Instructions

• Please read lecture notes and textbook. Ask questions during office hours if you need help.

• Please write neat solutions for the problems below. Show all your work. If you only write the answer with no work you will not be given any credit.

• You are required to write your answers clearly. You get 10% bonus points if you use latex to type your solutions. A chosen subset of the problems will be graded.

• Please submit your homework by 11:00 PM, Tuesday, Mar 10, 2020 (Shanghai time) via email. Late submission will not be accepted.

1. [Exercise 3.27, FoDS] Pick a photo you like and convert it into a matrix. If it is a color image, you may convert it into a gray-scale image. Perform the compact version of singular value decomposition of the matrix. Reconstruct the photo using only 1%, 5%, 10%, 25%, and 50% of the singular values.

   (a) Present the four pictures and comment your result.
   (b) What percentage of information is captured by these low-rank approximation under Frobenius norm?

   (Hint: When you input your image, it is possibly in uint8; you may first convert it into double format; after applying SVD and performing the reconstruction, convert it back to uint8.)

2. Principal component analysis: Download MNIST test dataset from [here](#) or from NYU Class. Note that the first column denotes the label of each sample \( \{0, 1, \cdots, 9\} \) and there are in total 10000 samples. Each row (except the first entry) is a vectorized image of size \( 28 \times 28 \). Apply PCA to this dataset.

   (a) Get the singular values of the centered dataset and plot them. Does it decay fast?
   (b) Let \( \alpha_i \in \mathbb{R}^k \) be the coefficient vector of the \( i \)th centered data \( x_i - \overline{x} \) projected on the first \( k \) components. Pick the first two principal components, i.e., \( k = 2 \). Plot the compressed data \( \alpha_i \) in 2D. Use different colors for each label.
   (c) Pick the first three principal components. Compute \( \alpha_i \in \mathbb{R}^k \) for \( k = 3 \). Plot the compressed data \( \alpha_i \) in 3D. Use different colors for each label.
   (d) Do you think PCA provides a satisfactory data visualization of MNIST dataset? Comment you results in (b) and (c).

3. High dimensional geometry

   (a) What is the volume of the largest hypercube contained in the \( \ell_2 \)-norm unit ball in \( \mathbb{R}^d \)?
   (b) Let \( x \) and \( y \) be two symmetric Bernoulli random vectors in \( \mathbb{R}^d \), i.e., \( \{x_i\}_{i=1}^d \) and \( \{y_i\}_{i=1}^d \) are independent symmetric Bernoulli random variables. Show that \( x \) and \( y \) are nearly orthogonal for large \( d \).