Midterm Solution: Mathematics for AI

Problem 1 - (15pt)

Prove or disprove the following statements.

(a) For any $n \times n$ matrix A, the largest singular value σ_1 is equal to or larger than all eigenvalues.
(b) For any $n \times n$ matrix A, there exists $n \times n$ positive semidefinite matrix P and $n \times n$ orthogonal matrix Q such that $A = PQ$.
(c) For any $m \times n$ matrix A, matrix $A^T A$ is symmetric and positive definite.

Solution.
(a) Prove.
Note that $\|Qx\| = \|x\|$ holds for any orthogonal matrix Q, since $\|Qx\|^2 = x^T Q^T Q x = x^T x = \|x\|^2$.
Let $U\Sigma V^T$ be the singular value decomposition of A.
Then,

$$\|Ax\| = \|U\Sigma V^T x\| = \|\Sigma V^T x\| \leq \sigma_1 \|V^T x\| = \sigma_1 \|x\|.$$

Therefore, for any eigenvalue λ and its corresponding eigenvector x,

$$|\lambda| \|x\| = \|\lambda x\| = \|Ax\| \leq \|\sigma_1 x\| = \sigma_1 \|x\|$$

$$|\lambda| \leq \sigma_1$$

(b) Prove.
Let $U\Sigma V^T$ be the singular value decomposition of A. Let $P = U\Sigma U^T$, $Q = UV^T$.
Then, for any $x \in \mathbb{R}^n$,

$$x^T P x = x^T U\Sigma U^T x = x^T U\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}} U^T x = x^T U\Sigma^{\frac{1}{2}}(\Sigma^{\frac{1}{2}})^T U^T x = \|(\Sigma^{\frac{1}{2}})^T U^T x\|_2^2 \geq 0$$

Here, $\Sigma^{\frac{1}{2}}$ is element-wise square root of Σ. Note that $\Sigma^{\frac{1}{2}}$ exists since Σ is a diagonal matrix and its every diagonal element is non-negative. Therefore, P is a positive semidefinite matrix.
Also, Q is an orthogonal matrix since $Q^T Q = (UV^T)^T UV^T = VU^T UV^T = VV^T = I$.

(c) Disprove.

Consider a zero matrix A, e.g.,

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

For any nonzero $x \in \mathbb{R}^2$, $x^T A^T A x = 0$. Therefore, $A^T A$ is not positive definite.
Note that the statements holds if positive semidefinite instead of positive definite.

Problem 2 - (15pt)

Consider the following real-world investigation problem known as Markowitz mean-variance portfolio optimization, which solves to minimize the variance by adjusting the ratio between two bi-variate Gaussian distribution investment model. This optimization problem is formulated as:

$$\begin{aligned} \min \quad & f(w_1, w_2) = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12} \\ \text{subject to} \quad & w_1 + w_2 = 1 \\ & w_1 \geq 0 \\ & w_1 \leq 1 \end{aligned}$$

Solve the above optimization problem with Lagrangian method where $\sigma_1^2 = 0.3, \sigma_2^2 = 0.2, \sigma_{12} = 0.1$.

Solution.
Krush-Kuhn-Tucker (KKT) condition describes the condition for optimal w_1, w_2: (1) gradient of the Lagrangian function is zero, (2) inequality constraints satisfied complementary slackness condition, and (3) all constraints are satisfied.
Lagrangian function of given optimization problem is as follows. (+ 5 points.)

$$L(w_1, w_2, \lambda_1, \lambda_2, \lambda_3) = 0.3w_1^2 + 0.2w_2^2 + 0.2w_1w_2 - \lambda_1w_1 + \lambda_2(w_1 - 1) + \nu(1 - w_1 - w_2)$$

Gradients are calculated as:

$$\frac{\partial L}{\partial w_1} = 0.6w_1 + 0.2w_2 - \lambda_1 + \lambda_2 - \nu, \quad \frac{\partial L}{\partial w_2} = 0.4w_2 + 0.2w_1 - \nu$$

$$\frac{\partial L}{\partial \lambda_1} = -w_1, \quad \frac{\partial L}{\partial \lambda_2} = w_1 - 1, \quad \frac{\partial L}{\partial \nu} = 1 - w_1 - w_2$$

The complementary slackness condition should be considered, so there exists four possible cases.
(+ 5 points.)

