Artificial Intelligence : HW 1

Jeong Min Lee

Department of Physics and Astronomy, Seoul National University, Seoul 08826, Korea

Email : jmleeluck@snu.ac.kr

September 2023

Abstract

This is for the assignment of "Artificial Intelligence" course in SNU 2023 Fall.

1 Linear Algebra

In this section, I used the following notation.

1. $\|X\|_2 = \|X\|$, that is, I denoted L^2 norm just simply by dropping lower index.
2. $< X, Y >= X^T Y$, that is, the bracket is simply dot product of $\mathbb{R}^n$
3. When the SVD is used, I defined $r = \min(m, n)$ where $A \in \mathbb{R}^{m\times n}$
4. If $A \in \mathbb{R}^{m\times n}$, the result of SVD is $A = U\Sigma V^T$ where $U \in O(m), V \in O(n)$ and $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_r, 0, ..., 0) \in \mathbb{R}^{m\times n}$. Here σ_i are the square root of eigenvalues of $A^T A$ and $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$.
5. For matrix $A \in \mathbb{R}^{m\times n}$, A_i denotes the column vectors while A^j row vectors

1.1

$$\|A\|_s = \max_{X\in\mathbb{R}^n} \frac{\|AX\|}{\|X\|} \text{ where } \|\cdot\|_s \text{ denotes the spectral norm.} \tag{1}$$

Claim1 : for the functions f, g in $\mathbb{R}$, $\max_{X\in\mathbb{R}^n}(f(X)\cdot g(X)) \leq \max_{X\in\mathbb{R}^n} f(X)\cdot\max_{X\in\mathbb{R}^n} g(X)$
proof :

Here we can confine our discussion to the continuous functions that always have maximum and minimum value by Maximum-Mimimum Theorem. Let M, N denote the $\max f, \max g$, respectively. $\exists X_1, X_2 \in \mathbb{R}^n$ s.t. $f(X_1) = M$ and $g(X_2) = N$. If $X_1 = X_2$, then $\max(fg) = MN$, trivially. If not, let $h = fg$. It is also trivial that function h is continuous. If $\max h > MN$, for some $X^* \in \mathbb{R}^n, f(X^*)g(X^*) > MN$. However, $f(X^*) \leq M$ and $g(X^*) \leq N$ by definition, therefore, contradict ■.

$$\begin{aligned}
\|AB\|_s &= \max_{X\in\mathbb{R}^n} \frac{\|ABX\|}{\|X\|} \\
&= \max_{X\in\mathbb{R}^n} \frac{\|ABX\|}{\|BX\|}\frac{\|BX\|}{\|X\|} \\
&\leq \max_{X\in\mathbb{R}^n} \frac{\|AB\|}{\|BX\|} \cdot \max_{X\in\mathbb{R}^n} \frac{\|BX\|}{\|X\|} = \|A\|_s \cdot \|B\|_s
\end{aligned}$$

Note that the inequaility is held due to the **Claim1** above.

1.2

The maximum singular value of A is the maximum eigenvalue of A^TA. Since A^TA is symmetry matrix regardless of what A is, there is a basis $\{X_1, X_2, ..., X_n\}$ of $\mathbb{R}^n$ where $A^TAX_i = \sigma_i X_i$ for $\sigma_i \in \mathbb{R}$ by Spectral Theorem of Real Symmetric Matrix.[1] It implies that $\forall X \in \mathbb{R}^n, X = \Sigma_{i=0}^n a_i X_i$. From now on, I'll use Einstein notation for simplicity. [2]

$$\begin{aligned}
\|X\|^2 &=< a_iX_i, a_jX_j >= a_ia_j\delta_{ij} = a_i^2 \\
\|AX\|^2 &=< AX, AX >=< X, A^TAX >=< a_iX_i, a_j\sigma_jX_j >= a_i^2\sigma_i^2 \\
\therefore \|AX\|^2/\|X\|^2 &= \frac{a_1^2\sigma_1 + \cdots a_n^2\sigma_n^2}{a_1^2 + \cdots a_n^2} \leq \sigma_{max}
\end{aligned}$$

