EECS 553 Homework 1 Solution (W24)

1. **Honor Code** (3 pts each)

   Grading Rubric:

   3 points each if correct, and 0 point if incorrect

   (a) True

   (b) True

   (c) False

2. **PD/PSD matrices**

   (a) (3 points)

   Grading Rubrics

   (1) 3 points if fully correct

   (2) 2 points for invoking correct definition of covariance/PSD matrix, but incorrect final result

   (3) 1 point for incorrect usage of eigenvalue decomposition or wrong definition of eigenvalue/vector

   (4) 0 if no effort

   Let $\boldsymbol{C}$ be the covariance matrix of a random vector $\boldsymbol{X}$. Then $\boldsymbol{C} = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$. For any vector $\boldsymbol{v}$, $\boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v} = \mathbb{E}[\boldsymbol{v}^T(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\boldsymbol{v}] = \mathbb{E}[(\boldsymbol{v}^T(X - \mathbb{E}[X]))^2] \geq 0$.

   (b) (6 points)

   Grading Rubrics

   (1) 2 points for identifying $\det(\boldsymbol{A}) = \prod \lambda_i$

   (2) 1 point for showing invertibility is equivalent to that all eigenvalues are nonzero.

   (3) 1 point for identifying eigenvalues of $\boldsymbol{A}^{-1}$.

   (4) 1 point for showing $\boldsymbol{A}^{-1} = \boldsymbol{U}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{U}$

   (5) 1 point for showing $\boldsymbol{A}^{-1}$ is also PD.

   By spectral theorem, $\boldsymbol{A} = \boldsymbol{U}^T\boldsymbol{\Lambda}\boldsymbol{U}$ where $\boldsymbol{\Lambda}$ is a real diagonal matrix and $\boldsymbol{U}$ is orthonormal. As $\det(\boldsymbol{U}) = \det(\boldsymbol{U}^T) = \pm 1$ and $\det(\boldsymbol{A}) = \det(\boldsymbol{U}^T)\det(\boldsymbol{\Lambda})\det(\boldsymbol{U}) = \det(\boldsymbol{U})^2\det(\boldsymbol{\Lambda}) = \det(\boldsymbol{\Lambda}) = \prod \lambda_i$ where $\{\lambda_i\}$ are eigenvalues of $\boldsymbol{\Lambda}$. $\boldsymbol{A}$ is invertible if and only if $det(\boldsymbol{A}) \neq 0$, which is equivalent to all of its eigenvalues are nonzero. $\boldsymbol{A}^{-1} = \boldsymbol{U}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{U}$. So the eigenvalues of $\boldsymbol{A}^{-1}$ are $\{\frac{1}{\lambda_i}\}$. If $\boldsymbol{A}$ is PD, then $\forall i$, $\lambda_i > 0$ and $\frac{1}{\lambda_i} > 0$. So $\boldsymbol{A}^{-1}$ is also PD, given that all of its eigenvalues are positive.

3. **Probability** (3 points each)

(a) (3 points)

Grading Rubrics

(1) 3 points for a correct proof.

(2) 2 points for applying the axiom of probability but the proof is not correct.

Let Ω be the sample space. $\Pr(X = x) = \Pr(\{\omega \in \Omega : X(w) = x\})$. $\Pr(\{\omega \in \Omega : X(w) = x\}) = \bigsqcup_y \{\omega \in \Omega : X(w) = x, Y(w) = y\}$. Note $\{\omega \in \Omega : X(w) = x, Y(w) = y\}$ and $\{\omega \in \Omega : X(w) = x, Y(w) = y'\}$ are disjoint when $y \neq y'$. Hence, $\Pr(X = x) = \Pr(\{\omega \in \Omega : X(w) = x\}) = \sum_y \Pr(\{\omega \in \Omega : X(w) = x, Y(w) = y\}) = \sum_y \Pr(X = x, Y = y) = \sum_y p(x, y)$.

(b) (3 points)

Grading Rubrics

(1) 3 points for a correct proof.

(2) 1 point for applying the definition of expectation.

(3) 1 point for applying the definition of conditional probability.

(4) 1 point for applying the definition of conditional expectation.

$\mathbb{E}[X] = \sum_x xp(x) = \sum_x x \sum_y p(x, y) = \sum_x x \sum_y p(x|y)p(y) = \sum_y p(y) \sum_x xp(x|y) = \sum_y p(y)\mathbb{E}[X|Y = y] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$

4. Gaussian level sets (3 pts each)

(a) **Grading Rubrics**

(1) Add 1 point to the student's score for each of the following that is satisfied:

- 1 points for drawing an ellipse with correct center
- 1 point for correct semi-minor and semi-major length or correct minor and major length
- 1 point for correct orientation

(2) 0 if no effort

Observe $\boldsymbol{U}$ is the rotation matrix that rotate vector counterclockwise $\frac{\pi}{6}$ with respect to the positive horizontal axis, with $\boldsymbol{U}^T$ as its inverse (the inverse of an orthogonal matrix is its transpose).

