# CX4240 Computing for Data Analysis - Homework 1

Name:
GTID:

Deadline: 11:59 pm EST, Feb 04

- Submit your answers as one single PDF file on Gradescope. **IMPORTANT: The solution to each problem/subproblem must be on a separate page. When submitting to Gradescope, please make sure to mark the page(s) corresponding to each problem/subproblem.**
- You will be allowed 2 total late days (48 hours) without penalty for the entire semester. Once those days are used, you will be penalized according to the following policy:
  - Homework is worth full credit before the due time.
  - It is worth 75% credit for the next 24 hours.
  - It is worth 50% credit for the second 24 hours.
  - It is worth zero credit after that.
- You are required to use Latex, or word processing software, to generate your solutions to the written questions. Handwritten solutions WILL NOT BE ACCEPTED.

## 1 Linear Algebra [35pts]

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. We say that $A$ is **positive semidefinite (PSD)** if

$$x^T A x \geq 0 \quad \text{for all } x \in \mathbb{R}^n.$$

(a) **[10 points]** Prove that if $A$ is PSD, then $cA$ is PSD for every scalar $c \geq 0$.

(b) **[10 points]** Suppose $A$ has an orthonormal eigenbasis $\{v_1, \ldots, v_n\}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. Show that for any $x \in \mathbb{R}^n$,

$$x^T A x = \sum_{i=1}^{n} \lambda_i (v_i^T x)^2.$$

(c) **[15 points]** Use part (b) to prove the following eigenvalue criterion:

$$A \text{ is PSD} \quad \iff \quad \lambda_i \geq 0 \ \text{ for all } i = 1, \ldots, n.$$

(*Hint: You may use without proof that symmetric matrices have an orthonormal eigenbasis.*)

**Solution:**

(a) Let $c \geq 0$ and assume $A$ is PSD. For any $x \in \mathbb{R}^n$,

$$x^T(cA)x \;=\; c\,x^T Ax \;\geq\; 0,$$

since $x^T Ax \geq 0$ and $c \geq 0$. Hence $cA$ is PSD.

(b) Since $A$ is symmetric, it has an orthonormal eigenbasis $\{v_1, \ldots, v_n\}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. Write

$$x = \sum_{i=1}^{n} \alpha_i v_i \quad \text{where } \alpha_i = v_i^T x.$$

Then

$$Ax = A\Big(\sum_{i=1}^{n} \alpha_i v_i\Big) = \sum_{i=1}^{n} \alpha_i A v_i = \sum_{i=1}^{n} \alpha_i \lambda_i v_i.$$

Therefore,

$$x^T Ax = \Big(\sum_{i=1}^{n} \alpha_i v_i\Big)^T \Big(\sum_{j=1}^{n} \alpha_j \lambda_j v_j\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \lambda_j\, v_i^T v_j = \sum_{i=1}^{n} \alpha_i^2 \lambda_i,$$

using $v_i^T v_j = \delta_{ij}$. Since $\alpha_i = v_i^T x$, we obtain

$$x^T Ax = \sum_{i=1}^{n} \lambda_i (v_i^T x)^2.$$

(c) ($\Rightarrow$) Assume $A$ is PSD. Fix $i$ and take $x = v_i$. Then by part (b),

$$0 \leq v_i^T A v_i = \sum_{j=1}^{n} \lambda_j (v_j^T v_i)^2 = \lambda_i,$$

so $\lambda_i \geq 0$ for all $i$.

($\Leftarrow$) Conversely, assume $\lambda_i \geq 0$ for all $i$. Then for any $x \in \mathbb{R}^n$, part (b) gives

$$x^T Ax = \sum_{i=1}^{n} \lambda_i (v_i^T x)^2 \geq 0,$$

since each term is nonnegative. Hence $A$ is PSD.

## 2 Probability and Statistics [35pts]

1. **[15 points]** A discrete random variable $X$ takes values in $\{0, 1, 2, 3\}$ with

$$\mathbb{P}(X = 0) = 0.1, \quad \mathbb{P}(X = 1) = 0.2, \quad \mathbb{P}(X = 2) = 0.3, \quad \mathbb{P}(X = 3) = 0.4.$$

Let $X_1, \ldots, X_5$ be i.i.d. copies of $X$, and define the sample mean

$$\overline{X} = \frac{1}{5} \sum_{i=1}^{5} X_i.$$

Compute:

(a) $\mathbb{E}[X]$ and $\mathrm{Var}(X)$;
(b) $\mathbb{E}[\overline{X}]$ and $\mathrm{Var}(\overline{X})$.
(c) Define $Y = 2X - 1$. Compute $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$.

