

Rounding proofs for decimal

Luna-Flow floating, design attachment

This note proves the facts the decimal design page relies on. Notation: a decimal format has precision p , adjusted-exponent range $[e_{\min}, e_{\max}]$, $E_{\text{tiny}} = e_{\min} - p + 1$ and $E_{\text{top}} = e_{\max} - p + 1$. For a real $x > 0$ and an integer exponent t write $x \cdot 10^{-t} = c + f$ with $c \in \mathbb{Z}_{\geq 0}$ and $0 \leq f < 1$; a rounding direction $\circ$ returns $c \cdot 10^t$ or $(c + 1) \cdot 10^t$. Negative arguments are rounded through $|x|$, with Ceiling and Floor exchanged.

1. Rounding digit and sticky information

Lemma 1 (midpoint test). Let $C = Q \cdot 10^s + R$ with $s \geq 1$ and $0 \leq R < 10^s$, and write $R = r \cdot 10^{s-1} + R'$ with $r \in \{0, \dots, 9\}$, $0 \leq R' < 10^{s-1}$. Then $2R > 10^s$ iff $r > 5$ or $(r = 5 \text{ and } R' > 0)$; $2R = 10^s$ iff $r = 5$ and $R' = 0$; and $2R < 10^s$ iff $r < 5$.

Proof. $2R - 10^s = 2(r - 5)10^{s-1} + 2R'$ and $0 \leq 2R' < 2 \cdot 10^{s-1}$. If $r \geq 6$ the first term is at least $2 \cdot 10^{s-1}$, so the sum is positive. If $r \leq 4$ the first term is at most $-2 \cdot 10^{s-1}$ and the sum is negative. If $r = 5$ the sum is $2R'$. $\square$

Hence the comparison of $2R$ with 10^s carries the rounding digit and the sticky bit together, and the comparison of $2R$ with a divisor D does the same for an integer quotient $N = QD + R$, because $f = \frac{R}{D}$.

Lemma 2 (no square-root ties). For integers $M \geq 0$ and $r = \lfloor \sqrt{M} \rfloor$, $\sqrt{M} \neq r + \frac{1}{2}$.

Proof. $\sqrt{M} = r + \frac{1}{2}$ would give $4M = (2r + 1)^2$, an even number equal to an odd one. $\square$

2. Double rounding with ZeroFiveUp

Let $\text{zfu}_k(x)$ denote rounding $x > 0$ with ZeroFiveUp to the grid 10^{t-k} , $k \geq 1$, and let $\circ$ be any rounding direction to the grid 10^t .

Lemma 3. $\circ(\text{zfu}_k(x)) = \circ(x)$ for every $x > 0$ and every direction $\circ$.

Proof. Let $g = 10^{t-k}$ and $x = (n + \varphi)g$ with $n \in \mathbb{Z}$, $0 \leq \varphi < 1$. If $\varphi = 0$ then $\text{zfu}_k(x) = x$. Otherwise $y = \text{zfu}_k(x)$ is ng or $(n + 1)g$, and the last digit of $\frac{y}{g}$ is in $\{1, 2, 3, 4, 6, 7, 8, 9\}$: if n ends in 0 or 5 the result is $n + 1$, ending in 1 or 6; otherwise it is n .

The decision of $\circ$ depends only on the position of its argument relative to the points $m \cdot 10^t$ ("grid points") and $(m + \frac{1}{2})10^t$ ("midpoints"), $m \in \mathbb{Z}$, and on whether the argument is one of them. Grid points are multiples of g whose last digit is 0; midpoints are multiples of g whose last digit is 5 if $k = 1$ and 0 if $k \geq 2$. So y is neither a grid point nor a midpoint, and neither is x , because x is not a multiple of g .

Let z be any grid point or midpoint; it is a multiple of g . If $z \leq ng$ then $z \leq ng \leq y$ and $z \leq ng < x$, and $z \neq y$, so both x and y are strictly above z . If $z \geq (n + 1)g$ then $x < z$ and $y \leq z$, $y \neq z$, so both are strictly below z . No multiple of g lies strictly between ng and $(n + 1)g$. Therefore x and y are on the same side of every grid point and midpoint, both are inexact at the grid 10^t , and the last digit of the truncation m is the same for both; every rounding direction makes the same decision. $\square$

The lemma fails for the half modes as the first rounding: rounding x just above a midpoint down onto the midpoint lets the tie rule of the second rounding decide, which may pick the other neighbour.

3. Overflow

Lemma 4 (overflow results). Let $N_{\max} = (10^p - 1)10^{E_{\text{top}}}$ and let $x > 0$ overflow, i.e. x rounded to p digits with unbounded exponent has adjusted exponent above $e_{\max}$. Then the IEEE 754 result is $+\infty$ for HalfEven, HalfUp, HalfDown, Up and Ceiling, and $N_{\max}$ for Down, Floor and ZeroFiveUp.