Thus $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T$. This is a handy thing to know, and students are encouraged to check it for themselves. A similar formula holds for the pseudoinverse when a square matrix is not full rank, or using the SVD to find the pseudoinverse of a non-square matrix.

Define $\boldsymbol{x}' = \boldsymbol{U}^T(\boldsymbol{x} - \mu)$. Then we can write that a vector $\boldsymbol{x} \in \mathcal{C}$ if $\boldsymbol{x}'$ satisfies

$$2 \geq (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = (\boldsymbol{x}')^T (\boldsymbol{\Lambda})^{-1} \boldsymbol{x}' = \frac{(x_1')^2}{4} + \frac{(x_2')^2}{3},$$

which is an Ellipse. In standard form, this ellipse can be written

$$\frac{(x_1')^2}{8} + \frac{(x_2')^2}{6} = 1$$

with lengths $2\sqrt{2}$ and $\sqrt{6}$ for the minor and major axis, respectively.

Thus $\mathcal{C}$ is simply the Ellipse specified above rotated counter clockwise $\frac{\pi}{6}$ and translated by vector $[-1 \ \ -1]^T$

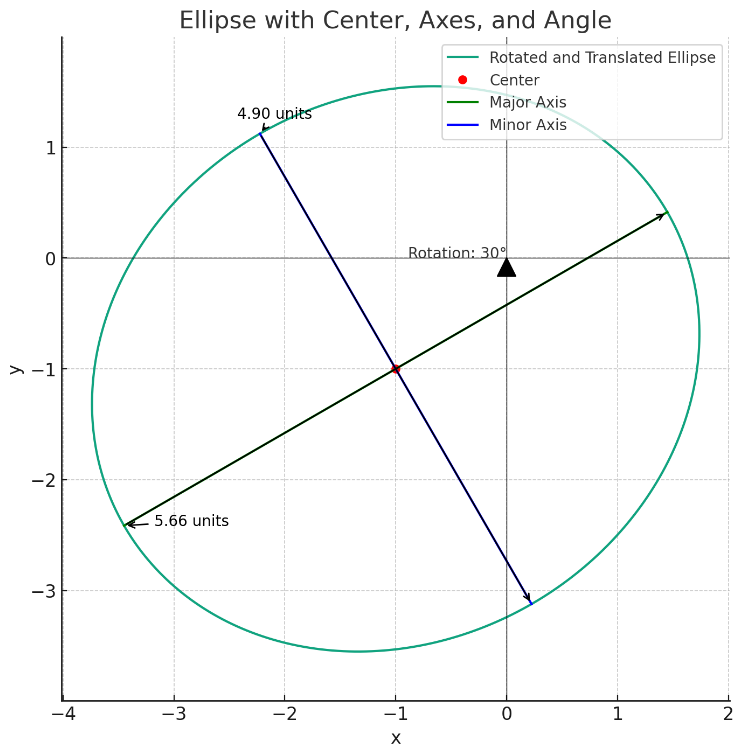

Figure 1: P3(a) plot

(b) **Grading Rubrics**

(1) Add 1 point to the student's score for each of the following that is satisfied:

- 1 point for deriving R as sum of square of two independent Gaussian random variables
- 1 point for identifying the correct chi-squared random variable
- 1 point for correct answer (full credit if the answer is within ± 0.01)

(2) 0 if no effort

[Method 1] Define $\boldsymbol{X'} = \boldsymbol{U}^T(\boldsymbol{X} - \mu)$, then by property[1] of Gaussian distribution $\boldsymbol{X'} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma'})$, where $\boldsymbol{\Sigma'} = \boldsymbol{U}^T\Sigma\boldsymbol{U} = \boldsymbol{\Lambda}$, as $\boldsymbol{\Lambda}$ is diagonal X_1' and X_2' are independent Gaussian random variables with the same mean 0 and variance 4 and 3 respectively.

Define $R = (\boldsymbol{X'})^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X'} = \frac{(X_1')^2}{4} + \frac{(X_2')^2}{3}$, notice that R follows a chi-square distribution with 2 degrees of freedom[2].

Observe $\{\boldsymbol{x} \in \mathcal{C}\} = \{(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq 2\}$.

Thus $\Pr(\boldsymbol{X} \in \mathcal{C}) = \Pr(R \leq 2) = F(2) = 0.6321$ where F denotes the cumulative distribution function (cdf) of the chi-square distribution with 2 degrees of freedom.

