NC Lecture Notebook

Contents

1 Introduction

Examples:

1. Compute the partial sums of the harmonic series

$$\sum_{k=1}^{n} \frac{1}{k}$$

   - $S(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n-1} + \frac{1}{n}$
   - $s(n) = \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + 1$

   Mathematically, $S(n) = s(n)$; Computationally, they're not.

   The difference growth with n

2. Let $f(x) = \sqrt{x}$ for $x > 0$, and we know $f'(x) = \dfrac{1}{2\sqrt{x}}$

   Define a function

$$\begin{aligned} y(k) &= \frac{f(16+k) - f(16)}{k} \\ &= \frac{\sqrt{16+k} - 4}{k} \end{aligned}$$

   then, $\lim_{k\to 0} y(k) = f'(16) = \frac{1}{8}$

   As k decrease to $k = 10^{-12}$, $y(k)$ is a good approximation for $f'(16) = \dfrac{1}{8}$, when k further decrease, y(k) starts to oscillates and the errors are visible; after k drops below 10^{-14}, the computed $y(k)$ is around 0.

$$\frac{\sqrt{16+k} - 4}{k} = \frac{1}{\sqrt{16+k} + 4}$$

   $\dfrac{\sqrt{16+k} - 4}{k}$ goes to 0.

3. Consider the function $y(x) = (x-1)^8$ and it's expanded form

$$y(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1$$

   Again, two mathematically identical function are nor the same computationally.

   The expected form even leads to negative values.

1.1 Computational Complexity : Polynomial Evaluation

$$p(x) = 2x^4 + 3x^3 - 3x^2 + 5x - 1$$

1. Straight Forward:

$$\begin{aligned} 2 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} & \qquad 4(x) \\ 3 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} & \qquad 3(x) \\ (-3) \cdot \frac{1}{2} \cdot \frac{1}{2} & \qquad 2(x) \\ 5 \cdot \frac{1}{2} & \qquad 1(x) \end{aligned}$$

$N_x = 10, N_+ = 4$

2. Storage

$$\left(\frac{1}{2}\right)^2 = \frac{1}{4} \qquad 1(x)$$
$$\left(\frac{1}{2}\right)^3 = \left(\frac{1}{2}\right)^2 \cdot \frac{1}{2} = \frac{1}{8} \qquad 1(x)$$
$$\left(\frac{1}{2}\right)^4 = \left(\frac{1}{2}\right)^3 \cdot \frac{1}{2} = \frac{1}{16} \qquad 1(x)$$

multiply by coefficient $4(x)$

$$N_x = 7, N_+ = 4$$

3. Horner's method:

$$\begin{aligned} P(x) &= x(2x^3 + 3x^2 - 3x + 5) - 1 \\ &= x(x(2x^2 + 3x - 3) + 5) - 1 \\ &= x(x(x(2x + 3) - 3) + 5) - 1 \end{aligned}$$

$$N_x = 4, N_+ = 4$$

$$\begin{aligned} N_+ &= 4 \\ N_x &= \begin{cases} \sum_{k=1}^{d} k & = \dfrac{d(1+d)}{2} \\ 2d - 1 \\ d \end{cases} \end{aligned}$$

1.2 Floating Point Arithmetic

Consider -321.416: (Decimal Representation)

$$-321.416 = -(3 \cdot 10^2 + 2 \cdot 10 + 1 \cdot 0 + 4 \cdot 10^{-1} + 1 \cdot 10^{-2} + 6 \cdot 10^{-3})$$

A similar representation is used in computer: floating - point arithmetic:

$$-.321416 \times 10^3$$

sign, fraction, base, exponent
In general,

$$\pm f \times \beta^e$$

where $\beta = 2$: binary number
$\beta = 10$: decimal number
$\beta = 16$: hexadecimal number f: fraction, digits from $0, 1, \cdots \beta - 1$
e: exponent, digits from $0, 1, \cdots \beta - 1$
Binary Number:

$$b_m \cdots b_2 b_1 b_0 . a_1 a_2 \cdots a_n$$

(all integer)

