Statistical Calculations Through Rational Arithmetic And Conversions

Timothy Hall*

Abstract
When making analytical calculations with floating point real numbers it is commonly the case that even the most basic operations, such as arithmetic and conversions, produce only approximate results. The cumulative effect from such incremental approximations may quickly become problematical, producing final results that significantly differ from the exact (correct) results. However, when making those same analytical calculations with rational numbers, the results are necessarily exact (correct) at every stage of the calculation. Only when the result need be converted to a floating point real number representation may the calculation error become non-zero. This eliminates the effect of intermediate calculation error, so that the implementing analyst may concentrate on input data error and precision management issues. In particular, to facilitate simulations and theoretical input data possibilities, irrational numbers may be symbolically represented in the calculations and in intermediate results until a controlled approximation is needed.

This paper presents an analytical framework for making error free statistical calculations using rational arithmetic and conversion operations. Several example calculations are provided to demonstrate how final result error is eliminated in all cases, and explicitly controlled when a floating point real number representation is needed.

Key Words: Rational Arithmetic, Exact Algorithms, Floating Point Calculations

1. Introduction

When making analytical calculations with floating point real numbers (regardless of parameter specifics, including the numerical base) it is commonly the case that even the most basic operations, such as arithmetic and conversions, produce only approximate results (with respect to the precision in which the operands are expressed). The cumulative effect from such incremental approximations may quickly become problematical. However, when making those same analytical calculations with rational numbers, the results are exact (regardless of precision concerns, which are actually moot until the results are converted to floating point real numbers). This eliminates the contribution of calculation error in the results, so that the implementing analyst may concentrate on data error and precision management issues. In particular, irrational numbers may be asymptotically approximated by rational numbers, so that the final approximation is within the required precision when converted to a floating point real number. In intermediate calculations, irrational numbers may be kept as symbolic values until a precise approximation is needed.

1.1 Preliminaries

Definition 1 *The* Greatest Common Divisor $\gcd$ *of two non-zero integers* u *and* v *is the largest positive integer* d *such that* $\frac{u}{d}$ *and* $\frac{v}{d}$ *are both non-zero integers.*

Definition 2 *Two non-zero integers* u *and* v *are* Relatively Prime *(symbolized* $u \perp v$*) if* $\gcd(u, v) = 1$.

Definition 3 *The* Reduced Form*,* $r\left(\frac{u}{v}\right)$*, of the ratio of two non-zero integers* u *and* v*, is* $\frac{u'}{v'}$ *where* $\frac{u'}{v'} = \frac{u}{v}$ *and* $u' \perp v'$.

*PQI Consulting, P. O. Box 425616, Cambridge, MA, USA 02142-0012 – info@pqic.com

Definition 4 *A* Rational Number *is the ratio of two non-zero numbers u and u'. Note that in this context, we have*

1. *$\frac{0}{u'}$ is **not** a rational number regardless of denominator, and*
2. *$\frac{-u}{u'}$ and $\frac{u}{-u'}$ represent two **different** rational numbers (whose **values** are equal as real numbers).*

Definition 5 *A rational number $\frac{u}{u'}$ is said to be* Rectified *if $u' > 0$. The unitary operator that converts a rational number into a rectified rational number is called* Rectification. *These definitions apply to all rational numbers regardless of the base in which u and u' are expressed.*

Definition 6 *A b-*Rational Number *is a rational number $\frac{u}{u'}$ where u and u' are expressed as integers base $b > 1$. A 10-rational number is identical to a rational number.*

Definition 7 *A* Reduced Rational *is a rational number $\frac{u}{u'}$ in reduced form, i.e., $r\left(\frac{u}{u'}\right) = \frac{u}{u'}$.*

Definition 8 *A b-*Reduced Rational *is a reduced rational where u and u' are relatively prime as base $b > 1$ non-zero integers. A 10-reduced rational is identical to a reduced rational.*

Definition 9 *A* (Reduced) Standard Rational *is a (reduced) rational $\frac{u}{u'}$ where $\operatorname{sgn}(u) + \operatorname{sgn}(u') \geq 0$.*

Definition 10 *A* (Reduced) b-Standard Rational *is a (reduced) standard rational $\frac{u}{u'}$ where u and u' are expressed as integers base $b > 1$. A (reduced) 10-standard rational is identical to a (reduced) standard rational.*

Throughout the remainder of this memorandum, any reference to integers will be considered relative to base 10 unless specifically stated otherwise.

