Midterm – Introduction to Computational Science

Claudio Maggioni

April 3, 2020

Question 1

Point a)

$$7.125_{10} = (1 + 2^{-1} + 2^{-2} + 2^{-5}) * 2^2{}_{10} = 0|110010000000|110_F$$
$$0.8_{10} = (1 + 2^{-1} + 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + 2^{-12}) * 2^{-1}{}_{10} \approx 0|100110011001|011_F$$
$$0.046875_{10} = (2^{-2} + 2^{-3}) * 2^{-3}{}_{10} = 0|001100000000|000_F$$

For the last conversion, we assume that the denormalized mode of this floating point representation implicitly includes an exponent of -3 (this makes the first bit in the mantissa of a denormalized number weigh 2^{-3}). This behaviour is akin to the denormalized implementation of IEEE 754 floating point numbers.

Point b)

$$1|011010111000|110_F = -(1 + 2^{-2} + 2^{-3} + 2^{-5} + 2^{-7} + 2^{-8} + 2^{-9}) * 2^2{}_{10} \approx -5.6796875$$
$$1|101010101010|010_F = -(1 + 2^{-1} + 2^{-3} + 2^{-5} + 2^{-7} + 2^{-9} + 2^{-11}) * 2^{-2}{}_{10} \approx -0.4166259766$$

Point c)

$$1|000000000000|001_F = 2^{-3}_{10} = 0.125_{10}$$

Point d)

$$1|111111111111|111_F =$$
$$= (1 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5} + 2^{-6} + 2^{-7} + 2^{-8} + 2^{-9} + 2^{-10} + 2^{-11} + 2^{-12}) * 2^3{}_{10} =$$
$$= 15.998046875_{10}$$

Point e)

With 12 independent binary choices (bits to flip), there are 2^{12} different denormalized numbers in this encoding.

Point f)

With 12 independent binary choices (bits to flip) and 3 extra bits for the exponent, there are $2^{15} - 1$ different denormalized numbers in this encoding. We subtract 1 in order to be consistent with the assumption in point *a)*, since 0.125_{10} would be representable both as $1|000000000000|001_F$ and as $1|100000000000|000_F$

Question 2

Point a)

$$\sqrt[3]{1+x}-1=(\sqrt[3]{1+x}-1)\cdot\frac{\sqrt[3]{(1+x)^2}+\sqrt[3]{1+x}+1}{\sqrt[3]{(1+x)^2}+\sqrt[3]{1+x}+1}=\frac{(1+x)-1}{\sqrt[3]{(1+x)^2}+\sqrt[3]{1+x}+1}=$$
$$\frac{x}{\sqrt[3]{(1+x)^2}+\sqrt[3]{1+x}+1}$$

Point b)

$$\frac{1-cos(x)}{sin(x)}=\frac{sin^2(x)cos^2(x)-cos(x)}{sin(x)}\cdot\frac{sin(x)}{cos(x)}\cdot\frac{cos(x)}{sin(x)}=(sin^2(x)cos(x)-1)\cdot\frac{cos(x)}{sin(x)}$$

Point c)

$$\frac{1}{1-\sqrt{x^2-1}}=\frac{1+\sqrt{x^2-1}}{(1-\sqrt{x^2-1})(1+\sqrt{x^2-1})}=\frac{1+\sqrt{x^2-1}}{1-(x^2-1)}=-\frac{1+\sqrt{x^2-1}}{x^2}$$

Point d)

$$x^3\cdot\left(\frac{x}{x^2-1}-\frac{1}{x}\right)=x^3\cdot\left(\frac{x^2-x^2+1}{x^3-x}\right)=\frac{x^2}{x^2-1}$$

Point e)

$$\frac{1}{x}-\frac{1}{x+1}=\frac{x+1-x}{x^2+x}=\frac{1}{x^2+x}$$

Question 3

Point a)

Consider the Taylor expansion with $a=x$ of $f(x+h)$ and $f(x-h)$:

$$f(x+h)\geq f(x)+\frac{f'(x)}{1}\cdot h+\frac{f''(x)}{2}\cdot h^2+\frac{f'''(x)}{6}\cdot h^3$$

$$f(x-h)\geq f(x)-\frac{f'(x)}{1}\cdot h+\frac{f''(x)}{2}\cdot h^2-\frac{f'''(x)}{6}\cdot h^3$$

Then, we can derive that:

$$\frac{f(x+h)-f(x-h)}{2h}\geq\frac{1}{2h}\cdot\left(2hf'(x)+\frac{2h^3f'''(x)}{6}\right)=f'(x)+\frac{h^2f'''(x)}{6}$$

So:

$$\left|f'(x)-\frac{f(x+h)-f(x-h)}{2h}\right|\leq\left|f'(x)-\left(f'(x)+\frac{h^2f'''(x)}{6}\right)\right|=\frac{h^2|f'''(x)|}{6}\Rightarrow C=\frac{f'''(x)}{6}$$

Point b)

In order to find a valid constant for the entire domain $[-10, 10]$ we find the constant for the value of x that maximizes $f'''(x)$, hence the sups.

