Solution of Homework 1

Note. This homework is due in class on January 30, 2026. Please show detailed steps to explain how you get your solution. Simply providing a final answer may not warrant full credit.

Question 1 [43503/53903] The decimal representation of the numbers in a floating point system is specified by

$$(\beta, t, L, U) = (10, 4, -2, 1).$$

(a) Explain what are β, t, L, and U.

(b) Find the largest possible number in the system.

(c) Find the smallest positive number in the system.

(d) Find the total different fractions and exponents in the system.

(e) Find the total different numbers in the system.

Solution. For a floating point system (β, t, L, U), parameter β denotes the base of number system, parameter t denotes precision (number of digits), parameter L denotes the lower bound on exponent, and parameter U denotes the upper bound on exponent.

- Largest number is 99.99
- Smallest positive number is 0.01
- Total different (normalized) fractions: 9000. Total different exponents: 4
- Total different numbers in system: 72001

Question 2 [43503/53903] How would you perform the following calculations to avoid cancellation? Justify your answers.

(a) Evaluate $\sqrt{x+1} - 1$ for $x \simeq 0$.

(b) Evaluate $\dfrac{1}{\cos^2 x - \sin^2 x}$ for $x \simeq \dfrac{\pi}{4}$.

Solution.

(a) For $x = 0$, we have $\sqrt{x+1} - 1 = 0$. This indicates that when x is close to 0 we will get cancellation. We can evaluate

$$\frac{x}{\sqrt{x+1}+1}.$$

(b) For $x = \frac{\pi}{4}$, we have $\cos^2 x - \sin^2 x = 0$. This indicates that when x is close to $\frac{\pi}{4}$ we will get cancellation. This can be avoided by noting that $\cos^2 x - \sin^2 x = \cos 2x$. We can evaluate

$$\frac{1}{\cos 2x}.$$

Question 3 [43503/53903] Let $\mathbf{S} \in \mathbb{C}^{m\times m}$ denote a skew-hermitian matrix, i.e., $\mathbf{S}^* = -\mathbf{S}$

(a) Show the eigenvalues of $\mathbf{S}$ are pure imaginary (or zero).

(b) Show that $\mathbf{I} - \mathbf{S}$ is nonsingular. Here, $\mathbf{I}$ denote an identity matrix.

(c) Show the matrix $\mathbf{Q} = (\mathbf{I} - \mathbf{S})^{-1}(\mathbf{I} + \mathbf{S})$ is unitary.

Remark: in (a) the zero eigenvalue associated to a marginal case, for this homework you can ignore it.

Solution. (a) Let $\lambda \in \mathbb{C}$ denote an eigenvalue of $\mathbf{S}$ with corresponding eigenvector $\boldsymbol{x} \neq \mathbf{0}$. We have $\mathbf{S}\boldsymbol{x} = \lambda\boldsymbol{x}$. Taking the inner product of both sides with vector $\boldsymbol{x}$, we get

$$\boldsymbol{x}^*(\mathbf{S}\boldsymbol{x}) = \boldsymbol{x}^*(\lambda\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_2^2.$$

Notice, the matrix $\mathbf{S}$ is skew-hermitian, i.e., $\mathbf{S}^* = -\mathbf{S}$. Then

$$\boldsymbol{x}^*(\mathbf{S}\boldsymbol{x}) = -\boldsymbol{x}^*(\mathbf{S}^*\boldsymbol{x}) = -(\mathbf{S}\boldsymbol{x})^*\boldsymbol{x} = -(\lambda\boldsymbol{x})^*\boldsymbol{x} = -\overline{\lambda}\|\boldsymbol{x}\|_2^2.$$

Hence, $\lambda\|\boldsymbol{x}\|_2^2 = -\overline{\lambda}\|\boldsymbol{x}\|_2^2$. Since $\boldsymbol{x} \neq \mathbf{0}$, it follows that $\lambda = -\overline{\lambda}$, which implies the real part of λ is 0. Therefore, all eigenvalues of $\mathbf{S}$ are purely imaginary.

(b) For any eigenvalue of $\mathbf{S}$, denoted by λ, from (a), we know λ is purely imaginary, i.e., $\lambda = i\mu$ for some $\mu \in \mathbb{R}$. All of the eigenvalues of $\mathbf{I} - \mathbf{S}$ are in form of $1 - \lambda = 1 - i\mu \neq 0$. Hence, zero is not an eigenvalue of $\mathbf{I} - \mathbf{S}$. Therefore, $\mathbf{I} - \mathbf{S}$ is nonsingular.

(c) Let us compute the conjugate transpose of $\mathbf{Q}$. Since, the matrix $\mathbf{S}$ is skew-hermitian, i.e., $\mathbf{S}^* = -\mathbf{S}$, using $(\mathbf{AB})^* = \mathbf{B}^*\mathbf{A}^*$ and $(\mathbf{A}^{-1})^* = (\mathbf{A}^*)^{-1}$, we have

$$\mathbf{Q}^* = (\mathbf{I} + \mathbf{S})^*((\mathbf{I} - \mathbf{S})^{-1})^* = (\mathbf{I} - \mathbf{S})(\mathbf{I} + \mathbf{S})^{-1}.$$

Thus, we get

$$\mathbf{Q}^*\mathbf{Q} = (\mathbf{I} - \mathbf{S})(\mathbf{I} + \mathbf{S})^{-1}(\mathbf{I} - \mathbf{S})^{-1}(\mathbf{I} + \mathbf{S}).$$

Since

$$(\mathbf{I} + \mathbf{S})(\mathbf{I} - \mathbf{S}) = \mathbf{I} - \mathbf{S}^2 = (\mathbf{I} - \mathbf{S})(\mathbf{I} + \mathbf{S}),$$

the matrices $\mathbf{I} + \mathbf{S}$ and $\mathbf{I} - \mathbf{S}$ commute, as do their inverses (notice both $\mathbf{I} + \mathbf{S}$ and $\mathbf{I} + \mathbf{S}$ are nonsingular). Therefore,

$$\mathbf{Q}^*\mathbf{Q} = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^{-1}(\mathbf{I} + \mathbf{S})^{-1}(\mathbf{I} + \mathbf{S}).$$

Hence, $\mathbf{Q}$ is unitary.

