MA311 (Scientific computing)-IITG

ASSIGNMENT-2

Due on: 9-08-2018, 6:00 PM

---

1. Represent the following numbers in the word form in both 32 and 64 bit binary IEEE standard floating point system.

   (a) $(12.345)_{10}$

   (b) $(35.12365)_6$

   (c) $(\frac{1}{10})_{10}$

2. Are these machine numbers in the 32- bit IEEE standard floating point system?

   - 10^{40}
   - $2^{-1} + 2^{-26}$
   - $\frac{1}{5}$
   - $\frac{1}{3}$
   - $\frac{1}{256}$

3. Plot all available machine number in a floating point system represented by $(\beta, t, L, U) = (2, 3, -2, 3)$. Determine the roundoff unit and machine epsilon for this floating point system (with usual roundoff procedure).

4. Machine epsilon ϵ is the smallest number of the form 2^{-n} such that $1 + \epsilon \neq 1$. By writing a code, compute the approximate value of the machine epsilon of your assigned machine or your PC.

5. Write a code for the following:

   (a) Sums up $1/n$ for $n = 1, 2, \cdots, 10000$;

   (b) Rounds each number $1/n$ to 5 decimal digits and then sums them up in 5-digit decimal arithmetic for $n = 1, 2, \cdots, 10, 000$;

   (c) Sums up the same rounded numbers (in 5-digit decimal arithmetic) in reverse order, i.e., for $n = 10000, \cdots, 2, 1$.

Compare the three results and explain your observations. For generating numbers with the requested precision, you may need to write a code.

6. Explain in detail how to avoid overflow when computing the ℓ_2 norm of a (possibly large in size) vector.

7. For small values of x, how good is the approximation $cosx \approx 1$? How small must x be to have $\frac{1}{2} \times 10^{-18}$ accuracy?

8. Prove that if x and y are machine numbers in 32 bit IEEE standard floating point system, and if $|y| \leq |x| 2^{-25}$, then $fl(x+y) = x$.

9. Suppose that x is a machine number in the range $-\infty < x < 0$. In IEEE standard arithmetic, what values are returned by the computer for the computations $-\infty + x, \infty * x, x/-\infty$, and $-\infty/x$.

10. Give examples of real numbers x and y for which $fl(x \odot y) \neq fl(fl(x) \odot fl(y))$. Illustrate all four arithmetic operations, using a machine with five decimal digits.

11. Let $x = 2^{12} + 2^{-12}$.

    (a) Find the machine numbers x_- and x_+ in 32-bit IEEE standard floating point system, that are just to the left and right of x, respectively.

    (b) For this number show that the relative error between x and $fl(x)$ is no greater than the corresponding unit roundoff error.

12. Let $f \in C[a, b]$ be a function whose derivative exists on (a, b). Suppose f is to be evaluated at x_0 in (a, b), but instead of computing the actual value $f(x_0)$, the approximate value $\tilde{f}(x_0)$, is computed, where $\tilde{f}(x_0) = f(x_0 + \epsilon)$.

    (a) Use the Mean value theorem to estimate the absolute error $|f(x_0) - \tilde{f}(x_0)|$ and the relative error $|f(x_0) - \tilde{f}(x_0)|/|f(x_0)|$, assuming $f(x_0) \neq 0$.

    (b) if $\epsilon = 5 \times 10^{-6}$ and $x_0 = 1$, find bounds for the absolute and relative errors for (a) $f(x) = e^x$ (b) $f(x) = \sin x$.