(i) $\lambda_1 = \lambda_2 = 0$. $(w_1 > 0, w_1 < 1)$
From $0.6w_1 + 0.2w_2 = \nu, 0.4w_2 + 0.2w_1 = \nu, w_1 + w_2 = 1$, we have $w_1 = \frac{1}{3}, w_2 = \frac{2}{3}$. The objective function $f(w_1, w_2)$ is $\frac{1}{6}$.

(ii) $\lambda_1 = 0, \lambda_2 \neq 0$. $(w_1 = 0)$
Since $w_1 = 0$, we can obtain $w_2 = 1, \nu = 0.4, \lambda_2 = 0.2$. The objective function $f(w_1, w_2)$ is 0.2.

(iii) $\lambda_1 \neq 0, \lambda_2 = 0$. $(w_1 = 1)$
Since $w_1 = 1$, we can obtain $w_2 = 0, \nu = 0.2, \lambda_1 = 0.4$. The objective function $f(w_1, w_2)$ is 0.3.

(iv) $\lambda_1 \neq 0, \lambda_2 \neq 0$. $(w_1 = 0, w_1 = 1)$
Then, w_1 should be both 0 and 1, which is contradiction.

Therefore, solution (w_1, w_2) for the optimization problem is $\left(\frac{1}{3}, \frac{2}{3}\right)$, and the optimal value is $\frac{1}{6}$.
(+ 5 points.)

Problem 3 - (15pt)

Let $\boldsymbol{\theta} = [\theta_0, \theta_1, ..., \theta_D], \mathbf{x}_t \in \mathbb{R}^D, y_t \in \mathbb{R}$ for all $t \in \{1, ..., N\}$. Consider a linear model of the form

$$f(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \sum_{i=1}^{D} \theta_i x_i$$

together with a sum-of-squares error function of the form

$$\text{error}_D(\boldsymbol{\theta}) = \frac{1}{2} \sum_{t=1}^{N} \{f(\mathbf{x}_t, \boldsymbol{\theta}) - y_t\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. Show that minimizing error_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of some regularization term.

Solution.

$$\begin{aligned}
\text{error}_{new}(\theta) =& \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left(\frac{1}{2} \sum_{t=1}^{N} \{f(x_t + \epsilon, \theta) - y_t\}^2 \right) \\
=& \frac{1}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left(\sum_{t=1}^{N} \left\{ f(x_t, \theta) + \theta^\top \epsilon - y_t \right\}^2 \right) \\
=& \frac{1}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left(\sum_{t=1}^{N} \{f(x_t, \theta) - y_t\}^2 + 2\theta^\top \epsilon \left(f(x_t, \theta) - y_t \right) + \theta^\top \epsilon \theta^\top \epsilon \right) \\
=& \frac{1}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left(\sum_{t=1}^{N} \{f(x_t, \theta) - y_t\}^2 + 2\theta^\top \epsilon \left(f(x_t, \theta) - y_t \right) + \theta^\top \epsilon \epsilon^\top \theta \right) \\
=& \text{error}_D(\theta) + 0 + \frac{1}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left(N \theta^\top \epsilon \epsilon^\top \theta \right) \\
=& \text{error}_D(\theta) + \frac{N}{2} \left(\sigma^2 \theta^\top I \theta \right) \qquad (\because \mathbb{E}_{\epsilon \sim \mathcal{N}(\mu, \Sigma)}[\epsilon \epsilon^\top] = \Sigma + \mu \mu^\top) \\
=& \text{error}_D(\theta) + \frac{N \sigma^2}{2} \theta^\top \theta
\end{aligned}$$

Problem 4 - (15pt)

