MATH 4701 Numerical Analysis

Problem Set #1

(1) A calculator uses 14 bit binary numbers with one bit for the sign of the number, followed by 8 bits for mantissa m, followed by 5 bits for the characteristic f, and it uses **chopping** for its termination.

(a) How does this calculator represent -37.45?

converting 37 into a binary expansion 100101. The extra three digits after the dot are .011. Therefore with only 9 digits we have the binary representation -1.00101011×2^5. Therefore the eight digits for mantissa are 00101011. With 5 bits for exponent we can cover exponents between $-15 \leq e \leq 16$ with binary codes between 0 and $2^5 - 1 = 31$. The exponent $e = 5$ will be represented by the code for $5 + 15 = 20$ which is 10100. Since the number is negative its sign code is 1. Therefore, this calculator will represent -37.45 as $1|00101011|10100$.

(b) How does this calculator represent 0.2?

0.2 has a periodic binary expansion $0.00110011001100110011... = 0.\overline{0011}$. Therefore with only 9 digits we have the binary representation 1.10011001×2^{-3}. Therefore the eight digits for mantissa are 10011001. With 5 bits for exponent we can cover exponents between $-15 \leq e \leq 16$ with binary codes between 0 and $2^5 - 1 = 31$. The exponent $e = -3$ will be represented by the code for $-3 + 15 = 12$ which is 01100. Since the number is positive its sign code is 0. Therefore, this calculator will represent 0.2 as $0|10011001|01100$.

(c) What is the largest number this calculator can express accurately?

The code for the largest number would be $0|11111111|11111$ which represents 1.11111111×2^{16} or 11111111100000000. Converted to decimal this number is $2^{16} + 2^{15} + ... + 2^8$ or $2^8(1 + 2 + 2^2 + 2^3 + ... + 2^8) = 2^8(2^9 - 1) = 256 \times 511 = 130816$.

(d) What is the smallest **positive** number this calculator can express accurately?

The code for the smallest positive number would be $0|00000000|00000$ which represents $1.00000000 \times 2^{-15} = 2^{-15}$.

(e) How many numbers in this calculator's number system are between 8 and 32?

Since the mantissa $m = 1.m_1m_2...m_8$ satisfies $1 \leq m < 2$, we have $8 \leq 2^3 \times m < 16$ and $16 \leq 2^4 \times m < 32$. Therefore, if $x \neq 32$, there are 2 choices for the exponent and only one choice for the sign. For each one of the 8 variable digits of mantissa there are 2 choices. Therefore, there are 2^8 choices for the mantissa. Hence, $2^9 = 2^8 \times 2$ numbers are in the interval $[8, 32)$ and $2^9 + 1 = 513$ in the interval $[8, 32]$.

(f) What interval of numbers are represented by this calculator with the code 0|11000110|01101.

The binary code 01101 corresponds with integer $8+4+1=13$. Since the five digit binary codes are between 0 and 31 they represent exponents between -15 and 16 and the code representing 13 corresponds with the exponent $13-15=-2$. Since this calculator uses chopping for its termination, the binary number $+1.11000110\times 2^{-2}=0.011100110$ is the smallest number represented by this code and the largest binary number represented by this code is $+1.11000110\bar{1}\times 2^{-2}=0.0111000111$. Since $\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{256}+\frac{1}{512}=\frac{227}{512}$ and $\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{256}+\frac{1}{512}+\frac{1}{1024}=\frac{455}{1024}$, the numbers in the interval $[\frac{227}{512},\frac{455}{1024})$ are exactly the numbers represented by the code 0|11000110|01101.

(g) Describe, step-by-step, how this calculator adds 1.10010111×2^{-2} to 1.10111001×2^{3}. Estimate the absolute error incurred by comparing the actual sum with the number this calculator expresses as the actual answer.

$$1.10010111\times 2^{-2}+1.10111001\times 2^{3}=0.0110010111+1101.11001=1110.0010110111$$

which is the real value of the sum but it will be chopped to 1110.00101 to be represented as 1.11000101×2^{3}. The absolute error is $|1110.0010110111-1110.00101|=0.0000010111=2^{-6}+2^{-8}+2^{-9}+2^{-10}=0.0224609375$.