Claim 11 *If u and v are non-zero integers, and $d = \gcd(u, v)$, then $r\left(\frac{u}{v}\right) = \frac{\frac{u}{d}}{\frac{v}{d}}$.*

Proof. Clearly $\frac{u}{v} = \frac{\frac{u}{d}}{\frac{v}{d}}$. Suppose $\frac{u}{d}$ and $\frac{v}{d}$ are not relatively prime. Then there is a prime $s \geq 2$ such that

$$\frac{u}{d} = u's \quad \text{and} \quad \frac{v}{d} = v's$$

for non-zero integers u' and v'. This means $ds > d > 0$ is a common divisor of u and v – a contradiction. ■

Corollary 12 *Every rational number $\frac{u}{u'}$ has a unique reduced form $r\left(\frac{u}{u'}\right)$.*

Proof. Suppose

$$\frac{v}{v'} = r\left(\frac{u}{u'}\right) = \frac{w}{w'}$$

are two reduced forms of $\frac{u}{u'}$. Then

$$\frac{vw'}{v'} = w$$

means v' is a divisor of w' (since v' is not a divisor of v by definition). So $w' = kv'$ for some integer k, which means

$$w = \frac{v(kv')}{v'} = kv$$

so that w and w' have k as a common factor – a contradiction. ■

Claim 13 *If u and v are non-zero integers, then* $\gcd(u,v) = \gcd(v,u)$.

Proof. We have

$$\frac{\frac{v}{\gcd(u,v)}}{\frac{u}{\gcd(u,v)}} = \frac{1}{r\left(\frac{u}{v}\right)} = r\left(\frac{v}{u}\right) = \frac{\frac{v}{\gcd(v,u)}}{\frac{u}{\gcd(v,u)}}$$

so that

$$\gcd(u,v) = \gcd(v,u)$$

■

Claim 14 *If u, u', v, and v' are non-zero integers, and $u \perp u'$ and $v \perp v'$, then $r\left(\frac{uv}{u'v'}\right) = r\left(\frac{u}{v'}\right) r\left(\frac{v}{u'}\right)$.*

Proof. Since $u \perp u'$ and $v \perp v'$, if uv and $u'v'$ have a common prime factor $s \geq 2$, then s must divide u and v' or divide u' and v (otherwise there is a contradiction). This means

$$1 = \frac{r\left(\frac{uv}{u'v'}\right)}{r\left(\frac{u}{v'}\right) r\left(\frac{v}{u'}\right)}$$

■

Claim 15 *If u, u', v, and v' are non-zero integers, and $u \perp u'$ and $v \perp v'$, then* $\gcd(uv, u'v') = \gcd(u,v')\gcd(u',v)$.

Proof. By Claim 11, we have

$$\begin{aligned}
\frac{\frac{uv}{\gcd(uv,u'v')}}{\frac{u'v'}{\gcd(uv,u'v')}} &= r\left(\frac{uv}{u'v'}\right) \\
&= r\left(\frac{u}{v'}\right) r\left(\frac{v}{u'}\right) \\
&= \left(\frac{\frac{u}{\gcd(u,v')}}{\frac{v'}{\gcd(u,v')}}\right)\left(\frac{\frac{v}{\gcd(u',v)}}{\frac{u'}{\gcd(u',v)}}\right) \\
&= \frac{\frac{uv}{\gcd(u,v')\gcd(u',v)}}{\frac{u'v'}{\gcd(u,v')\gcd(u',v)}}
\end{aligned}$$

so that

$$\gcd\left(uv, u'v'\right) = \gcd\left(u, v'\right)\gcd\left(u', v\right)$$

■

Algorithm 16 *If $\frac{u}{u'}$ and $\frac{v}{v'}$ are non-zero reduced rationals, then*

$$r\left(\left(\frac{u}{u'}\right)\left(\frac{v}{v'}\right)\right) = \frac{\left(\frac{u}{\gcd(u,v')}\right)\left(\frac{v}{\gcd(u',v)}\right)}{\left(\frac{u'}{\gcd(u',v)}\right)\left(\frac{v'}{\gcd(u,v')}\right)}$$

