Numerit Ass 1

Q1:

a.

Initial number	Times 2	Rest
0.1	0.2	0
0.2	0.4	0
0.4	0.8	0
0.8	1.6	1
0.6	1.2	1
0.2	0.4	0 …

So:

$(0.1)_2 = 0.0\overline{0011}$

$(\widetilde{0.1})_2 = (0.00011001100110011001100) = 0.1_{10} - 2^{-20} \cdot 0.1_{10} = 0.1_{10} \cdot (1 - 2^{-20})$

b. $\Delta(\widetilde{0.1}) = |\widetilde{0.1} - 0.1| = \left(0.\overbrace{0 \dots 0}^{23}\overline{1100}\right)_2 = 2^{-20} \cdot 0.1_{10}$

$\delta(\widetilde{0.1}) = \frac{\Delta(\widetilde{0.1})}{0.1} = \left(\frac{0.\overbrace{0 \dots 0}^{23}\overline{1100}}{0.0\overline{0011}}\right)_2 = \frac{0.0\overline{0011} \cdot 2^{-20}}{0.0\overline{0011} \cdot 2^{0}} = 2^{-20}$

c. $\delta(\widetilde{0.1}) \le 2^{1-d} \to 2^{-20} \le 2^{1-d} \to -20 \le 1 - d \to d \le 21 \to d = 21$

d. $n_1 = 8_h = (8 \cdot 60 \cdot 60 \cdot 10)_{10nth} = (288{,}000)_{10nth}$

$n_2 = 8_h + 2_s = n_1 + (2 \cdot 10)_{10nth} = (288{,}020)_{10nth}$

$\tilde{t}_2 - \tilde{t}_1 = \widetilde{0.1} \cdot n_2 - \widetilde{0.1} \cdot n_1 = \widetilde{0.1} \cdot 20$

$t = t_2 - t_1,\ \tilde{t} = \tilde{t}_2 - \tilde{t}_1 \to \Delta\tilde{t} = |t - \tilde{t}| = |0.1 \cdot 20 - \widetilde{0.1} \cdot 20| = |20 \cdot \Delta(\widetilde{0.1})| = 20 \cdot 2^{-20} \cdot 0.1 = 2^{-19}$

$\delta(\tilde{t}) = \frac{\Delta\tilde{t}}{t} = \frac{2^{-19}}{t_2 - t_1} = \frac{2^{-19}}{0.1 \cdot 20} = 2^{-20}$

e. Notice that at section d we didn't depend on the absolute number of hours, but on the difference between n2 and n1, which does't changed in this section. So we get the same answer.

f. $\tilde{t} = \tilde{t}_2 - \tilde{t}_1 = 0.1 \cdot n_2 - \widetilde{0.1} \cdot n_1 = 0.1 \cdot n_2 - 0.1_{10} \cdot (1 - 2^{-20}) \cdot n_1 = 0.1 \cdot (288{,}020 - (1 - 2^{-20}) \cdot 288{,}000) = 2.02746582$

$\Delta(\tilde{t}) = |t - \tilde{t}| = |2 - 2.02746582| = 0.02746582$

$\delta(\tilde{t}) = \frac{\Delta(\tilde{t})}{t} = \frac{0.027465}{2} = 0.0137329102$

g. $\tilde{t} = \tilde{t}_2 - \tilde{t}_1 = 0.1 \cdot n_2 - \widetilde{0.1} \cdot n_1 = 0.1 \cdot n_2 - 0.1_{10} \cdot (1 - 2^{-20}) \cdot n_1 = 0.1 \cdot (3{,}600{,}020 - (1 - 2^{-20}) \cdot 3{,}6000{,}000) = 2.343322754$

$\Delta(\tilde{t}) = |t - \tilde{t}| = |2 - 2.343322754\,| = 0.343322754$
$\delta(\tilde{t}) = \frac{\Delta(\tilde{t})}{t} = \frac{0.343322754}{2} = 0.171661377$

Q2 a. 32bit:

1. The smallest positive normalized number with 32bit in IEEE754 is when we have the lowest normalized exponent possible which is 1, and then by subtracting the bias we will get exponent of -126 and the mantissa will be all zeros.

   So the number will be: $\left(1.\overbrace{0\ldots0}^{23}\right)_2 \cdot 2^{-126}$ $=2^{-126}$

2. The smallest positive un-normalized number with 32bit in IEEE754 is when we have the exponent field at 0, which indicates the exponent to be -126 and that the number is un-normalized and starts with 0., and the mantissa will be all zeros and 1 at the last bit.

   So the number will be:$\left(0.\overbrace{0\ldots0}^{22}1\right)_2 \cdot 2^{-126}$$=2^{-149}$

b.

1. The smallest positive normalized number with <u>64</u>bit in IEEE754 is when we have the lowest normalized exponent possible which is 1, and then by subtracting the bias we will get exponent of -1022 and the mantissa will be all zeros.

   So the number will be: $\left(1.\overbrace{0\ldots0}^{52}\right)_2 \cdot 2^{-1022}$ $=2^{-1022}$

2. The smallest positive un-normalized number with 64bit in IEEE754 is when we have the exponent field at 0, which indicates the exponent to be -1022 and that the number is un-normalized and starts with 0., and the mantissa will be all zeros and 1 at the last bit.