The discussion above result in the following.

$$\|A\|_s^2 = \sigma_{max}^2 \tag{2}$$

Note that such X is exist. For example, let $\|A\|_s^2 = \frac{\|AX^*\|}{\|X^*\|}$. WLOG $\sigma_{max} = \sigma_1$, then $X^* = X_1$

1.3

From the result of SVD,

$$AV = U\Sigma \tag{3}$$

Since $V \in O(n)$, the column vectors of V compose the basis vector of $\mathbb{R}^n$. That is, $\forall X \in \mathbb{R}^n, X = a_iV_i$.(Like section 1.2, I used the Einstein notation.) Therefore, $AX = a_iAV_i = a_i\sigma_iU_i$.

$$\|AX\|^2 = a_i^2\sigma_i^2 \tag{4}$$

Since we consider only the normalized vectors, $a_i^2 = 1$. Therefore, to minimize 4, only σ_r has to survive. Thus, the solution of given problem is $X = a_rV_r$.

1.4

$$\begin{aligned}
\|AX - b\|^2 &= \|U\Sigma V^TX - b\|^2 \\
&= \|\Sigma V^TX - U^Tb\|^2 \quad (\because U \in O(m))
\end{aligned}$$

Let $V^TX = Y, U^Tb = Z$. If X is solution of $AX = b, \Sigma Y = Z$ or $\sigma_iY_i = Z_i$ for $i = 1, 2, .., r$. Since V conserves the norm, $\|X\| = \|Y\|$. Therefore, to minimize the $\|X\|$, $Y_{r+1}, ..., Y_n = 0$. Thus,

$$Y = \begin{cases} \begin{pmatrix} \sigma_1^{-1}Z_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r^{-1}Z_r & \cdots & 0 \end{pmatrix}, & \text{if } n > m \\ \begin{pmatrix} \sigma_1^{-1}Z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r^{-1}Z_r \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}, & \text{if } m > n \end{cases} \tag{5}$$

[1] The fact that some operator $T \in M_{n\times n} = \{A | A \text{ is } n \text{ by } n \text{ matrix}\}$ is normal is equivalent to the fact that the set composed by the eigenvectors of T is the orthonormal basis of vector space V where $\dim V = n$.

[2] Einstein notation is the notaion which is widely used in quantum mechanics devised by Einstein. In this notation, one can drop the Σ sign and just describe the expressions with the indices.

One can simplify the expression 5 by denoting $\Sigma^{-1} = \text{diag}(\sigma_1^{-1}, \sigma_2^{-1}, ..., \sigma_r^{-1}, 0, ..., 0)$.

$$Y = V^T X = \Sigma^{-1} U^T b \tag{6}$$

Therefore,

$$X = V\Sigma^{-1}U^T b = A^\dagger b \tag{7}$$

Note that

$$A^\dagger = (A^T A)^{-1} A^T = (V\Sigma^T \Sigma V^T)^{-1} V\Sigma^T U^T = V\Sigma^{-1}\Sigma^{-T} V^T V\Sigma^T U^T = V\Sigma^{-1}U^T$$

2 Probability

In this section, I used the following notation.

1. P(X) denotes the probability that the event X is happened

2.1

Let F be the event that the failed product is found and A_1, A_2, A_3 be the event of machine $M_1.M_2.M_3$ manufacutured the products, respectively. Then,

$$P(F) = 0.2 \times 0.03 + 0.3 \times 0.02 + 0.5 \times 0.01 = 0.017 \tag{8}$$

What we want to know is the probability of failure in each machine. It is simply expressed by the conditional probability. (Note that the event F and each manufacturing event A_i are statistically independent so that $P(A_i \cap F) = P(A_i) \times P(F)$)

$$P(A_1|F) = \frac{P(A_1 \cap F)}{P(F)} = \frac{0.03 \times 0.2}{0.017} = \frac{6}{17}$$
$$P(A_2|F) = \frac{P(A_2 \cap F)}{P(F)} = \frac{0.3 \times 0.02}{0.017} = \frac{6}{17}$$
$$P(A_3|F) = \frac{P(A_3 \cap F)}{P(F)} = \frac{0.01 \times 0.05}{0.017} = \frac{5}{17}$$

Note that these probabilities are normalized.