Suppose we have a sample of N pairs x_i, y_i drawn i.i.d. from the distribution characterized as follows:

$$\begin{aligned} x_i &\sim D(x), && \text{the data distribution with } x_i \in \mathbb{R} \\ y_i &= f(x_i) + \epsilon_i, && \text{the regression function with } y_i \in \mathbb{R} \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2), && \text{the noise distribution with } \epsilon_i \in \mathbb{R} \end{aligned}$$

We construct an estimator for f linear in the y_i for test data $x_0 \in \mathbb{R}$,

$$\hat{f}(x_0) = \sum_{i=1}^{N} \{w_i(x_0; \mathcal{X}) y_i\},$$

where the weights $w_i(x_0; \mathcal{X})$ do not depend on the y_i, but can depend on the training sequence of x_i, denoted here by $\mathcal{X}$. Show that the least squares estimator is a member of this class of estimators. Describe explicitly the weight $w_i(x_0; \mathcal{X})$ in this case.

Solution. Define $y^\top = (y_1, \ldots, y_n)$, and $X^\top = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}$.

We have $\hat{\theta} = \left(X^T X\right)^{-1} X^T y$, and then set

$$\hat{f}(x_0) = [x_0 \quad 1]\, \hat{\theta} = [x_0 \quad 1] \left(X^T X\right)^{-1} X^T y$$

In terms of the notation of the question,

$$w_i(x_0; \mathcal{X}) = [x_0 \quad 1] \left(X^\top X\right)^{-1} \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

for each i with $1 \leq i \leq n$.

More explicitly, $X^T X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$ which has determinant $(n-1)\sum_i x_i^2 - 2n \sum_{i<j} x_i x_j$.

This allows us to calculate $\left(X^T X\right)^{-1}$ and $w_i(x_0; \mathcal{X})$ even more explicitly.

Problem 5 - (20pt)

(a) Let

$$J(\mathbf{v}_2, \mathbf{z}_2) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - z_{i1}\mathbf{v}_1 - z_{i2}\mathbf{v}_2)^T(\mathbf{x}_i - z_{i1}\mathbf{v}_1 - z_{i2}\mathbf{v}_2),$$

where $x_i \in \mathbb{R}^D$, $v_i \in \mathbb{R}^D$, $\mathbf{z}_i = [z_{1i}, z_{2i}, \ldots, z_{ni}]^T \in \mathbb{R}^n$ and v_1 is first principal component (eigenvector with largest eigenvalue). Show that $\frac{\partial J}{\partial z_2} = 0$ yields $z_{i2} = \mathbf{v}_2^T\mathbf{x}_i$.

(b) Show that the value of $\mathbf{v}_2$ that minimizes

$$\tilde{J}(\mathbf{v}_2, \lambda_2, \lambda_{12}) = -\mathbf{v}_2^T\mathbf{C}\mathbf{v}_2 + \lambda_2(\mathbf{v}_2^T\mathbf{v}_2 - 1) + \lambda_{12}(\mathbf{v}_2^T\mathbf{v}_1 - 0) \quad (\text{lagrangian multiplier } \lambda_2, \lambda_{12} \geq 0)$$

is given by the eigenvector of $\mathbf{C}$ $(= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T)$ with the second largest eigenvalue.
Hint: orthonormal eigenvector $\mathbf{v}_1$ satisfy $\mathbf{C}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and $\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$.

Solution :

(a) Note that $\mathbf{v}_1^T\mathbf{v}_1 = \mathbf{v}_2^T\mathbf{v}_2 = 1$ and $\mathbf{v}_1^T\mathbf{v}_2 = 0$. Consequently,

$$J(\mathbf{v}_2, \mathbf{z}_2) = \frac{1}{N}\sum_i \left(\mathbf{x}_i^T\mathbf{x}_i + z_{i1}^2 + z_{i2}^2 - 2z_{i1}\mathbf{v}_1^T\mathbf{x}_i - 2z_{i2}\mathbf{v}_2^T\mathbf{x}_i\right)$$

$$\frac{\partial J}{\partial \mathbf{z}_{i2}} = \frac{1}{N}\left(2z_{i2} - 2\mathbf{v}_2^T\mathbf{x}_i\right) = 0.$$

