

Inclusion proofs for ball_float

Luna-Flow floating, design attachment

This note proves the facts the ball_float design page relies on. Notation: $x = [\underline{x}, \bar{x}]$ is a closed interval of the extended reals with $\underline{x} \leq \bar{x}$, $\underline{x} \neq +\infty$, $\bar{x} \neq -\infty$, read as the set of *real* numbers between the endpoints; $\emptyset$ is the empty interval. RD_p and RU_p round a real number to the nearest p -bit binary number below and above. F_p is the set of p -bit binary numbers (with unbounded exponent).

1. Directed rounding

Lemma 1 (directed rounding). For all reals $a \leq b$: (i) $\text{RD}_p(a) \leq a \leq \text{RU}_p(a)$; (ii) $\text{RD}_p(a) \leq \text{RD}_p(b)$ and $\text{RU}_p(a) \leq \text{RU}_p(b)$; (iii) for a finite set S , $\min_{s \in S} \text{RD}_p(s) = \text{RD}_p(\min S)$ and $\max_{s \in S} \text{RU}_p(s) = \text{RU}_p(\max S)$.

Proof. $\text{RD}_p(a) = \max\{f \in F_p : f \leq a\}$, which gives (i). If $a \leq b$, every $f \in F_p$ with $f \leq a$ also satisfies $f \leq b$, so the maximum over the larger set is not smaller: (ii). For (iii), let $m = \min S$. By (ii) $\text{RD}_p(m) \leq \text{RD}_p(s)$ for every $s \in S$, and the minimum is attained at $s = m$. The RU case is symmetric. $\square$

Corollary 2 (outward rounding preserves containment). If $L \leq \inf A$ and $U \geq \sup A$ for a set A of reals, then $A \subseteq [\text{RD}_p(L), \text{RU}_p(U)]$.

Proof. By Lemma 1(i), $\text{RD}_p(L) \leq L \leq \inf A$ and $\text{RU}_p(U) \geq U \geq \sup A$. $\square$

Lemma 3 (nested grids). If $q \leq p$ then $\text{RD}_q(\text{RD}_p(a)) = \text{RD}_q(a)$ and likewise for RU.

Proof. $F_q \subseteq F_p$: a q -bit significand is a p -bit significand padded with zeros. Let $r = \text{RD}_p(a)$. Every $f \in F_q$ with $f \leq a$ lies in F_p , hence $f \leq r$ by maximality of r ; conversely $f \leq r$ implies $f \leq a$. So $\{f \in F_q : f \leq r\} = \{f \in F_q : f \leq a\}$ and the maxima agree. $\square$

2. The fundamental theorem

For a real function f with domain D_f , an *interval extension* F is a map on intervals with $f(x \cap D_f) \subseteq F(x)$ for every x . Interval extensions are *inclusion isotone* if $x \subseteq y \Rightarrow F(x) \subseteq F(y)$.

Theorem 4 (fundamental theorem of interval arithmetic). Let e be an arithmetic expression in variables $\xi_1, \dots, \xi_n$ built from operations each of which is evaluated by an interval extension. Write f for the real function of e (defined where every subexpression is defined) and E for its interval evaluation. Then for all intervals $x_1, \dots, x_n$,

$$\{f(\xi) : \xi \in x_1 \times \dots \times x_n, \xi \in D_f\} \subseteq E(x_1, \dots, x_n). \quad (1)$$

Proof. Induction on e . A variable ξ_i evaluates to x_i and $\xi_i \in x_i$. For $e = g(e_1, \dots, e_k)$ with g evaluated by the extension G : take $\xi \in D_f$. Then each e_j is defined at ξ , and by the induction hypothesis $y_j := f_{e_j}(\xi) \in E_j$. Since $(y_1, \dots, y_k) \in D_g$ and G is an extension, $f(\xi) = g(y_1, \dots, y_k) \in g((E_1 \times \dots \times E_k) \cap D_g) \subseteq G(E_1, \dots, E_k) = E(x)$. $\square$

Nothing in the proof requires the two occurrences of a variable to take the same value, which is why $x - x \neq \{0\}$ (the *dependency problem*): E encloses the larger set where every occurrence varies independently.

3. Endpoint formulas

Lemma 5 (corners of a bilinear map). On a box $x \times y$ of bounded intervals, $h(\xi, \eta) = \xi\eta$ attains its minimum and maximum at corners.

Proof. For fixed η , h is affine in ξ , so its extrema over x are at $\underline{x}$ or $\bar{x}$; therefore $\max_{x \times y} h = \max_{\eta \in y} \max(\underline{x}\eta, \bar{x}\eta)$, and each $\eta \mapsto \underline{x}\eta$ is affine, attaining its maximum at $\underline{y}$ or $\bar{y}$. The minimum is analogous. $\square$

With Corollary 2 this gives $xy \subseteq [\text{RD min } S, \text{RU max } S]$ for $S = \{\underline{xy}, \underline{x}\bar{y}, \bar{x}y, \bar{x}\bar{y}\}$. The sign cases used by the code select from S : if $\underline{x}, \underline{y} \geq 0$ then $\xi\eta$ is increasing in both arguments on the box, so the minimum is $\underline{xy}$ and the maximum $\bar{x}\bar{y}$; the other three single-sign cases follow by $\xi\eta = (-\xi)(-\eta) = -((- \xi)\eta)$; when both intervals contain 0 in their interior, the products $\underline{xy}$ and $\bar{x}\bar{y}$ are ≥ 0 and the other two ≤ 0 , so the minimum is among the latter and the maximum among the former. For unbounded intervals the corner lemma holds in the limit, with $0 \cdot (\pm\infty)$ contributing 0: a zero endpoint multiplies an unbounded side to 0 because the real points near that endpoint give products near 0.

