Leoza Kabir — CSC 349A: Numerical Analysis — 2018-09-30
V00840048 — Assignment 2
Instructor: Rich Little

Question #1 - 8 marks.

Consider a base 5 normalized, floating-point number system. Assume that a hypothetical computer using this system has the following floating-point representation: sm f1 f2 f3 f4 se e1 e2 where sm is the sign of the mantissa, se is the sign of the exponent (1 for negative, 0 for positive), fi are the digits of the mantissa, and ej are the digits of the exponent.

(a) Consider the base 5 number, given using the above representation, 02003004. What exact decimal value does it represent?

Leoza Kabir
V00840048
Date: Sep 27, 2018
CSC 349A
Assignment#2

1 a) 02003004 = $(0.2003 \times 5^4)_5 = (2003)_5$
$2\times5^3 + 0\times5^2 + 0\times5^1 + 3\times5^0$
$= 250 + 0 + 0 + 3 = (253)_{10}$

(b) What decimal value does 11004003 represent?

b) 11004003 = $(-0.1004 \times 5^3)_5 = (-100.4)_5$
$= 1\times5^2 + 0\times5^1 + 0\times5^0 + 4\times5^{-1}$
$= 25 + 0 + 0 + 4/5$
$= 129/5 = (25.8)_{10}$

(c) What is the smallest positive, non-zero, number that can be represented in this system? Give the answer in the above form (i.e. as 8 base-5 digits.) and in decimal.

c)

Sm	f_1	f_2	f_3	f_4	Se	e_1	e_2
0	1	0	0	0	1	4	4

$4\times5^1 + 4\times5^0 = 20 + 4 = 24$

$(0.1000 \times 5^{-44}) = 1\times5^{-45} = (3.51844 \times 10^{-32})_{10}$

(d) What is the size of the gap between any two consecutive numbers in the interval 25(10) and 125(10) in this floating-point representation system? Your answer should be in decimal.

d) $[5^2, 5^3]$

Size of interval $= 5^3 - 5^2 = 100$

# of sub intervals $= (b-1)b^{k-1} = (5-1)(5)^{4-1} = 500$

Size of gap $= b^{t-k} = b^{3-4} = 5^{3-4} = 5^{-1} = 1/5 = 0.2$

Question #2 - 6 Marks.

The polynomial P(x) = x 2 − 83.12x + 3.123 has two roots, at approximately 0.0375892 and 83.0824. The roots of a quadratic polynomial ax2 + bx + c can be computed by (i) −b ± √ b 2 − 4ac 2a or equivalently (ii) −2c b ± √ b 2 − 4ac. Using floating-point arithmetic, one of these formulas is often much more accurate than the other. For example, if (−b + √ b 2 − 4ac)/(2a) is used to compute one of the roots of P(x) = x 2 − 83.12x + 3.123 = 0 with base b = 10 , precision k = 4, idealized chopping arithmetic, the results are as follows: fl(b 2) = fl(()6908.9344) = 6908 or 0.6908 × 104 . fl(4a) = 4 or 0.4000 × 101 . fl(4ac) = fl(4 × 3.123) = fl(12.492) = 12.49 fl(b 2 − 4ac) = fl(()6908 − 12.49) = fl(()6895.51) = 6895 fl(√ b 2 − 4ac) = fl(√ 6895) = fl(83.036136 · · ·) = 83.03 fl(−b + √ b 2 − 4ac) = fl(83.12 + 83.03) = fl(166.15) = 166.1 fl(2a) = 2 fl −b + √ b 2 − 4ac 2a = fl 166.1 2 = 83.05 or 0.8305 × 102 2 which is very accurate. The relative error is about 0.00039 or 0.039%. On the other hand, it can be shown (similar to the above) that fl −2c b + √ b 2 − 4ac = 69.40 or 0.6940 × 102 which (using the exact value of 83.0824 · · ·) has a large relative error of 0.165 or 16.5%.

(a) Use base b = 10, precision k = 4, idealized chopping arithmetic and each of the mathematically equivalent formulas −2c b − √ b 2 − 4ac and −b − √ b 2 − 4ac 2a to compute an approximation to one root of P(x) = 1.2x 2 − 78.99x + 1.234 = 0. As above, specify each step of the computation. Note that many of the computations for the two formulas are identical, and need only be done once. Use your calculator to do this, not MATLAB.

