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).

  11 bits for exponents gives $2048 = 2^{11}$ distinct powers (orders of binary magnitude) that can be represented;

  52 bits for mantissa gives $4,503,599,627,370,496 = 2^{52}$ distinct numbers per order of magnitude (that seems like 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**:

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

- **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

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

2. Properties/Tricks/Hints/Etc.

Relative errors for floating-point form:

- k-digit chopping: 10^{-k+1}
- k-digit rounding: 0.5 x 10^{-k+1}

3. Summary

Machine numbers are the approximations we may use for all real numbers. It's odd to imagine that we're going to use a bounded finite set of rational numbers to stand for all real numbers, but that's the case.

Each is generally stored as a binary number, including information about sign, exponent (characteristic), 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).