Question 4 [43503/53903] Let $\|\cdot\|$ denote any norm on $\mathbb{C}^m$ and also the induced matrix norm on $\mathbb{C}^{m\times m}$. Show that $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$, where $\rho(\mathbf{A})$ is the spectral radius of $\mathbf{A}$, i.e., the largest absolute eigenvalue of $\mathbf{A}$.

Solution. Let λ be an eigenvalue of the matrix $\mathbf{A} \in \mathbb{C}^{m\times m}$ and let $\boldsymbol{\xi}$ be the associated eigenvector. We have $\mathbf{A}\boldsymbol{\xi} = \lambda\boldsymbol{\xi}$. After taking norm $\|\cdot\|$, we get

$$|\lambda|\|\boldsymbol{\xi}\| = \|\lambda\boldsymbol{\xi}\| = \|\mathbf{A}\boldsymbol{\xi}\|.$$

Notice, the eigenvector $\boldsymbol{\xi} \neq \mathbf{0}$. Then,

$$|\lambda| = \frac{\|\mathbf{A}\boldsymbol{\xi}\|}{\|\boldsymbol{\xi}\|} \leq \sup_{\boldsymbol{x}\in\mathbb{C}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \|\mathbf{A}\|.$$

Because the inequality above holds for any eigenvalue λ of $\mathbf{A}$, it must also hold for the largest absolute eigenvalue, which is the spectral radius, i.e., $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$.

Question 5 [53903] In this question, we need to implement our own code. Consider a floating point system defined by the parameters

$$(\beta, t, L, U) = (2, 3, -2, 4).$$

(a) Plot the decimal representation of all numbers that can be represented in this system, and zoom in to examine numbers close to zero.

(b) Describe and explain your observations regarding the distribution of the floating point numbers. How is the spacing between consecutive numbers changes near zero, any patterns in the density of representable numbers, and the symmetry of the distribution?

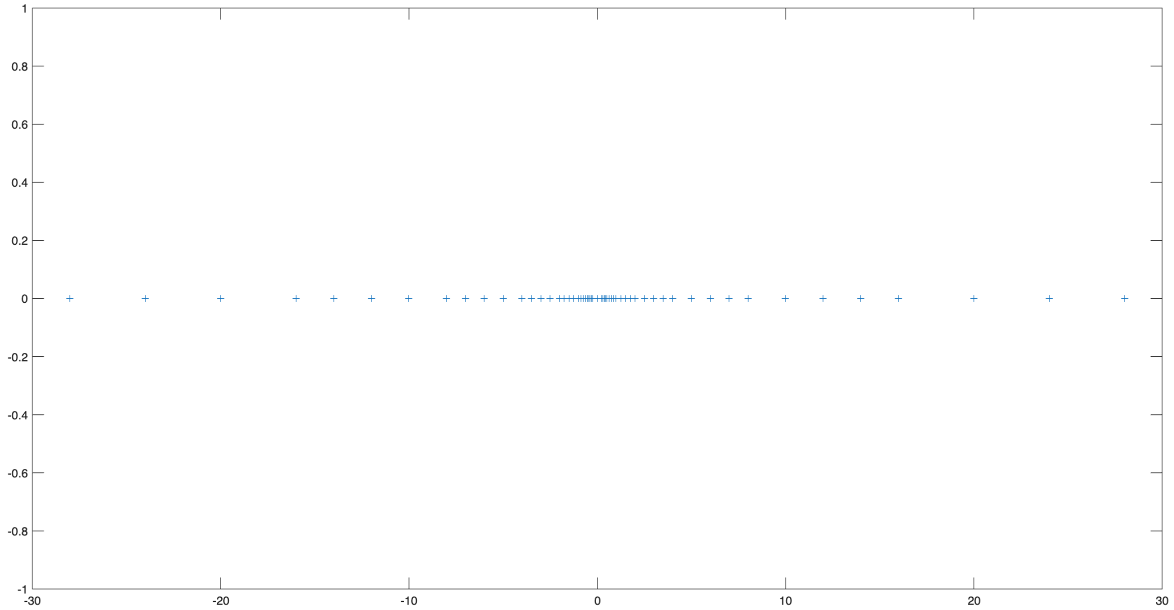
Solution. (a) See the following for MATLAB code.

```
x = [];

% Generate all positive numbers of the system (2,3,-2,4)
for i = 1:0.25:1.75,
    for j = -2:4,
        x = [x i*2^j];
    end
end

x=[x -x 0];     % Add all negative numbers and 0
x = sort(x);    % Sort
y = zeros(1,length(x));
plot(x,y,+)
```

(b) See the reading material page 25-26.


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-30
-20
-10
0
10
20
30