2.2

Let X be the length of the remaining stick of first break and Y be the length of the stick of second break. Since the stick is uniform, the probability density function(PDF) of X is

$$f_X(x) = \begin{cases} 1/L & \text{if } 0 \le x \le L \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Futhermore, for a given X = x, the PDF of Y is

$$f_{Y|X=x}(y|x) = \begin{cases} 1/x & \text{if } 0 \le y < x \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

From the Bayes' theorem,

$$f_{X,Y}(x) = \begin{cases} 1/Lx & \text{if } 0 \le y < x \le L \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Then, the expectation value $E(Y|X = x)$ is

$$E(Y|X = x) = \int y f_{Y|X}(y|x) dy = x/2$$

Therefore,

$$E(Y) = E(E(Y|X = x)) = \frac{1}{L}\int_0^L x/2 \, dx = L/4$$

There is another solution that uses the PDF of Y. From the summation rule of PDF,

$$\begin{aligned} f_Y(y) &= \int f_{X,Y}(x, y) dx \\ &= \int_y^L \frac{1}{Lx} dx = \frac{1}{L} \ln \frac{L}{y} \text{ where } 0 \le y \le L \end{aligned}$$

Therefore, from the integration by part,

$$E(Y) = \int_0^L y \frac{1}{L} \ln \frac{L}{y} dy = L/4 \tag{12}$$

2.3

To solve this problem, I'll prove the following claim.

Claim2 : for $\forall x \ge 0, \ln x \le x - 1$

Proof :

At $x = 1, y = x - 1$ is tangent line of $\ln x$. From the fact that for $x > 1$, $\frac{d}{dx} \ln x = 1/x < 1 = \frac{d}{dx}(x - 1)$, the amount of increment of $y = x - 1$ is larger than that of $\ln x$. This can be verifed by the following graph ■.

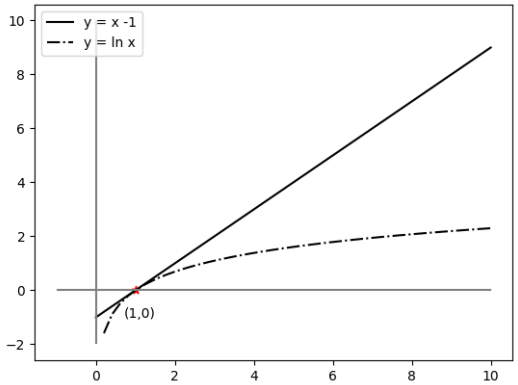

Figure 1: The graph of $y = x - 1$ and $y = \ln x$

Therefore, from the **Claim2** and the normality of the probability mass function,

$$\begin{aligned} -KL(p||q) &= \int p \ln \frac{q}{p} dx \\ &\leq \int p(\frac{q}{p} - 1) dx \\ &= \int q - p dx \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

Therefore, the KL divergence is always greater than or equal to zero.

3 Optimization

I used following notation in this section.

1. $D(X_0, r) = \{X \in \mathbb{R}^n | \ \|X - X_0\| < r\}$. This set is called by ball in Mathematics.
2. $\{\hat{e}_i\}_{i \in \{1,2,..,n\}}$ is the standard basis of $\mathbb{R}^n$

