ECE C147/C247 Winter 2023 Homework #1

Neural Networks & Deep Learning

UCLA ECE

Yaman Yucel , UID = 605704529

1) Linear algebra Refresher

a) Q real orthogonal matrix. => $QQ^T = QQ^{-1} = I$

$Q^TQ = Q^{-1}Q = I$

i) $Q^T = A \Rightarrow A^TA \overset{?}{=} I$

$(Q^T)^T Q^T = QQ^T \overset{\checkmark}{=} I$, Q^T is an orthogonal matrix

Since $Q^T = Q^{-1} \Rightarrow Q^{-1}$ is an orthogonal matrix.

ii) $Qx = \lambda x$

$(Qx)^T Qx = |\lambda|^2 x^Tx$

$\Downarrow$

$x^TQ^TQx = x^Tx$

Then, $x^Tx = \lambda^2 x^Tx$

$\lambda^2 = 1 \Rightarrow \boxed{|\lambda| = 1}$

iii)

$QQ^T = I$

$\Downarrow$

$\det(I) = \det(QQ^T)$

$\det(I) = 1$

$\Rightarrow \det(QQ^T) = 1$

$\det(Q)\det(Q^T) = 1$

$\Rightarrow \det(Q)^2 = 1 \rightarrow \boxed{\det(Q) = \mp 1}$

iv) $\boxed{\|Qx\|^2} = (Qx)^T(Qx) = x^TQ^TQx = x^Tx = \boxed{\|x\|^2}$

Therefore, Q is a length-preserving transformation

b) $A = U\Sigma V^T \rightarrow$ compact form

$AA^T = U\Sigma V^T V\Sigma U^T$, $A^TA = V\Sigma U^T U\Sigma V^T$

$= U\Sigma^2 U^T$ $\qquad = V\Sigma^2 V^T$

i)

$\Rightarrow$ eigenvectors of AA^T are left singular vectors of AA^T.

$\Rightarrow$ eigenvectors of A^TA are right singular vectors of A^TA.

ii)

$\Rightarrow$ eigenvalues of A^TA and AA^T are squared singular values of A.

$$\boxed{\begin{aligned} eig(AA^T) &= \sigma(A)^2 \\ eig(A^TA) &= \sigma(A)^2 \end{aligned}}$$

c)

i) $A = I_{n\times n} \Rightarrow \lambda_i = 1$ for all $1 \le i \le n$

$\Rightarrow$ eigenvalues are not distinct, $\boxed{\text{FALSE}}$

ii) $Ax_1 = \lambda_1 x_1$

$Ax_2 = \lambda_2 x_2$

$A(x_1+x_2) = \lambda_1 x_1 + \lambda_2 x_2 \overset{?}{=} \lambda(x_1+x_2)$

This is only true if $\lambda_1 = \lambda_2 = \lambda$, therefore in general it is $\boxed{\text{FALSE}}$

Ex $\lambda_1 = 2$, $\lambda_2 = 3$

$x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $\Rightarrow$ $2\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

$\begin{bmatrix} 2 \\ 3 \end{bmatrix} \neq \lambda \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \lambda$ can not be found.

iii) $x^T A x \geq 0 \Rightarrow (x^T x)\lambda \geq 0$

$x^T x = \|x\|^2 \geq 0$, then $\lambda \geq 0$

λ must be non-negative

$\boxed{\text{TRUE}}$

iv) $A = I_{n \times n}$

$\text{rank}(A) = n$, $\text{eig}(A) = 1$

The rank of A is n, but number of distinct non-zero eigenvalues is 1. Therefore, rank can exceed the number of distinct non-zero eigenvalues.