(2) Use three-digit rounding arithmetic to perform the following calculations. Compute the absolute error and relative error with exact value determined to at least five digits.

(a) $\frac{4}{5}+\frac{1}{3}$

$fl\big(fl(\frac{4}{5})+fl(\frac{1}{3})\big)=fl\big(0.8+0.333\big)=fl\big(1.133\big)=1.13$
Absolute Error$=|\frac{4}{5}+\frac{1}{3}-1.13|=\frac{1}{300}\approx 0.00333$
Relative Error$=\frac{|\frac{4}{5}+\frac{1}{3}-1.13|}{|\frac{4}{5}+\frac{1}{3}|}=\frac{1}{340}\approx 0.00294$

(b) $(\frac{4}{5})(\frac{1}{3})$

$fl\big(fl(\frac{4}{5})fl(\frac{1}{3})\big)=fl\big((0.8)(0.333)\big)=fl\big(0.2664\big)=0.266$
Absolute Error$=|(\frac{4}{5})(\frac{1}{3})-0.266|=\frac{1}{1500}\approx 0.00067$
Relative Error$=\frac{|(\frac{4}{5})(\frac{1}{3})-0.266|}{|(\frac{4}{5})(\frac{1}{3})|}=\frac{1}{400}=0.0025$

(c) $133+0.921$

$fl\big(fl(133)+fl(0.921)\big)=fl\big(133.921\big)=134$
Absolute Error$=|133.921-134|=0.079$
Relative Error$=\frac{|133.921-134|}{|133.921|}=\frac{79}{133921}\approx 0.00059$

(d) $(\frac{2}{9})(\frac{9}{7})$

$fl\big(fl(\frac{2}{9})fl(\frac{9}{7})\big) = fl\big((0.222)(1.29)\big) = fl\big(0.28638\big) = 0.286$
Absolute Error$= |(\frac{2}{9})(\frac{9}{7}) - 0.286| = \frac{1}{3500} \approx 0.00029$
Relative Error$= \frac{|(\frac{2}{9})(\frac{9}{7})-0.286|}{|(\frac{2}{9})(\frac{9}{7})|} = \frac{1}{1000} = 0.001$

(e) $\frac{\sqrt{13}+\sqrt{11}}{\sqrt{13}-\sqrt{11}}$

$fl\left(\frac{fl\big(fl(\sqrt{13})+fl(\sqrt{11})\big)}{fl\big(fl(\sqrt{13})-fl(\sqrt{11})\big)}\right) = fl\left(\frac{fl\big(3.61+3.32\big)}{fl\big(3.61-3.32\big)}\right) = fl\left(\frac{6.93}{0.290}\right) = fl(23.89655...) = 23.9$
Absolute Error$= \left|\frac{\sqrt{13}+\sqrt{11}}{\sqrt{13}-\sqrt{11}} - 23.9\right| \approx 0.05826$
Relative Error$= \frac{\left|\frac{\sqrt{13}+\sqrt{11}}{\sqrt{13}-\sqrt{11}}-23.9\right|}{\left|\frac{\sqrt{13}+\sqrt{11}}{\sqrt{13}-\sqrt{11}}\right|} \approx 0.00243$

(f) $-10\pi + 6e$

$fl\Big(fl\Big(fl\big(fl(-10)fl(\pi)\big)+fl\big(fl(6)fl(e)\big)\Big)\Big) = fl\Big(fl\Big(fl\big(-10\times3.14\big)+ fl\big(6\times 2.72\big)\Big)\Big) = fl\Big(fl\Big(-31.4+16.3\Big)\Big) = fl\Big(-15.1\Big) = -15.1$
Absolute Error$= \Big|-10\pi + 6e - (-15.1)\Big| \approx 0.00624$
Relative Error$= \frac{\Big|-10\pi+6e-(-15.1)\Big|}{\Big|-10\pi+6e\Big|} \approx 0.00041$

(3) (a) Use four digit rounding arithmetic to find the roots of $\frac{x^2}{3} - \frac{123x}{4} + \frac{1}{6} = 0$ using the quadratic formula $x = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$. What are the absolute and relative errors for each root? (record the actual values of the roots accurate within 5 decimal places)

