# Rounding and exactness proofs for `binary`

Companion to the `binary` design page of Luna-Flow `floating`

This note proves the claims that the design page states with a sketch. The notation follows that page. $F = F(p, e_{\min}, e_{\max})$ is the set of finite values of a context, $\eta = 2^{e_{\min}-p+1}$ the subnormal quantum, $\mathrm{top}(x) = \lfloor \log_2 |x| \rfloor$, $\mathrm{bits}(c) = \lfloor \log_2 c \rfloor + 1$ for an integer $c > 0$, and $\nu_2(c)$ the number of trailing zero bits of $c$. A rounding function $\circ$ is one of RD, RU, RZ, RA, RNE, RNA, and every one of them satisfies

$$(R1)\quad x \in F \Rightarrow \circ(x) = x, \qquad (R2)\quad x \le y \Rightarrow \circ(x) \le \circ(y).$$

## 1. Round and sticky bits

**Lemma 1.** Let $m > 0$ be an integer, $\sigma \ge 1$, $q = \lfloor m2^{-\sigma} \rfloor$, $f = m2^{-\sigma} - q \in [0, 1)$, $g$ = bit $\sigma - 1$ of $m$ and $t = [\nu_2(m) < \sigma - 1]$. Then $g = \left[f \ge \frac{1}{2}\right]$, $f = \frac{1}{2} \Leftrightarrow g \wedge \neg t$, $f > \frac{1}{2} \Leftrightarrow g \wedge t$, and $f > 0 \Leftrightarrow g \vee t$.

*Proof.* Write $m = q2^{\sigma} + g2^{\sigma-1} + \ell$ with $0 \le \ell < 2^{\sigma-1}$. Then $f = \frac{g}{2} + \ell 2^{-\sigma}$ with $0 \le \ell 2^{-\sigma} < \frac{1}{2}$, so $f \ge \frac{1}{2} \Leftrightarrow g = 1$. The low part $\ell$ is nonzero exactly when some bit below $\sigma - 1$ is set, that is when $\nu_2(m) < \sigma - 1$, so $\ell \ne 0 \Leftrightarrow t$. The remaining equivalences follow by cases on $g$ and $t$. □

Consequently the increment rules of the design page (for example $g \wedge (t \vee q_0)$ for RNE) are exactly the definitions of the rounding functions applied to $q + f$, and the result is inexact iff $g \vee t$.

## 2. One rounding onto the subnormal grid

**Lemma 2.** Let $r = m2^{e} > 0$ and $\sigma = \max(\mathrm{bits}(m) - p, (e_{\min} - p + 1) - e, 0)$. Then the points of $F$ nearest to $r$ from below and above are $q2^{e+\sigma}$ and $(q+1)2^{e+\sigma}$ (the latter possibly equal to $2^{\mathrm{top}(r)+1}$), where $q = \lfloor m2^{-\sigma} \rfloor$.

*Proof.* The exponent $e + \sigma$ is the quantum of $F$ at $r$: in the normal range the quantum is $2^{\mathrm{top}(r)-p+1}$ and $\mathrm{top}(r) = e + \mathrm{bits}(m) - 1$, so the precision shift gives it; below $2^{e_{\min}}$ the quantum is $\eta$, which the range shift gives. Taking the maximum covers both and never makes the quantum smaller than either constraint requires. $q2^{e+\sigma}$ is the largest multiple of the quantum not above $r$ and has at most $p$ significant bits. □

Rounding to $p$ bits first and then onto the subnormal grid can differ. In binary16 ($p = 11$, $\eta = 2^{-24}$) take $r = 2^{-25}(1 + 2^{-11})$. At 11 bits $r$ is a tie between $2^{-25}$ and $2^{-25}(1 + 2^{-10})$ and RNE chooses $2^{-25} = \frac{\eta}{2}$, which is again a tie and rounds to 0. Directly, $r > \frac{\eta}{2}$ rounds to $\eta$. The finalizer uses Lemma 2 and returns $\eta$.