Proof. By Claim 15, we have

$$\begin{aligned}
r\left(\left(\frac{u}{u'}\right)\left(\frac{v}{v'}\right)\right) &= r\left(\frac{uv}{u'v'}\right) \\
&= \frac{\frac{uv}{\gcd(uv,u'v')}}{\frac{u'v'}{\gcd(uv,u'v')}}
\end{aligned}$$

$$= \frac{\frac{uv}{\gcd(u,v')\gcd(u',v)}}{\frac{u'v'}{\gcd(u,v')\gcd(u',v)}}$$
$$= \frac{\left(\frac{u}{\gcd(u,v')}\right)\left(\frac{v}{\gcd(u',v)}\right)}{\left(\frac{u'}{\gcd(u',v)}\right)\left(\frac{v'}{\gcd(u,v')}\right)}$$

■

Algorithm 17 *If $\frac{u}{u'}$ and $\frac{v}{v'}$ are non-zero reduced rationals, then*

$$r\left(\frac{\left(\frac{u}{u'}\right)}{\left(\frac{v}{v'}\right)}\right) = \frac{\left(\frac{u}{\gcd(u,v)}\right)\left(\frac{v'}{\gcd(u',v')}\right)}{\left(\frac{u'}{\gcd(u',v')}\right)\left(\frac{v}{\gcd(u,v)}\right)}$$

Proof. By Claim 15, we have

$$r\left(\frac{\left(\frac{u}{u'}\right)}{\left(\frac{v}{v'}\right)}\right) = r\left(\frac{uv'}{u'v}\right)$$
$$= \frac{\frac{uv'}{\gcd(uv',u'v)}}{\frac{u'v}{\gcd(uv',u'v)}}$$
$$= \frac{\frac{uv}{\gcd(u,v)\gcd(u',v')}}{\frac{u'v'}{\gcd(u,v)\gcd(u',v')}}$$
$$= \frac{\left(\frac{u}{\gcd(u,v)}\right)\left(\frac{v'}{\gcd(u',v')}\right)}{\left(\frac{u'}{\gcd(u',v')}\right)\left(\frac{v}{\gcd(u,v)}\right)}$$

■

Claim 18 *If u and v are non-zero integers, then $\gcd(u,v) = \gcd(v, u-qv)$, for any integer $q \geq 1$.*

Proof. ($\Longrightarrow$) Let $d = \gcd(u,v)$, and suppose s is the greatest common divisor of v and $u - qv$ for an integer q. Then $v = k_1 s$ and $u - qv = k_2 s$, for some integers $k_1, k_2 \geq 2$, which means

$$u = qv + k_2 s = (qk_1 + k_2)\, s$$
$$v = k_1 s$$

so that s is a common divisor of u and v. Hence, $s \leq d$. However, d is a common divisor of u and v, so that $u = k_3 d$ and $v = k_4 d$, for some integers $k_3, k_4 \geq 2$, which means

$$u - qv = (k_3 - qk_4)\, d$$
$$v = k_4 d$$

so that d is a common divisor of v and $u - qv$ for an integer q. Hence, $d \leq s$.

Therefore, $s = d$.

($\Longleftarrow$) Let $d = \gcd(v, u - qv)$ for some integer q, and suppose s is the greatest common divisor of u and v. Then $u = k_1 s$ and $v = k_2 s$, for some integers $k_1, k_2 \geq 2$, which means

$$u - qv = (k_1 - qk_2)\, s$$

$$v = k_2 s$$

so that s is a common divisor of v and $u - qv$ for an integer q. Hence, $s \leq d$. However, d is a common divisor of v and $u - qv$ for an integer q, so that $v = k_3 d$ and $u - qv = k_4 d$, which means

$$\begin{aligned} u &= qv + k_4 d = (qk_3 + k_4)\, d \\ v &= k_3 d \end{aligned}$$

so that d is a common divisor of u and v. Hence, $d \leq s$.

