DSA5103 Assignment 1

Instructions While all languages are acceptable, it is recommended that you code using Python or MATLAB. You must write your own code. Your submission should be in softcopy including the code. Due on Feb 9, 11:59pm.

1. **Convexity**

   Show by definition that f is convex

   (1) $f(x) = x^2, x \in \mathbb{R}$, [1 mark]

   Solution. For any $x_1, x_2, \lambda \in [0,1]$, it holds that

   $$\begin{aligned}&\lambda f(x_1) + (1-\lambda)f(x_2) - f(\lambda x_1 + (1-\lambda)x_2)\\ &= \lambda x_1^2 + (1-\lambda)x_2^2 - (\lambda x_1 + (1-\lambda)x_2)^2 = \lambda(1-\lambda)(x_1 - x_2)^2 \geq 0.\end{aligned}$$

   By definition, f is convex.

   Alternative solution. The Hessian of f is positive definite: $H_f(x) = 2 > 0$. Therefore, f is convex.

   (2) $f(x) = x_1^2 + x_2^2 + 2x_1 + 4, x = (x1; x2) \in \mathbb{R}^2$. [1 mark]

   Solution. For any $x = (x_1; x_2), y = (y_1; y_2), \lambda \in [0,1]$, it holds that

   $$\begin{aligned}&\lambda f(x) + (1-\lambda)f(y) - f(\lambda x + (1-\lambda)y)\\ &= \lambda(1-\lambda)(x_1 - y_1)^2 + \lambda(1-\lambda)(x_2 - y_2)^2 \geq 0.\end{aligned}$$

   By definition, f is convex.

   Alternative solution. The Hessian of f is positive definite: $H_f(x) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \succ 0$. Therefore, f is convex.

2. **Matrix inverse**

   (1) Generate an $n \times n$ random matrix X from whatever distribution (e.g., every entry X_{ij} is uniformly distributed on $[0,1]$). [1 mark]

   Solution. In MATLAB

```
n = 10;
X = rand(n,n);
```

(2) Prove by definition that $A = I + X^T X$ is symmetric and positive definite (Here I denotes the identity matrix). [1 mark]

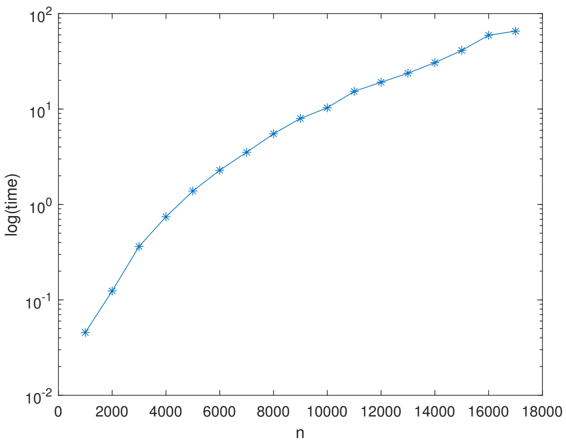
Solution. A is symmetric since $A^T = (I + X^T X)^T = I^T + X^T X = A$. For any nonzero vector y, $y^T Ay = y^T(I + X^T X)y = y^T y + (Xy)^T(Xy) = \|y\|_2^2 + \|Xy\|_2^2 > 0$. Therefore, A is positive definite.

(3) For different values of $n = 1000, 2000, 3000, 4000, 5000, 6000, \ldots$, measure the time of inverting the matrix A: $B = A^{-1}$ (e.g., measure by "tic" "toc" function). Plot a figure of $\log_{10}(\text{time})$ against n. [2 marks]

Solution. (Note that here one should invert A: $A^{-1} = (I + X^T X)^{-1}$ instead of X^{-1})

In MATLAB, I allow one minute (alternatively, one may use 30 minutes or 1 hour) for this inverse computation.

```
maxtime = 60; % 1 min
n_list = 1000:1000:20000;
inv_time = zeros(1,length(n_list));
for i = 1:length(n_list)
    n = n_list(i);
    X = rand(n,n);
    A = eye(n) + X'*X;
    tic;
    B = inv(A);
    inv_time(i) = toc;
    fprintf('n=%d,time=%3.1fsec\n',n,inv_time(i));
    if inv_time(i) > maxtime
        break;
    end
end
semilogy(n_list(1:i),inv_time(1:i),'-*');
xlabel('n');
ylabel('log(time)');
```



(4) What is the largest n that your device can handle? How long it takes to invert this large $n \times n$ matrix A? [1 mark]

Solution. If I only allow one minuter for this inverse computation, the largest $n = 16000$. It takes 59.4 seconds for inverting A.