**Solution:**

(a) **Compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$.**
First compute the mean:

$$\mathbb{E}[X] = \sum_{x \in \{0,1,2,3\}} x \, \mathbb{P}(X = x) = 0 \cdot 0.1 + 1 \cdot 0.2 + 2 \cdot 0.3 + 3 \cdot 0.4.$$

Thus,

$$\mathbb{E}[X] = 0 + 0.2 + 0.6 + 1.2 = 2.$$

Next compute the second moment:

$$\mathbb{E}[X^2] = \sum_{x \in \{0,1,2,3\}} x^2 \, \mathbb{P}(X = x) = 0^2 \cdot 0.1 + 1^2 \cdot 0.2 + 2^2 \cdot 0.3 + 3^2 \cdot 0.4.$$

Thus,

$$\mathbb{E}[X^2] = 0 + 0.2 + 1.2 + 3.6 = 5.$$

Therefore, the variance is

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 = 5 - 2^2 = 1.$$

(b) **Compute $\mathbb{E}[\overline{X}]$ and $\mathrm{Var}(\overline{X})$.**
Since $X_1, \ldots, X_5$ are i.i.d. copies of $X$,

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{5} \sum_{i=1}^{5} X_i\right] = \frac{1}{5} \sum_{i=1}^{5} \mathbb{E}[X_i] = \frac{1}{5} \cdot 5 \, \mathbb{E}[X] = \mathbb{E}[X] = 2.$$

For the variance, using independence and $\mathrm{Var}(cZ) = c^2\mathrm{Var}(Z)$,

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{5}\sum_{i=1}^{5} X_i\right) = \frac{1}{25}\,\mathrm{Var}\left(\sum_{i=1}^{5} X_i\right) = \frac{1}{25}\sum_{i=1}^{5}\mathrm{Var}(X_i) = \frac{1}{25}\cdot 5\,\mathrm{Var}(X).$$

Thus,

$$\mathrm{Var}(\overline{X}) = \frac{5}{25}\cdot 1 = \frac{1}{5}.$$

(c) **Define $Y = 2X - 1$. Compute $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$.**
By linearity of expectation,

$$\mathbb{E}[Y] = \mathbb{E}[2X - 1] = 2\mathbb{E}[X] - 1 = 2 \cdot 2 - 1 = 3.$$

For the variance, adding/subtracting a constant does not change variance and scaling multiplies variance by the square of the scale:

$$\mathrm{Var}(Y) = \mathrm{Var}(2X - 1) = \mathrm{Var}(2X) = 4\,\mathrm{Var}(X) = 4 \cdot 1 = 4.$$

2. **[20 points]**

Let $X \sim \mathcal{N}(\mu, \Sigma)$ and consider the affine transformation

$$Y = AX + b,$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.

(a) Derive the distribution of $Y$ by computing its mean and covariance matrix.

(b) Now let's consider the one-dimensional case and let $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X' \sim \mathcal{N}(\mu_2, \sigma_2)$, where $X, X' \in \mathbb{R}$. The Kullback-Leibler (KL) divergence between two distributions with densities $p(x)$ and $q(x)$ is defined as

$$D_{\mathrm{KL}}(p \,\|\, q) = \int_{\mathbb{R}^n} p(x) \log \frac{p(x)}{q(x)}\, dx.$$

Derive the closed-form expression for

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \sigma_1) \,\|\, \mathcal{N}(\mu_2, \sigma_2)).$$

**Solution:**

(a) The mean for $Y$ is

$$\mathbb{E}[Y] = \mathbb{E}[AX + b] = A\mathbb{E}[X] + b = A\mu + b,$$

and the covariance is

$$\mathrm{Cov}(Y) = \mathrm{Cov}(AX + b) = \mathrm{Cov}(AX) = A\,\mathrm{Cov}(X)\,A^\top = A\Sigma A^\top.$$

Therefore, $Y$ is Gaussian $N(A\mu + b, A\Sigma A^\top)$.

