

MAT360 Section Summary:
1.2: Roundoff Errors and Computer Arithmetic

1. Definitions

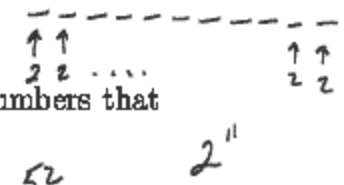
- **long real:** 8 byte real (64 bits):
 - first bit for the sign (positive or negative);
 - 11 bits for the **characteristic** (exponent); and the remaining
 - 52 bits for the **mantissa**, which is the rational representation of the number in the interval from 0 to 1.

“To save storage and provide a unique representation for each floating-point number, a normalization is imposed”, so that the decimal representation of the binary number is

$$(-1)^s 2^{e-1023} (1 + f)$$

(where f is the decimal expansion of the mantissa).

11 bits for exponents gives $2^{11} - 1$ distinct numbers that can be represented;



Quadrillions!

52 bits for mantissa gives $9,007,199,254,740,991 = 2^{52}$ distinct numbers (that's pretty many...).

The largest number that can be represented using this normalized scheme is about 10^{308} , and the smallest about 10^{-308} . Calculations resulting in numbers larger than 10^{308} result in **overflows**, which usually mean “expect junk” (if not an impolite crash); numbers smaller than 10^{-308} result in **underflows**, which generally cause no trouble (they're set to zero).

- **k -digit decimal machine numbers:**

$$\pm 0.d_1 d_2 \dots d_k \times 10^n, \quad 1 \leq d_1 \leq 9, \quad 0 \leq d_i \leq 9$$

- **chopping to a k -digit decimal number:** simply truncating an

$$\pm 0.d_1d_2 \dots d_k d_{k+1} d_{k+2} \dots \times 10^n \approx \pm 0.d_1d_2 \dots d_k \times 10^n$$

-
- **rounding to a k -digit decimal number:** add 5 in the $k+1$ place, then chop.

- **floating-point form:** the form $fl(y)$ of a number y that results from chopping or rounding.

- **roundoff error:** the error that results from replacing a number with its floating-point form.

-
- **absolute error:** $|p - p^*|$

- **relative error:**

$$\frac{|p - p^*|}{|p|}$$

- p^* is said to approximate p to t **significant digits** (or figures) if t is the largest non-negative integers for which

$$\frac{|p - p^*|}{|p|} \leq 5 \times 10^{-t}$$

2. Properties/Tricks/Hints/Etc.

Relative errors for floating-point form:

$$15. \quad p = \pi \quad p^* = 3.1416 \quad 2 \quad \pi = 3.1415926 \dots$$

$$|p - p^*| = .000007346 \dots$$

$$\frac{|p - p^*|}{|p|} = .000002339 \dots$$

2b

$$p=e \quad \left| \frac{e-p^*}{e} \right| \leq 10^{-4} \quad (\Leftrightarrow) \quad |e-p^*| \leq e \times 10^{-4}$$

- chopping: 10^{-k+1}
- rounding: $0.5 \times 10^{-k+1}$

$$\mathbb{I} \sim [2.71901, 2.71955]$$

3. Summary

$$-e \times 10^{-4} \leq e-p^* \leq e \times 10^{-4}$$

$$-e \times 10^{-4} - e \leq -p^* \leq e \times 10^{-4} - e$$

$$e + e \times 10^{-4} \geq p^* \geq e - e \times 10^{-4}$$

$$p^* \in [e - e \times 10^{-4}, e + e \times 10^{-4}] = \mathbb{I}$$

Machine numbers are the approximations we may use for all real numbers. Each is generally stored as a binary number, including information about sign, exponent, and mantissa (with a fixed number of digits dedicated to distinguishing adjacent numbers).

By replacing the infinite number of numbers within the interval of 10^{-308} and 10^{308} by the finite number of machine numbers between those values, we're obviously making some errors. Those errors get compounded as we perform arithmetic operations. Two very dangerous operations are

- The subtraction of nearly equal numbers, resulting in the cancellation significant digits;
- Division by very small numbers (or multiplication by very large numbers).

These two problems can be seen clearly in two standard mathematical computations:

- The quadratic formula (e.g. example 5) and
- Polynomial evaluation (e.g. example 6).

Bad things can happen when

Two k -digit numbers:

$$x_1 = 0.d_1 \dots d_{k-1} d_k \times 10^n$$

$$x_2 = 0.d_1 \dots d_{k-1} e_k \times 10^n$$

exactly the same except for the last digit:

What happens when you subtract?

$x_1 - x_2$ will be stored as a k -digit number, only one of which contains any information!

Where does this happen? Well, in the quadratic formula, for example:

$$\text{roots: } \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad f(x) = ax^2 + bx + c$$

If $b > 0$, then

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

is a dangerous calculation if

$$b^2 \gg 4ac$$

(i.e. $\sqrt{b^2 - 4ac} \sim b$, so the numerator

is essentially $-b + b$...

$$\#4d \quad i) \left(\frac{1}{3} + \frac{3}{11} \right) - \frac{3}{20} = \frac{301}{640} \approx .4560404$$

$$ii) (.3 + .27\bar{27}) - .150$$

$$\approx (.333 + .272) - .150$$

$$= .605 - .150$$

$$= .455$$

$$\begin{aligned} \text{iii) } &\approx (.333 + .273) - .150 \\ &= .456 \end{aligned}$$