3. **Principal component analysis** on the iris data set (see `http://archive.ics.uci.edu/ml/datasets/Iris` for data description.) It contains 3 species (Setosa/Versicolor/Virginica) of 50 instances each ($n = 150$), and $p = 4$ features (sepal length, sepal width, petal length, petal width)

(1) Download the iris data set from the above link, or load it directly in Python

```
from sklearn import datasets
iris = datasets.load_iris()
```

or in MATLAB

```
iris = iris_dataset;
```

(2) Transform each feature to the same scale (subtracting the mean and dividing by the standard deviation). Report in Table 1 the first three samples after feature scaling. [1 mark]

sepal length	sepal width	pedal length	petal width	species
				setosa
				setosa
				setosa

Table 1

Solution.

sepal length	sepal width	pedal length	petal width	species
-0.8977	1.0286	-1.3368	-1.3086	setosa
-1.1392	-0.1245	-1.3368	-1.3086	setosa
-1.3807	0.3367	-1.3935	-1.3086	setosa

(3) Compute the $p \times p$ covariance matrix and report the its numerical values

$$\Sigma = \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix}.$$

[1 mark]

Solution.

$$\Sigma = \begin{bmatrix} 1 & -0.1094 & 0.8718 & 0.8180 \\ -0.1094 & 1 & -0.4205 & -0.3565 \\ 0.8718 & -0.4205 & 1 & 0.9628 \\ 0.8180 & -0.3565 & 0.9628 & 1 \end{bmatrix}$$

(4) Compute eigenvalue decomposition of $\Sigma = QDQ^T$. Report the numerical values of Q and D

$$Q = \begin{bmatrix} \quad & \quad \end{bmatrix}, \quad D = \begin{bmatrix} \quad & \quad \end{bmatrix}.$$

[2 marks]

Solution. (Note that the eigenvectors, i.e., columns of Q may differ by a negative sign.)

$$Q = \begin{bmatrix} 0.5224 & 0.3723 & 0.7210 & -0.2620 \\ -0.2634 & 0.9256 & -0.2420 & 0.1241 \\ 0.5813 & 0.0211 & -0.1409 & 0.8012 \\ 0.5656 & 0.0654 & -0.6338 & -0.5235 \end{bmatrix}, D = \begin{bmatrix} 2.9108 & & & \\ & 0.9212 & & \\ & & 0.1474 & \\ & & & 0.0206 \end{bmatrix}$$

(5) Give the first principal component. Recast the data onto the first principal component and plot the results. [2 marks]

Solution. The first principal direction is

$$\begin{bmatrix} 0.5224 \\ -0.2634 \\ 0.5813 \\ 0.5656 \end{bmatrix} \text{ or } - \begin{bmatrix} 0.5224 \\ -0.2634 \\ 0.5813 \\ 0.5656 \end{bmatrix}$$

and the recasted data is plotted in Figure 1 (or symmetric with respect to the original point).

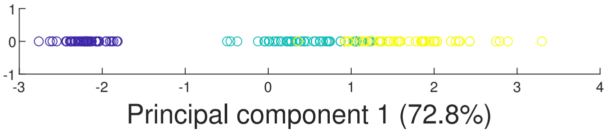


Figure 1: Setosa (blue), Versicolor (green), Virginica (yellow)

(6) Give the first two principal components. Recast the data onto the first two principal components and plot the results. [2 marks]

Solution. The second principal direction is

$$\begin{bmatrix} 0.3723 \\ 0.9256 \\ 0.0211 \\ 0.0654 \end{bmatrix} \text{ or } - \begin{bmatrix} 0.3723 \\ 0.9256 \\ 0.0211 \\ 0.0654 \end{bmatrix}$$

and the recasted data is plotted in Figure 2 (Principal components may be plotted in a symmetric way with respect to the original point since the principal directions may differ by a negative sign).

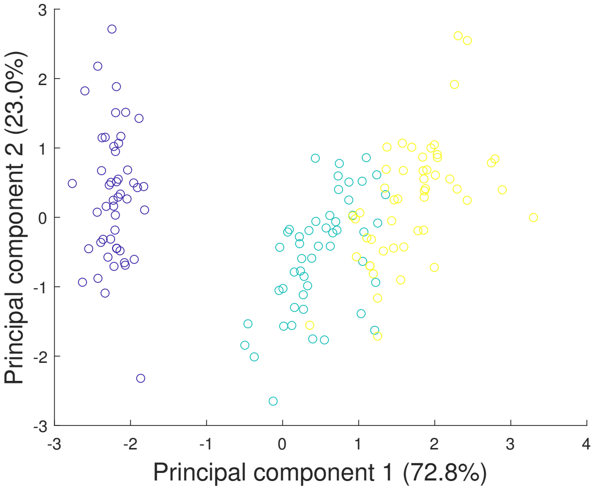


Figure 2: Setosa (blue), Versicolor (green), Virginica (yellow)

(Lastly, the iris data from different libraries may differ slightly. Therefore, the answers in Q3 may differ slightly.)