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

Dr. Martin Köhler, Dr. Pavan Laxmipathy Veluvali
Website: http://www.mpi-magdeburg.mpg.de/scicomp2425

Winterterm 2024/2025

Scientific Computing 1
Handout 5
October 16, 2024

Rounding Errors and Error Propagation

- **Relative rounding errors in $\mathbb{M}(p,t,e_{\min},e_{\max})$:**

$$\frac{|\gamma(x)-x|}{|x|} \;<\; \frac{1}{2}p^{1-t} \quad \forall x \in [-x_{\max},-x_{\min}] \cup [x_{\min},x_{\max}].$$

- **unit roundoff:**

$$\mathbf{u} := \frac{1}{2}p^{1-t}$$

- **machine epsilon:**

$$\mathsf{eps} := \min\{|\tilde{x}-1| \mid \tilde{x} \in \mathbb{M}(p,t,e_{\min},e_{\max}), \tilde{x} > 1\} = p^{1-t} = 2\mathbf{u}$$

- **standard model of the floating point arithmetic:**

$$x \circledcirc y = (x \triangle y)(1+\delta), \quad |\delta| \leqslant \mathbf{u} \quad \forall \triangle \in \{+,-,\cdot,/\}$$

  (is also expected to hold for $\sqrt{}$).

- **error propagation:** Let $\tilde{x} := \gamma(x)$, $\tilde{y} := \gamma(y)$.

  addition: $x,y \in \mathbb{R}$, $\operatorname{sign}(x) = \operatorname{sign}(y) \Longrightarrow$

$$\frac{|(\tilde{x} \oplus \tilde{y}) - (x+y)|}{|x+y|} \leqslant 2\mathbf{u} + \mathbf{u}^2$$

  subtraction: $x,y \in \mathbb{R}$, $\operatorname{sign}(x) = \operatorname{sign}(y) \Longrightarrow$

$$\frac{|(\tilde{x} \ominus \tilde{y}) - (x-y)|}{|x-y|} \leqslant \left(\frac{2|y|}{|x-y|}+2\right)\mathbf{u} + \left(\frac{2|y|}{|x-y|}+1\right)\mathbf{u}^2.$$

  cancellation: $x \approx y \rightsquigarrow$ large error

  multiplication and division:

$$\frac{|\tilde{x} \odot \tilde{y} - x \cdot y|}{|x \cdot y|} \leqslant 3\mathbf{u} + \mathcal{O}(\mathbf{u}^2).$$