Question 1

[15 marks]

Consider the normalized floating-point system $\mathbb{R}_3(3,1)$ with *limited* exponent range $-1 \le e \le 1$.

a. What is the smallest positive (nonzero) number representable? Give your answer in both base-3 and base-10.

Base-3: $(0.100)_3 \times 3^{-1}$

Base-10: $1 \times 3^{-2} = \frac{1}{9} = 0.\bar{1}$

b. What is the largest positive number representable? Give your answer in both base-3 and base-10.

Base-3: $(0.222)_3 \times 3^{1}$

Base-10: $2 \cdot 3^0 + 2 \cdot 3^{-1} + 2 \cdot 3^{-2} = 2 + \frac{2}{3} + \frac{2}{9} = 2 + \frac{6+2}{9} = 2 + \frac{8}{9} = 2.\bar{8}$

c. Assuming round-to-nearest, what is the tightest upper bound on the relative error $|fl(x) - x|/|x|$ when $x \in \mathbb{R}$ is stored as $fl(x) \in \mathbb{R}_3(3,1)$ in this floating-point system? Give your answer in base-10.

$$\frac{|fl(x)-x|}{|x|} = \varepsilon < \frac{1}{2} b^{1-t} = \frac{1}{2} \cdot 3^{1-3} = \frac{1}{2} \cdot \frac{1}{9} = \frac{1}{18}$$

d. What is the floating-point representation of $(407)_{10}$ in this system? (*Hint*: Does the representation exist?)

There is no representation of $(407)_{10}$ in this system since the largest positive number representable is $2.\bar{8}$

e. What is the floating-point representation of $(0.567)_{10}$ in this system? Give your answer in base-3. Recall that there are only *three* base-3 digits in the mantissa.

0.567	3	1.701	1	0.701
0.701	3	2.103	2	0.103
0.103	3	0.309	0	0.309
0.309	3	0.927	0	0.927
0.927	3	2.781	2	0.781
0.781	3	2.343	2	1

So $(0.567)_{10} \approx (0.120)_3$ in this system

f. List all possible normalized, non-zero mantissas in this system. In total, how many floating-point numbers are representable? Recall that the exponent range is *limited*.

Normalized so: combinations excluding 0.000 a, i.e

$0.\underline{2}\times\underline{3}\times\underline{3} = 18$ possible numbers

cannot be 0, so 1 or 2

$\times 2$ because of sign

$\times 3$ because of e^{-1}, e^{0} or e^{1} multiplied

For a total of $18 \times 2 * 3 + 1 = 109$ representable numbers in this system

$(0.000)_3$

CONTINUED ...

Question 2

[10 marks]

A floating-point operation, or *flop*, is an operation of the form $ax + b$. Show how to convert a $(k+1)$-digit base b ($b \neq 10$) positive integer

$$d_k d_{k-1} \ldots d_1 d_0$$

into its base 10 equivalent in k flops or less.

To convert to base 10, we can just do:

$$d_0 \times b^0 + d_1 \times b^1 + \ldots + d_k \times b^k$$

$$= d_0 + b(d_1 + b(d_2 + \ldots + b(\underbrace{d_{k-1} + bd_k}_{1 \text{ flop}}))\ldots))$$

2 flops

$k-1$

k

Thus, we can clearly see that to convert a $(k+1)$-digit in base b, we would just need k flops

CONTINUED ...

Question 3

[15 marks]

Consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} 2 & 5 & 10 \\ 8 & 32 & 8 \\ 1 & 8 & 13 \end{bmatrix}, \quad b = \begin{bmatrix} 7 \\ -16 \\ 6 \end{bmatrix}.$$

a. Compute the $PA = LU$ factorization of A. Use exact arithmetic. Show all intermediate calculations, including Gauss transforms and permutation matrices.

$$P_{12}A = \begin{bmatrix} 8 & 32 & 8 \\ 2 & 5 & 10 \\ 1 & 8 & 13 \end{bmatrix} \quad L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{2}{8} & 1 & 0 \\ -\frac{1}{8} & 0 & 1 \end{bmatrix} \quad L_1 P_{12} A = \begin{bmatrix} 8 & 32 & 8 \\ 0 & -3 & 8 \\ 0 & 0 & 11 \end{bmatrix}$$

So Let $P = P_{12}$ $\quad L = L_1^{-1}$

$$L_1 P_{12} A = U \Rightarrow P_{12} A = L_1^{-1} U \Rightarrow PA = LU$$

b. Use the factorization computed in (a) to solve the system.

