DS5220 - HW1. SOLUTIONS

KAVANA VENKATESH.

1. Given an arbitrary matrix $A \in R^{n \times n}$,

Let us assume that A is the sum of matrices B and C.

$$A = B + C. \text{ —— } ①.$$

Taking the transpose of matrices on either side,

$$A' = B' + C'.$$

For matrix B to be symmetric, B' must be equal to B. i.e,

$B' = B$ & also for matrix C to be anti-symmetric, C' must be equal to −C. i.e,

$$C' = -C.$$

Substituting these conditions in the above equation, we get

$$A' = B - C \longrightarrow ②.$$

Adding equations ① and ②,

$$A + A' = 2B.$$

$$\Rightarrow \boxed{B = \frac{1}{2}(A+A')}$$

Subtracting ② from ①, we get

$$A - A' = 2C.$$

$$\Rightarrow \boxed{C = \frac{1}{2}(A-A')}.$$

Hence, we can write matrix A as,

$$A = B + C.$$

$$\Rightarrow \boxed{A = \underbrace{\frac{1}{2}(A+A')}_{B} + \underbrace{\frac{1}{2}(A-A')}_{C}}$$

Hence, we have proved that any arbitrary matrix, A can be expressed as the sum of a symmetric and an anti-symmetric matrix.

2. Given $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$

we need to show that $\|A\|_F^2 = Tr(A^T A)$.

By the definition of Forbenices norm,

LHS, $\|A\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2$

$$= (a_{11}^2 + a_{12}^2 + \cdots + a_{1n}^2) + (a_{21}^2 + a_{22}^2 + \cdots + a_{2n}^2) + \cdots + \cdots (a_{m1}^2 + a_{m2}^2 + \cdots + a_{mn}^2).$$

Expanding the RHS,

$$Tr(A^T A) = Tr\left(\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & \cdots & a_{22} & \cdots \\ \vdots & & \vdots & \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & & a_{mn} \end{pmatrix} \right)$$

$$= (a_{11}^2 + a_{21}^2 + \cdots + a_{m1}^2) + (a_{21}^2 + a_{22}^2 + \cdots + a_{m2}^2) + \cdots (a_{1n}^2 + a_{2n}^2 + \cdots + a_{mn}^2). = LHS$$

The above expression is written in such a way as trace is obtained as the sum of product of first row with first col, second row with second column and so on.

Since LHS = RHS, we have proved that,

$$\boxed{\|A\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}}$$

3. Given $U \in \mathbb{R}^{n \times n}$ is an orthogonal matrix.

$$\Rightarrow U^T U = U U^T = I_n.$$

we need to show that,

$$\|Ux\|_2^2 = \|x\|_2^2.$$ we know that,

By definition of l_2 norm,

$$\|Ux\|_2^2 = (Ux)^T (Ux).$$

$$= x^T U^T \cdot Ux.$$

$U^T U = I_n$ as U is orthogonal

$$\Rightarrow \|Ux\|_2^2 = x^T \cdot I_n \cdot x$$

$$\|Ux\|_2^2 = x^T x.$$

$$= \|x\|_2^2 \quad \text{(Definition of } l_2 \text{ norm).}$$

Hence, we have proved that,

$$\boxed{\|Ux\|_2^2 = \|x\|_2^2.}$$

4) 4.1 Given $x \in \mathbb{R}^n$, $a \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times n}$ and $A \in \mathbb{R}^{n\times n}$.

We need to show that,

$$\frac{\partial}{\partial x}(a^T x) = \vec{a}$$

$$a^T x = (a_1 \; a_2 \cdots a_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$= (a_1 x_1 + a_2 x_2 + \cdots + a_n x_n)$$

Now, let us take the derivative.

$$\frac{\partial (a^T x)}{\partial x} = \begin{pmatrix} \frac{\partial (a^T x)}{\partial x_1} \\ \frac{\partial}{\partial x_2}(a^T x) \\ \vdots \\ \frac{\partial (a^T x)}{\partial x_n} \end{pmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1}(a_1 x_1 + a_2 x_2 + \cdots + a_n x_n) \\ \frac{\partial}{\partial x_2}(a_1 x_1 + a_2 x_2 + \cdots + a_n x_n) \\ \vdots \\ \frac{\partial}{\partial x_n}(a_1 x_1 + a_2 x_2 + \cdots + a_n x_n) \end{bmatrix}$$

$$\boxed{\frac{\partial(a^T x)}{\partial x} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \vec{a}}$$

Hence, proved.

