Homework 2, CS 4220, Spring 2021
Instructor: Austin R. Benson
Due Friday, March 12, 2021 at 3:44pm ET on CMS (before lecture)

## Policies

**Submission.** Submit your write-up as a single PDF on CMS: `https://cmsx.cs.cornell.edu`.
**Coding questions.** You can use any programming language for the coding parts of the assignment. Include your code in your write-up.
**Typesetting.** Your write-up must be typeset with LaTeX.
**Collaboration.** Please discuss and collaborate on the homework, but you have to write your own solutions and code.
**Resources and attribution.** Feel free to use any resources that might be helpful, and provide attribution for any key ideas. We only ask that you work on the problems in earnest. Please do not hunt for solutions with a search engine.

## Problems

1. *Well-posed, ill-conditioned problems are close to ill-posed problems.*
   Let $A \in \mathbb{R}^{n \times n}$ be nonsingular. Show that the inverse of the condition number $\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$ is the relative distance to the nearest singular matrix in the 2-norm, i.e.,

$$\frac{1}{\kappa_2(A)} = \min_{\text{rank}(B) < n} \frac{\|A - B\|_2}{\|A\|_2}.$$

2. *Fast LU updates.*
   Let $A \in \mathbb{R}^{n \times n}$, and suppose that we have already computed a factorization $PA = LU$. Let $w, v \in \mathbb{R}^n$ and $z \in \mathbb{R}$, and consider the matrix

$$M = \begin{bmatrix} A & v \\ w^T & z \end{bmatrix}.$$

   Design an algorithm to compute the factorization $\bar{P}M = \bar{L}\bar{U}$ in $O(n^2)$ time.

   When might this be useful?

3. *Sparse substitution.*
   Suppose that $L$ and $U$ are both nonsingular sparse $n \times n$ matrices stored in compressed sparse column format, where $L$ is lower triangular and $U$ is upper triangular.

   Given $b$, show how to solve $LUx = b$ in $O(\text{nnz}(L) + \text{nnz}(U))$ time.

4. *Regularized least squares.*
   Let $A \in \mathbb{R}^{m \times n}$. Show that the solution to the $l_2$-regularized least squares problem

$$\arg\min_x \|Ax - b\|_2^2 + \beta \|x\|_2^2 \tag{1}$$

   for $\beta > 0$ is equivalent to the solution of the least squares problem

$$\arg\min_x \left\| \begin{bmatrix} A \\ \sqrt{\beta} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2. \tag{2}$$

Equation (1) tries to find an approximate least squares solution $x$ that is also small in the sense of the 2-norm. The formulation works even if $m < n$. This is a type of regularization that we will discuss more in class and use in the next question.

5. *Matrix completion.*
   Let $A \in \mathbb{R}^{m \times n}$ with $m > n$. Suppose that we want a low-rank approximation to $A$, i.e., $A \approx WZ^T$ for $W \in \mathbb{R}^{m \times k}$, $Z \in \mathbb{R}^{n \times k}$. We saw how to use the SVD for this problem, but we will explore a QR-based method in this question.

   (a) Suppose that you are given the factor $W$. Explain how to efficiently solve

   $$\operatorname*{arg\,min}_{Z \in \mathbb{R}^{n \times k}} \|A - WZ^T\|_F^2 \tag{3}$$

   with a QR factorization of $W$. Similarly, suppose $Z$ is fixed and explain how to solve

   $$\operatorname*{arg\,min}_{W \in \mathbb{R}^{m \times k}} \|A - WZ^T\|_F^2 \tag{4}$$

   given a QR factorization of $Z$.

   This leads to an alternating algorithm (often called *alternating least squares*) where we repeatedly (i) solve for $Z$ in eq. (3) and (ii) update $W$ via eq. (4) given the new factor $Z$.

   (b) Now suppose that $A$ is sparse and that we only care about the error in our low-rank factorization for the non-zero entries in $A$. This scenario arises if the "zeros" in $A$ correspond to *unknown* values as opposed to true zeros.

   Let $\Omega = \{(i, j) \mid A_{ij} \neq 0\}$. Our objective is

   $$\operatorname*{arg\,min}_{W \in \mathbb{R}^{m \times k}, Z \in \mathbb{R}^{n \times k}} \|A - WZ^T\|_{F,\Omega}^2 + \beta\|W\|_F^2 + \beta\|Z\|_F^2, \tag{5}$$

   where $\|M\|_{F,\Omega}^2 = \sum_{(i,j) \in \Omega} M_{ij}^2$. This problem is often called *matrix completion* because one can "complete" the matrix $A$ at unknown entries $(i, j)$ with estimates $A_{ij} \approx w_i^T z_j$, where $w_i$ and $z_j$ are the $i$th and $j$th rows of $W$ and $Z$. Design an alternating least squares algorithm similar to the one in part (a) for finding an approximate solution to eq. (5).

   (c) Implement your algorithm from part (b) using built-in library least squares solvers.

   The file at `https://drive.google.com/uc?id=196W2kDoZXRPjzbTjM6uvTidn6aTpsFnS` is a dataset of 1,378,033 ratings of 25,475 books from 18,892 users on Goodreads. The dataset is described at `https://cseweb.ucsd.edu/~jmcauley/datasets.html#spoilers`. Download the data and write code to read the data to create a sparse matrix $A \in \mathbb{R}^{25{,}475 \times 18{,}892}$, where $A_{ij}$ is the rating user $j$ gave to book $i$ (if $j$ rated $i$), and $A_{ij} = 0$ otherwise.

   Run your algorithm on this matrix with $k = 32$ and $\beta = 10^{-4}$ for 50 alternating steps to find low-rank factors $W$ and $Z$.

   (d) The rows of $W$ give $k$-dimensional vectors corresponding to the books. We will call these vectors *embeddings* and explore some of their basic structure.

   You can use the "book_id" key in the dataset to look up the identity of a book, using the URL `https://www.goodreads.com/book/show/BOOK_ID`, where BOOK_ID is the integer key. For example, book_id=119322 corresponds to the book at `https://www.goodreads.com/book/show/119322`.

   List the five books whose embeddings are closest to the embedding of the book with book_id=2, where closeness is measured by 2-norm distance. You should find that the books are related in some sense.