$\boxed{\text{TRUE}}$, $\text{rank}(A) = r$, then there are at most r distinct non-zero eigenvalues

v) $Ax = \lambda x$

$Ay = \lambda y$

$x + y \neq 0$

$A(x+y) = \lambda x + \lambda y = \lambda(x+y)$

$Az = \lambda z$, z is an eigenvector, $\boxed{\text{TRUE}}$

2) Probability refresher

a) $Pr(X = H50) = 0.5 = p(H50)$

$Pr(X = H60) = 0.5 = p(H60)$

$Pr(Y = H \mid X = H50) = 0.5 = p(H \mid H50)$

$Pr(Y = H \mid X = H60) = 0.6 = p(H \mid H60)$

$Pr(Y = T \mid X = H50) = 0.5 = p(T \mid H50)$

$Pr(Y = T \mid X = H60) = 0.4 = p(T \mid H60)$

i) $$P(H50 \mid T) = \frac{P(T \mid H50) \cdot P(H50)}{P(T \mid H50) \cdot P(H50) + P(T \mid H60) P(H60)}$$

$$= \frac{\cancel{0.5} \times 0.5}{0.5 \times \cancel{0.5} + 0.4 \times \cancel{0.5}} = \frac{0.5}{0.9} = \frac{5}{9} = 0.56$$

ii) $P(T, H, H, H \mid H50) = (0.5)^4$ (i) flipping is an independent event

$P(T, H, H, H \mid H60) = 0.4\,(0.5)^3$

$$P(H50 \mid T, H, H, H) = \frac{P(H50)\,(0.5)^4}{P(H50)\,(0.5)^4 + P(H60)\,(0.4)\,(0.6)^3}$$

$$= \frac{(0.5)^4}{(0.5)^4 + (0.4)(0.6)^3} = 0.4197$$

iii)

$Pr(X = H50) = 1/3 = p(H50)$
$Pr(X = H60) = 1/3 = p(H60)$
$Pr(X = H55) = 1/3 = P(H55)$

$Pr(Y = H \mid X = H50) = 0.5 = p(H \mid H50)$
$Pr(Y = H \mid X = H60) = 0.6 = p(H \mid H60)$
$Pr(Y = H \mid X = H55) = 0.55 = p(H \mid H55)$

$Pr(Y = T \mid X = H50) = 0.5 = p(T \mid H50)$
$Pr(Y = T \mid X = H60) = 0.4 = p(T \mid H60)$
$Pr(Y = T \mid X = H55) = 0.45 = p(T \mid H55)$

$$P(9H, 1T) = \frac{1}{3}(0.5)^{10} + \frac{1}{3}(0.55)^{9}(0.45) + \frac{1}{3}(0.60)^{9}(0.40)$$
$$(3.255 \times 10^{-4}) + 6.908 \times 10^{-4} + 0.001343$$
$$= 0.00236$$

$$P(H50 \mid 9H, 1T) = \frac{3.255 \times 10^{-4}}{0.00236} = \underline{0.13793} \cong 0.14$$

$$P(H55 \mid 9H, 1T) = \frac{6.908 \times 10^{-4}}{0.00236} = \underline{0.29271} \cong 0.30$$

$$P(H60 \mid 9H, 1T) = \frac{0.001348}{0.00236} = \underline{0.56935} \cong 0.56$$

which makes sense since we got a lot of heads than tails

b) $Pr(X = \text{Science}) = 0.15 = P(S) = 0.15$

$Pr(X = \text{Healthcare}) = 0.21 = P(H) = 0.21$

$Pr(X = LA) = 0.24 = P(L) = 0.24$

$Pr(X = \text{Engi}) = 0.4 = P(E) = 0.4$

$P(Y = \text{Liked} \mid X = \text{Science}) = 0.9$

$P(Y = \text{Liked} \mid X = \text{Hea..}) = 0.18$

$P(Y = \text{Liked} \mid X = LA) = 0$

$P(Y = \text{Liked} \mid X = \text{Engi}) = 0.10$

$$P(S \mid L) = \frac{P(L \mid S)\, P(S)}{P(S)P(L \mid S) + P(H)\, P(L \mid H) + \underbrace{\cdots}_{\text{2 more terms}}}$$

$$= \frac{0.15 \times 0.9}{0.15 \times 0.9 + 0.21 \times 0.18 + 0 \times 0.24 + 0.10 \times 0.4}$$

$$= \frac{0.135}{0.2128} = 0.63439 \cong \boxed{0.63}$$

c) $p(P \mid preg) = 0.99 \Rightarrow p(N \mid preg) = 0.01$

$p(P \mid notpreg) = 0.10$

$p(notpreg) = 0.99$ $\qquad$ $p(notpreg) = 0.01$

$$P(preg \mid P) = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.1 \times 0.99} = \frac{0.099}{0.1089} = \boxed{0.091}$$