Therefore, $s = d$. ■

Claim 19 *If u and v are non-zero even integers, then* $\gcd(u, v) = 2\gcd\left(\frac{u}{2}, \frac{v}{2}\right)$.

Proof. Let $d = \gcd(u, v) \geq 2$. Then $u = k_1 d$ and $v = k_2 d$ for some integers $k_1, k_2 \geq 1$. Since u and v are even, then $\frac{u}{2} = k_1 \frac{d}{2}$ and $\frac{v}{2} = k_2 \frac{d}{2}$. Hence, $\frac{d}{2}$ is a divisor of $\frac{u}{2}$ and $\frac{v}{2}$. So $\frac{d}{2} \leq \gcd\left(\frac{u}{2}, \frac{v}{2}\right)$. However, $\gcd\left(\frac{u}{2}, \frac{v}{2}\right)$ is a common divisor of $\frac{u}{2}$ and $\frac{v}{2}$, so that $2\gcd\left(\frac{u}{2}, \frac{v}{2}\right)$ is a common divisor of u and v, which means $2\gcd\left(\frac{u}{2}, \frac{v}{2}\right) \leq d$. Therefore, $d = 2\gcd\left(\frac{u}{2}, \frac{v}{2}\right)$. ■

Claim 20 *If u is an even non-zero integer and v is a non-zero integer, then* $\gcd(u, v) = \gcd\left(\frac{u}{2}, v\right)$.

Proof. Let $d = \gcd(u, v) \geq 2$. Then $u = k_1 d$ and $v = k_2 d$ for some integers $k_1, k_2 \geq 1$. Since u is even, then $\frac{u}{2} = \frac{k_1}{2} d$ and $v = k_2 d$ mean d is a divisor of $\frac{u}{2}$ and v. So $d \leq \gcd\left(\frac{u}{2}, v\right)$. However, $s = \gcd\left(\frac{u}{2}, v\right)$ is a common divisor of $\frac{u}{2}$ and v, so that $\frac{u}{2} = k_3 s$ and $v = k_4 s$ for some integers $k_3, k_4 \geq 2$, i.e., $u = (2k_3)\, s$ and $v = k_4 s$, so that s is a common divisor of u and v. Hence, $s \leq d$. Therefore, $s = d$. ■

Definition 21 *The* Least Common Multiple lcm *of two non-zero integers u and v is the smallest positive integer d such that $\frac{d}{u}$ and $\frac{d}{v}$ are both non-zero integers.*

Claim 22 *For non-zero integers u and v, we have* $\operatorname{lcm}(u, v) = \frac{uv}{\gcd(u,v)}$.

Proof. Let

$$u = \prod_{i_k \in \{i_1, i_2, \ldots, i_n\}} p_{i_k}^{u_k} \quad \text{and} \quad v = \prod_{j_r \in \{j_1, j_2, \ldots, j_m\}} p_{j_r}^{v_r}$$

be the prime factorizations of u and v, respectively, for n-many prime numbers p_{i_k} with integer $u_k \geq 0$ and m-many prime numbers p_{j_r} with integer $v_r \geq 0$.

Now let

$$\begin{aligned} A &= \{i_1, i_2, \ldots, i_n\} \cup \{j_1, j_2, \ldots, j_m\} \\ B &= \{i_1, i_2, \ldots, i_n\} \cap \{j_1, j_2, \ldots, j_m\} \end{aligned}$$

Then

$$uv = \left(\prod_{i_s \in B} p_{i_s}^{u_s + v_s} \right) \left(\prod_{i_k \in A,\, j_r \notin B} p_{i_k}^{u_k} p_{j_r}^{0} \right) \left(\prod_{i_k \notin B,\, j_r \in A} p_{i_k}^{0} p_{j_r}^{v_r} \right)$$

and

$$\gcd(u, v) = \prod_{i_s \in B} p_{i_s}^{\min\{u_s, v_s\}}$$

so that

$$\frac{uv}{\gcd(u,v)} = \left(\prod_{i_s \in B} p_{i_s}^{\max\{u_s, v_s\}}\right)\left(\prod_{i_k \in A,\, j_r \notin B} p_{i_k}^{u_k} p_{j_r}^{0}\right)\left(\prod_{i_k \notin B,\, j_r \in A} p_{i_k}^{0} p_{j_r}^{v_r}\right) = \operatorname{lcm}(u,v)$$

