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

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

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

  2 2 .... 2 2

  11 bits for exponents gives $\underbrace{2047 = 2^{11} - 1}$ distinct numbers that can be represented;

  +1 = 2048 $\qquad 2^{11}$

  Quadrillions!

  52 bits for mantissa gives $9,007,199,254,740,99\cancel{1}2 = 2^{\cancel{53}\,52}\cancel{-1}$ 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:**

  $$\pm 0.d_1 d_2 \ldots d_k \text{ x } 10^n,\ 1 \le d_1 \le 9,\ 0 \le d_i \le 9$$

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

$$\pm 0.d_1 d_2 \ldots d_k d_{k+1} d_{k+2} \ldots \text{ x } 10^n \approx \pm 0.d_1 d_2 \ldots d_k \text{ x } 10^n$$

- **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

$$\frac{|p - p^*|}{|p|} \quad 5 \text{ x } 10^{-t}$$

2. Properties/Tricks/Hints/Etc.

Relative errors for floating-point form:

1b. $p = \pi$ $\quad p^* = 3.1416$

$\pi = 3.1415924\ldots$

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

$$\frac{|p - p^*|}{|p|} = .000002338\ldots$$

2b $p = e$ $\quad \dfrac{|e - p^*|}{e} \le 10^{-4} \iff |e - p^*| \le e \times 10^{-4}$

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

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

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

$I \sim [2.71801,\ 2.71855]$

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

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

3. Summary

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:

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

$x_2 \quad 0.d_1 \cdots d_{k-1} e_k \times 10^n$ 3

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!

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

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

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 i) $\left(\frac{1}{3} + \frac{3}{11}\right) - \frac{3}{20} = 301/660 \simeq .4560606$

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

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

$= .605 - .150$

$= .455$

iii) $= (.333 + .273) - .150$

$= .456$