(b) Let $p = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q = \mathcal{N}(\mu_2, \sigma_2^2)$ on $\mathbb{R}$. Their densities are

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right), \qquad q(x) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

Then

$$\log \frac{p(x)}{q(x)} = \log \frac{\sigma_2}{\sigma_1} - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}.$$

Taking expectation under $p$ gives

$$D_{\mathrm{KL}}(p\|q) = \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] = \log \frac{\sigma_2}{\sigma_1} - \frac{1}{2\sigma_1^2}\mathbb{E}_p[(X-\mu_1)^2] + \frac{1}{2\sigma_2^2}\mathbb{E}_p[(X-\mu_2)^2].$$

Use $\mathbb{E}_p[(X-\mu_1)^2] = \sigma_1^2$ and

$$\mathbb{E}_p[(X-\mu_2)^2] = \mathbb{E}_p[(X-\mu_1+\mu_1-\mu_2)^2] = \mathbb{E}_p[(X-\mu_1)^2] + (\mu_1-\mu_2)^2 = \sigma_1^2 + (\mu_1-\mu_2)^2.$$

Substitute:

$$D_{\mathrm{KL}}\left(\mathcal{N}(\mu_1, \sigma_1^2) \,\|\, \mathcal{N}(\mu_2, \sigma_2^2)\right) = \log \frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{1}{2\sigma_2^2}\left(\sigma_1^2 + (\mu_1-\mu_2)^2\right).$$

Equivalently,

$$D_{\mathrm{KL}} = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2} - 1 + 2\log\frac{\sigma_2}{\sigma_1}\right) = \frac{1}{2}\left(\frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{\sigma_2^2} - 1 + \log\frac{\sigma_2^2}{\sigma_1^2}\right).$$

## 3 Optimization [30pts]

Suppose we want to minimize the function:

$$F(x, y) = \frac{10x^2 + y^2}{2}.$$

The actual minimum is $F = 0$ at $(x^*, y^*) = (0, 0)$. Solve the following questions in *vector* notation.

1. **[10 points]**

Give the expression of the gradient vector $\nabla F$ at point $(x, y)$.

**Solution:**

$$F(\mathbf{z}) = \frac{1}{2}(10x^2 + y^2), \quad \mathbf{z} = \begin{bmatrix} x \\ y \end{bmatrix}$$

This can be written in quadratic form as

$$F(\mathbf{z}) = \frac{1}{2}\mathbf{z}^\top A\mathbf{z}, \quad A = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}.$$

Since $A$ is symmetric, the gradient is

$$\nabla F(\mathbf{z}) = A\mathbf{z} = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10x \\ y \end{bmatrix}.$$

2. **[10 points]** Let the initial point be $(x_0, y_0) = (1, 1)$. Perform gradient descent with step size $s = 0.5$ for two iterations. For each iteration, explicitly show:

   - the gradient computation, and
   - the updated solution.

   Based on your results, discuss whether the resulting solution sequence will converge to the optimal solution.

**Solution:**
The gradient descent update rule is

$$\mathbf{z}_{k+1} = \mathbf{z}_k - s\nabla F(\mathbf{z}_k) = (I - sA)\mathbf{z}_k.$$

For $s = 0.5$,

$$I - sA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.5 \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -4 & 0 \\ 0 & 0.5 \end{bmatrix}.$$

Let $\mathbf{z}_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$.

**Step 1:**

$$\nabla F(\mathbf{z}_0) = \begin{bmatrix} 10x_0 \\ y_0 \end{bmatrix}, \quad \mathbf{z}_1 = \mathbf{z}_0 - 0.5\nabla F(\mathbf{z}_0) = \begin{bmatrix} -4x_0 \\ 0.5y_0 \end{bmatrix}.$$

**Step 2:**

$$\nabla F(\mathbf{z}_1) = \begin{bmatrix} -40x_0 \\ 0.5y_0 \end{bmatrix}, \quad \mathbf{z}_2 = \begin{bmatrix} 16x_0 \\ 0.25y_0 \end{bmatrix}.$$

Therefore, the solution sequence is **not convergent**.

3. **[10 points]** Again, let $(x_0, y_0) = (1, 1)$. Perform gradient descent with step size $s = 0.1$ for two iterations. As before, clearly demonstrate:

   - the gradient computation, and
   - the updated solution at each step.

   Determine whether the resulting sequence is convergent.

**Solution:**
For $s = 0.1$,

$$I - sA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.1 \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0.9 \end{bmatrix}.$$

**Step 1:**

$$\mathbf{z}_1 = \mathbf{z}_0 - 0.1\nabla F(\mathbf{z}_0) = \begin{bmatrix} 0 \\ 0.9y_0 \end{bmatrix}.$$

**Step 2:**

$$\nabla F(\mathbf{z}_1) = \begin{bmatrix} 0 \\ 0.9y_0 \end{bmatrix}, \quad \mathbf{z}_2 = \begin{bmatrix} 0 \\ 0.81y_0 \end{bmatrix}.$$

The solution sequence is **convergent** since the magnitude is decreasing.

## References

**Solution:**
Please mention any AI tools, people, post or blog etc. you used.