_j^{>})^T \rangle_F$$

$$J'_{\Delta\theta}(\theta) = 2\sum_{j=1}^{L} \langle w_j^{>}(xw^T - y)^T x w_j^{<}, \Delta w_j \rangle_F$$

↳ $\nabla_\theta J(\theta) = (2w_1^{>}(xw^T - y)^T x w_L^{<}, \ldots, 2w_j^{>}(xw^T - y)^T x w_j^{<},$

$\ldots 2w_1^{>}(xw^T - y)^T x$

5) $\nabla_\theta S(\theta)$

$$\nabla_\theta S(\theta) = (2w_1^2 (xw^T - Y)^T x w_1^<, \dots, 2w_j^2 (xw^T - Y)^T x w_j^<, \dots$$
$$2w_1^2 (xw^T - Y)^T x)$$

Q2: $RSS'_{\delta w}(w) = 2 \langle xw^T - Y, x\delta w^T \rangle_F$

Q3: $\alpha'_{\delta\theta}(\theta) = \sum_{j=1}^{L} w_j^2 \delta w_j w_j^<$

Q4:

$$\nabla_\theta S(\theta) = (2w_1^2 (xw^T - Y)^T x w_1^<, \dots, 2w_j^2 (xw^T - Y)^T x w_j^<,$$
$$\dots 2w_1^2 (xw^T - Y)^T x$$

$$\nabla_{w_j} S(\theta) = 2w_j^2 (xw^T - Y)^T x w_j^< \quad : w^T = (w_1^T w_2^T \dots w_L^T)$$

$$\nabla_{w_j} S(\theta) = 2w_j^2 (xw^T - Y)^T x w_j^<$$