Sum and difference are monotone in each argument (increasing, resp. increasing in ξ and decreasing in η), so $x + y = [\underline{x} + \underline{y}, \bar{x} + \bar{y}]$ and $x - y = [\underline{x} - \bar{y}, \bar{x} - \underline{y}]$ exactly; Corollary 2 makes the rounded versions enclosures.

4. Far addends

The endpoint sum $A + s$ with $|s| \ll |A|$ is not formed exactly. Let $t(v)$ be the exponent of the leading bit of $v \neq 0$ (so $2^{t(v)} \leq |v| < 2^{t(v)+1}$), $e(A)$ the exponent of the last stored bit of A , $M = \max(65536, \text{prec}(A), \text{prec}(s))$ and $c = \min(e(A), t(A) - M)$. If $t(s) < c - 2$, the code replaces s by $s' = \text{sign}(s)2^{c-2}$ when s pushes the sum in the requested rounding direction, and by $s' = 0$ otherwise.

Lemma 6 (soundness). For upward sums $A + s' \geq A + s$; for downward sums $A + s' \leq A + s$.

Proof. $|s| < 2^{t(s)+1} \leq 2^{c-2}$. Upward, $s > 0$: $s' = 2^{c-2} > s$. Upward, $s < 0$: $s' = 0 > s$. The downward cases are mirror images. $\square$

Lemma 7 (no change after rounding). If $p \leq M$ and $s \neq 0$, then $\text{RD}_p(A + s) = \text{RD}_p(A + s')$ and $\text{RU}_p(A + s) = \text{RU}_p(A + s')$.

Proof. Near $|A|$ (in the binade of A and the one below it) the p -bit numbers are multiples of $2^{t(A)-p}$, and $t(A) - p \geq t(A) - M \geq c$; A itself is a multiple of $2^{e(A)}$, hence of 2^c . So the open interval $I = (A - 2^c, A + 2^c)$ contains no p -bit number other than possibly A . Consequently RU_p is constant on $(A - 2^c, A]$ (its value is $\text{RU}_p(A)$) and on $(A, A + 2^c)$, and RD_p is constant on $[A, A + 2^c)$ and on $(A - 2^c, A)$. Since $|s|, |s'| \leq 2^{c-2}$, both sums lie in I . Upward with $s > 0$: both lie in $(A, A + 2^c)$. Upward with $s < 0$: $s' = 0$, so the sums are A and a point of $(A - 2^c, A)$, both in $(A - 2^c, A]$. The downward cases are symmetric. $\square$

Lemma 6 is what the inclusion property needs and holds for every precision; Lemma 7 says the surrogate costs no tightness when the result precision is at most M . Alignment of A and s' shifts by at most $t(A) - c + 2 \leq M + 2$ bits plus the width of A , instead of the full exponent gap.

5. Certified elementary endpoints

Lemma 8 (acceptance test). Let $L \leq y \leq U$ with rational L, U . If $\text{RD}_p(L) = \text{RD}_p(U)$ then $\text{RD}_p(L) = \text{RD}_p(y)$; if $\text{RU}_p(L) = \text{RU}_p(U)$ then $\text{RU}_p(U) = \text{RU}_p(y)$.

Proof. By Lemma 1(ii), $\text{RD}_p(L) \leq \text{RD}_p(y) \leq \text{RD}_p(U) = \text{RD}_p(L)$. $\square$

The trigonometric and arctangent kernels accept a work precision only when both tests pass, so their endpoints are the correctly directed roundings of the exact values; otherwise the work precision w grows to $w + \max(32, \frac{w}{2})$, at most 12 times. The exponential and logarithm kernels skip the test and return $\text{RD}_p(L)$ and $\text{RU}_p(U)$, which are enclosures by Corollary 2 and, with 64 guard bits, almost always the correct directed roundings.

Series tails. For $0 \leq x \leq \frac{1}{8}$ the terms $t_k = \frac{x^k}{k!}$ satisfy $\frac{t_{k+1}}{t_k} = \frac{x}{k+1} \leq \frac{1}{16}$, so the tail after t_n is at most $t_{n+1} \sum_{j \geq 0} 16^{-j} = \frac{16}{15} \cdot t_{n+1} \leq 2t_{n+1}$, the bound the code adds. For $\ln y = 2 \sum_{k \geq 0} \frac{z^{2k+1}}{2k+1}$ with $z = \frac{y-1}{y+1} \in [0, \frac{1}{3}]$ ($1 \leq y \leq 2$), let t' be the first omitted term. Consecutive terms have ratio at most $z^2 \leq \frac{1}{9}$, so the omitted tail is at most $t' \sum_{j \geq 0} 9^{-j} = \frac{9}{8}t'$; after doubling it is at most $\frac{9}{4}t' \leq 3t'$, the bound the code adds.