Winterterm 2012/2013

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/mpcsc/lehre/2012_WS_SC/

Scientific Computing 1
Handout 5
November 7, 2012

Floating Point Numbers and Rounding

- **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:**

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

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

$$xy = (x\nabla y)(1 + \delta), \quad |\delta| \leq \mathbf{u} \quad \forall \nabla \in \{+, -, \cdot, /\}$$

  (for $\sqrt{\ }$ as well).

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

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

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

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

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

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

  multiplication and division:

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