Each digit b_i, a_j takes 0 or 1.
This number in base 10, is

$$b_m \cdot 2^m + b_{m-1} \cdot 2^{m-1} + \cdots + b_1 \cdot 2^1 + b_0 \cdot 2^0 + a_1 \cdot 2^{-1} + a_2 \cdot 2^{-2} + \cdots + a_n 2^{-n}$$

Note:

$$\begin{aligned}(0.1101)_2 &= (1.101)_2 \cdot 2^{-1} \\ &= (0.001101)_2 \cdot 2^3\end{aligned}$$

To convert between binary ($\beta = 2$) and decimal ($\beta = 10$)
Example:

1. $x = (1.1011)_2$ convert x to a decimal number:

$$\begin{aligned}x &= 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} \\ &= 1 + \frac{1}{2} + 0 + \frac{1}{8} + \frac{1}{16} \\ &= \frac{27}{16}\end{aligned}$$

2.

$$\begin{aligned}x &= (1.1010 \cdots 10)_2 \\ &= (1.1\bar{0})_2 \\ &= 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-3} + 1 \cdot 2^{-5} + \cdots\end{aligned}$$

Recall geometric series:

$$1 + r + r^2 + \cdots = \frac{1}{1-r}$$

if $|r| < 1$

$$\begin{aligned}x &= 1 + \frac{1}{2} + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^5 + \cdots \\ &= 1 + \frac{1}{2}\left(1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots\right) \\ &= 1 + \frac{1}{2} \cdot \frac{1}{1 - \frac{1}{4}} \\ &= 1 + \frac{2}{3} \\ &= \frac{5}{3}\end{aligned}$$

Alternatively,

$$\begin{aligned}x &= (1.1\bar{0})_2 \\ &= 1 \cdot 2^0 + (0.1\bar{0})_2 \\ &= 1 + (10.1\bar{0})_2 \cdot 2^{-2}\end{aligned}$$

$$\begin{aligned}
y &= (0.\bar{1}\bar{0})_2 \\
&= (10.\bar{1}\bar{0})_2 \cdot 2^{-2} \\
&= \{(10)_2 \cdot 2^{-2} + (0.\bar{1}\bar{0})_2 \cdot 2^{-2}\} \\
y &= (2 + y) \cdot 2^{-2} \\
4y &= 2 + y \\
y &= \frac{2}{3} \\
x &= 1 + y \\
&= \frac{5}{3}
\end{aligned}$$

3. Convert 14.8125 to a binary number:

   We are looking for

$$14.8125 = (b_m b_{m-1} \cdots b_1 b_0 . a_1 a_2 \cdots a_n)_2$$

   Fractional part:

$$\begin{aligned}
0.8125 &= (.a_1 a_2 \cdots a_n)_2 \\
&= a_1 \cdot 2^{-1} + a_2 \cdot 2^{-2} + \cdots + a_n \cdot 2^{-n}
\end{aligned}$$

   - $*2$

$$\begin{aligned}
1.6250 &= a_1 + a_2 \cdot 2^{-1} + \cdots a_n \cdot 2^{-(n-1)} \\
a_1 &= 1 \\
0.6250 &= a_2 \cdot 2^{-1} + \cdots a_n \cdot 2^{-(n-1)}
\end{aligned}$$

   - $*2$

$$\begin{aligned}
1.2500 &= a_2 + a_3 \cdot 2^{-1} \cdots a_n \cdot 2^{-(n-2)} \\
a_2 &= 1
\end{aligned}$$

   - $*2$

$$\begin{aligned}
0.25 \cdot 2 &= 0.50 & a_3 &= 0 \\
0.50 \cdot 2 &= 1 & a_4 &= 1
\end{aligned}$$

$$\therefore 0.8125 = (.1101)_2$$

   Collect Integer part ordered from radix point:

   Integer part:

$$\begin{aligned}
14 &= (b_m \cdots b_2 b_1 b_0)_2 \\
&= b_m \cdot 2^m + b_{m-1} \cdot 2^{m-1} + \cdots b_1 \cdot 2^1 + b_0
\end{aligned}$$

Divided by 2:

$$
\begin{aligned}
\frac{14}{2} &= 7R0 \\
&= (b_m \cdot 2^{m-1} + \cdots b_1)Rb_0 \\
\frac{7}{2} &= 3R1 && b_1 \\
\frac{3}{2} &= 1R1 && b_2 \\
\frac{1}{2} &= 0R1 && b_3 \\
\therefore 14 &= (1110)_2
\end{aligned}
$$

$$\therefore 14.8125 = (1110.1101)_2$$

1.2.1 Floating point number

$$\pm f \times \beta^e$$

f(fraction) : the number of digits in f determines the precision.
e(exponent):the number of digits in e determines the range of representable numbers.
We follow IEEE 754 floating point standard:

1. Normalized form: $f = 1.b_m b_{m-1} \cdots b_1 b_0$, $(0.0101010\cdots)$
2. Advantage: leading 1 needs not be stored
   - 32-bit single precision:
     sign : 1 -bit
     exponent : 8-bits
     fraction: 23 -bits
   - 64 - bit double precision:
     sign : 1 -bit
     exponent : 11-bits
     fraction: 52-bits
3. The represented number is
   $$(-1)^s \cdot (1+f) \cdot 2^{e-e_0}$$
   - e: unsigned, e^0: exponent bias
   - $e - e^0$: can be either positive or negative (negative represent small number)

   Let's focus on "e" or equivalently 2^{e-e_0}

   - Single Precision:

$$
\begin{aligned}
e &\in [e_{min}, e_{max}] \\
e_{min} &= (0\cdots 01)(8 \text{ bit}) = 1 \\
e_{max} &= (11\cdots 10) \\
&= 1 \cdot (2 + 2^2 + \cdots + 2^7) \\
&= 2 \cdot \left(\frac{1-2^7}{1-2}\right) \\
&= 254 \\
\Longrightarrow 2^{e-e_0} &\in [2^{-126}, 2^{127}] && e_0 = 127 \\
&\approx [10^{-38}, 10^{38}]
\end{aligned}
$$

-
$$\begin{aligned}
e &\in [e_{min}, e_{max}] \\
e_{min} &= 1 \\
e_{max} &= 2^1 + 2^2 + \cdots + 2^{10} \\
&= 2046 \\
e_0 &= 1023 \\
2^{e-e_0} &\in [2^{-1022}, 2^{1023}] \\
&\approx [10^{-308}, 10^{308}]
\end{aligned}$$

1.2.2 fraction f and precision

Using double-precision on an example:

- How to store a number
- How to do calculation

Consider

$$\begin{aligned}
x_1 &= \frac{27}{16} = (1.1011)_2 \\
x_2 &= \frac{5}{3} = (1.\bar{1}\bar{0})_2 \\
x_3 &= \frac{2}{3} = (.\bar{1}\bar{0})_2 = (1.\bar{0}\bar{1})_2 \cdot 2^{-1} \\
x_4 &= 1 && = (1.0)_2 \\
x_5 &= 1 \times 2^{-52} \\
x_6 &= 1 \times 2^{-53}
\end{aligned}$$

$$\begin{aligned}
x_1 &: 1.\boxed{101100\cdots0}(52bits) \\
x_4 &: 1.\boxed{00\cdots0}(52bits) \\
x_5 &: 1.\boxed{00\cdots0}(52bits) \times 2^{-52} \\
x_6 &: 1.\boxed{00\cdots0}(52bits) \times 2^{-53}
\end{aligned}$$

Now x_2, x_3:

$$\begin{aligned}
x_2 &: 1.\boxed{101010\cdots10}10\cdots \\
x_3 &: 1\boxed{0101\cdots01}0101\cdots \times 2^{-1}
\end{aligned}$$

We follow: IEEE rounding to the nearest rule: $x \to fl(x)$