$PAx = Pb \Rightarrow LUx = Pb$ Let $Ux = d$ $Ld = Pb$

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{8} & 1 & 0 \\ \frac{1}{8} & 0 & 1 \end{bmatrix} d = \begin{bmatrix} -16 \\ 7 \\ 6 \end{bmatrix}$$

$1d_1 = -16$ $\frac{2}{8}d_1 + d_2 = 7 \Rightarrow \boxed{d_2 = 11}$

$\frac{1}{8}d_1 + d_3 = 6 \Rightarrow \boxed{d_3 = 8}$

So,
$$\begin{bmatrix} 8 & 32 & 8 \\ 0 & -3 & 8 \\ 0 & 0 & 11 \end{bmatrix} x = \begin{bmatrix} -16 \\ 11 \\ 8 \end{bmatrix}$$

$11x_3 = 8 \Rightarrow \boxed{x_3 = \frac{8}{11}}$

$-3x_2 + 8x_3 = 11 \Rightarrow \boxed{x_2 = \frac{-19}{11}}$

$8x_1 + 32x_2 + 8x_3 = -16 \Rightarrow \boxed{x_1 = \frac{46}{11}}$

c. Instead of first computing the $PA = LU$ factorization, we could have solved the system above by processing the left and right-hand sides simultaneously with Gauss transforms and permutations. Would this alternate approach incur any extra cost? **Explain.** We are solving one system only.

$\Rightarrow L_1P_{12}Ax = L_1P_{12}b$

Solving with $PA = LU$ factorization cost $\frac{n^3}{3} + O(n^2)$ as shown in class whereas here

Question 7

[15 marks]

Consider the data points $\{(0,3),(1,7),(2,37),(3,141)\}$.

a. Set up the Vandermonde system for determining the monomial-basis form of the polynomial which interpolates these data points. Do not solve the system.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \end{bmatrix} a = \begin{bmatrix} 3 \\ 7 \\ 37 \\ 141 \end{bmatrix}$$

b. Derive the Newton form of the interpolating polynomial. Show all of your work, including the divided-difference table.

X_i	$y[X_i]$	$y[X_{i+1}, X_i]$	$y[X_{i+2} \ldots X_i]$	$y[X_{i+3} \ldots X_i]$
0	3			
		$\frac{7-3}{1-0} = 4$		
1	7		$\frac{30-4}{2-0} = 13$	
		$\frac{37-7}{2-1} = 30$		$\frac{37-13}{3-0} = 8$
2	37		$\frac{104-30}{3-1} = 37$	
		$\frac{141-37}{3-2} = 104$		
3	141			

So polynomial is $p(x) = 3 + 4(x) + 13(x)(x-1) + 8(x)(x-1)(x-2)$

CONTINUED ...

c. Derive the Lagrange form of the interpolating polynomial. Verify it is the same polynomial as in (b).

$$p(x) = \sum_{i=0}^{n} l_i(x)\, y_i = l_0(x)\, y_0 + l_1(x)\, y_1 + l_2(x)\, y_2 + l_3(x)\, y_3$$

$$l_0(x) = \prod_{\substack{j=0 \\ j\neq 0}}^{n} \frac{x - x_j}{x_i - x_j} = \left(\frac{x-1}{0-1}\right)\left(\frac{x-2}{0-2}\right)\left(\frac{x-3}{0-3}\right) = \frac{1}{-6}(x-1)(x-2)(x-3)$$

$$l_1(x) = \left(\frac{x-0}{1-0}\right)\left(\frac{x-2}{1-2}\right)\left(\frac{x-3}{1-3}\right) = \frac{1}{2}(x)(x-2)(x-3)$$

$$l_2(x) = \left(\frac{x-0}{2-0}\right)\left(\frac{x-1}{2-1}\right)\left(\frac{x-3}{2-3}\right) = -\frac{1}{2}(x)(x-1)(x-3)$$

$$l_3(x) = \left(\frac{x-0}{3-0}\right)\left(\frac{x-1}{3-1}\right)\left(\frac{x-2}{3-2}\right) = \frac{1}{6}(x)(x-1)(x-2)$$

$$p(x) = -\frac{1}{2}(x-1)(x-2)(x-3) + \frac{7}{2}(x)(x-2)(x-3) - \frac{37}{2}(x)(x-1)(x-3) + \frac{141}{6}(x)(x-1)(x-2)$$

d. Briefly discuss the relative efficiency of the methods in (a), (b), and (c). Which method is best if we need to include additional data points? Explain.

Divided difference since if we add additional data points we do not need to erase anything from our divided-difference table but just require to add the corresponding additional calculation unlike the other methods in which we would need to recalculate everything from scratch

CONTINUED ...

e. **Construct the linear spline (i.e., the piecewise linear interpolant) which interpolates all four data points $\{(0,3),(1,7),(2,37),(3,141)\}$.**

First line: $y = 3 + \left(\frac{7-3}{1-0}\right)(x-0) = 3 + 4x$

Second line: $y = 7 + \left(\frac{37-7}{2-1}\right)(x-1) = 7 + 30(x-1) = -23 + 30x$

Third line: $y = 37 + \left(\frac{141-37}{3-2}\right)(x-2) = 37 + (104)(x-2)$

$= -171 + 104x$

So spline is

$$f(x) = \begin{cases} 3+4x, & \text{if } 0 \le x \le 1 \\ -23+30x, & \text{if } 1 < x \le 2 \\ -171+104x, & \text{if } 2 < x \le 3 \end{cases}$$

we could extend this to cover all of $\mathbb{R}$

END OF EXAM