4.2) we need to prove that, $\frac{\partial x^T A x}{\partial x} = (A + A^T)x$.

Given, $A = \begin{pmatrix} a_{11} & a_{12} \cdots & a_{1n} \\ a_{21} & a_{22} \cdots & a_{2n} \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{nn} \end{pmatrix}$, $A^T = \begin{pmatrix} a_{11} & a_{21} \cdots & a_{n1} \\ a_{12} & a_{22} \cdots & a_{n2} \\ \vdots & \vdots & \vdots \\ a_{1n} & a_{2n} & a_{nn} \end{pmatrix}$

and $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

$$\Rightarrow x^T A x = (x_1\, x_2 \cdots x_n) \begin{pmatrix} a_{11} & a_{12} \cdots & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$= \begin{pmatrix} (a_{11}x_1 + a_{21}x_2 + \cdots + a_{n1}x_n) \cdots \\ (a_{1n}x_1 + a_{2n}x_2 + \cdots + a_{nn}x_n) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_n \end{pmatrix}$$

$$= \left(\sum_{p=1}^{n} a_{p1} x_p \cdots \sum_{p=1}^{n} a_{pn} x_p \right) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

$$= x_1 \times \sum_{i=1}^{n} a_{i1} x_i + x_2 \sum_{i=1}^{n} a_{i2} x_i + \cdots + x_n \sum_{i=1}^{n} a_{in} x_n.$$

$$x^T A x = \sum_{j=1}^{n} x_j \sum_{i=1}^{n} a_{ij} x_i.$$

Taking the derivative of $x^T A x$,

$$\frac{\partial \, x^T A x}{\partial x} = \begin{pmatrix} \frac{\partial (x^T A x)}{\partial x_1} \\ \frac{\partial (x^T A x)}{\partial x_2} \\ \vdots \\ \frac{\partial (x^T A x)}{\partial x_n} \end{pmatrix}$$

→ To simplify calculation, we shall calculate the derivative of pth row.

$$\frac{\partial (x^T A x)}{\partial x_p} = \frac{\partial}{\partial x_p} \sum_{j=1}^{n} x_j \sum_{i=1}^{n} a_{ij} x_i.$$

$$= \frac{\partial}{\partial x_p} \Big(x_1 \sum_{i=1}^{n} a_{i1} x_i + \cdots + x_p \sum_{i=1}^{n} a_{ip} x_i + \cdots + x_n \sum_{i=1}^{n} a_{in} x_i \Big).$$

$$= x_1 a_{p1} + \cdots + \Big(x_p \frac{\partial}{\partial x_p} \sum_{i=1}^{n} a_{ip} x_i + \sum_{i=1}^{n} a_{ip} x_i \cdot \frac{\partial}{\partial x_p} x_p \Big) + \cdots + x_n a_{pn}$$

$$= x_1 a_{p1} + \cdots + \left(x_p a_{pp} + \sum_{i=1}^{n} a_{ip} x_i \cdot 1 \right) + \cdots + x_n a_{pn}.$$

$$= x_1 a_{p1} + \cdots + x_n a_{pn} + \cdots + \sum_{i=1}^{n} a_{ip} x_i$$

$$= \sum_{j=1}^{n} x_j a_{pj} + \sum_{i=1}^{n} a_{ip} x_i.$$

$$= \sum_{j=1}^{n} a_{pj} x_j + \sum_{i=1}^{n} a_{ip} x_i.$$

$$= (p^{th} \text{ row of } A)x + (\text{transpose of } p^{th} \text{ col } A)x.$$

$$\frac{\partial (x^T A x)}{\partial x_p} = \left((p^{th} \text{ row of } A) + (\text{transpose of } p^{th} \text{ col of } A) \right) x.$$

we shall generalise the result obtai-ned for p to the whole matrix.

we can generalise this as shown below.

$$\frac{\partial (x^T A x)}{\partial x} = \begin{pmatrix} (1^{st} \text{ row of } A + \text{transpose of 1st col of } A)x \\ (2^{nd} \text{ row of } A + \text{transpose of 2nd col of } A)x \\ \vdots \\ (n^{th} \text{ row of } A + \text{transpose of } n^{th} \text{ col of } A)x. \end{pmatrix}$$

$$= \left(\begin{pmatrix} 1^{st} \text{ row of } A \\ 2^{nd} \text{ row of } A \\ \vdots \\ n^{th} \text{ row of } A \end{pmatrix} + \begin{pmatrix} \text{transpose of } 1^{st} \text{ col of } A \\ \text{transpose of } 2^{nd} \text{ col of } A \\ \vdots \\ \text{transpose of } n^{th} \text{ col of } A \end{pmatrix} \right) x$$

$$\boxed{\frac{\partial (x^T A x)}{\partial x} = (A + A^T)\, x}$$

(Hence proved).

