CMPT 726/410 Fall 2023 Assignment 1

Your Name Here (Your Student ID #)

November 5, 2023

Due: Thursday, November 23, 2023 at 11:59 pm Pacific Time

Important: Make sure to download the zip archive associated with this assignment, which contains essential data and starter code that are required to complete the programming component of the assignment.

This assignment is designed to be substantially more challenging than the quizzes and requires thorough understanding of the course material and extensive thought, so **start early!** If you are stuck, come to office hours. Make sure to check the discussion board often for updates, clarifications, corrections and/or hints.

Partial credit will be awarded to reasonable attempts at solving each question, even if they are not completely correct.

There will be office hours dedicated to assignment-related questions. Times and Zoom links will be posted on Canvas. If you have a question that you would like to ask during office hours, make sure to post a brief summary of your question in the office hours thread on the Canvas discussion board. During office hours, questions will be answered in the order they appeared on the office hours thread. We may not be able to get to all questions, so please start early and plan ahead.

Requests for extensions will not be considered under any circumstances. Make sure you know how to submit your assignment to Crowdmark and leave sufficient time to deal with any technical difficulties, Internet outages, server downtimes or any other unanticipated events. It is possible to update your assignment after submission, so we recommend uploading a version of your assignment at least several hours before the deadline, even if it is incomplete.

Submission Instructions

Carefully follow the instructions below when submitting your assignment. Not following instructions will result in point deductions.

1. You should typeset your assignment in LaTeX using this document as a template. We recommend using Overleaf to compile your LaTeX. Include images in the section they go with and **do not** put them in an appendix.
2. At the beginning of the assignment (see above), please copy the following statement and sign your signature next to it. (macOS Preview and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) We want to make it *extra* clear so that no one inadvertently cheats.

   "I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."
3. At the start of each problem—not subproblem; you only need to do this twice—please state: (1) who you received help from on the problem, and (2) who you provided help to.
4. For all theory questions, make sure to **show your work** and include all key steps in your derivations and proofs. If you apply a non-trivial theorem or fact, you must state the theorem or fact you used, either by name (e.g.: Jensen's inequality) or by writing down the general statement of the theorem or fact. You may use theorems and facts covered in lecture without proof. If you use non-trivial theorems and facts not covered in lecture, you must prove them. All conditions should be checked before applying a theorem, and if a statement follows from a more general fact, the reason why it is a special case of the general fact should be stated. The direction of logical implication must be clearly

stated, e.g.: if statement A implies statement B, statement B implies statement A, or statement A is equivalent to statement B (i.e.: statement A is true if and only if statement B is true). You should **number** the equations that you refer to in other parts and refer to them by their numbers. You must **highlight the final answer or conclusion** in your PDF submission, either by changing the font colour to red or drawing a box around it.

5. For all programming questions, starter code is provided as a Jupyter notebook. You should work out your solution in the notebook. We recommend uploading your notebook to a hosted service like Google Colab to avoid the installation of dependencies and minimize setup time. When you are done, take a screenshot of your code and output and include it in your PDF. In addition, you should download the Jupyter notebook with your solutions (`File` → `Download` → `Download .ipynb`) and also download your code (`File` → `Download` → `Download .py`). You need to **submit both** as a zip archive named "`CMPT726-410_A1_⟨Last Name⟩_⟨Student ID⟩.zip`", whose directory structure should be the same as that of the zip archive for the starter code. Do **NOT** include any data files we provided. Please include a short file named **README** listing your name and student ID. The PDF should not be included in the zip archive and should be submitted **separately**. Please make sure that your code doesn't take up inordinate amounts of time or memory. If your code cannot be executed, your solution cannot be verified.

6. Assignment submissions will be accepted through Crowdmark. The submission portal will open a week before the assignment is due. Make sure to set aside at least **one hour** to submit the assignment. Crowdmark will ask you to split the PDF by each part of each question. Make sure to upload the correct pages for each part, and double check when you are done.

Python Configuration

We recommend using Google Colab to complete the parts of this assignment that require coding. However, if you would like to set up your own Python environment on your computer, follow the instructions below.

