Unit 2 Error Analysis

Numerical Analysis

Mar. 24, 2015

Computer Numbers

- Numbers are represented in computers using finite number of bits.
- In `C` and `C++` we have the following `int` types with various numbers of bits:

Table 2.1.1. Integer Numbers

Type	Bits	Range
char	8	-128 127
short	16	-32,768 32,767
int	32	-2,147,483,648 2,147,483,647
long	64	-9,223,372,036,854,775,808 9,223,372,036,854,775,807
long long	128	-170,141,183,460,469,231,731,687,303,715,884,105,728 170,141,183,460,469,231,731,687,303,715,884,105,727

- Note implementation on each machine can be different
  - Use `#include <limits.h>` to find out the range of each type

Round-off Errors

- When using `int` types, care should be taken to avoid overflow problem
- For example,

```
short i,j;

i=32767;     // largest short number
i++;         // i is now -32768

j=-32768;    // smallest short number
j++;         // j is now 32767
```

- All types of `int` can have overflow problem
- C and C++ will not notify the user when overflow happens
  - It is the programmer's responsibility to guard against this problem

Floating Point Numbers

- Computers use floating point numbers to represent real numbers
- Floating point numbers have the following form

$$\underbrace{\pm 0.dddddd}_{\text{mantissa}}\, e\, \underbrace{\pm bbbb}_{\text{exponent}} \tag{2.1.1}$$

  where mantissa and exponent are usually in binary format with fixed numbers of bits.
- Typical real numbers representable in C and C++ are:

Table 2.1.2. Floating Point Numbers

Type	Bits mantissa	Bits exponent	$\lvert x\rvert_{\min}$	$\lvert x\rvert_{\max}$	ϵ
float	24	8	1.17549e-38	3.40282e+38	1.19209e-07
double	53	11	2.22507e-308	1.79769e+308	2.22045e-16
long double	65	15	3.3621e-4932	1.18973e+4932	1.0842e-19

- Note that numbers in between x and $x(1-\epsilon)$, or between x and $x(1+\epsilon)$ cannot be represented by C or C++.
- Special numbers such as $+\infty$, $-\infty$ and NaN are included in computer floating numbers today.
  - NaN: `not a number`.

Floating Point Numbers, II

- Any number in C or C++ has a finite number of bits.
  - The number of significant digits of any floating numbers are limited.
- Not all real numbers can be represented in a computer.
- Real numbers are approximated in computers.
  - Round-off errors exist in any computer arithmetic.

Example 2.1.3. 4-digit floating point numbers

Assuming a computer's floating number can have 4 significant digits, then

real number	4-digit representation	abs. error δ_{abs}	rel. error δ
1/7 (0.142857142857)	0.1429	4.28571e-05	3e-4
2/7 (0.285714285714)	0.2857	-1.42857e-05	-5e-5
3/7 (0.428571428571)	0.4286	2.85714e-05	6.66667e-5
4/7 (0.571428571429)	0.5714	-2.85714e-05	-5e-5
5/7 (0.714285714286)	0.7143	1.42857e-05	2e-5
6/7 (0.857142857143)	0.8571	-4.28571e-05	-5e-5

Note that the absolute error, δ_{abs}, $|\delta_{abs}| \le$ 5e-5.

Computer Number Approximation

- In C and C++, any real number, y, can be approximated only if $y \in [-|x|_{\max}, |x|_{\max}]$ where $|x|_{\max}$ are given in the table (2.1.2).
- Given any number, $-|x|_{\min} < y < |x|_{\min}$, then y is approximated by 0.
- If $y < -|x|_{\max}$ then y is represented by $-\infty$.
- If $y > |x|_{\max}$ then y is represented by ∞.
- When $-|x|_{\max} \le y \le |x|_{\max}$ then y is represented by $fl(y)$ such that

$$fl(y) = y(1+\delta) \quad \text{with } |\delta| \le \epsilon. \tag{2.1.2}$$

  where ϵ is given in table (2.1.2).
- Given two real numbers x and y, their sum is approximated by

$$fl(x+y) = (x+y)(1+\delta_{x+y}), \quad \text{with } |\delta_{x+y}| \le \epsilon. \tag{2.1.3}$$

  But, sum generated from the individual approximations is (assuming $x+y \neq 0$)

$$\begin{aligned} fl(x) + fl(y) &= x(1+\delta_x) + y(1+\delta_y) \\ &= (x+y)(1 + \frac{x\delta_x + y\delta_y}{x+y}) \end{aligned}$$

Addition/Subtraction Errors

- If $xy > 0$ then it can be shown that

$$0 \le \left| \frac{x\delta_x + y\delta_y}{x+y} \right| \le |\delta_x| + |\delta_y| \le 2\epsilon.$$

- Thus, the error in approximating the sum of two real numbers can accumulate after addition.
- It is also possible that the error decreases after addition (cancellation effect).
- If $xy < 0$ and $x + y \approx 0$, then $\dfrac{x\delta_x + y\delta_y}{x+y}$ can be a large number.
- Thus, the round-off error can be large after arithmetic operations.