■

Corollary 23 *If $u \perp v$, then $\operatorname{lcm}(u,v) = uv$.*

Proof. When $u \perp v$, we have $\gcd(u,v) = 1$. ■

Claim 24 *If u, u', v, and v' are non-zero integers, then $r\left(\frac{u}{u'} \pm \frac{v}{v'}\right) = r\left(\frac{\frac{v'}{\gcd(u',v')}u \pm \frac{u'}{\gcd(u',v')}v}{\frac{u'v'}{\gcd(u',v')}}\right)$.*

Proof. Let $d = \operatorname{lcm}(u', v') > 0$. Then $\frac{d}{u'}$ and $\frac{d}{v'}$ are non-zero integers, and from Claim 22, we have

$$\frac{u}{u'} \pm \frac{v}{v'} = \frac{\frac{d}{u'}u \pm \frac{d}{v'}v}{d} = \frac{\frac{v'}{\gcd(u',v')}u \pm \frac{u'}{\gcd(u',v')}v}{\frac{u'v'}{\gcd(u',v')}}$$

■

Note that $\frac{u'v'}{\gcd(u',v')}$ may be calculated as $\left(\frac{u'}{\gcd(u',v')}\right)v'$ or as $u'\left(\frac{v'}{\gcd(u',v')}\right)$ depending on which formulation uses the most convenient[1] integers.

Corollary 25 *If u, u', v, and v' are non-zero integers, and $u \perp u'$ and $v \perp v'$, then*

$$r\left(\frac{u}{u'} \pm \frac{v}{v'}\right) = \frac{\frac{v'}{\gcd(u',v')}u \pm \frac{u'}{\gcd(u',v')}v}{\frac{u'v'}{\gcd(u',v')}}$$

Proof. Suppose $p > 1$ is a common prime factor of $\frac{u'v'}{\gcd(u',v')}$ and $\frac{v'}{\gcd(u',v')}u \pm \frac{u'}{\gcd(u',v')}v$. Then p must be a prime factor of either u' or v' but not both (for otherwise it would be part of the $\gcd(u',v')$). Without loss of generality, say p is a prime factor of u', which means p is not a prime factor of v'. We have that $\frac{v'}{p\gcd(u',v')}u \pm \frac{u'}{p\gcd(u',v')}v$ is an integer, and since p is a prime factor of u', then $\frac{u'}{p\gcd(u',v')}v$ is an integer, so that $\frac{v'}{p\gcd(u',v')}u$ is also an integer. However, p is not a prime factor of v', so p must be a prime factor of u; this is a contradiction since $u \perp u'$. The same approach applies if p is a prime factor of v' and not of u'. ■

Corollary 26 *If u, u', v, and v' are non-zero integers, and $u' \perp v'$, then $r\left(\frac{u}{u'} \pm \frac{v}{v'}\right) = r\left(\frac{uv' \pm u'v}{u'v'}\right)$.*

Proof. We have $\gcd(u',v') = 1$, which means

$$\frac{u}{u'} \pm \frac{v}{v'} = \frac{\frac{u'v'}{u'}u \pm \frac{u'v'}{v'}v}{u'v'} = \frac{uv' \pm u'v}{u'v'}$$

■

Corollary 27 *If u, u', v, and v' are non-zero integers, and $u \perp u'$ and $v \perp v'$ and $u' \perp v'$, then $r\left(\frac{u}{u'} \pm \frac{v}{v'}\right) = \frac{uv' \pm u'v}{u'v'}$.*

[1] The smallest integers in absolute value should be used.

