Homework 1 Foundations of Computational Math 2 Spring 2025

The solutions are due on Canvas by 11:59 PM on February 7, 2025

Problem 1.1 (20 points)

Consider the floating point system

$$\mathcal{F}(\beta, t, L, U) = \mathcal{F}(12, 10, -15, 16).$$

1.1.a. How many bits does it take to store a floating point number in this floating point system?

1.1.b. What is the minimum positive normalized floating point number in the system?

1.1.c. What is the maximum positive normalized floating point number in the system?

1.1.d. Encode the decimal number 157 in the system. Use the bias (excess) encoding for the exponent.

1.1.e. Give and example of a subnormal (denormalized) number in the system.

Problem 1.2 (20 points)

Define the function $f(x) = x + 1$ on the domain $x < -1$. Let $x_0 \in \mathbb{R}$, $x_0 < -1$, and $x_1 = x_0(1 + \delta)$ where $\delta \in \mathbb{R}$ with $|\delta| < 1$.

(1.2.a) Determine the relative error between $f(x_1)$ and $f(x_0)$, and the relative condition number $\kappa_{rel}(x_0)$.

(1.2.b) Suppose $|\delta| < 10^{-7}$. Determine the region of values for x_0 for which the relative error between $f(x_1)$ and $f(x_0)$ is no more than 10^{-4}.

Problem 1.3 (35 points)

1.3.a

The backward error analysis of the sum of n floating point numbers presented in the notes can be generalized to yield the following result.

Lemma. *Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ with components, ξ_i and η_i, $1 \leq i \leq n$, respectively, that are floating point numbers. Computing the inner product $x^T y$ on a standard floating point arithmetic machine yields*

$$fl(x^T y) = (x + \Delta x)^T y = x^T (y + \Delta y)$$
$$\Delta x \in \mathbb{R}^n \text{ and } \Delta y \in \mathbb{R}^n$$
$$|\Delta x| \leq \omega_n |x| \text{ and } |\Delta y| \leq \omega_n |y|$$
$$\omega_n = \frac{nu}{1 - nu}$$

(Inequalities using absolute values of matrices and vectors should be interpreted componentwise.)

You need not prove this lemma but use it to show that the matrix vector product $y = Ax$ with $A \in \mathbb{R}^{m \times n}$ can be computed on a standard floating point arithmetic machine in such a way so as to satisfy

$$\hat{y} = (A + \Delta A)x$$
$$|\Delta A| \leq \omega_n |A|$$

1.3.b

Suppose $A \in \mathbb{R}^{n \times n}$ is a banded matrix, i.e., a matrix with all of its nonzero elements on the main diagonal, i.e., $\alpha_{i,i} \neq 0$, the first superdiagonal, i.e., $\alpha_{i,i+1} \neq 0$, through the k-th superdiagonal, i.e., $\alpha_{i,i+k} \neq 0$, the first subdiagonal, i.e., $\alpha_{i-1,i} \neq 0$, through the k-th subdiagonal, i.e., $\alpha_{i,i+k} \neq 0$. All elements not on these diagonals are 0. For $k = 4$ and $n = 15$ the pattern is

$$\begin{pmatrix}
* & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & * & * & * & * & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & * & * & * & * & * & * & * & * & * & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & * & * & * & * & * & * & * & * & * & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & * & * & * & * & * & * & * & * & * & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * & * & * & * & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & *
\end{pmatrix}$$

Does the matrix-vector product $y = Ax$ when A is banded have a backward error and backward stability result similar to that derived for dense A, i.e.,

$$\hat{y} = (A + \Delta A)x$$
$$|\Delta A| \leq \omega_n |A|,$$

but in this case with ΔA being a banded matrix with the same nonzero/zero structure as A? If so derive it and comment on the differences with the result above. If not prove why it cannot.