## 3. Addition with far-apart operands

**Lemma 3.** Let $H, L$ be nonzero dyadic operands of arbitrary length with $\mathrm{top}(H) - \mathrm{top}(L) > p + 3$, $H = c_H 2^{e_H}$ with $c_H$ odd. Let $T = e_H - p - 3$, let $L'$ be $L$ truncated toward zero to a multiple of $2^T$, and $\delta = [L \ne L']$. Then $H + L$ and the value

$$S = H + L' + \mathrm{sign}(L) \cdot \delta \cdot 2^{T-1}$$

have the same rounding $\circ(H + L) = \circ(S)$ and the same inexactness, for every context of precision $p$.

*Proof.* Since $|L| < 2^{\mathrm{top}(L)+1} \le 2^{\mathrm{top}(H)-p-3}$, we have $|H + L| > 2^{\mathrm{top}(H)} - 2^{\mathrm{top}(H)-p-3} \ge 2^{\mathrm{top}(H)-1}$, so $\mathrm{top}(H + L) \ge \mathrm{top}(H) - 1$. The rounding quantum at $H + L$ is therefore at least $2^{\mathrm{top}(H)-p}$ and the round bit sits at $2^{\mathrm{top}(H)-p-1}$ or higher. Because $c_H$ is odd, $e_H \le \mathrm{top}(H)$, so $T \le \mathrm{top}(H) - p - 3$: every bit of $L - L'$ and the substitute bit $2^{T-1}$ lie strictly below the round-bit position. Both $H + L$ and $S$ agree on all positions from $2^T$ upward, hence on $q$ and $g$ of Lemma 1, and $L - L'$ and $\mathrm{sign}(L)\delta 2^{T-1}$ are both zero or both nonzero with the same sign. When $L$ and $H$ have the same sign this gives the same $t$. When they have opposite signs, write $|L| = |L'| + \varepsilon$ with $0 \le \varepsilon < 2^T$; then $|H| - |L| = (|H| - |L'| - 2^T) + (2^T - \varepsilon)$, and for $\varepsilon > 0$ the last term lies strictly between 0 and $2^T$, which is exactly what $|H| - |L'| - 2^T + 2^{T-1}$ represents, the form

the code uses. If the exponents also force a subnormal shift, the rounding position only moves up, and the argument is unchanged. The values agree in sign, and by Lemma 1 the rounding and the inexact flag agree.□

# 4. The IEEE remainder

**Lemma 4.** If $x, y \in F$, $y \neq 0$, $n$ is the integer nearest to $\frac{x}{y}$ (ties to even) and $r = x - ny$, then $r \in F$.

*Proof.* Write $x = M_x 2^{q_x}$, $y = M_y 2^{q_y}$ with $|M_x|, |M_y| < 2^p$ and $q_x, q_y \geq e_{\min} - p + 1$. By the choice of $n$, $|r| \leq \frac{|y|}{2}$. If $n = 0$ then $r = x \in F$. Otherwise $\left|\frac{x}{y}\right| \geq \frac{1}{2}$, so $|r| \leq \frac{|y|}{2} \leq |x|$. In all cases $r$ is an integer multiple of $2^{\min(q_x, q_y)}$. If $q_x \geq q_y$, $r = k2^{q_y}$ with $|k| \leq \frac{|M_y|}{2} < 2^{p-1}$. If $q_x < q_y$, $r = k2^{q_x}$ with $|k| \leq |M_x| < 2^p$. In both cases $r$ has a representation with fewer than $p$ bits on an admissible quantum, and $|r| \leq \max(|x|, |y|) \leq \Omega$, so $r \in F$. □

**Lemma 5.** For integers $X \geq 0$ and $Y > 0$ let $R = X \bmod 2Y$. Then $X \bmod Y = R - Y[R \geq Y]$ and $\left\lfloor \frac{X}{Y} \right\rfloor \equiv [R \geq Y] \pmod 2$.