Thus, $z_{i2} = \mathbf{v}_2^T\mathbf{x}_i$

(b)

$$\frac{\partial \tilde{J}}{\partial \mathbf{v}_2} = -2\mathbf{C}\mathbf{v}_2 + 2\lambda_2\mathbf{v}_2 + \lambda_{12}\mathbf{v}_1 = 0$$

Multiplying by $\mathbf{v}_1^T$ yields

$$-2\mathbf{v}_1^T\mathbf{C}\mathbf{v}_2 + 2\lambda_2\mathbf{v}_1^T\mathbf{v}_2 + \lambda_{12}\mathbf{v}_1^T\mathbf{v}_1$$

Since $\mathbf{C}\mathbf{v}_1 = \lambda_1\mathbf{v}_1, \mathbf{C} = \mathbf{C}^T, \mathbf{v}_1^T\mathbf{v}_2 = 0, \mathbf{v}_1^T\mathbf{v}_1 = 1$, we have

$$-2\lambda_1\mathbf{v}_1^T\mathbf{v}_2 + 2\lambda_2\mathbf{v}_1^T\mathbf{v}_2 + \lambda_{12} = 0 \Rightarrow \lambda_{12} = 0$$

Plugging this back to the original equation gives $-2\mathbf{C}\mathbf{v}_2 + 2\lambda_2\mathbf{v}_2 = 0$. Thus, $\mathbf{v}_2$ is the eigen vector corresponding to the second largest eigen value.

Problem 6 - (20pt)

Consider a mixture of K Gaussians, $p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$,
and define the log-likelihood as

$$l(\theta) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta)$$

Define the posterior responsibility that cluster k has for data point n as follows:

$$r_{nk} \triangleq p(z_n = k|\mathbf{x}_n, \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

(a) Show that the gradient of the log-likelihood w.r.t. $\boldsymbol{\mu}_k is$

$$\frac{d}{d\boldsymbol{\mu}_k} l(\boldsymbol{\theta}) = \sum_n r_{nk} \Sigma_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

(b) One way to handle the constraint that $\sum_{k=1}^{K} \pi_k = 1$ is to reparameterize using the softmax function, i.e., $\pi_k \triangleq \frac{e^{w_k}}{\sum_{k'=1}^{K} e^{w_{k'}}}$. Show that the gradient of the log-likelihood w.r.t w_k is

$$\frac{d}{dw_k} l(\theta) = \sum_n r_{nk} - \pi_k$$

Hint: use the chain rule and the fact that

$$\frac{d\pi_j}{dw_k} = \begin{cases} \pi_j(1 - \pi_j), & \text{if} \quad j = k \\ -\pi_j \pi_k, & \text{if} \quad j \neq k \end{cases}$$

Solution :

$$l(\boldsymbol{\theta}) = \sum_n \log p\left(\mathbf{x}_n \mid \boldsymbol{\theta}\right) = \sum_n \log \left(\sum_k \pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right)$$

(a)

$$\frac{\partial l}{\partial \boldsymbol{\mu}_k} = \sum_n \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{k'} \pi_{k'} \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}\right)} \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right) = \sum_n r_{nk} \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)$$

(b)

$$\begin{aligned} \frac{\partial l}{\partial w_k} &= \sum_j \frac{\partial l}{\partial \pi_j} \frac{\partial \pi_j}{\partial w_k} = \sum_j \left(\sum_n \frac{r_{nj}}{\pi_j} \frac{\partial \pi_j}{\partial w_k} \right) \\ &= \sum_{j \neq k} \left(\sum_n r_{nj} \right) \frac{1}{\pi_j} \left(-\pi_j \pi_k\right) + \sum_n r_{nk} \frac{1}{\pi_k} \pi_k \left(1 - \pi_k\right) \\ &= \sum_n r_{nk} - \sum_{k'} r_{nk'} \pi_k = \sum_n r_{nk} - \pi_k. \end{aligned}$$