4.3). we now need to prove $\frac{\partial}{\partial x}(Tr(A^T x)) = A$

$$A^T x = \begin{pmatrix} a_{11} & a_{21} \cdots & a_{n1} \\ a_{12} & a_{22} & a_{n2} \\ \vdots & \vdots & \vdots \\ a_{1n} & a_{2n} & a_{nn} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & x_{1n} \\ x_{21} & x_{22} \cdots & x_{2n} \\ \vdots & & \\ x_{n1} & x_{n2} \cdots & x_{nn} \end{pmatrix}$$

Now, $Tr(A^T x)$ can be calculated as

((first row of A x first col of x) + (second row of A * second col of x) + ... +(n^{th} row of A x n^{th} col of x))

$$\Rightarrow Tr(A^T x) = (a_{11}x_{11} + a_{21}x_{21} + \quad + a_{n1}x_{n1}) + (a_{12}x_{12} + a_{22}x_{22} + \cdots + a_{n2}x_{n2}) + \cdots + (a_{1n}x_{1n} + a_{2n}x_{2n} + \cdots + a_{nn}x_{nn})$$

Scanned with CamScanner

$$\frac{\partial (Tr(A^T x))}{\partial x} = \begin{pmatrix} \frac{\partial Tr(A^T x)}{\partial x_{11}} & \frac{\partial Tr(A^T x)}{\partial x_{12}} & \cdots & \frac{\partial Tr(A^T x)}{\partial x_{1n}} \\ \frac{\partial Tr(A^T x)}{\partial x_{21}} & \frac{\partial Tr(A^T x)}{\partial x_{22}} & \cdots & \frac{\partial Tr(A^T x)}{\partial x_{2n}} \\ \vdots & & & \vdots \\ \frac{\partial Tr(A^T x)}{\partial x_{n1}} & \frac{\partial Tr(A^T x)}{\partial x_{n2}} & \cdots & \frac{\partial Tr(A^T x)}{\partial x_{nn}} \end{pmatrix}$$

$\rightarrow$ we can write the above eq. as,

$$\frac{\partial Tr(A^T x)}{\partial x_{11}} = \frac{\partial}{\partial x_{11}} (a_{11} x_{11} + a_{21} x_{21} + \cdots + a_{n1} x_{n1})$$

$$= a_{11}$$

$$\Rightarrow \frac{\partial Tr(A^T x)}{\partial x} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & & a_{nn} \end{pmatrix}$$

$$\Rightarrow \boxed{\frac{\partial Tr(A^T x)}{\partial x} = A}$$

(Hence proved)

4.4) To prove: $\frac{\partial}{\partial x} \|y - Ax\|_2^2 = 2A^T(Ax - y)$.

By definition of L_2 - norm.

$$\|y - Ax\|_2^2 = (y - Ax)^T \cdot (y - Ax)$$

$$= (y^T - x^T A^T) \cdot (y - Ax)$$

$$= y^T y - y^T Ax - x^T A^T y + x^T A^T A x.$$

$$= y^T y - y^T Ax - (Ax)^T y + x^T A^T A x.$$

$$= y^T y - y^T Ax - y^T (Ax) + x^T A^T A x.$$

$$= y^T y - 2y^T Ax + x^T A^T A x.$$

$$\Rightarrow \frac{\partial}{\partial x} \|y - Ax\|_2^2 = \frac{\partial}{\partial x} (y^T y - 2y^T Ax + x^T A^T A x)$$

$$= -2y^T A + \frac{\partial}{\partial x} (Ax)^T Ax$$

$$= -2y^T A + \frac{\partial}{\partial x} \|(Ax)\|^2$$

$$= -2y^T A + 2A^2 x.$$

$$= -2y^T A + 2A^T Ax.$$

$$= 2A^T Ax - 2A^T y.$$

Hence, we have proved that,

$$\boxed{\frac{\partial}{\partial x} \| y - Ax \|^2 = 2A^T (Ax - y)}$$

5. By definition, a function is convex if and only if its second derivative is non-negative. So, we shall use this fact to check the convexity of the given functions.

5.1). $f(x) = e^{ax}$.

$$f'(x) = a e^{ax}.$$

$$f''(x) = a^2 e^{ax} \cdot \forall \; x \in (-\infty, +\infty).$$

We can see that $f''(x)$ is +ve ranging from 0 to ∞. So, the given function is convex.