We are more likely to choose someone who is not pregnant since pregnant population is low.

d) $Ax + b = \begin{bmatrix} a_1^T x + b_1 \\ \vdots \\ a_n^T x + b_n \end{bmatrix}$

$E[Ax+b] = \begin{bmatrix} E[a_1^T x + b_1] \\ \vdots \\ E[a_n^T x + b_n] \end{bmatrix}$

$E[a_i^T x + b_i] = E[a_i^T x] + b_i$

$E[a_i^T x] = E[\sum_{j=1}^{n} (a_i)_j x_j] = \sum_{j=1}^{n} (a_i)_j E[x_j]$

$= \sum_{j=1}^{n} (a_i)_j E[x_j] + b_i = a_i^T E[x] + b_i$

$E[Ax+b] = \begin{bmatrix} \vdots \\ a_i^T E[x] + b_i \\ \vdots \end{bmatrix}$

$\boxed{= A\,E[x] + b}$ where $E[x] = \begin{bmatrix} E[x_1] \\ \vdots \\ E[x_n] \end{bmatrix}$

e) $\mathrm{cov}(Ax+b) = E[(Ax+b - E[Ax+b])(\quad)^T]$

$= E[(Ax+\cancel{b} - A\,E[x] - \cancel{b})(\quad)^T]$

$= E[(A(x-E[x]))(A(x-E[x]))^T]$

$= E[(A(x-E[x])((x-E[x])^T A^T)]$

$= A\,E[(x-E[x])(x-E[x])^T]A^T$

$\boxed{= A\,\mathrm{cov}(x)\,A^T}$

3) $\nabla_x x^T A y$

a)

$x^T A y = x^T b = b^T x \Rightarrow$

$\nabla_x b^T x = b = \boxed{Ay}$

b) $\nabla_y x^T A y$

$x^T A y = b^T y$

$\nabla_y x^T A y = b = \boxed{A^T x}$

c) $x^T A y = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} x_i y_j$

$$\frac{\partial \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} x_i y_j}{\partial a_{ij}} = x_i y_j \Rightarrow \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_1 \\ x_2 y_1 & & \\ \vdots & & \\ x_n y_1 & & x_n y_1 \end{bmatrix}$$

$= \boxed{x y^T}$

d) $\nabla_x x^T A x = A^T x + Ax$

$\nabla_x b^T x = b$

$\nabla_x (x^T A x + b^T x) = A^T x + Ax + b$

$\boxed{= (A^T + A) x + b}$

e) $\nabla_A \operatorname{tr}(AB)$

$$\operatorname{tr}(AB) = \sum_i \sum_j a_{ij} b_{ji}$$

$$\frac{\partial f}{\partial a_{ij}} = b_{ji} \Rightarrow \nabla_A \operatorname{tr}(AB) = \boxed{B^T}$$

f) $\underbrace{\operatorname{tr}(BA + A^T B + A^2 B)}_{} =$

$$\operatorname{tr}(BA) + \operatorname{tr}(A^T B) + \operatorname{tr}(A^2 B)$$

$$\nabla_A \operatorname{tr}(BA) = \nabla_A \operatorname{tr}(AB) = B^T$$ → cook book (100)

$$\underset{\operatorname{tr}(C) = \operatorname{tr}(C^T)}{\nabla_A \operatorname{tr}(A^T B) =} \nabla_A \operatorname{tr}(B^T A) \underset{\operatorname{tr}(AB) = \operatorname{tr}(BA)}{=} \nabla_A \operatorname{tr}(A B^T) = B$$

cookbook (103)

$$\nabla_A \operatorname{tr}(A^2 B) = (AB + BA)^T$$ → cook book (107)