1. Ensure you have Python 3 installed. If you don't, we recommend miniconda as a package manager. To ensure you're running Python 3, open a terminal in your operating system and execute the following command: `python --version` **Do not proceed until you're running some (recent) version of Python 3.**

2. Install `scikit-learn` and `scipy`:
   `conda install -y -c conda-forge scikit-learn scipy matplotlib`

3. To check whether your Python environment is configured properly for this homework, ensure the following Python script executes without error. Pay attention to errors raised when attempting to import any dependencies. Resolve such errors by manually installing the required dependency (e.g. execute `conda install -c conda-forge numpy` for import errors relating to the numpy package).

```
import sys
if sys.version_info[0] < 3:
    raise Exception("Python 3 not detected.")

import numpy as np
import matplotlib.pyplot as plt
from scipy import io
```

4. Please use only the packages listed above.

Collaboration and Academic Integrity

While collaboration is encouraged, all of your submitted work must be your own. Concretely, that means you may discuss general approaches to problems with other students, but you must write your solutions on your own. For more information, see the course syllabus under "Academic Integrity." If you got help from or gave help to other students, please note their names at start of every question. Additionally, sign the declaration at the bottom of this page.

Warning: Copying others' solutions, seeking help from others not in this course, posting questions online or entering questions into automated question answering systems are considered cheating. Consequences are severe and could lead to suspension or expulsion. If you become aware of such instances, you must report them here: https://forms.gle/9fw3oMLyhD1A81qy5.

Page Numbers

Declaration

I certify that all solutions are entirely in my own words and that I have not looked at solutions other than my own. I have given credit to all external sources I consulted.

Your Name Here (Your Student ID #)

Question 1 Linear Algebra and Calculus [35 points]

List your collaborators on this question below.

Part 1(a) [5 points]

Let A be an $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$. **Prove $\lambda_1^k, \lambda_2^k, ..., \lambda_n^k$ are eigenvalues of A^k.** Justify each step that is not a simple algebraic manipulation.

Part 1(b) [5 points]

Let A be an $m \times m$ matrix. Consider a decomposition of $A = BCD^\top$, where C is a diagonal matrix. **If** $D^\top B = -2I$**, where** I **is the identity matrix, derive an expression for** A^n **in terms of** B**,** C **and** D **for any** $n \geq 1$**.** Simplify the expression as much as possible and show your work.

Part 1(c) [5 points]

Let A be an 2×2 matrix. Suppose the singular value decomposition of A is $I\Sigma V^\top$, where I is the identity matrix. Let $\vec{v}_1$ and $\vec{v}_2$ denote the right-singular vectors of A, which are associated with the singular values 5 and 2 respectively. Consider a vector $\vec{\beta} = 3\vec{v}_1 - \vec{v}_2$. **What is $A\vec{\beta}$?** Show your work.

Part 1(d) [5 points]

If A, B and BC are symmetric matrices, prove that $(ABC)^\top = BCA$. Justify each step that is not a simple algebraic manipulation.

Part 1(e) [5 points]

Let A be a symmetric matrix. **Prove that A can be expressed as $B - C$, where B and C are in the form $UD_1U^\top$ and $UD_2U^\top$ respectively, with U being orthogonal and D_1, D_2 being diagonal with strictly positive entries.** Justify each step that is not a simple algebraic manipulation.

Part 1(f) [5 points]

Derive a quadratic form $\begin{bmatrix} x \\ y \end{bmatrix}^\top A \begin{bmatrix} x \\ y \end{bmatrix}$ **that is equivalent to the quadratic function** $ax^2 + 2bxy + cy^2$**, where the off-diagonal entries of the matrix** A **should differ by exactly 10.** Show your work.

Part 1(g) [5 points]

Write down the second-order Taylor expansion of $f(x) = \sin x + xy + e^y$ **around** $x = 0$ **in matrix form.** Show your work.

Question 2 Probability [35 points]

List your collaborators on this question below.

Part 2(a) [5 points]

Find a 2×3 matrix A such that $\|A\|_F = \sqrt{8}$ and $\|A\|_2 = \sqrt{5}$, where $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_2$ denotes the spectral norm. Explain each step of your reasoning.

