Homework 1

Jimin Chae
4190.408 001, Artificial Intelligence

September 29, 2025

Problem 1. Linear algebra

Problem 1.1 : Matrix norm

The spectral norm (or L2-norm) of a matrix A is defined as:

$$||A||_2 = \max_{||x||_2=1} ||Ax||_2$$

Squaring both sides, we get:

$$||A||_2^2 = \left(\max_{||x||_2=1} ||Ax||_2\right)^2 = \max_{||x||_2=1} ||Ax||_2^2$$

The squared L2-norm $||Ax||_2^2$ can be expressed as a dot product:

$$||Ax||_2^2 = (Ax)^T(Ax) = x^T A^T Ax$$

Let the Singular Value Decomposition (SVD) of A be $A = U\Sigma V^T$, where U and V are orthogonal matrices, and Σ is a diagonal matrix with the singular values σ_i of A on its diagonal.

Substituting the SVD into the expression for $A^T A$:

$$A^T A = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T$$

Since U is an orthogonal matrix, $U^T U = I$ (the identity matrix). Thus,

$$A^T A = V(\Sigma^T\Sigma)V^T$$

Now, we can rewrite the optimization problem:

$$||A||_2^2 = \max_{||x||_2=1} x^T V(\Sigma^T\Sigma)V^T x$$

Let's perform a change of variable with $y = V^T x$. Since V is orthogonal, the norm is preserved: $||y||_2 = ||V^T x||_2 = ||x||_2 = 1$. Also, $x = Vy$.

$$||A||_2^2 = \max_{||y||_2=1} (Vy)^T V(\Sigma^T\Sigma)V^T(Vy) = \max_{||y||_2=1} y^T V^T V(\Sigma^T\Sigma)V^T Vy$$

Since $V^T V = I$, the expression simplifies to:

$$||A||_2^2 = \max_{||y||_2=1} y^T(\Sigma^T\Sigma)y$$

The matrix $\Sigma^T\Sigma$ is a diagonal matrix with diagonal entries $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$. The quadratic form is then:

$$y^T(\Sigma^T\Sigma)y = \sum_{i=1}^{n} \sigma_i^2 y_i^2$$

We need to maximize $\sum_{i=1}^n \sigma_i^2 y_i^2$ subject to the constraint $\sum_{i=1}^n y_i^2 = 1$. This expression is maximized when all the weight (y_i^2) is placed on the largest coefficient, which is σ_{max}^2. This occurs when y is the basis vector corresponding to σ_{max}, and the maximum value is σ_{max}^2.

Therefore, we have proven that:

$$||A||_2^2 = \sigma_{max}^2$$

Problem 1.2 : Ax = 0

The objective function to minimize is $||Ax||_2^2$. From the derivation in Problem 1, we know that:

$$||Ax||_2^2 = x^T A^T Ax$$

Using the SVD of $A = U\Sigma V^T$ and substituting $y = V^T x$ (which implies $||y||_2 = 1$ for $||x||_2 = 1$), the objective function becomes:

$$||Ax||_2^2 = y^T(\Sigma^T\Sigma)y = \sum_{i=1}^{n} \sigma_i^2 y_i^2$$

The problem is thus transformed into minimizing $\sum_{i=1}^{n} \sigma_i^2 y_i^2$ subject to the constraint $||y||_2^2 = \sum_{i=1}^{n} y_i^2 = 1$. The minimum value is achieved when all the weight is placed on the smallest coefficient, which is σ_{min}^2.

Let the smallest singular value of A be $\sigma_k = \sigma_{min}$. The minimum value is σ_{min}^2, which is attained when y is the k-th standard basis vector (i.e., $y_k = 1$ and all other elements are zero).

$$y_{sol} = e_k$$

To find the solution x, we use the relation $x = Vy$:

$$x_{sol} = Vy_{sol} = Ve_k$$

The product Ve_k is simply the k-th column vector of the matrix V. The columns of V are the right singular vectors of A.

