max planck institut informatik

Numerical Algebraic Computation **Winter 2013/2014**
Michael Sagraloff **To be handed in on October, 30th.**

Exercise 1

1.1 Arithmetic with Floating Point Numbers

a) Finish the proof of Theorem I.3., that is, show that the rule with respect to multiplication applies.

b) Give a proof for Theorem I.4.1.

1.2 Interesting Properties

a) Show that, for two floating point numbers a, $b \in \mathbb{F}$ with $\frac{1}{2} \leq \frac{a}{b} \leq 2$, it holds that $a \ominus b = a - b$.

b)* Show that, for any two floating point numbers a, $b \in \mathbb{F}$ with $|a| \geq |b|$, it holds that

$$a + b = (a \oplus b) + ((a \ominus (a \oplus b)) \oplus b).$$

Hence, notice that errors with respect to addition of two floating point numbers can be exactly computed with floating point arithmetic.

1.3 Interval Arithmetic with Fixed Point Precision

For a given non-negative integer L, let $\mathbb{F}_L$ be the set of all rational values $x \in \mathbb{Q}$ of the form $x = \sum_{i=-L}^{i_0} s_i \cdot 2^i$, with $s_i \in \{0, 1\}$ and an arbitrary but finite i_0. We define $\mathrm{fl}^+ : \mathbb{R} \mapsto \mathbb{F}$ ($\mathrm{fl}^- : \mathbb{R} \mapsto \mathbb{F}$) as the rounding mode to the nearest value in $\mathbb{F}$ that is larger (smaller) than or equal to the input. For a polynomial expression $E(x)$ in x, we aim to compute an interval $\mathfrak{B}(E, L) = [\mathrm{down}(E), \mathrm{up}(E)]$ which contains the value $E(x)$. We achieve this by iteratively using the following definitions:

$$\begin{aligned}
\mathfrak{B}(c, L) &:= [\mathrm{fl}^-(c), \mathrm{fl}^+(c)] \quad \text{if } c \text{ is a constant} \\
\mathfrak{B}(x, L) &:= [\mathrm{fl}^-(x), \mathrm{fl}^+(x)] \\
\mathrm{down}(E_1 + E_2) &:= \mathrm{down}(E_1) + \mathrm{down}(E_2) \\
\mathrm{up}(E_1 + E_2) &:= \mathrm{up}(E_1) + \mathrm{up}(E_2) \\
\mathrm{down}(E_1 \cdot E_2) &:= \mathrm{fl}^-(\min\{\mathrm{down}(E_1)\mathrm{down}(E_2), \mathrm{up}(E_1)\mathrm{up}(E_2), \mathrm{up}(E_1)\mathrm{down}(E_2), \mathrm{down}(E_1)\mathrm{up}(E_2)\}) \\
\mathrm{up}(E_1 \cdot E_2) &:= \mathrm{fl}^+(\max\{\mathrm{down}(E_1)\mathrm{down}(E_2), \mathrm{down}(E_1)\mathrm{up}(E_2), \mathrm{up}(E_1)\mathrm{down}(E_2), \mathrm{up}(E_1)\mathrm{up}(E_2)\})
\end{aligned}$$

Prove the following Lemma:

Suppose that, according to the above rules, we evaluate $f(x) = \sum_{i=0}^{d} a_i x^i \in \mathbb{R}[x]$ at some $c \in \mathbb{R}$ using Horner's evaluation scheme (i.e. $f(x) = a_0 + x \cdot (a_1 + x \cdot (a_2 + \cdots x \cdot (a_{d-1} + x \cdot a_d) \cdots))$). Then, it holds that

$$|f(c) - \mathrm{down}(f(c), L)| \leq 2^{-L+1}(d+1)^2 2^{\tau + d\Gamma} \tag{1}$$

$$|f(c) - \mathrm{up}(f(c), L)| \leq 2^{-L+1}(d+1)^2 2^{\tau + d\Gamma}, \tag{2}$$

where τ and Γ are arbitrary non-negative integers with $\max_i |a_i| \leq 2^\tau$ and $|c| \leq 2^\Gamma$. In particular, $\mathfrak{B}(f(c), L)$ has a width of at most $2^{-L+2}(d+1)^2 2^{\tau + d\Gamma}$.