CS 391L: Machine Learning **Summer 2023**

Homework 3 - Theory

Lecture: Prof. Adam Klivans

Keywords: SVD, PCA

Instructions: Please either typeset your answers (LaTeX recommended) or write them very clearly and legibly and scan them, and upload the PDF on edX. Legibility and clarity are critical for fair grading.

1. [**10 points**] Let A be an $m \times d$ matrix, and let $X = AA^T$. Assume that X has d distinct, non-zero eigenvalues. Assume that $m \gg d$. In order to find the eigendecomposition of X, we will need to find the eigendecomposition of an $m \times m$ matrix. Since m is much larger than d, this is slow. Give an algorithm for finding the eigenvectors and eigenvalues of X that only requires computing the eigendecomposition of a $d \times d$ matrix. You can use simple matrix operations and assume that you have an eigendecomposition "black box" subroutine, but avoid using the SVD as a black box.

2. In this problem we explore some relationships between SVD, PCA and linear regression.

   (a) [**2 points**] True or false: linear regression is primarily a technique of *supervised* learning, i.e. where we are trying to fit a function to labeled data.

   (b) [**2 points**] True or false: PCA is primarily a technique of *unsupervised* learning, i.e. where we are trying to find structure in unlabeled data.

   (c) [**2 points**] True or false: SVD is primarily an operation on a *dataset* whereas PCA is primarily an operation on a *matrix*.

   (d) [**2 points**] A common problem in linear regression is *multicollinearity*, where the input variables are themselves linearly dependent. For example, imagine a healthcare data set where height is measured both in inches and centimetres. This is a problem because there may now be multiple w satisfying $y = w \cdot x$. Explain how you could use a preprocessing step to solve this problem.