Example 2.1.4. 4-digit Subtraction

Using 4-digit computer as the last example.

$$\begin{aligned} x &= \frac{100}{7} = 14.29, \quad \delta_x = 3e-4, \\ y &= \frac{99}{7} = 14.14, \quad \delta_y = -2.0202e-4, \\ x - y &= \frac{100}{7} - \frac{99}{7} = 0.15, \quad \delta_{x-y} = 5e-2. \end{aligned}$$

Thus, the error in subtraction can increase significantly.

Multiplication Errors

- Given two real numbers x and y, the ideal computer product is

$$fl(x \times y) = xy(1 + \delta_{xy}) \qquad \text{with } |\delta_{xy}| \le \epsilon. \tag{2.1.4}$$

But, computer generates

$$\begin{aligned} fl(x \times y) &= x(1+\delta_x) \cdot y(1+\delta_y) \\ &= xy(1 + \delta_x + \delta_y + \delta_x\delta_y) \\ &\approx xy(1 + \delta_x + \delta_y) \end{aligned} \tag{2.1.5}$$

where $\delta_x\delta_y \ll 1$. Thus, the relative error can increase in multiplication.

$$\delta_{xy} \approx \delta_x + \delta_y. \tag{2.1.6}$$

Example 2.1.5. 4-digit Multiplication.

Using 4-digit computer,

$$\begin{aligned} x &= \frac{10}{7} = 1.429, \qquad \delta_x = 3e-4, \\ y &= \frac{30}{7} = 4.286, \qquad \delta_y = 6.66667e-5, \\ x \times y &= \frac{300}{49} = 6.125, \qquad \delta_{xy} = 4.16667e-4, \\ \text{compared to } &\frac{300}{49} = 6.122, \qquad \delta = -7.33333e-5. \end{aligned}$$

Division Errors

- For 2 real numbers x and y, the quotient is

$$fl(\frac{x}{y}) = \frac{x}{y}(1 + \delta_{x/y}), \qquad \text{with } |\delta_{x/y}| \le \epsilon. \tag{2.1.7}$$

  Computer arithmetic generates

$$\begin{aligned} fl(\frac{x}{y}) &= \frac{x(1+\delta_x)}{y(1+\delta_y)} \\ &= \frac{x}{y}\frac{1+\delta_x}{1+\delta_y} \approx \frac{x}{y}(1+\delta_x)(1-\delta_y) \\ &\approx \frac{x}{y}(1+\delta_x-\delta_y) \end{aligned} \tag{2.1.8}$$

- Since δ_x and δ_y can have different signs, the errors can also accumulate.
- The approximation formulas above are valid only if δ is very small.

Associative and Distributive Laws

Example 2.1.6. 4-digit Associative Law.

Using 4-digit computer as the last example.

$$\begin{aligned} x &= \frac{100}{7} = 14.29, \\ y &= -\frac{99}{7} = -14.14, \\ z &= \frac{1}{7} = 0.1429, \\ (x+y)+z &= 0.15 + 0.1429 = 0.2929, \\ x+(y+z) &= 14.29 - 14 = 0.29. \end{aligned}$$

Thus, $(x+y)+z \neq x+(y+z)$.

- It can also be shown that distribution law is not observed in general, that is

$$(x+y)z \neq xz + yz.$$

- Thus, the order of operations is important in numerical analysis.

More Significant numbers

Example 2.1.7. 6-digit Associative Law.

Using 6-digit computer as the last example.

$$
\begin{aligned}
x &= \frac{100}{7} = 14.2857, \\
y &= -\frac{99}{7} = -14.1429, \\
z &= \frac{1}{7} = 0.142857, \\
(x+y)+z &= 0.1428 + 0.142857 = 0.285657, \\
x+(y+z) &= 14.2857 - 14 = 0.2857.
\end{aligned}
$$

Again, $(x+y)+z \neq x+(y+z)$.

- With longer mantissa, the differences become smaller.

Round-off Errors

- Due to the finite length approximation of real numbers by the computer number system, errors exist.
  - With longer mantissa, the approximation error is smaller
  - With longer exponent, the range of the computer number is larger.
- Using longer number system can reduce approximation errors.
- Round-off error can increase significantly after arithmetic operations
  - The order of arithmetic operation is important to the resulting errors.
- Again, using longer number system can reduce arithmetic errors.
  - But need to trade-off the execution time.

Sources of Errors

- Numerical analysis applies some mathematical models to approximate physical problems, then applies numerical methods to solve the numerical equations.
- Possible sources of errors:
  - Inaccurate model,
    - For example, in high doped source/drain region we need to apply concentration dependent mobility model for more accurate device simulations.
  - Inaccurate model parameters,
    - If the concentration dependent mobility model has wrong parameters, then the solution is not vary accurate either.
  - Data errors,
  - Truncation errors in approximating the model equations,
    - Example, $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$
    - Finite series sum results in errors.
  - Discretization errors in approximating the model is space/time dimensions,
    - Example the finite number of N_x and N_y when discretize the conductor for equivalent resistance calculation.
  - Round-off errors in solving the equations.
- To get accurate results, it is important to analyze all possible error sources.

Summary

- Computer number systems
  - Integer types
  - Floating point types
- Round-off errors
  - Arithmetic error propagation
- Sources of errors