If we evaluate the two roots, with no rounding after every calculation, we get

$$x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a} = \frac{\frac{123}{4} + \sqrt{\frac{15129}{16} - \frac{4}{18}}}{\frac{2}{3}} = \frac{\frac{123}{4} + \sqrt{\frac{136129}{144}}}{\frac{2}{3}} \approx 92.24458$$

$$x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a} = \frac{\frac{123}{4} - \sqrt{\frac{15129}{16} - \frac{4}{18}}}{\frac{2}{3}} = \frac{\frac{123}{4} - \sqrt{\frac{136129}{144}}}{\frac{2}{3}} \approx 0.00542$$

Rounding after each calculation to four digits with $a = \frac{1}{3} = 0.3333$, $b = -\frac{123}{4} = -30.75$, and $c = \frac{1}{6} = 0.1667$ we have $b^2 = 945.6$, $4ac = 0.2222$, and $b^2 - 4ac = 945.4$ and $\sqrt{b^2-4ac} = 30.75$. Therefore, roots are approximated to be $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a} = \frac{30.75+30.75}{0.6666} = \frac{61.5}{0.6666} = 92.26$ and $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a} = \frac{30.75-30.75}{0.6666} = 0$.
Absolute error for x_1 is $|92.24458-92.26| = 0.01542$ and absolute error for x_2 is $|0.00542 - 0| = 0.00542$.
Relative error for x_1 is $\frac{0.01542}{92.24458} \approx 0.00017$ and relative error for x_2 is $\frac{0.00542}{0.00542} = 1$.

(b) Use four digit rounding arithmetic to find the roots of $\frac{x^2}{3} - \frac{123x}{4} + \frac{1}{6} = 0$ using the quadratic formula $x = \frac{-2c}{b \pm \sqrt{b^2-4ac}}$. What are the absolute and relative errors for each root?

Rounding after each calculation to four digits and, using preliminary calculations in the previous part, $x_1 = \frac{-2c}{b+\sqrt{b^2-4ac}} = \frac{-0.3334}{-30.75+30.75}$ which is undefined and $x_2 = \frac{-2c}{b-\sqrt{b^2-4ac}} = \frac{-0.3334}{-30.75-30.75} = 0.0054$.
Absolute error for x_2 is $|0.00542 - 0.0054| = 0.00002$ and the relative error for x_2 is $\frac{0.00002}{0.00542} \approx 0.0037$.

(4) Consider the polynomial $P(x) = x^3 - 2.14x^2 + 1.16x + 7.25$. Using three digit chopping arithmetic calculate $P(4.58)$ first the normal way and then using Nested Arithmetic. In each case calculate the absolute and relative errors.
$P(4.58) = (4.58)^3 - 2.14(4.58)^2 + 1.16(4.58) + 7.25 = 63.745216$

Using the normal way and chopping after each calculation to three digits we get:
$x = 4.58 \Rightarrow x^2 = (4.58)^2 = 20.9$, $x^3 = x(x^2) = (4.58)(20.9) = 95.7$, $2.14x^2 = (-2.14)(20.9) = 44.7$, $1.16x = (1.16)(4.58) = 5.31$,
$x^3 - 2.14x^2 + 1.16x + 7.25 = ((x^3 - 2.14x^2) + 1.16x) + 7.25 = ((95.7 - 44.7) + 5.31) + 7.25 = (51.0 + 5.31) + 7.25 = 56.3 + 7.25 = 63.5$.
Therefore, the absolute error in this computation is $|63.745216 - 63.5| = 0.245216$ and the relative error is $\frac{0.245216}{63.745216} \approx 0.00385$.

Using Nesting, calculation of $P(4.58)$ goes as follows:

$$P(x) = ((x-2.14)x+1.16)x+7.25 \Rightarrow P(4.58) = ((4.58-2.14)4.58+1.16)4.58+7.25 =$$
$$= ((2.44)(4.58)+1.16)4.58+7.25 = (11.1+1.16)4.58+7.25 = (12.2)(4.58)+7.25 = 55.8+7.25 = 63.0$$

Therefore, the absolute error in this computation is $|63.745216 - 63| = 0.745216$ and the relative error is $\frac{0.745216}{63.745216} \approx 0.01169$.