Therefore, the solution x that minimizes $||Ax||_2^2$ subject to $||x||_2 = 1$ is the **right singular vector of A corresponding to its smallest singular value (σ_{min}).**

Problem 1.3 : Ax = b

The pseudo-inverse A^+ provides the minimum-norm solution to the least squares problem $\min_x ||Ax - b||_2^2$. The solution to this problem satisfies the normal equations:

$$A^T Ax = A^T b$$

First, express $A^T A$ and $A^T b$ using the SVD of A:

$$A^T A = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T = V(\Sigma^T\Sigma)V^T$$

$$A^T b = (U\Sigma V^T)^T b = V\Sigma^T U^T b$$

Substitute these into the normal equations:

$$V(\Sigma^T\Sigma)V^T x = V\Sigma^T U^T b$$

Multiply by V^T from the left. Since V is orthogonal, $V^T V = I$.

$$(\Sigma^T\Sigma)V^T x = \Sigma^T U^T b$$

Let r be the rank of A. Then $\Sigma^T\Sigma$ is an $n \times n$ diagonal matrix with diagonal entries $\sigma_1^2, ..., \sigma_r^2, 0, ..., 0$. Since it may contain zeros on the diagonal, it might not be invertible. We use its pseudo-inverse, $(\Sigma^T\Sigma)^+$, which is a diagonal matrix with entries $1/\sigma_i^2$ for non-zero σ_i and 0 otherwise.

$$V^T x = (\Sigma^T\Sigma)^+\Sigma^T U^T b$$

The term $(\Sigma^T\Sigma)^+\Sigma^T$ simplifies to Σ^+. Σ^+ is the $n \times m$ pseudo-inverse of Σ, formed by taking the reciprocal of the non-zero singular values and keeping the zero entries.

$$V^T x = \Sigma^+ U^T b$$

Finally, to solve for x, multiply by V from the left:

$$x = V\Sigma^+ U^T b$$

The solution to the least squares problem is given by $x = A^+ b$. By comparing this with our derived equation, we can identify the pseudo-inverse of A as:

$$A^+ = V\Sigma^+ U^T$$

Problem 2. Probability

Problem 2.1 : Bayes' Theorem

There are three machines M_1, M_2, M_3 accounting for 20%, 30%, and 50% of the total production, respectively. And the probability of defective products from each machine is 3%, 2%, and 1%, respectively.

Thus, the probablilty of defective products is :

$$\begin{aligned}
\Pr(F) &= \Pr(M_1 \cap F) + \Pr(M_2 \cap F) + \Pr(M_3 \cap F) \\
&= \Pr(M_1)\Pr(F|M_1) + \Pr(M_2)\Pr(F|M_2) + \Pr(M_3)\Pr(F|M_3) \\
&= 0.2 \times 0.03 + 0.3 \times 0.02 + 0.5 \times 0.01 \\
&= 0.017
\end{aligned}$$

If one randomly chosen product is defective, the probability for each machine is :

$$\begin{aligned}
\Pr(M_1|F) &= \frac{\Pr(M_1 \cap F)}{\Pr(F)} \\
&= \frac{\Pr(M_1)\Pr(F|M_1)}{\Pr(F)} \quad \text{(by Bayes' Theorem)} \\
&= \frac{0.2 \times 0.03}{0.017} \\
&= \frac{6}{17}
\end{aligned}$$

$$\begin{aligned}
\Pr(M_2|F) &= \frac{\Pr(M_2 \cap F)}{\Pr(F)} \\
&= \frac{\Pr(M_2)\Pr(F|M_2)}{\Pr(F)} \quad \text{(by Bayes' Theorem)} \\
&= \frac{0.3 \times 0.02}{0.017} \\
&= \frac{6}{17}
\end{aligned}$$

$$\begin{aligned}
\Pr(M_3|F) &= \frac{\Pr(M_3 \cap F)}{\Pr(F)} \\
&= \frac{\Pr(M_3)\Pr(F|M_3)}{\Pr(F)} \quad \text{(by Bayes' Theorem)} \\
&= \frac{0.5 \times 0.01}{0.017} \\
&= \frac{5}{17}
\end{aligned}$$

