Math 125 Hw 1

1a) Cancellation at $x=0$ as $f(x)=\frac{0}{0}$

To fix: $\frac{\sqrt{25+x}-5}{x}\left(\frac{\sqrt{25+x}+5}{\sqrt{25+x}+5}\right)$

$= \frac{x}{x\sqrt{25+x}+5} = \frac{1}{\sqrt{25+x}+5} = g(x)$

$\lim_{x\to 0} \frac{1}{\sqrt{25+x}+5} = \frac{1}{10}$ $\quad \lim_{x\to 0} \frac{\sqrt{25+\cancel{x}}-5}{x} = \lim_{x\to 0} \frac{1}{2\sqrt{25+x}} = \frac{1}{10}$

b) Cancellation at $x=0$ $f(x)=\frac{0}{0}$

To fix, note for small values of x, $\sin x = x$.

So, $f(x) = \frac{1-\cos x}{\sin^2 x} = \frac{1-\cos x}{1-\cos^2 x} = \frac{1-\cos x}{(1-\cos x)(1+\cos x)} = \frac{1}{1+\cos x}$

$\lim_{x\to 0} \frac{1}{1+\cos x} = \frac{1}{2}$ $\quad \lim_{x\to 0} \frac{1-\cos x}{x^2} = \lim_{x\to 0} \frac{\sin x}{2x} = \lim_{x\to 0} \frac{\cos x}{2} = \frac{1}{2}$

c) Catastrophic cancellation for $x = \pm 2\pi n$ where $n \in \mathbb{N}$.

$f(\Rightarrow \frac{0}{0}$

To fix $\lim_{x\to\pm 2\pi n} \frac{1-\sec(x)}{1-\sec^2 x} \Rightarrow \lim_{x\to 0} \frac{1-\cancel{\sec x}}{(1-\cancel{\sec x})(1+\sec x)}$ $\quad \lim_{x\to\pm 2\pi n} \frac{1}{1+\sec x} = \frac{1}{2}$

So use $\frac{1}{1+\sec x}$ which is same form, via trig identities, and $\lim_{x\to\pm 2\pi n} \frac{1-\sec x}{\tan^2 x} = \lim_{x\to\pm 2\pi n} \frac{1}{1+\sec x} = \frac{1}{2}$

2a) $\det(A) = \det(L)\det(U)$

$\det(A) = (L_1 \cdot L_2 \cdot \ldots L_k)(U_1 \cdot U_2 \ldots U_k)$

Where $L_1 \ldots L_k$ and $U_1 \ldots U_k$ are diagonal element.

To fix under/overflow, take log:

$\log(\det A) = \log(L_1 \cdot L_2 \ldots L_k \cdot U_1 \cdot U_2 \cdot \ldots \cdot U_k)$

Scanned with CamScanner

$$\log(\det A) = \sum_{i=1}^{K} \log L_i + \log U_i$$

This handles under/over flow, since it scales the numbers, like $\log 10^{-26} = -26$, and $\log 10^{26} = 26$. The original number can be easily recovered by doing $10^{\log \det A}$

2b) Let $M = \max(x_1, x_2 \ldots x_n)$

$$\vec{x} = \vec{x} \cdot \frac{M}{M} = \begin{bmatrix} \frac{M}{M} x_1 \\ \vdots \\ \frac{M}{M} x_n \end{bmatrix}$$

So $$\|\vec{x}\| = \sqrt{\left(\frac{M}{M}x_1\right)^2 + \left(\frac{M}{M}x_2\right)^2 + \ldots \left(\frac{M}{M}x_n\right)^2}$$

$$= M\sqrt{\left(\frac{x_1}{M}\right)^2 + \left(\frac{x_2}{M}\right)^2 + \ldots \left(\frac{x_n}{M}\right)^2}$$

- This handles overflow, however, underflow isn't really an issue with finding $\|\vec{x}\|$.

3a) $101110 = 46$ in base 10

b) $0.375 = 0.011$ $\quad .375 = \frac{1}{4} + \frac{1}{8}$

$1.25 = 1.010$

$1.25 + .375 = 1.625$

In binary:

$$\begin{array}{r} \overset{1}{0}.011 \\ +\ 1.010 \\ \hline 1.101 \end{array} = 1.625 = \text{base } 10$$

Scanned with CamScanner

M125

3c) For float: $\boxed{1}\,.\,\boxed{23}\,\boxed{8}$

sign fraction exponent

$$x = (-1)^s 2^{e-127}(1+f)$$

For smallest:

$e=1$, as All $e \neq 0$'s, so smallest $= 2^{1-127}(1+0) = 2^{-126}$

$f=0$

$= 1.175 \times 10^{-38}$

Largest:

All ones, minus the first one, so $e = 254$ as all 1's $= 255$

$$\text{Largest} = 2^{254-127} \cdot (1 + 2^{-1} + 2^{-2} + 2^{-3} + \dots 2^{-23})$$

$$= 2^{127}\left(1 + \sum_{i=1}^{23}\left(\tfrac{1}{2}\right)^i\right)$$

$$= 2^{127}\left(1 + \frac{1-(\frac{1}{2})^{24}}{1-\frac{1}{2}}\right)$$

$$= 2^{127} \cdot 1.9999998 = 3.403 \times 10^{38}$$

$$\boxed{\begin{array}{l}\text{Max float: } 3.403 \times 10^{38} \\ \text{Min float: } 1.175 \times 10^{-38}\end{array}}$$

d) π doesn't have an exact representation as it is irrational

e) $\forall$ float $\in [1,2]$, we multiply by 2^k.

Our domain is now $[1 \cdot 2^k, 2 \cdot 2^k] = [2^k, 2^{k+1}]$

Since we multiply all floats by 2^k, no new floats are introduced. However, there are the same amount of floats, and as for $k > 0$, $2^{k+1} - 2^k > 1$, then the gaps between large numbers is bigger than the gaps between small numbers.

Scanned with CamScanner

4 To show $\frac{|y-fl(y)|}{|y|} \leq B^{-m}$, First let's determine an upper band for $|y-fl(y)|$

$$y = \pm B^e \sum_{k=\infty}^{8} d_k B^{-k} = \pm\left(B^e \sum_{k=\infty}^{m+1} d_k B^{-k} + B^e \sum_{k=m}^{8} d_k B^{-k}\right)$$

Finite geometric convergence

$$B_e \sum_{k=\infty}^{m+1} d_k B^{-k} \leq B^e \sum_{k=\infty}^{m+1} (B-1) B^{-k}$$

$$\leq B^e (B-1) \sum_{k=\infty}^{m+1} B^{-k}$$

$$\leq B^e (B-1) \cdot \frac{B^{-(m+1)}}{\frac{B-1}{B}}$$

$$B_e \sum_{k=\infty}^{m+1} d_k B^{-k} \leq \frac{B^e (B^{-m-1})}{B} = B^e B^{-m}$$

since in denominator

For an upper band on $|y|$, $|y| \geq B_e \sum_{k=0}^{\infty} d_k B^{-k} \geq B^e$

as minimally, $|y| = B^e B^{-k}$ as $d_1 \ldots d_k = 0$

So $\frac{|y-fl(y)|}{|y|}$ Note for $|y-fl(y)|$, $B^e \sum_{k=m}^{c} d_k B^{-k}$ cancel so we just use our bands

$$\frac{|B^e B^{-m}|}{|B^c|} \leq B^{-m}, \rightarrow B^{-m} \leq B^{-m}$$

$\square$

5 a) $p(x) = x^5(3x^2 + 2x^3 - 2x^6 + 9x^9)$

$= x^5(x^2(3 + 2x - 2x^4 + 9x^7))$

$= x^5(x^2(3 + x(2 - 2x^3 + 9x^6)))$

$= x^5(x^2(3 + x(2 + x^3(-2 + 9x^3))))$

$\boxed{p(x) = x^3 \cdot x^2 (x^2 (3 + x(2 + x^3(-2 + 9x^2 \cdot x))))}$

Store x^2 and x^3 upon initial computations for reuse.

5 b) $C_K = a_K$

$C_{K-1} = a_{K-1} + a_K x_0$

$C_{K-2} = a_{K-2} + (a_{K-1} + a_K x_0) x_0$

$C_{K-3} = a_{K-3} + (a_{K-2} + a_{K-1} x_0 + a_K x_0^2) x_0$

$\downarrow$ Following this pattern

$C_{K-K} = a_{K-K} + (a_{K-(K+1)} + a_{K-(K+2)} x_0 + \dots + a_K x_0^{K-1}) x_0$

$\downarrow$

$C_0 = a_0 + (a_1 + a_2 x_0 + \dots + a_K x_0^{K-1}) x_0$

$C_0 = a_0 + a_1 x_0 + a_2 x_0^2 + \dots + a_K x_0^K$

Since P is a polynomial, the most efficient way to factor is Horner's Scheme which takes 2K operations for a degree K polynomial.

c) See Code for specifics. My code got value of 6891205, which matches calculator. Code at end of PDF on last page.

d) The most efficient method would be the following steps.