[1] if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for any matrix $\boldsymbol{A}$, $\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{x} - \mathbf{b}) \sim \mathcal{N}(\boldsymbol{A}(\boldsymbol{\mu} - \mathbf{b}), \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$

[2] A random variable Y follows a chi-square distribution if $Y = X_1^2 + X_2^2$ for X_1, X_2 that are independent Gaussian random variables with mean 0 and variance 1

[Method 2] Though the following may look slightly different, it is equivalent to Method 1 just with a different choice of $\boldsymbol{X}'$.

You can identify $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$, where $\boldsymbol{\Lambda}^{-1/2} = diag(\frac{1}{2}, \frac{1}{\sqrt{3}})$. Then define $\boldsymbol{X}' = \boldsymbol{A}(\boldsymbol{X} - \mu)'$ where $\boldsymbol{A} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^T$. Invoke the linear transformation property of Gaussian distribution, we can conclude $\boldsymbol{X}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix. Hence, $R = (\boldsymbol{X}')^T\boldsymbol{X}'$ follows a Chi-square distribution with 2 degrees of freedom. Thus $\Pr(\boldsymbol{X} \in \mathcal{C}) = \Pr(R \leq 2) = F(2) = 0.6321$ where F denotes the cumulative distribution function (cdf) of the chi-square distribution with 2 degrees of freedom.

5. Unconstrained Optimization (3 points each)

(a) (3 points)

Grading Rubrics

(1) Add 1 point to the student's score for each of the following that is satisfied:

- 1 point for $g(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) = f(\boldsymbol{A}(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) + \boldsymbol{b})$, or if proving in reverse order of what is shown below $tg(\boldsymbol{x}_1) + (1-t)g(\boldsymbol{x}_2) = tf(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}) + (1-t)f(\boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{b})$
- 1 point for correctly factoring $f(t(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}) + (1-t)(\boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{b}))$
- 1 point for correctly applying the definition of convexity via $f(t(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}) + (1-t)(\boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{b})) \leq tf(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}) + (1-t)f(\boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{b})$

(2) 0 if no effort

We will prove this directly using the definition of convexity. Starting with the term we wish to upper bound, we have

$$
\begin{aligned}
g(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) &= f(\boldsymbol{A}(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) + \boldsymbol{b}) \\
&= f(t\boldsymbol{A}\boldsymbol{x}_1 + (1-t)\boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{b}) \\
&= f(t(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}) + (1-t)(\boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{b})) \\
&\leq tf(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}) + (1-t)f(\boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{b}) \\
&= tg(\boldsymbol{x}_1) + (1-t)g(\boldsymbol{x}_2),
\end{aligned}
$$

where the inequality follows from the convexity of f. This verifies the convexity of g.

(b) **Grading Rubrics**

(1) Add points to the student's score for each of the following that is satisfied:

- 2 point for writing the quadratic function in terms of summation
- 2 point for evaluating the entries of the hessian matrix
- 1 point for concluding f is convex if A is PSD, and strictly convex if A is PD

(2) 0 if no effort

(3) Students receive full credits if they directly apply the rules of vector calculus.

Let $\boldsymbol{x} = [x_1, \ldots, x_d]^T$, A_{ij} denote the (i,j)-th entry of matrix $\boldsymbol{A}$, and b_i denote the i-th entry of $\mathbf{b}$ then the quadratic function $f(\boldsymbol{x})$ can be written explicitly as:

$$
f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \mathbf{b}^T\boldsymbol{x} + c = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} A_{ij}x_ix_j + \sum_{i=1}^{d} b_ix_i + c\,.
$$

Applying the definition of the Hessian matrix, the (k, ℓ)-th entry of $\nabla^2 f(x)$ is given by:

$$\begin{aligned}
\left[\nabla^2 f(\boldsymbol{x})\right]_{k,l} &= \frac{\partial^2 f(\boldsymbol{x})}{\partial x_k \partial x_\ell} \\
&= \frac{\partial^2}{\partial x_k \partial x_\ell} \left\{ \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} A_{ij} x_i x_j + \sum_{i=1}^{d} b_i x_i + c \right\} \\
&= \frac{\partial}{\partial x_k} \left\{ \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} A_{ij} \frac{\partial}{\partial x_\ell} x_i x_j + \sum_{i=1}^{d} \frac{\partial}{\partial x_\ell} b_i x_i \right\} \\
&= \frac{\partial}{\partial x_k} \left\{ \sum_{i=1}^{d} A_{i\ell} x_i + b_\ell \right\} \\
&= A_{k\ell} \,,
\end{aligned}$$

thus the Hessian of f is $\boldsymbol{A}$. The function f is convex when $\boldsymbol{A}$ is positive semi-definite, and strictly convex if $\boldsymbol{A}$ is positive definite.