

DEN: Errors in calculations**Definition.**

Let x be a number and \hat{x} its estimate. Then we define

absolute error $E_x = x - \hat{x}$

and **relative error** $\varepsilon_x = \frac{|E_x|}{|x|}$ if $x \neq 0$.

By an **error estimate** we mean any number e_x satisfying $|E_x| \leq e_x$.

Definition.

By a **floating point representation** of a number x with respect to base β , with precision of p significant digits we mean the best approximation $f(x)$ of x that can be written as

$$f(x) = d_1.d_2d_3 \cdots d_p \times \beta^e,$$

where $d_1 \in \{1, \dots, \beta - 1\}$ and $d_2, \dots, d_p \in \{0, 1, \dots, \beta - 1\}$.

The number e is called the exponent, the part $d_1.d_2 \cdots d_p$ is called the significand or the mantissa.

Fact.

Assume that a number x was represented as \hat{x} in floating point representation with base β and precision p . Then the relative error is bounded as follows:

$$\varepsilon_x \leq \frac{1}{2}\beta \cdot \beta^{-p}$$

Fact.

Consider real numbers x, y and their estimates \hat{x}, \hat{y} . Then the following are true:

$$\begin{aligned} |E_{x+y}| &\leq |E_x| + |E_y| & \varepsilon_{x+y} &\leq \max(\varepsilon_x, \varepsilon_y) \text{ for } x, y > 0; \\ |E_{x-y}| &\leq |E_x| + |E_y| & \varepsilon_{x-y} &\leq \frac{|x|+|y|}{|x-y|} \max(\varepsilon_x, \varepsilon_y); \\ |E_{x \cdot y}| &\leq |y| \cdot |E_x| + |\hat{x}| \cdot |E_y| & \varepsilon_{x \cdot y} &\leq \varepsilon_x + (1 + \varepsilon_x)\varepsilon_y; \\ |E_{x/y}| &\leq \frac{1}{|y|} (|E_x| + |E_y| \frac{|\hat{x}|}{|\hat{y}|}) & \varepsilon_{x/y} &\leq \varepsilon_x + \varepsilon_y \frac{1+\varepsilon_x}{1-\varepsilon_y} \text{ for } \varepsilon_y < 1; \\ |E_{1/x}| &\leq \frac{|x|}{|\hat{x}|} |E_x| & \varepsilon_{1/x} &\leq \frac{1}{1-\varepsilon_x} \varepsilon_x \text{ for } \varepsilon_x < 1. \end{aligned}$$

Fact.

Assume that there is a rounding error $\varepsilon > 0$ on input, then for $x, y > 0$ we (almost) have

$$\begin{aligned} \varepsilon_{ax+y} &\leq \varepsilon, \\ \varepsilon_{x-y} &\leq \frac{x+y}{|x-y|} \varepsilon, \\ \varepsilon_{x \cdot y} &\leq 2\varepsilon, \\ \varepsilon_{x/y} &\leq 2\varepsilon, \\ \varepsilon_{1/x} &= \varepsilon. \end{aligned}$$