$$\boxed{\nabla_A f = B^T + B + (AB + BA)^T}$$

$$= B^T + B + B^T A^T + A^T B^T$$

g) $\|A + \lambda B\|_F^2 = \operatorname{Tr}((A + \lambda B)(A + \lambda B)^T)$

$$(A + \lambda B)(A^T + \lambda B^T)$$

$$(AA^T + \lambda A B^T + \lambda B A^T + \lambda^2 B B^T) = g(A)$$

$$\nabla_A \operatorname{Tr}(g(A)) = \nabla_A \operatorname{Tr}(AA^T) + \nabla_A \operatorname{Tr}(\lambda A B^T)$$
$$+ \nabla_A \operatorname{Tr}(\lambda B A^T) + \nabla_A \operatorname{Tr}(\cancel{\lambda^2} B B^T) \;0$$

$$\Rightarrow \nabla_A \mathrm{Tr}(AA^T) = 2A \quad (115)$$

$$\Rightarrow \nabla_A \mathrm{Tr}(\lambda AB^T) = \lambda \nabla_A \mathrm{Tr}(AB^T) = \lambda B$$

$$= \nabla_A \mathrm{Tr}(\lambda BA^T) = \lambda \nabla_A \mathrm{Tr}(BA^T) =$$

$$\lambda \nabla_A \mathrm{Tr}(AB^T) = \lambda B$$

$$\Rightarrow 2A + 2\lambda B \Rightarrow \boxed{2(A+\lambda B)}$$

4) $\hat{y} = Wx$

$$X = \begin{bmatrix} x^{(1)T} \\ \vdots \\ x^{(n)T} \end{bmatrix}, \quad Y = \begin{bmatrix} y^{(1)T} \\ \vdots \\ y^{(n)T} \end{bmatrix}$$

$$\hat{Y} = XW^T$$

$$J(W) = \frac{1}{2} \sum_{i=1}^{n} \| y^{(i)} - Wx^{(i)} \|^2$$

$$= \frac{1}{2} \| Y - \hat{Y} \|_F^2$$

$$= \frac{1}{2} \mathrm{Tr}\left[(Y - XW^T)(Y^T - WX^T) \right]$$

$$= \frac{1}{2} \Big(\mathrm{Tr}(YY^T) - \mathrm{Tr}(XW^TY^T) - \mathrm{Tr}(YWX^T) +$$

$$\mathrm{Tr}(XW^TWX^T) \Big)$$

$\nabla_W Tr(YY^T) = 0$

$\nabla_W Tr(XW^TY^T) \underset{\uparrow}{=} \nabla_W Tr(YWX^T) = Y^TX$ (cookbook 101)

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

$\nabla_W Tr(\underset{A}{X}\underset{B}{W^T}\underset{C}{W}\underset{D}{X^T}) = \nabla_W Tr(\underset{C}{W}\underset{D}{X^T}\underset{A}{X}\underset{B}{W^T}) = WX^TX + WX^TX$

$Tr(ABCD) = Tr(CDAB)$ cookbook (111)

Cyclic permutation

set derivative to zero

Total

$\Rightarrow \frac{1}{\cancel{2}}(-\cancel{2}Y^TX + \cancel{2}WX^TX) = 0$

$\Rightarrow -Y^TX + WX^TX = 0$

$WX^TX = Y^TX$

$$\boxed{W = Y^TX(X^TX)^{-1}}$$

5)

$$L(\theta) = \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - \theta^T x^{(i)})^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

$$Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}, \quad X = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{bmatrix}$$

$$L(\theta) = \frac{1}{2}(Y - X\theta)^T(Y - X\theta) + \frac{\lambda}{2}\theta^T\theta$$

$$= \frac{1}{2}\left(Y^TY - Y^TX\theta - \theta^TX^TY + \theta^TX^TX\theta + \lambda\theta^T\theta\right)$$

$$(+\lambda\theta^TI\theta)$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{1}{\cancel{2}}\left[0 - \cancel{2}X^TY + (X^TX + X^TX)\theta + \cancel{2}\lambda I\theta\right]$$

set derivative to 0

$$= -X^TY + (X^TX + \lambda I)\theta \overset{\downarrow}{=} 0$$

$$= (X^TX + \lambda I)\theta = X^TY$$

$$\boxed{\theta^* = (X^TX + \lambda I)^{-1}X^TY}$$

perturbation

6) <u>Linear Regression</u>