*Proof.* $X = Q(2Y) + R$ with $0 \leq R < 2Y$. If $R < Y$ then $X = (2Q)Y + R$; if $R \geq Y$ then $X = (2Q + 1)Y + (R - Y)$ with $0 \leq R - Y < Y$. □

The implementation computes $R$ without forming $X$ when $X = c_x 2^{q_x - q_y}$ is huge, as $((c_x \bmod 2Y) \cdot (2^{q_x - q_y} \bmod 2Y)) \bmod 2Y$ with the power obtained by square-and-multiply. With $\rho = X \bmod Y$, the nearest integer to $\frac{X}{Y}$ is $\left\lfloor \frac{X}{Y} \right\rfloor$ if $2\rho < Y$, $\left\lfloor \frac{X}{Y} \right\rfloor + 1$ if $2\rho > Y$, and on a tie the even one of the two; Lemma 5 supplies the parity, so the remainder is $\rho$ or $\rho - Y$ with the sign of $x$ adjusted accordingly.

# 5. nextUp

**Lemma 6.** Let $x$ be finite (any number of bits), $\pi = \min(e_{\min} - p, e(x), \mathrm{top}(x) - p)$ for $x \neq 0$ and $\pi = e_{\min} - p$ for $x = 0$, where $e(x)$ is the stored exponent. Then $\mathrm{RU}(x + 2^{\pi - 2})$ is the least element of $F$ greater than $x$ (or $+\infty$ beyond $\Omega$).

*Proof.* Let $x^+$ be the least element of $F \cup \{+\infty\}$ greater than $x$, and assume first $x^+$ finite. Every point of $F$ is a multiple of $\eta = 2^{e_{\min} - p + 1}$, every point of $F$ of magnitude at least $2^{\mathrm{top}(x) - 1}$ is a multiple of $2^{\mathrm{top}(x) - p}$ (this covers the binade of $x$ and, for negative $x$, the binade just below it in magnitude), and $x$ is a multiple of $2^{e(x)}$. Both $x$ and $x^+$ are therefore multiples of $2^\pi$, so $x^+ - x \geq 2^\pi$ and $x < x + 2^{\pi - 2} < x^+$. Since RU maps every point of $(x, x^+]$ to $x^+$, the claim follows; if $x = \Omega$ the same step rounds upward to $+\infty$. For $x = 0$, $x^+ = \eta > 2^{e_{\min} - p - 2}$. □

# 6. Certification by enclosures

**Lemma 7.** If $L \leq y \leq U$ and $\circ(L) = \circ(U)$, then $\circ(y) = \circ(L)$.

*Proof.* By (R2), $\circ(L) \leq \circ(y) \leq \circ(U) = \circ(L)$. □

The enclosures are produced by evaluating every operation with RD for the lower and RU for the upper end. Each such operation is monotone in its arguments on the region where it is used (sums; products and quotients of positive quantities; squaring of positive quantities), so the computed interval contains the exact value; series tails are added as explicit upper bounds. For the exponential series on $\left[0, \frac{1}{8}\right]$ the tail after the term $t_n$ is bounded by $t_n \sum_{i \geq 1} 8^{-i} = \frac{t_n}{7}$, and the code adds $2t_n$. For $\ln v = 2 \sum_{k \geq 0} \frac{z^{2k+1}}{2k+1}$ with $z \in \left[0, \frac{1}{3}\right]$ the tail from index $m$ is at most $\frac{z^{2m+1}}{(2m+1)(1-z^2)}$.

**Lemma 8 (termination).** If $y$ is not a breakpoint of $\circ$ (a point of $F$ for directed rounding, a midpoint of adjacent points of $F$ for nearest rounding) and the enclosure width tends to 0 as the working precision grows, then for some working precision $\circ(L) = \circ(U)$.

