Do all four of the following problems. Show all your work. Answers without supporting work may receive no credit.

Students may discuss the exam only with the instructor and the teaching assistant. No discussion of the exam contents, directly or indirectly, is permitted among students or with any third parties. Any book or internet resource may be used, as long as the book or the website are cited, along with the material taken from it.

You may use any numerical software available, unless you are specifically instructed in the problem statement to write your own code. All codes that are written by you should be turned in with the exam as an executable program (e.g. as a Matlab script). Numerical results without the code that produced them will receive no credit.

I attest that I have completed this exam without unauthorized assistance from any person, materials, or device:

Full name: 

Signature: 
Problem 1. The following are hexadecimal representations of IEEE standard floating-point single-precision numbers:

(a) c1f80cdb       (b) 3fddb3d7
(c) 3e0a9555       (d) bf13c468

Find their single-precision representations as ordinary decimal (base-10) numerals. The calculations are similar to those used in the in-class program hexadecimal_demo.m for double-precision numbers except that for single-precision the binary (base-2) representations of the numbers must be used. You may convert the above hexadecimal representations to binary by using the code hexconvert.m available on the course website, which will convert the string of eight hexadecimal digits to 32-bit strings of 0’s and 1’s. You must find the sign $\sigma$, the exponent $E$, and the fraction $F$ of each of the numbers and use them to calculate the decimal representation in single precision.
Problem 2. (a) Calculate the condition number $K_f(x)$ for evaluating the function $f(x) = \sin x$ and plot $K_f(x)$ over the range $0 < x < 2\pi$. For what $x$-values is the given function difficult to evaluate accurately in this range?

(b) Evaluate $f(x) = \sin x$ at $x = \pi - 10^{-8}$ directly in double precision arithmetic. Do you believe that the result is accurate? If not, find a method to calculate the function at this value to double precision accuracy and use this accurate evaluation to estimate the error in the direct evaluation.

(c) Repeat part (a) for the function $f(x) = \cos x$.

(d) Repeat part (b) for the function $f(x) = \cos x$ and $x = (\pi/2) - 10^{-8}$.
Problem 3 (a) The problem of solving the linear system

\[ Ax = b \]

in finite precision arithmetic gets some error \( \delta x \) in its solution from any error \( \delta b \) in the input vector, which are related by

\[ A(\delta x) = \delta b. \]

Prove that if \( A \) is invertible, then the ratio of the relative errors in the worst case is given by

\[ \sup_{b \neq 0, \delta b \neq 0} \frac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|} = \|A\| \cdot \|A^{-1}\|, \]

where the supremum gives the largest error ratio for all choices of \( b, \delta b \) and the matrix norms are those induced by the given vector norm.

(b) Consider the linear system \( Ax = b \) for the specific choices

\[
A = \begin{pmatrix}
17 & -864 & 716 & -799 \\
1 & -50 & 0 & 0 \\
0 & 1 & -50 & 0 \\
0 & 0 & 1 & -50
\end{pmatrix}, \quad b = \begin{pmatrix}
1501 \\
0 \\
-49 \\
51
\end{pmatrix}.
\]

Solve this problem by using Gaussian elimination numerically in double precision arithmetic with the book algorithm ALG061 and state the approximate solution \( \hat{x} \).

(c) Comparing with the exact solution \( x = (50, 1, 1, -1)^T \) of the problem in (b), calculate the relative error \( \|\delta x\|_1/\|x\|_1 \) of the approximate solution in (b), using the vector 1-norm. Is \( \hat{x} \) accurate to double precision?

(d) Calculate \( \|A\|_1 \cdot \|A^{-1}\|_1 \) for the matrix \( A \) in part (b), using the induced matrix 1-norm. Is the result consistent with the relative error found in part (c) using double-precision arithmetic? Explain why or why not.
Problem 4. (a) It was shown in the class lectures that the spectral radius $\rho(A)$ of a square matrix $A$ is not a matrix norm. Explain why.

(b) Show in addition that $\rho(A)$ is not sub-multiplicative; that is, find explicit matrices $A, B$ such that $\rho(AB) > \rho(A)\rho(B)$.

(c)* Show that the spectral radius can, however, be obtained as the limit

$$\rho(A) = \lim_{k \to \infty} \|A^k\|^{1/k}$$

for any consistent matrix norm.

Hint: You should be able to show easily that $\rho(A) \leq \|A^k\|^{1/k}$ for all $k \geq 0$. Show in addition that, for any $\epsilon > 0$,

$$\|A^k\|^{1/k} \leq \rho(A) + \epsilon, \quad \text{for all } k > N_{A,\epsilon}$$

with $N_{A,\epsilon}$ depending upon both $A$ and $\epsilon$.

(d) Based on the result in (c), write a code to approximate

$$\rho(A) \doteq \rho_k(A) := \|A^k\|^{1/k}_{\infty}$$

with matrix $\infty$-norm and with $k$ chosen so that $(\rho_{k-1}(A) - \rho_k(A))/\rho_k(A) < TOL$. Use your code to approximate $\rho(A)$ for the specific matrix

$$A = \begin{pmatrix}
21 & -3 & -7 & 1 & -12 & -9 & 12 \\
-8 & -2 & 17 & -24 & 26 & 15 & -16 \\
14 & -8 & 1 & -7 & -10 & -2 & 10 \\
8 & -12 & 8 & -14 & -8 & 4 & 8 \\
-18 & -12 & 22 & -24 & 10 & 20 & -8 \\
6 & -10 & 18 & -31 & 16 & 13 & -6 \\
-32 & -18 & 45 & -54 & 38 & 41 & -26
\end{pmatrix}$$

with $TOL = 2 \times 10^{-5}$. Compare with the result using the Matlab command

$$\text{>> rho(A)=max(abs(eig(A)))}$$

Comment: In Chapter 9 we shall consider a closely related but more accurate method called the “power method” to evaluate the leading eigenvalue.