Binary to Decimal

- Convert the following from binary to decimal format. Express fractional values as rational numbers.

- $x_2 = 1001$

  $$x_{10} = 9$$

- $x_2 = 1010.1101$

  $$x_{10} = 10\frac{13}{16}$$

Reformulation

1. Calculate the exact value (calculator)
2. Calculate the computed value
3. Reformulate and re-compute
4. Use 3 significant digits

$$\sqrt{x^2 + 4} - 2 \quad \text{for } x = .13$$

- exact: $.422 \times 10^{-2}$
- computed: $.000$
- reformulated: $.423 \times 10^{-2}$
- relative error: before = 100%, after = $.2\%$

Reformulation

- Indicate when the loss of precision occurs and reformulate.

- $1 - \sin^2 x - \cos^2 x$
  - David: demonstrate for $x = \frac{\pi}{6}$

- $1 - \sin x$

  occurs when $|x|$ is close to $\frac{\pi}{2}$

  $$(1 - \sin x)\left(\frac{1+\sin x}{1+\sin x}\right) = \frac{1-\sin^2 x}{1+\sin x} = \frac{\cos^2 x}{1+\sin x}$$

- $\sqrt{x+2} - \sqrt{x}$

  occurs when $|x|$ is significantly larger than 2

  $$\left(\sqrt{x+2} - \sqrt{x}\right)\left(\frac{\sqrt{x+2}+\sqrt{x}}{\sqrt{x+2}+\sqrt{x}}\right) = \frac{2}{\sqrt{x+2}+\sqrt{x}}$$

Reformulation – Taylor series

- Identify where loss of precision will be a problem and reformulate. Truncate terms greater than $O(x^3)$.

- $1 - \cos x$
  - Compute values for the (1) exact, (2) computed, (3) reformulated, and (4) relative error for $x = 0.05$
  - exact: $.125 \times 10^{-2}$
  - computed: $.1 \times 10^{-2}$
  - reformulated: $.125 \times 10^{-2}$
  - relative error: before = 20%, after = 0%

Reformulation – Taylor series

- Identify where loss of precision will be a problem and reformulate. Truncate terms smaller than $O(x^3)$.

- $e^x - \cos x$

occurs when $|x|$ is small

$$\approx x + x^2 + \frac{x^3}{6}$$

- e^{x-y}

occurs when x is close to y

$$= \frac{e^x}{e^y} \approx \frac{1+x+\frac{x^2}{2}+\frac{x^3}{6}}{1+y+\frac{y^2}{2}+\frac{y^3}{6}}$$

Reformulation – in practice

- Compute the following as accurately as possible using three significant digits

- $\sqrt{102} - \sqrt{100}$

exact: $.995 \times 10^{-1}$

$$= \frac{\epsilon}{\sqrt{x+\epsilon}+\sqrt{x}} = \frac{2}{\sqrt{102}+\sqrt{100}} \approx \frac{2.00}{10.1+10.0} \approx .995 \times 10^{-1}$$

- $1 - \sin(1.6)$

exact: $.426 \times 10^{-3}$

$$= \frac{\cos^2 x}{1+\sin x} \approx \frac{(-.0291)^2}{1.00+1.00} \approx \frac{.847\times 10^{-3}}{2} \approx .427 \times 10^{-3}$$

Bisection Method

- Find the solution to the following equation using 3 iterations of the bisection method and 4 significant digits. Use a table with the values: $a \quad f(a) \quad b \quad f(b) \quad c \quad f(c)$

$$\cos x = x \text{ on the interval } [0.5, 1.0]$$

$$x - \cos x = 0$$

$\boldsymbol{a}$	$\boldsymbol{f(a)}$	$\boldsymbol{b}$	$\boldsymbol{f(b)}$	$\boldsymbol{c}$	$\boldsymbol{f(c)}$
0.5	-.3776	1.0	.4597	.75	.01831
0.5	-.3776	.75	.01831	.625	-.1860
.625	-.1860	.75	.01831	.6875	

- What bound can you put on your error?

Newton's Method

- Find the solution to the following equation using 4 iterations of Newton's method and 4 significant digits. Use a table with the values: x_{n-1} $f(x_{n-1})$ $\frac{d}{dx}f(x_{n-1})$ x_n

$$x^3 + 3x^2 = 6x + 8$$
$$f(x) = x^3 + 3x^2 - 6x - 8$$
$$f'(x) = 3x^2 + 6x - 6$$

x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
-6	-80	66	-4.788
-4.788	-20.26	34.04	-4.193
-4.193	-3.813	21.58	-4.016
-4.016	-.2934	18.29	-4.000

Determinants (3×3)

- Compute the determinant of the following matrices and give the number of unique solutions.