Function 1

$$f(x) = e^{-x^2} \qquad f'''(x) = -4e^{-x^2}(2x^3 - 3x) \qquad C = \sup_{x\in[-10,10]} \frac{f'''(x)}{6} = 2\sqrt{1-\sqrt{\frac{2}{3}}e^{\sqrt{\frac{3}{2}}-\frac{3}{2}}}$$

Function 2

$$f(x) = x^2 \qquad f'''(x) = 0 \qquad C = \sup_{x\in[-10,10]} \frac{f'''(x)}{6} = 0$$

Function 3

$$f(x) = sin(x) \qquad f'''(x) = -cos(x) \qquad C = \sup_{x\in[-10,10]} \frac{f'''(x)}{6} = \frac{1}{6}$$

Question 4

Point a)

$$A_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 4 & 4 & 4 \\ 3 & 7 & 10 & 10 & 10 \\ 4 & 10 & 16 & 20 & 20 \\ 5 & 13 & 22 & 30 & 35 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

$$l_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}, u_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}, A_2 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 2 & 2 & 2 \\ 0 & 4 & 7 & 7 & 7 \\ 0 & 6 & 12 & 16 & 16 \\ 0 & 8 & 17 & 25 & 30 \end{bmatrix}$$

$$l_2 = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, u_2 = \begin{bmatrix} 0 & 2 & 2 & 2 & 2 \end{bmatrix}, A_3 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 3 & 3 \\ 0 & 0 & 6 & 10 & 10 \\ 0 & 0 & 9 & 17 & 22 \end{bmatrix}$$

$$l_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}, u_3 = \begin{bmatrix} 0 & 0 & 3 & 3 & 3 \end{bmatrix}, A_4 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 8 & 13 \end{bmatrix}$$

$$l_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 2 \end{bmatrix}, u_4 = \begin{bmatrix} 0 & 0 & 0 & 4 & 4 \end{bmatrix}, A_5 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}$$

$$l_5 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, u_5 = \begin{bmatrix} 0 & 0 & 0 & 0 & 5 \end{bmatrix}, L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 3 & 2 & 1 & 0 & 0 \\ 4 & 3 & 2 & 1 & 0 \\ 5 & 4 & 3 & 2 & 1 \end{bmatrix}, U = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 & 2 \\ 0 & 0 & 3 & 3 & 3 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}$$

Point b)

$$l_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}, e_1^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}, L_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ -3 & 0 & 1 & 0 & 0 \\ -4 & 0 & 0 & 1 & 0 \\ -5 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$l_2 = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, e_2^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}, L_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 \\ 0 & -3 & 0 & 1 & 0 \\ 0 & -4 & 0 & 0 & 1 \end{bmatrix}$$

$$l_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}, e_3^T = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}, L_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 1 & 0 \\ 0 & 0 & -3 & 0 & 1 \end{bmatrix}$$

$$l_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 2 \end{bmatrix}, e_4^T = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}, L_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 \end{bmatrix}$$

Point c)

$$Ly = b$$

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 3 & 2 & 1 & 0 & 0 \\ 4 & 3 & 2 & 1 & 0 \\ 5 & 4 & 3 & 2 & 1 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 3 & 2 & 1 & 0 & 0 \\ 4 & 3 & 2 & 1 & 0 \\ 5 & 4 & 3 & 2 & 1 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

$$y_1 = \frac{1}{1} = 1$$

$$y_2 = \frac{2 - 2 \cdot 1}{1} = 0$$

$$y_3 = \frac{3 - 3 \cdot 1 - 2 \cdot 0}{1} = 0$$

$$y_4 = \frac{4 - 4 \cdot 1 - 3 \cdot 0 - 2 \cdot 0}{1} = 0$$
$$y_5 = \frac{5 - 5 \cdot 1 - 4 \cdot 0 - 3 \cdot 0 - 2 \cdot 0}{1} = 0$$
$$y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Point d)

$$Ux = y$$
$$U = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 & 2 \\ 0 & 0 & 3 & 3 & 3 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix} y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$
$$x_5 = \frac{0}{5} = 0$$
$$x_4 = \frac{0 - 4 \cdot 0}{4} = 0$$
$$x_3 = \frac{0 - 3 \cdot 0 - 3 \cdot 0}{3} = 0$$
$$x_2 = \frac{0 - 2 \cdot 0 - 2 \cdot 0 - 2 \cdot 0}{2} = 0$$
$$x_1 = \frac{0 - 1 \cdot 0 - 1 \cdot 0 - 1 \cdot 0 - 1 \cdot}{1} = 0$$
$$x = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Question 5

Point a)

$$f(x) = x \qquad K_{abs} = |f'(x)| = 1 \qquad K_{rel} = \left| \frac{1 \cdot x}{x} \right| = 1$$

Point b)

$$f(x) = \sqrt[3]{x} \qquad K_{abs} = |f'(x)| = \frac{1}{3\sqrt[3]{x^2}} \qquad K_{rel} = \left| \frac{1}{3\sqrt[3]{x^2}} \cdot \frac{x}{\sqrt[3]{x}} \right| = \frac{1}{3}$$

Point c)

$$f(x) = \frac{1}{x} \qquad K_{abs} = |f'(x)| = \frac{1}{x^2} \qquad K_{rel} = \left| \frac{-x}{x^2} \cdot \frac{1}{\frac{1}{x}} \right| = 1$$

Point d)

$$f(x) = e^x \qquad K_{abs} = |f'(x)| = e^x \qquad K_{rel} = \left| \frac{xe^x}{e^x} \right| = |x|$$

Point e)

Cases *a),b)* and *c)* are well-conditioned for any x since their K_rel is not defined by x. Case *d)* is well-conditioned only for xs whose absolute value is in the order of magnitude of 1 or less, since K_rel in this case is exactly $|x|$.