CSC352 HW6

Alex Zhang

Feb 2023

Question 1

Given that $\tilde{x} = x(1+\varepsilon_x)$, and $\tilde{y} = y(1+\varepsilon_y)$, we can simplify the inequality,

$$\begin{aligned}
\left|\frac{xy - \tilde{x}\tilde{y}}{xy}\right| &\leq (2+\varepsilon)\varepsilon \\
\left|\frac{xy - (xy + xy\varepsilon_x + xy\varepsilon_y + xy\varepsilon_x\varepsilon_y)}{xy}\right| &\leq (2+\varepsilon)\varepsilon \\
\left|\frac{-xy\varepsilon_x - xy\varepsilon_y - xy\varepsilon_x\varepsilon_y}{xy}\right| &\leq (2+\varepsilon)\varepsilon \\
|-\varepsilon_x - \varepsilon_y - \varepsilon_x\varepsilon_y| &\leq (2+\varepsilon)\varepsilon \\
|\varepsilon_x + \varepsilon_y + \varepsilon_x\varepsilon_y| &\leq \varepsilon + \varepsilon + \varepsilon^2
\end{aligned}$$

Csae 1: $\varepsilon = \left|\frac{x-\tilde{x}}{x}\right|$
Since $\tilde{x} = x(1+\varepsilon_x)$, $\varepsilon_x = \left|\frac{\tilde{x}-x}{x}\right| = \left|\frac{x-\tilde{x}}{x}\right|$. Indicate $\varepsilon = \varepsilon_x$. Because $\varepsilon_x \geq \varepsilon_y$, $\varepsilon_x^2 \geq \varepsilon_x\varepsilon_y$, we can get that,

$$\varepsilon_x + \varepsilon_x + \varepsilon_x^2 \geq |\varepsilon_x + \varepsilon_y + \varepsilon_x\varepsilon_y|$$

which is the same as the simplified inequality.

Case 2: $\varepsilon = \left|\frac{y-\tilde{y}}{y}\right|$
Without loss of generality, we can apply the same proof on ε_y using ε_x's and it will have the same result.
$\blacksquare$

Question 3

(a)

The solution for $\mathbf{x}$ is $\begin{bmatrix} 0.9999 \\ 1 \\ 0.9999 \end{bmatrix}$.

(b)

The solution for $\mathbf{x}$ is $\begin{bmatrix} -238 \\ 490 \\ -266 \end{bmatrix}$.

(c)

I think it is ill-conditioned because I changed one entry by subtracting one, but my results vary from an absolute value about 200 to 400.

(d)

The value of condition number is about 65886, which is large. I think my assumption about the ill-conditioned holds true because the condition number

Question 4

(a)

The first component is fraction f, the second one is exponent e, and the thrid one is the sign of this number. In double precision, every number can be represented as,

$$\pm(1+f)\cdot 2^e$$

(b)

Following is the floating point represeation of decimal number -12,

$$(-1)^1(1+0.5)\cdot 2^3$$

(c)

The biggest possible floating point number shoule be,

$$(-1)^0(1+(1-2^{-52}))\cdot 2^{1023}=(2-2^{-52})\cdot 2^{1023}$$

The smallest possible positive floating point number will be,

$$(-1)^0(1)\cdot 2^{-1022}=2^{-1022}$$

(d)

By definition, machine epsilon is the distance from 1 to the next larger floating point number. In term of floating point representation,

$$eps=|(1+\min f)\cdot 2^0-1|=\min f$$

The value will be $\min f = 2^(-52) \approx 2.22e-16$