Otto-von-Guericke-University Magdeburg
Max Planck Institute for Dynamics of Complex Technical Systems
Computational Methods for Systems and Control Theory

Dr. Jens Saak, Dipl.-Math. Martin Köhler
Website: http://www.mpi-magdeburg.mpg.de/csc/teaching/18ws/sc1/

Winterterm 2018/2019

Scientific Computing 1
Handout 4
November 15, 2018

Floating Point Numbers and Rounding

- **Normalized floating point representation (with base p)** for $x \in \mathbb{R}$:

$$x = (-1)^j \sum_{i=1}^{\infty} \frac{\alpha_i}{p^i} p^b,$$

  with $\alpha_i \in \{0, 1, \dots, p-1\}$, $\alpha_1 \neq 0$ and exponent b.

- **Set of normalized floating point numbers of length t with base p and range of exponent $\{e_{\min}, e_{\min}+1, \dots, e_{\max}\} \subset \mathbb{Z}$:**

$$\mathbb{M}(p, t, e_{\min}, e_{\max}) := \{\pm 0.\alpha_1 \dots \alpha_t \cdot p^b | \alpha_i \in \{0, 1, \dots, p-1\},\ \alpha_1 \neq 0,\ e_{\min} \leq b \leq e_{\max}\} \cup \{0\}.$$

  $x \in \mathbb{M}(p, t, e_{\min}, e_{\max})$ is called **machine number** or **computer number**.

- **Rounding function**

$$\gamma : \mathbb{R} \to \mathbb{M}(p, t, e_{\min}, e_{\max})$$

  for $x \in Z := [-x_{\max}, -x_{\min}] \cup \{0\} \cup [x_{\min}, x_{\max}]$ determined by

$$\gamma(x) = \arg\min\{|x - \tilde{x}| \mid \tilde{x} \in \mathbb{M}(p, t, e_{\min}, e_{\max})\}$$

  where

$$x_{\min} := \min\{|x| \mid x \in \mathbb{M}(p, t, e_{\min}, e_{\max}) \backslash \{0\}\},$$
$$x_{\max} := \max\{|x| \mid x \in \mathbb{M}(p, t, e_{\min}, e_{\max})\}.$$

  $\Longrightarrow$ for $x = \pm \sum_{i=1}^{\infty} \frac{\alpha_i}{p^i} \cdot p^b \in Z$ with $\alpha_1 \neq 0$ holds:

$$\gamma(x) = \begin{cases} \pm \sum_{i=1}^{t} \frac{\alpha_i}{p^i} \cdot p^b, & \alpha_{t+1} < \frac{p}{2} \\ \pm \left(\sum_{i=1}^{t} \frac{\alpha_i}{p^i} + \frac{1}{p^t} \right) \cdot p^b, & \alpha_{t+1} > \frac{p}{2} \end{cases}$$

  - for $\alpha_{t+1} = \frac{p}{2}$: round up or **Round-to-even**;
  - for $|x| < x_{\min}$ (**Underflow**): round to 0, $\operatorname{sign}(x)\, x_{\min}$ respectively, or **gradual underflow** (allow *denormalized* floating point numbers, i.e., $\alpha_1 = 0$);
  - for $|x| > x_{\max}$ (**Overflow**): $\gamma(x) = \operatorname{sign}(x)\, x_{\max}$ or $\gamma(x) = \infty$.