

## CS 323 Homework - due in lecture on 2/2/12

I. On a machine which rounds and has floating point numbers characterized by  $\beta = 10$ ,  $n = 4$ ,  $m = -30$ ,  $M = 30$  ...

(i) How would the following numbers be approximated: (a)  $1.6726231 \times 10^{-24}$  (the mass of a proton in grams) (b)  $5.9742 \times 10^{27}$  (the mass of the earth in grams)?

(ii) What are the absolute and relative errors in the above approximations? (Note the connection between relative error and significant digits.)

II. (i) Convert to binary form:  $(23.625)_{10}$ ,  $(.8)_{10}$ .

(ii) How will the above two numbers be approximated on a binary computer 6-bit precision and which rounds?

(iii) What are the absolute and relative errors in the approximations in (ii)? (Express errors in base 10.)

III. On a 3-bit binary computer which chops and accumulates sums from left to right, what are the computed values of the following expressions?

(i)  $5 + 4 + 3 + 2 + 1$

(ii)  $1 + 2 + 3 + 4 + 5$

(iii)  $\overbrace{1 + \dots + 1}^{100 \text{ times}}$

IV. The IEEE double precision standard, used by MATLAB, has normalized floating point numbers of the form  $\pm(1.d_1 \dots d_{52})_2 \times 2^e$ ,  $e \in [-1022, +1023]$ . A real number  $x$  lying between consecutive floating point numbers is approximated by the closer of the two, and in the event of a tie, the one with  $d_{52} = 0$  is selected.

On an IEEE double precision standard computer...

(i) What is the distance between 1 and the next larger floating point number?

(ii) For which integers  $k$  in the range  $[-100, +100]$  is the computed value of  $1 + 2^k$  equal to:

(a)  $1 + 2^k$ ?

(b)  $2^k$ ?

(c) 1?

V. The standard formula for the roots of  $ax^2 + bx + c = 0$  is

$$(1) \quad x_{\pm} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

By multiplying (1) by  $\frac{-b \mp \sqrt{b^2 - 4ac}}{-b \mp \sqrt{b^2 - 4ac}}$ , the following equivalent formula is obtained:

$$(2) \quad x_{\pm} = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}.$$

Using 3-decimal digit rounding arithmetic, compute both roots of  $x^2 - 16x + 1 = 0$  using (1); then repeat using (2). Explain any discrepancy in the two pairs of computed roots.

VI. (i) Approximate  $\ln 1.2$  using the quadratic Taylor polynomial for  $f(x) = \ln x$  about  $x = 1$ .

(ii) Use the error formula for Taylor polynomials to bound the error in your approximation. Compare with the actual error (using the exact value of  $\ln 1.2$ ).