Notes for section 1.2

Math. 481a, Spring 2026

Assume that machine numbers are represented in the following decimal floating-point form (called *k-digit decimal machine numbers*):

$$\pm 0.d_1 d_2 \dots d_k \times 10^n, \quad 1 \le d_1 \le 9, \quad 0 \le d_i \le 9, \quad \text{for} \quad i = 2 \dots k.$$

For any positive real number y (within the machine range),

$$y = 0.d_1 d_2 \dots d_k d_{k+1} d_{k+2} \cdots \times 10^n,$$

the floating-point form $fl(y)$ is obtained by terminating the mantissa of y at k decimal digits. The are two ways of doing it.

Chopping

$$fl(y) = 0.d_1 d_2 \dots d_k \times 10^n.$$

Rounding

$$fl(y) = \begin{cases} 0.d_1 d_2 \dots d_k \times 10^n, & \text{if } d_{k+1} < 5 \\ 0.d_1 d_2 \dots d_k \times 10^n + 10^{n-k}, & \text{if } d_{k+1} \ge 5 \end{cases}$$

The error that results replacing a number with its floating-point form is called **round-off error**.

Definition 1. If p^* is an approximation of p, then $|p - p^*|$ is the **absolute error**. If $p \neq 0$, then $|p - p^*|/|p|$ is the **relative error**.

The absolute error is not necessarily a good measure of accuracy. The relative error takes into account the size of the value and thus is a more meaningful measure of accuracy.

Definition 2. The number p^* approximates p to t **significant digit** if t is the largest nonnegative integer for which

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

Proposition 1. If $fl(y)$ is a k-digit rounding approximation to y then

$$\frac{|y - fl(y)|}{|y|} \le 5 \times 10^{-k}.$$

Proof. Let $y = 0.d_1 d_2 \dots d_k d_{k+1} d_{k+2} \cdots \times 10^n$ and is in the machine range.
Case when $d_{k+1} < 5$.
We have,

$$\frac{|y - fl(y)|}{|y|} = \frac{0.d_{k+1} \cdots \times 10^{n-k}}{0.d_1 \cdots \times 10^n} \le \frac{0.5 \times 10^{n-k}}{0.d_1 \cdots \times 10^n} \le \frac{0.5 \times 10^{n-k}}{0.1 \times 10^n} = 5 \times 10^{-k}.$$

Case when $d_{k+1} \geq 5$.

We have,

$$\frac{|y - fl(y)|}{|y|} = \frac{(1 - 0.d_{k+1}\ldots) \times 10^{n-k}}{0.d_1 \cdots \times 10^n} \leq \frac{(1 - 0.5) \times 10^{n-k}}{0.d_1 \cdots \times 10^n} \leq \frac{0.5 \times 10^{n-k}}{0.1 \times 10^n} = 5 \times 10^{-k}.$$

□

Proposition 2 (see, Exercise 28, page 28)**.** If $fl(y)$ is a k-digit chopping approximation to y then

$$\frac{|y - fl(y)|}{|y|} \leq 10^{-k+1}.$$

Proof. Let $y = 0.d_1 d_2 \ldots d_k d_{k+1} d_{k+2} \cdots \times 10^n$ and is in the machine range. We have,

$$\frac{|y - fl(y)|}{|y|} = \frac{0.d_{k+1} \cdots \times 10^{n-k}}{0.d_1 \cdots \times 10^n} = \frac{0.d_{k+1}\ldots}{0.d_1 \ldots} \times 10^{-k} \leq \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}.$$

□

The conclusions

(1) Floating-point forms obtained through k-digit rounding provide k **significant digits** accuracy.

(2) Floating-point forms obtained through k-digit chopping provide k **significant digits** accuracy only when $d_{k+1} < 5$.

Cautionary notes

(1) Subtractions of nearly equal numbers very often produces additional errors. Indeed, suppose that nearly equal numbers x and y $(x > y)$ have the k-digit representations:

$$fl(x) = 0.d_1 d_2 \ldots d_p \alpha_{p+1} \alpha_{p+2} \ldots \alpha_k \times 10^n,$$

and

$$fl(y) = 0.d_1 d_2 \ldots d_p \beta_{p+1} \beta_{p+2} \ldots \beta_k \times 10^n,$$

The floating-point form of $x - y$ is

$$fl[fl(x) - fl(y)] = 0.\sigma_{p+1} \sigma_{p+2} \ldots \sigma_k \times 10^{n-p},$$

where

$$0.\sigma_{p+1} \sigma_{p+2} \ldots \sigma_k = 0.\alpha_{p+1} \alpha_{p+2} \ldots \alpha_k - 0.\beta_{p+1} \beta_{p+2} \ldots \beta_k.$$

The floating-point form of $x - y$ will have at most $k - p$ significant digits; however, in most systems $x - y$ will be assigned k digits, **with the last p being either zero or randomly assigned**. Any subsequent calculations will retain only $k - p$ significant digits.

(2) The additional error is also being produced when dividing by a number with small magnitude, or equivalently, when multiplying by a number with large magnitude.