   So the number will be:$\left(0.\overbrace{0\ldots0}^{51}1\right)_2 \cdot 2^{-1022}$$=2^{-1022}$

Q3

a. $\Delta(\tilde{x} - \tilde{y}) \leq^{?} \Delta(\tilde{x}) + \Delta(\tilde{y})$

$$\Delta(\tilde{x} - \tilde{y}) = |(x-y) - (\tilde{x} - \tilde{y})|$$

$$= |(x-\tilde{x}) + (\tilde{y} - y)| \overset{\text{משולש}}{\leq} |x - \tilde{x}| + |\tilde{y} - y|$$
$$= |x - \tilde{x}| + |y - \tilde{y}| = \Delta(\tilde{x}) + \Delta(\tilde{y})$$

b. $\delta\left(\frac{\tilde{x}^2}{\tilde{y}^2}\right) \lesssim^? 2(\delta\tilde{x} + \delta\tilde{y})$

We saw in class that for x,y we get $\delta(\tilde{x} \cdot \tilde{y}) \lesssim \delta(\tilde{x}) + \delta(\tilde{y})$.

We saw in practical session that for x,y we get $\delta\left(\frac{\tilde{x}}{\tilde{y}}\right) \lesssim \delta(\tilde{x}) + \delta(\tilde{y})$

We will use both of those inequalities to prove that $\delta\left(\frac{\tilde{x}^2}{\tilde{y}^2}\right) \lesssim^? 2(\delta\tilde{x} + \delta\tilde{y})$

$$\delta\left(\frac{\tilde{x}^2}{\tilde{y}^2}\right) = \delta\left(\frac{\tilde{x}}{\tilde{y}} \cdot \frac{\tilde{x}}{\tilde{y}}\right) \lesssim \delta\left(\frac{\tilde{x}}{\tilde{y}}\right) + \delta\left(\frac{\tilde{x}}{\tilde{y}}\right) \lesssim \delta(\tilde{x}) + \delta(\tilde{y}) + \delta(\tilde{x}) + \delta(\tilde{y}) = 2(\delta\tilde{x} + \delta\tilde{y})$$

Q4 $f(x,y) = Z = e^{\alpha(x-y)}$

$f_x = \alpha \cdot Z, f_y = -\alpha \cdot Z$

a. $\Delta\tilde{Z} \approx |\nabla f(x,y)| \cdot |(\Delta\tilde{x}, \Delta\tilde{y})| = |\alpha \cdot Z| \cdot \Delta\tilde{x} + |\alpha \cdot Z| \cdot \Delta\tilde{y} = |\alpha \cdot Z| \cdot (\Delta\tilde{x} + \Delta\tilde{y})$

$\delta(\tilde{Z}) = \frac{\Delta(\tilde{Z})}{|Z|} = \frac{|\alpha \cdot Z| \cdot \Delta\tilde{x} + |\alpha \cdot Z| \cdot \Delta\tilde{y}}{|Z|} = \frac{|\alpha Z| \cdot (\Delta\tilde{x} + \Delta\tilde{y})}{|Z|} = |\alpha| \cdot (\Delta\tilde{x} + \Delta\tilde{y})$

b. Let $\Delta X = \Delta Y = 2$:

We want $\delta(\tilde{Z}) \leq 0.05$

From a:

$$\delta(\tilde{Z}) = |\alpha| \cdot (\Delta\tilde{x} + \Delta\tilde{y}) = |\alpha|(2+2) = 4|\alpha| \leq 0.05 \rightarrow -0.05 \leq 4\alpha \leq 0.05$$
$$\rightarrow \boldsymbol{-0.0125 \leq \alpha \leq 0.0125}$$

Q5.

<u>Program output</u>:

err 70 = 0.0

err 7000 = 27.001

err72 - err70 = 5.551115123125783e-17

err8002 - err8000 = 0.008000000000002672

a. err 70 = 0.0

err 7000 = 27.001

- שגיאת הצובר גדולה יותר ככל שעובר הזמן מכיוון שאנחנו תומכים רק ב-3 ספרות עשרוניות, אז כאשר מגיעים למצב שיש באקומולטור 1 ואנחנו מנסים להוסיף לו 0.00x עבור $1 \le x \le 9$ אז לאחר החיתוך המנטיסה לא משתנה, ולכן השגיאה גדלה.

b. err72 - err70 = 5.551115123125783e-17

err8002 - err8000 = 0.008000000000002672

- ההפרשים הנ"ל כל כך שונים זה מזה מכיוון שלאחר 70 איטרציות השגיאה היא 0, ולאחר 72 איטרציות השגיאה היא בערך 10^{-17}, כלומר השגיאה מתחילה להצטבר החל מהצעד ה72.

בתיאוריה, ההפרש בין err_72 ל- err_70 הוא 0, כיוון שבצעדים אלו כמעט ואין שגיאה, ובין err_8002 ל- err_8000 הוא בערך 0.008 כי זה ההפרש האבסולוטי בין המספרים האמיתיים שיחושבו בצעדים ה-8002 וה- 8000.

קוד התכנית מצורף בקובץ נפרד.