*Proof.* $\circ$ is constant on each open interval between consecutive breakpoints. $y$ lies in the interior of such an interval, at a positive distance $d$ from its ends; once the width is below $d$, both $L$ and $U$ lie in the same interval. □

Breakpoints are dyadic. By Lindemann–Weierstrass, $e^a$ is transcendental for algebraic $a \neq 0$, hence so are $\ln a$ ($a \neq 1$), $\sin a$, $\cos a$, $\tan a$ and the inverse functions at nonzero algebraic $a$ other than their trivial points; by Niven's theorem $\sin(\pi r)$, $\cos(\pi r)$ and $\tan(\pi r)$ are dyadic for dyadic $r$ only at multiples of $\frac{1}{2}$ and, for $\tan$, at odd multiples of $\frac{1}{4}$. The implementation special-cases these points, so Lemma 8 applies to every other input; the refinement budget limits how long it may take.

## 7. Integer powers

**Lemma 9.** Let an addition chain $1 = a_0, a_1, \ldots, a_k = n$ be evaluated with one rounding per step, $\hat{x}_{a_j} = \hat{x}_{a_i}\hat{x}_{a_l}(1+\delta_j)$ with $|\delta_j| \le u$. Then $\hat{x}_n = x^n \prod_j (1+\delta_j)^{m_j}$ with $\sum_j m_j \le n-1$.

*Proof.* Let $E(a)$ be the total exponent of the error factors in $\hat{x}_a$. Then $E(1) = 0$ and $E(a_j) = E(a_i) + E(a_l) + 1$ for $a_j = a_i + a_l$. By induction, $E(a_j) \le (a_i - 1) + (a_l - 1) + 1 = a_j - 1$. □

Therefore $\hat{x}_n = x^n(1+\theta)$ with $|\theta| \le \gamma_{n-1} = (n-1)\frac{u}{1-(n-1)u}$. With $u = 2^{-w}$ and $w \ge p + \mathrm{bits}(n) + 4$, $(n-1)u < 2^{\mathrm{bits}(n)-w}$ is tiny, and since $|\hat{x}_n| < 2^{\mathrm{top}+1}$ the error is below $2^{\mathrm{bits}(n)}(1+\gamma)$ units of $2^{\mathrm{top}-w+1}$, which is covered by the radius $2^{\mathrm{bits}(n)+2}$ units used by the code. For negative exponents the reciprocal of the base adds the factor $(1+\delta_0)^n$, so $E \le 2n-1$, covered by the radius $2^{\mathrm{bits}(n)+3}$.

## 8. The number-theoretic transform

**Lemma 10.** Split $A$ and $B$ into $n_a$ and $n_b$ base-$2^{16}$ digits and let $c_k = \sum_{i+j=k} a_i b_j$. If the transform length is at most $2^{23}$, then $0 \le c_k < p_1 p_2$ for $p_1 = 998244353$, $p_2 = 754974721$, so $c_k$ is determined by $c_k \bmod p_1$ and $c_k \bmod p_2$.

*Proof.* Each $c_k$ has at most $\min(n_a, n_b) \le 2^{23}$ terms, each below $2^{32}$, so $c_k < 2^{55}$. And $p_1 p_2 = 753649251896000513 > 2^{59} > 2^{55}$. The Chinese remainder theorem gives a unique residue in $[0, p_1 p_2)$. Both primes admit transforms of length $2^{23}$ because $2^{23} \mid p_1 - 1 = 119 \cdot 2^{23}$ and $2^{23} \mid p_2 - 1 = 45 \cdot 2^{24}$. □

## 9. Shortest decimal output

**Lemma 11.** Let $I$ be an interval containing $x > 0$, and let $d_n$ be $x$ truncated to $n$ significant decimal digits. If $d_n \in I$ then $d_{n+1} \in I$; the same holds for the $n$-digit decimals rounded away from zero.

*Proof.* Truncation to more digits only removes less, so $d_n \le d_{n+1} \le x$. An interval containing $d_n$ and $x$ contains every point between them. The upward case is symmetric. □

With $I$ the set of reals that round to $x$, a candidate is accepted iff it lies in $I$, so acceptance is monotone in $n$ and the least accepted digit count can be found by bisection.