$$\begin{vmatrix} 3 & -1 & 2 \\ 1 & 2 & 1 \\ 6 & -2 & 4 \end{vmatrix}$$

$$\begin{aligned} &= (3 \cdot 2 \cdot 4) + (-1 \cdot 1 \cdot 6) + (2 \cdot 1 \cdot -2) \\ &\quad -(3 \cdot -2 \cdot 1) - (1 \cdot -1 \cdot 4) - (6 \cdot 2 \cdot 2) \\ &= 24 - 6 - 4 + 6 + 4 - 24 = 0 \end{aligned}$$

Determinants $(n \times n)$

$$\begin{vmatrix} 2 & 1 & 3 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 3 & 2 & 2 \\ 0 & 1 & 2 & 2 \end{vmatrix}$$

$$= \begin{vmatrix} 2 & 1 & 3 & 1 \\ 0 & 1 & 2 & 2 \\ 0 & 3 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{vmatrix} = (-1) \begin{vmatrix} 2 & 1 & 3 & 1 \\ 0 & 1 & 2 & 2 \\ 0 & 3 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{vmatrix}$$

$$= (-1) \begin{vmatrix} 2 & 1 & 3 & 1 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & -4 & -4 \\ 0 & 0 & 0 & 2 \end{vmatrix} = (-1) \cdot 2 \cdot 1 \cdot (-4) \cdot 2 = 16$$

Lost bits

- Calculate the number of bits of precision lost in the following calculations

- $\sqrt{1.01} - \cos x$ where $x = \frac{1}{64}$

$$\frac{-\log\left(1 - \frac{\cos\frac{1}{64}}{\sqrt{1.01}}\right)}{\log 2} = \frac{-\log\left(1 - \frac{.999878}{1.004988}\right)}{\log 2} = 7.6$$

between 7 and 8 bits are lost

Algorithm Design

- Describe the convergence of the bisection method.
  - Under what conditions will the bisection method fail to find a root? How do you test for them?

- The IEEE half-precision floating point format is specified as:
  - 1 sign bit
  - 5 bit exponent with a -15 bias (15 is subtracted from the exponent)
    - 00000_2 and 11111_2 are used to store special values (infinity, NaN)
  - 10 bit mantissa with an implicit leading 1 for normalization

1. What are the largest and smallest **positive** values that can be represented?

small: $(.10000000000) \times 2^{(00001)_2} = 0.5 \times 2^{-14} = .30517578 \times 10^{-4}$
large: $(.11111111111) \times 2^{(11110)_2} = \frac{2047}{2048} \times 2^{15} \approx 32752$

2. What are the largest and smallest gaps between values?

$$(.10000000001) \times 2^{(00001)_2} = \frac{1025}{2048} \times 2^{-14} = .3054738 \times 10^{-4}$$
$$.3054738 \times 10^{-4} - .30517578 = \pm .29802 \times 10^{-7}$$

$$(.11111111110) \times 2^{(11110)_2} = \frac{2046}{2048} \times 2^{15} \approx 32736$$
$$32752 - 32736 = 16$$

Algorithm Design

- When using the Mean Value Theorem to reformulate an expression, such as $\ln(x+\epsilon) - \ln(x)$, how does the size of ϵ affect the relative error of the calculation?

  - The accuracy will improve. The derivative slope of the line between x and $(x+\epsilon)$ becomes a better approximation of the derivative at the mid-point $\theta = \frac{2x+\epsilon}{2}$.
  - This is in direct contrast to dealing with catastrophic cancellation, where the relative error will increase.

Horner’s Algorithm

- Use synthetic division to compute the value and derivative of the polynomial

$$2x^3 - 3x^2 + x + 4 \quad \text{at } x = 2$$

2	2	-3	1	4
		4	2	6
	2	1	3	10

$f(2)$

$f'(2)$

$$q(x) = 2x^2 + x + 3$$

$$f'(2) = q(2) = 2(2)^2 + (2) + 3 = 13$$

Scaled Partial Pivoting

- Solve the following matrix using scaled partial pivoting. Make sure to keep track of a scale vector and index vector. Represent the matrix as it would be represented using the algorithm.

$$A = \begin{bmatrix} -1 & -5 & -7 \\ 4 & -5 & 4 \\ 1 & 5 & -1 \end{bmatrix} \qquad s = \begin{bmatrix} 7 \\ 5 \\ 5 \end{bmatrix} \quad l = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$l = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \qquad A_1 = \begin{bmatrix} 0 & \frac{-25}{4} & -6 \\ 4 & -5 & 4 \\ 0 & \frac{25}{4} & -2 \end{bmatrix}$$

$$l = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \qquad A_1 = \begin{bmatrix} 0 & \frac{-25}{4} & -6 \\ 4 & -5 & 4 \\ 0 & 0 & -8 \end{bmatrix}$$

- What is the determinant?

$$(-1)(-8)\left(\frac{-25}{4}\right)(4) = -200$$