5.2) $f(x) = -\log(x)$, $x \in (0, +\infty)$

$$f'(x) = - \cancel{\log} \, 1/x$$

$$f''(x) = \frac{1}{x^2}.$$

We see that $f''(x)$ is +ve for all values of $x \in (0, \infty)$. So, the function is convex.

5.3) $f(x) = e^{g(x)}$, given $g(x)$ is convex.

$\Rightarrow g''(x)$ is +ve for all values of x.

$f'(x) = g'(x)\, e^{g(x)}$

$f''(x) = g''(x) \cdot e^{g(x)} + g'(x) \cdot g'(x) \cdot e^{g(x)}$

$= \left[g''(x) + (g'(x))^2\right] \cdot e^{g(x)}$

Let us evaluate each term.

Since $g''(x)$ is convex $\Rightarrow$ it is +ve for all values of x.

And $(g''(x))^2$ is always positive for any value of x. and

$e^{g(x)}$ is positive for all values of $g(x)$, (which ranges from 0 to ∞ as, $e^{-\infty} = 0$ & $e^{\infty} = \infty$).

So, the given function is convex.

6.1) Given the modified regression problem,

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{N} S_i (y_i - \theta^T x_i)^2,$$

$$S_i \in [0, 1]$$

Given $\{(x_1, y_1), \dots (x_N, y_N)\}$ where

$x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.

In order to obtain the minimum value we need to differentiate the above expression and equate it to zero.

→ The above function can be written as, By definition of L2 norm,

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{N} S_i (y_i - \theta^T x_i)^2$$

$$= \min_{\theta \in \mathbb{R}^d} S_i \| Y - X\theta \|_2^2 \quad \text{where } Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

and $X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}$

Taking the derivative of the above expression wrt θ, we have,

$$\frac{\partial}{\partial \theta} S_i \| Y - X\theta \|_2^2 = \left(\sum_{i=1}^{N} S_i \right) * 2 \cdot (-X^T)(Y - X\theta) \quad \downarrow ①$$

To find the value of θ, let us equate it to zero.

$$\Rightarrow -2 S_i X^T (Y - X\theta) = 0.$$

$$\Rightarrow X^T Y - X^T X \theta = 0.$$

$$X^T X \theta = X^T Y$$ multiplying both sides by $(X^T X)^{-1}$

$$\Rightarrow \theta^* = (X^T X)^{-1} X^T Y$$

$$\boxed{\theta^* = (X^T X)^{-1} X^T Y.}$$

6.2). Let us now solve the above problem using gradient descent.

Given $\text{cost}(\theta) = \sum_{i=1}^{N} S_i (y_i - \theta^T x_i)$

$$\triangleq 2 S_i (-X^T)(Y - X\theta)$$

from ①.

- Initialize $\theta^0 \in \mathbb{R}^d$ with a random vector such that,

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$$ is $\mathcal{N}(0, 0.01^2)$ a gaussian distribution.

- Updating the values of θ until convergence for $t = 1, 2 ..$ (convergence/ max iterations) using the formula,

$$\theta^t = \theta^{(t-1)} - \rho \left.\frac{\partial\, cost}{\partial \theta}\right|_{\theta^{(t-1)}}$$

- Get the value of θ which is the optimum value called θ^*.

→ Thus, the above eq. is written as,

$$\theta^t = \theta^{t-1} - \rho\left(\sum_{i=1}^{N} S_i \times 2 \times (-X^T)(Y - X\theta)\right).$$

$$\theta^t = \theta^{t-1} - \left(2\rho \sum_{i=1}^{N} S_i\right) \cdot X^T(X\theta - Y).$$

The criteria for convergence: Let us take a small value of the order of 10^{-3} or 10^{-5} to stop the convergence in the 3rd step or @ $t >$ maximum iterations.

7.1). Given $f(x_1, x_2) = x_1^2 + (x_2 - 2)^2.$

Expression of gradient descent of $f(x_1, x_2)$ is,

$$\begin{bmatrix} \frac{\partial}{\partial x_1}(f(x_1, x_2)) \\ \frac{\partial}{\partial x_2}(f(x_1, x_2)) \end{bmatrix} = \begin{pmatrix} \frac{\partial}{\partial x_1}(x_1^2 + (x_2-2)^2) \\ \frac{\partial}{\partial x_2}(x_1^2 + (x_2-2)^2) \end{pmatrix}$$

$$= \begin{pmatrix} 2x_1 \\ 2x_2 - 2 \end{pmatrix}.$$