1.) $x+1$

2.) $(x+1)^2$

3.) $((x+1)^2)^2 = (x+1)^4$

4.) $((x+1)^4)^2 = (x+1)^8$

This doesn't Contradict Horner's Scheme as Horner's Scheme is most efficient w/ no structure in $p(x)$.

6 a) If backwards stable, $\tilde{f}(x,y) = f(\tilde{x},\tilde{y})$

$fl(x) \ominus fl(y) = [x(1+\epsilon_1) + y(1+\epsilon_2)](1+\epsilon_3)$

Let $\tilde{x} = x(1+\epsilon_1)(1+\epsilon_3)$, $\tilde{y} = y(1+\epsilon_2)(1+\epsilon_3)$

We must prove $\frac{|\tilde{x}-x|}{|x|} \leq \epsilon_{mc}$ and $\frac{|\tilde{y}-y|}{|y|} \leq \epsilon_{m}$.

$$\frac{|\tilde{x}-x|}{|x|} = \frac{|x(1+\epsilon_1)(1+\epsilon_3)-x|}{|x|} = \frac{|x+x\epsilon_1+x\epsilon_3+x\epsilon_1\epsilon_3-x|}{|x|}$$

$$= \frac{|x\epsilon_1+x\epsilon_3+x\epsilon_1\epsilon_3|}{|x|} = \epsilon_1+\epsilon_3+\epsilon_1\epsilon_3$$

Since ϵ is small,

$$|\epsilon_1+\epsilon_3+\epsilon_1\epsilon_3| \leq |\epsilon_1|+|\epsilon_3|+|\epsilon_1\epsilon_3|$$
$$\leq 3(\epsilon_{machine})$$
$$\leq O(\epsilon_{machine})$$

For y, it's a similar process.

$$\frac{|\tilde{y}-y|}{|y|} = \frac{|y(1+\epsilon_2)(1+\epsilon_3)-y|}{|y|} = \frac{|y\epsilon_2+y\epsilon_3+y\epsilon_2\epsilon_3|}{|y|}$$

$$= |\epsilon_2+\epsilon_3+\epsilon_2\epsilon_3| \leq |\epsilon_2|+|\epsilon_3|+|\epsilon_2\epsilon_3|$$
$$\leq 3\epsilon_{machine}$$
$$\leq O(\epsilon_{machine})$$

As a result, $\tilde{f}(x,y) = f(\tilde{x},\tilde{y})$ and subtraction is backwards stable.

b) $xy^T = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} [y_1 \cdots y_n]$

$$= \begin{bmatrix} x_1y_1(1+\epsilon_1) & \cdots & x_1y_n(1+\epsilon_j) \\ \vdots & & \\ x_ny_1(1+\epsilon_n) & \cdots & x_ny_n(1+\epsilon_m) \end{bmatrix}$$

Each multiplication has it's own ϵ.

Normally, $x \cdot y^T = \begin{bmatrix} x_1y_1 & \cdots & x_1y_m \\ \vdots & & \\ x_ny_1 & \cdots & x_ny_m \end{bmatrix}$ = rank 1.

However by introducing $(1+\epsilon)$ we don't preserve ranks anymore, as xy^T can no longer be rank 1, as each product is multiplied by a unique $(1+\epsilon)$ term.

7 a) Number converted to binary is:
0.099999990463256835988
- I'm calling it 0.0999999 for simplicity

$$\left|0.0999999 - \frac{1}{10}\right| = \boxed{0.0000001}$$

b) $\frac{.0000001 \text{ error}}{1/10 \text{ sec}} = \frac{.000001 \text{ error}}{\text{sec}} \cdot \frac{60 \text{ sec}}{\text{min}} \cdot \frac{60 \text{ min}}{\text{hr}} \cdot \frac{100 \text{ hr}}{\text{operation}}$

$= 0.036$ hrs off $= \boxed{129.6 \text{ seconds}}$

c) $\frac{3750 \text{ miles}}{1 \text{ hr}} \cdot \frac{1 \text{ hr}}{60 \text{ min}} \cdot \frac{1 \text{ minute}}{60 \text{ sec}} = 1.04$ miles

$1.04 \text{ miles} \times 129.6 \text{ sec} = \boxed{134.784 \text{ miles}}$

Scanned with CamScanner

```python
#This function implements Horner's method in Python
def horner(coeff, x):
    result = coeff[0]
    for i in range(1, len(coeff)):
        result = result*x+coeff[i]
    return result

#p(x) = 7x^3 -11x^2 +12x + 5 is polynomial in question
#Horner's method means p(x) = 5 + x(12 + x(-11+7x)))

coeff = [7, -11, 12, 5] #Coefficients of p(x) for use
x = 100 #Initial value
print(horner(coeff, x)) #Prints result
```