3.1

The definition of differenciable function f is that the function $y = f(X)$ for $X \in \mathbb{R}^n$ has a tagential plane (whose slope is finite) at every points $(X_0, f(X_0)) \in \mathbb{R}^{n+1}$. From the defition of differenciable function, $\exists \vec{a} \in \mathbb{R}^n$ s.t. $y = f(X_0) + < \vec{a}, (X - X_0) >$. Let g be the hyperplane. Then, $\lim_{X \to X_0} g(X) - f(X) = 0$. That is,

$$\lim_{X \to X_0} \frac{f(X) - f(X_0) - < \vec{a}, X - X_0 >}{|X - X_0|} = 0 \tag{13}$$

By replacing $X - X_0$ to hY for $Y \in \mathbb{R}^n$,the equation 13 reduces to the following.

$$\lim_{h \to 0} \frac{f(X_0 + hY) - f(X_0) - < \vec{a}, hY >}{h|Y|} = 0 \tag{14}$$

or by denoting $\hat{Y} = Y/|Y|$,

$$\lim_{h \to 0} \frac{f(X_0 + h\hat{Y}) - f(X_0)}{h} =< \vec{a}, \hat{Y} > \tag{15}$$

Therefore, for arbitrary unit vector Y,

$$D_Y f =< \vec{a}, Y > \tag{16}$$

From the relation between directional derivative and partial derivative,$D_{\hat{e}_i} f = \frac{\partial f}{\partial x_i}$, $D_{\hat{e}_i} f =< \vec{a}, \hat{e}_i >= a_i$.[3] Note that the relation above is derived by equation 16. Thus, the vector $\vec{a}$ is nothing else but the gradient of $f, \nabla f_{\blacksquare}$.

[3]Actually, I know that this is a definition of partial derivative, referring to **HJ Kim,Calculus 2+, pg 438.**

3.2

Let X^* be the local minimum of function f. Then, for arbitrarily small ϵ, $f(X^*) \leq f(X)$ for $X \in D(X^*, \epsilon)$ by definition of local minimum. From the definition of convex function, $f(tX^* + (1-t)Y) \leq tf(X^*) + (1-t)f(Y) \leq f(Y)$ for $Y \in D(X^*, \epsilon)$. This implies that the value of every points in the disk centered by local minimum of X^* evaluated by function f has smaller than one evaluated in local minimum. To show that X^* is global minimum, suppose $\exists Y \in \mathbb{R}^n \setminus D(X^*, \epsilon)$ s.t. $f(Y) < f(X^*)$. Then, from the definition of convex function, $f(tY + (1-t)X^*) \leq tf(Y) + (1-t)f(X^*) < f(Y)$ for $\forall t \in [0, 1]$. However, this is contradiciton since the the last inequaility is not held in t = 1. This implies that $\forall Y \in \mathbb{R}^n \setminus D(X^*, \epsilon), f(Y) \leq f(X^*)$. Therefore, $\forall Y \in \mathbb{R}^n, f(Y) \leq f(X^*)$.

4 Orthogormal Procustrates Analysis

In this section, I'll prove that the following proposition : the solution of $\Omega^* = \min_{\Omega \in \mathbb{R}^{3\times 3}} \|A\Omega - B\|$ is $\Omega = V^T U$ where U, V is the left singular vector and right singular vector of $B^T A$.

To prove this, let $B^T A = U\Sigma V^T$. Then,

$$\begin{aligned} \|A\Omega - B\|^2 &= \text{tr}\left((A\Omega - B)^T (A\Omega - B)\right) \\ &= \text{tr}\left(A^T A\right) + \text{tr}\left(B^T B\right) - 2\text{tr}\left(\Omega^T A^T B\right) \end{aligned}$$

Minimizing the $\|A\Omega - B\|^2$ is equivalent to maximize $\text{tr}\left(\Omega^T B^T A\right)$.

$$\begin{aligned} \text{tr}\left(\Omega^T A^T B\right) &= \text{tr}\left(\Omega^T U\Sigma V^T\right) \\ &= \text{tr}\left(V^T \Omega^T U\Sigma\right) \geq \text{tr}\left(I\Sigma\right) \end{aligned}$$

Therefore, $V^T \Omega^{*T} U = I$, or $\Omega^* = V^T U$.