2a) $fl(b^2) = fl(78.99^2) = fl(6239.4201) = 6239$ or 0.6239×10^4

$fl(4a) = fl(4 \times 1.2) = fl(4.8) = 4.8$ or 0.4800×10

$fl(4ac) = fl(4.8 \times 1.234) = fl(5.9232) = 5.923$

$fl(b^2 - 4ac) = fl(6239 - 5.923) = fl(6233.0768) = 6233$

$fl(\sqrt{b^2-4ac}) = fl\sqrt{6233} = fl(78.9493124) = 78.94$

$fl(b - \sqrt{b^2-4ac}) = fl(-7899 - 78.94) = -157.9$

$fl(-2c) = fl(-2 \times 1.234) = fl(-2.468) = -2.468$

$fl\left(\frac{-2c}{b - \sqrt{b^2-4ac}}\right) = fl\left(\frac{-2.468}{-157.9}\right) = 0.01563$

$fl(\sqrt{b^2-4ac}) = 78.94$

$fl(-b - \sqrt{b^2-4ac}) = fl(-(-78.99) - 78.94) = 0.05 = 0.0500$

$fl(2a) = fl(2 \times 1.2) = fl(2.4) = 2.4$ or 0.2400×10

$fl\left(\frac{-b - \sqrt{b^2-4ac}}{2a}\right) = fl\left(\frac{0.05}{2.4}\right) = fl(0.020833) = fl(0.0208) = 0.02083$

(b) Compute the relative errors of each of the approximations in (a) using the fact that the exact value of the root is 0.01562594 · · · . Give at least 2 significant digits.

b) Relative error for $\dfrac{-2c}{b-\sqrt{b^2-4ac}}$ method

$$|E_t| = \left|\frac{0.01562594 - 0.01563}{0.01562594}\right| = 0.000259829 \approx 0.026\%$$

Relative error for $\dfrac{-b-\sqrt{b^2-4ac}}{2a}$ method

$$|E_t| = \left|\frac{0.01562594 - 0.02083}{0.01562594}\right| = 0.3330398 \approx 33.30\%$$

(c) One of the two zeros of a quadratic polynomial ax2 + bx + c can be computed using either the formula (i) −b + √ b 2 − 4ac 2a or (ii) −2c b + √ b 2 − 4ac For each of the specified polynomials in the table below, place an X in the appropriate box to indicate which of these formulas is more accurate in precision k = 4 floatingpoint arithmetic. Put exactly one X in each row of the table. (No justification for your answers is required. It is NOT necessary to do any floating-point computation to answer this question.)

c)

Polynomial	(i) is more accurate	(ii) is more accurate
$0.01x^2 - 125x + 0.05$		X
$-0.3x^2 + 125x + 0.025$	X	

Leoza Kabir
V00840048
Instructor: Rich Little

CSC 349A: Numerical Analysis
Assignment 2

2018-09-30

Question #3 - 6 Marks

(a) Determine the second order (n = 2) Taylor series expansion for f(x) = √ x + 3 expanded about a = 1 including the remainder term. Leave your answer in terms of factors (x − 1) (that is, do not simplify). Show all your work.

3)a) $f(x) = \sqrt{x+3}$ $\qquad f(1) = \sqrt{1+3} = \sqrt{4} = 2$

$f'(x) = (x+3)^{1/2} = \frac{1}{2}(x+3)^{-1/2} = \frac{1}{2(x+3)}$ $\qquad f'(1) = \frac{1}{2\sqrt{1+3}} = \frac{1}{2\sqrt{4}} = 1/4$

$f''(x) = \frac{1}{2}(x+3)^{-1/2} = -\frac{1}{4}(x+3)^{-3/2} = \frac{-1}{4(x+3)^{3/2}}$ $\qquad f''(1) = \frac{-1}{4(1+3)^{3/2}} = \frac{-1}{4\times 8} = -1/32$

$f'''(x) = -\frac{1}{4}(x+3)^{-3/2} = \frac{3}{8}(x+3)^{-5/2} = \frac{3}{8(x+3)^{5/2}}$ $\qquad f'''(1) = \frac{3}{8(1+3)^{5/2}} = \frac{3}{256}$

$\frac{1}{4} \times \left(-\frac{3}{2}\right)$

$\frac{3}{8}$

$$f(x) \approx f(1) + f'(1)(x-1) + \frac{f''(1)(x-1)^2}{2!} + \frac{f'''(\xi_0)(x-1)^3}{3!}$$

$$= 2 + \frac{1}{4}(x-1) - \frac{\frac{1}{32}}{2!}(x-1)^2 + \frac{\frac{3}{8(\xi_0+3)^{5/2}}}{3!}(x-1)^3$$

$$= 2 + \frac{1}{4}(x-1) - \frac{1}{64}(x-1)^2 + \frac{1}{16(\xi_0+3)^{5/2}}(x-1)^3$$

(b) Use the polynomial approximation in (a) (without the remainder term) to approximate f(1.12) = √ 4.12. Use either hand computation, your calculator or MATLAB. Give an exact answer. 3

b) $\sqrt{4.12} = f(1.12)$

$\approx 2 + \frac{1}{4}(1.12-1) - \frac{1}{64}(1.12-1)^2 = 2.029775$

Leoza Kabir
V00840048
Instructor: Rich Little

c) Determine a good upper bound for the truncation error of the Taylor polynomial approximation in (a) for all values of x such that 1 ≤ x ≤ 1.2 by bounding the remainder term.

c) $R_2 = \frac{1}{16(\xi_0+3)^{5/2}}(x-1)^3$. $\xi_0 = [1, 1.2]$

$$= \frac{(1.2-1)^3}{16(1+3)^{5/2}} = 0.000015625.$$

$$|E_t| \leq |R_3| = 0.000015625$$