Math 528: Mathematical Foundations of Data Science and Machine Learning

Homework 2: DUE Monday 2/16/2026 by 11:59 PM

Remember that you MAY use AI/Large-language models to help you, but you MUST document the prompts used and responses provided. You may use any common coding language (MATLAB, Python, Julia, R, or C++).

Formatting Requirements

For this homework assignment, the following should be provided:

(1) If using an AI tool, **the prompts and response should be documented at the end of the homework assignment. Failure to document this when using AI tools will result in a grade of a 0.**
(2) The answers and responses to questions may be typed. Plots and figures corresponding to answers to the question should be provided in the homework.

Problem A

Download the script "HW2_a" from Blackboard in either MATLAB or Python, as well as the zip file "HW2_data." In this problem, you will assess the efficiency and applicability of the five main factorization/decompositions we have discussed: LU, QR, Cholesky, Eigen decomposition, and the SVD. Each datafile contains a matrix, A, which varies in size. We will focus on solving $Ax = b$.

(i) For each dataset (1-7), first determine the properties of A (size, symmetric, positive definite). The code will help you do this. Document this in your homework response.
(ii) Based on what you find in (i), what decompositions will you be able to apply to each dataset? Change the variable "methods" in the code to reflect this for each dataset.
(iii) Use the code to solve the linear system, starting with $\boldsymbol{x} = \mathbf{1}$, i.e. the ones vector. Provide details on the error in finding $\boldsymbol{x}$, the residual $\boldsymbol{Ax} - \boldsymbol{b}$, and the time it takes to compute (all provided by the code). Comment on the general behavior of certain algorithms and which might be more effective.
(iv) Now, make $\boldsymbol{x}$ an nx5 vector (can be random numbers or the 1's vector). How does the performance of the algorithms change? Focus on datasets 5, 6, and 7.

Problem B

We will now revisit image analyzing using the SVD. Download the folder "mnist_4" from Blackboard, which contains 12 images of a handwritten digit "4". Each image is 28x28. Use the Blackboard code "HW2_b" for your analysis.

(i) Load each image into MATLAB or Python. This image we can treat as a 28x28 matrix A. Compute the SVD of each image. Plot the singular values (remember that they are on the diagonal), and comment on whether the singular values quickly or slowly decay.

(ii) Recall again that the SVD can be rewritten as

$$\boldsymbol{A} = \boldsymbol{U\Sigma V}^{\top} \Rightarrow \boldsymbol{A} = \sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{\sigma}_i \boldsymbol{v}_i^{\top}$$

where here $n = 28$. Using the MATLAB or Python script provided, investigate how many terms in the sum are necessary to reconstruct each digit. Comment on whether there is some "global minimum" number of singular values needed. Provide plots that illustrate how the digits are reconstructed fairly well for only $k < n$ components.

Problem C

Download the script "HW2_c", which loads the same digit images but now stores them as a $28^2 = 784$ dimensional vector, and then constructs a matrix, X with each image along the columns. Consider the matrix

$$A = X\,X'$$

which is the outer product of the two matrices. This should be 784x784. We will revisit this later as the **sample covariance** of our data. The following operations can then be performed on A

- Taking the average over the second dimension of A provides an "average" dataset. If we cast this average back into a 28x28 matrix, we should get the **average image, $\bar{\boldsymbol{X}}$.**
- We can center the data in $\boldsymbol{A}$ so that the mean is zero by subtracting, i.e. $\boldsymbol{A}^* = \boldsymbol{A} - \bar{\boldsymbol{X}}$**.**
- This matrix is now **symmetric positive definite.** This means that the SVD of this image, $\boldsymbol{A}^* = \boldsymbol{U\Sigma V}^{\top}$**,** contains similar information to the eigen decomposition of the original matrix $\boldsymbol{X}$.

(i) Use the code to visualize the "mean" image. What do you see?

(ii) Visualize the components of $\boldsymbol{U}$**.** What do you see?