Part 2(b) [5 points]

You are given samples of a 2-dimensional random vector $\vec{X}$ in the `x.npy` file. In this part, we want to try to fit a distribution to this data. First, plot the samples as a scatter plot to check how the samples are distributed. Subsequently, you have to fit a Gaussian distribution to the data. Given samples $\{\vec{x}_i\}_{i=1}^N$ of a random vector $\vec{X}$, the true mean $\vec{\mu} := E[\vec{X}]$ and true covariance $\Sigma := E[(\vec{X} - E[\vec{X}])(\vec{X} - E[\vec{X}])^\top]$ of $\vec{X}$ can be estimated with the sample mean $\widehat{\vec{\mu}}$ and the (biased) sample covariance $\widehat{\Sigma}$, the formulas of which are below:

$$\widehat{\vec{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \vec{x}_i \tag{1}$$

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \left(\vec{x}_i - \widehat{\vec{\mu}}\right)\left(\vec{x}_i - \widehat{\vec{\mu}}\right)^\top \tag{2}$$

Find the sample mean and the sample covariance matrix by implementing the equations above on the given samples. You are not permitted to use the built-in Numpy functions for computing the sample mean and covariance, but you may use built-in Numpy functions for other elementary operations, such as summation, matrix multiplication, transposes, etc. **Then plot the contours of Gaussian distribution with the estimated parameters on the scatter plot of the samples. Include a screenshot of your code, a screenshot of the output and the plot of Gaussian contours in your PDF submission.**

Part 2(c) [5 points]

Then consider two other random vectors $\vec{Y} = A\vec{X} + \vec{b}$, and $\vec{Z} = C\vec{X} + \vec{d}$. **Find A and $\vec{b}$ such that the probability density of $\vec{Y}$ is the same as the probability density $\vec{X}$ rotated about the origin by 30 degrees counterclockwise and then translated by $[-1, -1]$, and C and $\vec{d}$ such that the probability density of $\vec{Z}$ is the same as the probability density $\vec{X}$ translated by $[-1, -1]$ and then rotated about the origin by 30 degrees counterclockwise. Plot the probability density of $\vec{Y}$ and $\vec{Z}$ as contour plots.** Show your work and justify why your answers achieve the desired effect.

Part 2(d) [5 points]

Suppose the random vector $\vec{X}$ is distributed according to a multivariate Gaussian with mean $\vec{0}$ and covariance matrix $AA^\top$. In other words, $\vec{X} \sim \mathcal{N}(\vec{0}, AA^\top)$. **Find a unit vector $\vec{b}$ such that the variance of the random variable $\left|\vec{b}^\top \vec{X}\right|$ is minimized, i.e.,** $\arg\min_{\vec{b}:\|\vec{b}\|_2=1} \mathrm{Var}\left(\left|\vec{b}^\top \vec{X}\right|\right)$**.** Show your work and justify why your answer is correct.

Hint 1: For any random variable U, $\mathrm{Var}(|U|) = \mathbb{E}[|U|^2] - \mathbb{E}[|U|]^2 = \mathbb{E}[U^2] - \mathbb{E}[|U|]^2 = \mathrm{Var}(U) + \mathbb{E}[U]^2 - \mathbb{E}[|U|]^2$
Hint 2: For any random variable U, $\mathrm{Var}(\sigma U) - \mathbb{E}[|\sigma U|]^2 = \sigma^2 \mathrm{Var}(U) - \sigma^2 \mathbb{E}[|U|]^2 = \sigma^2(\mathrm{Var}(U) - \mathbb{E}[|U|]^2)$

Part 2(e) [5 points]

40 people play a game in which they have to choose an integer between 1 and 100 inclusive, independently of others. **For a particular pair of people, what is the probability that they choose the same integer?** Show your work.

Part 2(f) [5 points]

Consider the same game as in the previous part. **Find the expected number of pairs of people that choose the same integer.** Show your work.

Hint: This expectation is of a sum of appropriately defined Bernoulli random variables.

Part 2(g) [5 points]

Consider the same game as in the previous two parts. **Find the probability that the number 50 is chosen by at least two people.** Show your work.