Please provide complete and well-written solutions to the following exercises.

Due September 22, 1159PM PST, to be uploaded as a single PDF document to blackboard (under the Assignments tab).

Homework 5

Exercise 1. In Google's PageRank Algorithm, and in many other applications of linear algebra in mathematics, computer science, statistics, data science, etc., we often need to find the largest eigenvalue of a symmetric matrix. Later in the course, we will discuss efficient (fast) ways to find the largest eigenvalue of a matrix. In this exercise, we will examine a naïve, slow, inefficient algorithm for this task.

Let's first generate a large symmetric random matrix. In applications, matrices often have many zero entries. For example, in order to rank n websites on the internet according to their relevance, Google considers a matrix A with an entry A_{ij} equal to 1 when website $1 \leq i \leq n$ and website $1 \leq j \leq n$ share a hyperlink (i.e. one of these websites has a hyperlink pointing to the other one). (Otherwise, A_{ij} is set equal to zero.) With so many websites, we might need to take n of size about 10^9. One website will often not have many hyperlinks, so many entries of A will be zero. With this in mind, let's start with a large matrix with a lot of zeros (with $n = 100$). Define

```
A=round(.52*rand(100,100));
A=A+A';
```

(The last command ensures that A is a symmetric matrix; also we will not be concerned with entries of 2 appearing in A, for the purpose of this exercise.)

Since A is a real symmetric matrix, an eigenvalue $x \in \mathbf{R}$ of A satisfies the equation

$$\det(A - xI) = 0,$$

where I is the identity matrix (with the same dimensions as A). We can define such a function in Matlab with the command

```
f = @(x) det(A-x*eye(100));
```

- Using any method you wish to use, try to find the largest zero of the function f we just defined. That is, find the largest value of $x \in \mathbf{R}$ such that $f(x) = 0$. (For example, you could try either divide and conquer, or Newton's method.) Are you able to find a zero of f? If not, why?
- Does the answer you found agree with the output of the command `eigs(A)` ?

Exercise 2. In modern applications of linear algebra, the vectors and matrices we encounter are very large. Dealing with a large number of dimensions is difficult. Thankfully, there are methods for reducing the number of dimensions, while preserving some of the structure of the vectors. To illustrate this fact, make a few vectors, x, y with the commands

```
x=round(rand(1000,1));
y=round(rand(1000,1));
```

The lengths of these vectors can be examined with the command `norm(x)`, which might be around 22 (so the Euclidean distance from the vector x to the origin should be around 22.)

To reduce the number of dimensions, let's define the matrix

```
A=(1/sqrt(10))*randn(10,1000);
```

(One entry of the `rand` command outputs a floating point number with a probability of being in the interval (a,b) approximately equal to $b-a$ when $0<a<b<1$. In contrast, one entry of the `randn` command outputs a floating point number with a probability of being in the interval (a,b) approximately equal to $\frac{1}{\sqrt{2\pi}}\int_a^b e^{-s^2/2}ds$ when $-\infty<a<b<\infty$ are floating point numbers (in double precision).)

- Find the absolute error between the norms of x and Ax. Are these quantities close?
- Find the absolute error between the norms of $x-y$ and $Ax-Ay$. Are these quantities close?

The matrix A can almost preserve the lengths and distances between many vectors. Note that Ax is a 10-dimensional vector, and 10 is much smaller than 1000 (the original dimension of the vectors). We will not further explain in this course why A has the properties we just observed.

Exercise 3. Let

$$A=\begin{pmatrix} 0 & 1 & 1 & 0 & 2\\ 2 & 3 & 0 & 0 & 0\\ 4 & 5 & 1 & 0 & 2\\ -6 & 0 & 1 & 2 & 0\\ 3 & 0 & 4 & 0 & -1\end{pmatrix}.$$

Using the LU decomposition of A, solve the equation

$$Ax=b,$$

using Matlab code (without using any built-in linear algebra solvers) where $b=(2,4,3,1,5)^T$. (Note that solving a linear system such as $Ly=b$ when L is lower triangular should be relatively straightforward, by e.g. first solving for y_1, then y_2, and so on.)

Exercise 4. Write a computer program on your own that finds the LU factorization of the matrix

$$A=\begin{pmatrix} 6 & 0 & -4 & 0\\ 0 & 7 & 0 & -1\\ -4 & 0 & 6 & 0\\ 0 & -1 & 0 & 7\end{pmatrix}.$$

(Hint: note that A is symmetric. You can use the command `eigs` to check also that A is positive definite.)