Problem 2.2 : Gaussian Distribution

Probability Density Function : $X \sim \mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Thus, MGF of X is :

$$\begin{aligned}
M_X(t) &= \mathbb{E}[e^{tX}] \\
&= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(x^2-2x\mu+\mu^2-2tx\sigma^2\right)} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(x-\mu-\sigma^2 t\right)^2} e^{\mu t+\frac{1}{2}\sigma^2 t^2} dx \\
&= e^{\mu t+\frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(x-\mu-\sigma^2 t\right)^2} dx \\
&= e^{\mu t+\frac{1}{2}\sigma^2 t^2}
\end{aligned}$$

For two independent random variables $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$, the MGF of $Z = X + Y$ is :

$$\begin{aligned}
M_Z(t) &= M_{X+Y}(t) \\
&= \mathbb{E}[e^{t(X+Y)}] \\
&= \mathbb{E}[e^{tX}]\,\mathbb{E}[e^{tY}] \\
&= e^{\frac{1}{2}t^2} e^{\frac{1}{2}t^2} \\
&= e^{t^2} \\
&= e^{0\times t+\frac{1}{2}\times 2\times t^2}
\end{aligned}$$

Thus, $Z \sim \mathcal{N}(0,2)$ by the uniqueness of MGF.

Problem 2.3 : KL Divergence

The KL divergence between two distributions p and q is defined as :

$$D_{KL}(p||q) = -\int_{-\infty}^{\infty} p(x) \log \frac{q(x)}{p(x)} dx$$

And then, for $f(x) = -\log x$, it is convex because $\frac{d^2}{dx^2} f(x) = \frac{1}{x^2} > 0$.
Thus, by Jensen's Inequality, we have :

$$f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$$

Let's say $Z = \frac{q(x)}{p(x)}$ is a random variable and $p(x)$ is the probability distribution of Z. Then, we can rewrite the inequality as :

$$-\log \mathbb{E}[Z] \le \mathbb{E}[-\log Z]$$

This time, $\mathbb{E}[Z] = 1$ because :

$$\mathbb{E}[Z] = \int_{-\infty}^{\infty} \frac{q(x)}{p(x)} p(x) dx = \int_{-\infty}^{\infty} q(x) dx = 1$$

And we know that $D_{KL}(p||q) = \mathbb{E}[-\log \frac{q(x)}{p(x)}] = \mathbb{E}[-\log Z]$. Therefore,

$$0 \le D_{KL}(p||q)$$

Problem 3. Optimization

Problem 3.1

a. The $f : \mathbb{R}^2 \to \mathbb{R}$ is differentiable, 1st order Taylor expansion at x is :

$$f(x + hv) \approx f(x) + \nabla_x f(x)^T hv$$

Thus, for any $v \in \mathbb{R}^n$, we have :

$$D_v f(x) = \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h} = \nabla_x f(x)^T v$$

b. The direction derivative of f at x in the direction u is :

$$D_u f(x) = \lim_{h \to 0} \frac{f(x + hu) - f(x)}{h} = \nabla_x f(x)^T u$$

The inner product of $\nabla_x f(x)$ and u is the amount of change of f at x in the direction u.

$$-\|\nabla_x f(x)\| \cdot \|u\| \leq \nabla_x f(x)^T u \leq \|\nabla_x f(x)\| \cdot \|u\| \quad \text{by Cauchy-Schwarz Inequality}$$

Thus, the direction that yields the largest decrease of f at x is $u^\star = -\frac{\nabla_x f(x)}{\|\nabla_x f(x)\|}$.

Problem 3.2

The $f : \mathbb{R}^n \to \mathbb{R}$ is convex, so we have, for all $0 \leq t \leq 1$ and all $x_1, x_2 \in X$:

$$f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2)$$

Thus, for $x \in X$ and a local minimizer x^*, we have :

$$f(tx^* + (1 - t)x) \leq tf(x^*) + (1 - t)f(x)$$

By the definition of local minimizer, we have :

$$f(x^*) \leq f(tx^* + (1 - t)x)$$

Thus, we have :

$$f(x^*) \leq tf(x^*) + (1 - t)f(x) \implies f(x^*) \leq f(x)$$

Therefore, x^* is a global minimizer.