Proof. The overflowing rounded value is at least $10^{e_{\max}+1}$, so $x > N_{\max}$. For Down and Floor the rounding never increases $|x|$ relative to the truncation, and the largest representable value not above x is $N_{\max}$. For ZeroFiveUp the truncation is at least $N_{\max}$, whose last digit is 9, so no increment happens; saturation gives $N_{\max}$. For Up and Ceiling the rounding of any $x > N_{\max}$ is above $N_{\max}$, hence infinite. For the half modes, x rounds with unbounded exponent either to $N_{\max}$ (no overflow) or to a value of at least $10^{e_{\max}+1}$; the latter happens only for $x \geq N_{\max} + \frac{1}{2} \cdot 10^{E_{\text{top}}}$, i.e. when x is at least as far from $N_{\max}$ as the midpoint toward the next power of ten. IEEE 754 §7.4 delivers $\pm\infty$ for every overflow under a round-to-nearest attribute, which is the saturation of that upward rounding. The negative case is symmetric with Ceiling and Floor exchanged. $\square$

4. Clamping

Lemma 5 (fold-down fits). If a finite non-zero $x = C \cdot 10^Q$ with d digits has $Q > E_{\text{top}}$ and adjusted exponent $a = Q + d - 1 \leq e_{\max}$, then $C \cdot 10^{Q-E_{\text{top}}}$ has at most p digits.

Proof. Its digit count is $d + Q - E_{\text{top}} = (a - Q + 1) + Q - (e_{\max} - p + 1) = a - e_{\max} + p \leq p$. $\square$

5. Certified rounding of enclosures

Lemma 6 (certification). Let $L \leq v \leq U$ and let $\circ$ be a rounding to a decimal context. If $\circ(L)$ and $\circ(U)$ are the same representation, then $\circ(v)$ is that representation. If moreover v is not representable and $\circ(L)$, $\circ(U)$ carry the same flags, $\circ(v)$ carries them too.

Proof. Rounding is monotone, so $\circ(L) \leq \circ(v) \leq \circ(U)$; equal outer values force the middle one in value. L and U cannot straddle zero, because then they would round to zeros of different signs or to values of different signs; so all three have one sign, and the member of the cohort is the one the finalizer chooses for that value, which is the same. For the flags: let r be the common result. Not both $L = r$ and $U = r$, since then $v = r$ would be representable; so one endpoint differs from r and is inexact, and because the flags agree both are. $v \neq r$, so v is inexact too, and rounded accompanies inexact. overflow, subnormal and underflow are monotone predicates of $|v|$ on a one-signed interval (adjusted exponent above $e_{\max}$, below $e_{\min}$, and inexactness), so if they agree at both ends they hold at v . clamped depends only on the result. $\square$

If v is itself representable, a directed rounding sends $L < v$ and $U > v$ to different neighbours and the test never succeeds; the implementation therefore decides exact cases before entering the loop.

6. Kernel bounds

Lemma 7 (Comba columns). For $n \leq 18$ limbs of base 10^9 , a column sum of n limb products plus the incoming carry is below 2^{64} .

Proof. The carry into a column is the previous column sum divided by 10^9 , below $18 \cdot 10^9 + 1 < 2 \cdot 10^{10}$ by induction. The sum is at most $18(10^9 - 1)^2 + 2 \cdot 10^{10} < 1.8 \cdot 10^{19} < 2^{64} \approx 1.8447 \cdot 10^{19}$. With $n = 19$, $19(10^9 - 1)^2 > 1.89 \cdot 10^{19} > 2^{64}$. $\square$

Lemma 8 (two-prime NTT). Let $p_1 = 998244353 = 119 \cdot 2^{23} + 1$ and $p_2 = 754974721 = 45 \cdot 2^{24} + 1$. For digit sequences in base 10^4 of lengths n_a, n_b with $n_a + n_b - 1 \leq 2^{23}$, every coefficient z of the integer convolution is recovered from $r_i = z \bmod p_i$ by $z = r_1 + p_1((r_2 - r_1)p_1^{-1} \bmod p_2)$.

Proof. Both $p_i - 1$ are divisible by 2^{23} , so $\mathbb{Z}/p_i\mathbb{Z}$ contains a primitive root of unity of every order $2^j \leq 2^{23}$ and the cyclic convolution of the zero-padded sequences equals the integer convolution reduced modulo p_i . Each z is a sum of at most $m = \min(n_a, n_b)$ products below 10^8 , so $0 \leq z < m \cdot 9999^2 < 2^{23} \cdot 10^8 < p_1 p_2 \approx 7.54 \cdot 10^{17}$. The formula gives the unique integer in $[0, p_1 p_2)$ congruent to r_1 modulo p_1 and to r_2 modulo p_2 (Chinese remainder theorem), which is z . $\square$