Exercise 5. This exercise examines the eigenvalues of random matrices. We can make a large random real symmetric matrix with entries in $\{-1, 1\}$ with the following program.

```
n=1000;
iterations=100;
eiglist=zeros(n,iterations);
for i=1:iterations
    A= zeros(n,n);
    A( ((1:n)')*ones(1,n) < (ones(1,n)')*(1:n) )= 2*round(rand(1,n*(n-1)/2)) -1;
    A =  A+A';
    A( ((1:n)')*ones(1,n) == (ones(1,n)')*(1:n) )= 2*round(rand(1,n )) -1;
    eiglist(:,i)=eig(A);
end
eiglist=reshape(eiglist,1,n*iterations);
hist(eiglist/sqrt(n),100);
```

We generate several random matrices and then make a histogram of a list of the eigenvalues of all of the matrices.

What curve does the histogram resemble?

What numbers $a, b \in \mathbf{R}$ tend to satisfy $a \leq \lambda \leq b$ for all eigenvalues λ of the given matrices? (Choose the largest a and smallest b possible.)

For a random non-symmetric matrix, its eigenvalues will be complex, so we can instead plot these eigenvalues as dots in the complex plane.

```
n=1000;
A=2*round(rand(n,n)) -1;
eiglist=eig(A)/sqrt(n);
plot(real(eiglist),imag(eiglist),'.');
```

In what region of the complex plane do the eigenvalues of the matrix tend to reside?

If you modify this program so that it makes a list of the eigenvalues of 100 different matrices (of the same type used in the program), how does the plot change?

Exercise 6. Let A be an $m \times n$ matrix with nonnegative entries. A **nonnegative matrix factorization** for A with k classes is a factorization of the form

$$A = WH,$$

where W is an $m \times k$ matrix, H is a $k \times n$ matrix, and both W, H have nonnegative entries. Sometimes writing a factorization in this way is impossible. (If a factorization exists like this, then A must have rank at most k.) However, we can still try to find W, H that approximately satisfy $WH \approx A$. This is exactly what the Matlab function `nnmf` does. (More specifically, Matlab tries to find W, H that minimize `norm(A-W*H)`, and it uses randomness to do this, so we will ask Matlab to output the best factorization among 1000 attempts.)

Nonnegative matrix factorization is used in several machine learning applications, e.g. to cluster data into similar groups, in recommendation algorithms, etc. To illustrate this, let's consider the matrix A whose entries are the numbers in the following table

	apple	banana	bell pepper	crab	broccoli	carrot	pear	shrimp
calories	130	110	25	100	45	30	100	100
sodium	0	0	40	330	80	60	0	240
potassium	260	450	220	300	460	250	190	220
carbohydrates	34	30	6	0	8	7	26	0
vitamin A	2	2	4	0	6	110	0	4
vitamin C	8	15	190	4	220	10	10	4

- Verify that A has rank 6 with the command `rank(A)`, so that we know for sure we cannot write $A = WH$ exactly with $k = 3$.
- Find an approximate nonnegative matrix factorization of A with 3 classes with the Matlab command `[W,H]=nnmf(A,3,'replicates',1000)`. Do the matrices W, H satisfy $A = WH$? If not, check the value `norm(A-W*H)` and compare it with `norm(A)`.
- Each of the three rows of H corresponds to a different class of food. The largest entry in a column of H sorts the food into a given class. For example, the first row of H seems to correspond to "fruits," since the columns for apple, banana, and pear all have largest values in their top entries. (Carrot also seems to have a largest value here even though it is not a fruit.) What classes of foods do the other two rows of H seem to represent, and which food items are in those classes according to H?
- Each column of W also corresponds to a different class of food (like the rows of H). The largest entry in a row of W indicates which food characteristics are most important for being in each class. For example, calories, potassium, carbohydrates and vitamin A have their largest entries in the first column of W, so these four characteristics are the most significant contributions to being in the class of "fruits" in this table. (Carrot is the only one with a large value of vitamin A so it is unclear why exactly it got sorted in to the class of "fruits.") What food characteristics are most important for the other two classes of foods, according to W?
- When $k = 4$ instead of 3, is the carrot still in the same class as the apple, banana and pear?