Proof. Suppose $u'v'$ and $uv' + u'v$ had a common prime factor $p > 1$. Since $u' \perp v'$, then either p is a prime factor of u' or v' but not both. Without loss of generality, say p is a prime factor of u', which means p is not a prime factor of v'. Now $\frac{uv'}{p} + \frac{u'v}{p}$ is an integer. Since p is a factor of u', then $\frac{uv'}{p}$ must also be an integer. However, since $u \perp u'$, then p must be a prime factor of v'; this is a contradiction. The same approach applies if p is a prime factor of v' and not of u'. ■

These results may be summarized in the following theorem.

Theorem 28 *If $\frac{u}{u'}$ and $\frac{v}{v'}$ are reduced rationals, then*

1. $r\left(\frac{u}{u'} \pm \frac{v}{v'}\right) = \frac{\frac{v'}{\gcd(u',v')}u \pm \frac{u'}{\gcd(u',v')}v}{\frac{u'v'}{\gcd(u',v')}}$
2. $r\left(\left(\frac{u}{u'}\right)\left(\frac{v}{v'}\right)\right) = \frac{\left(\frac{u}{\gcd(u,v')}\right)\left(\frac{v}{\gcd(u',v)}\right)}{\left(\frac{u'}{\gcd(u',v)}\right)\left(\frac{v'}{\gcd(u,v')}\right)}$
3. $r\left(\frac{\left(\frac{u}{u'}\right)}{\left(\frac{v}{v'}\right)}\right) = \frac{\left(\frac{u}{\gcd(u,v)}\right)\left(\frac{v'}{\gcd(u',v')}\right)}{\left(\frac{u'}{\gcd(u',v')}\right)\left(\frac{v}{\gcd(u,v)}\right)}$

and, in addition, if $u \perp v'$ and $u' \perp v$, then

2. $r\left(\left(\frac{u}{u'}\right)\left(\frac{v}{v'}\right)\right) = \frac{uv}{u'v'}$

and, in addition, if $u' \perp v'$, then

1. $r\left(\frac{u}{u'} \pm \frac{v}{v'}\right) = \frac{uv' \pm u'v}{u'v'}$
3. $r\left(\frac{\left(\frac{u}{u'}\right)}{\left(\frac{v}{v'}\right)}\right) = \frac{\left(\frac{u}{\gcd(u,v)}\right)v'}{u'\left(\frac{v}{\gcd(u,v)}\right)}$

and, in addition, if $u \perp v$, then

3. $r\left(\frac{\left(\frac{u}{u'}\right)}{\left(\frac{v}{v'}\right)}\right) = \frac{u\left(\frac{v'}{\gcd(u',v')}\right)}{\left(\frac{u'}{\gcd(u',v')}\right)v}$

and, in addition, if $u' \perp v'$ and $u \perp v$, then

3. $r\left(\frac{\left(\frac{u}{u'}\right)}{\left(\frac{v}{v'}\right)}\right) = \frac{uv'}{u'v}$

2. Rational Numbers As A Superset Of Floating Point Numbers

In general, and except for extreme analytical circumstances, every floating point number (for a particular choice of base and number of digits) may be expressed as either a non-zero rational number or as 0; however, not every rational number may be expressed as a floating point number (regardless of base, number of digits, or any other consideration). In this respect, the rational numbers may be viewed as a superset of the floating point numbers.

Theorem 29 states this relationship in precise terms. The following preliminary development is in support of that theorem.

Given an N digit floating point real number x in normal form with sign bit s, base $b > 1$, n digit exponent p such that $N - 1 \geq n \geq 1$ and $0 \leq p \leq b^n - 1$, offset $b^{n-1} - 1$, and $(N - n - 1)$ digit mantissa $m \geq 0$, we have

$$x = (-1)^s \, b^{p - b^{n-1} + 1} \left(1 + \frac{m}{b^{N-n-1}}\right)$$

Note this means

$$1 - b^{n-1} \leq p - b^{n-1} + 1 \leq b^{n-1} \, (b - 1)$$

and

$$0 \leq m \leq b^{N-n-1} - 1$$

For example, the IEEE-754 64-bit base 2 (binary) floating point real number x in normal form has $N = 64$, $b = 2$, $n = 11$, $k = 2^{10} - 1 = 1023$, and $N - n - 1 = 52$, so that

$$x = (-1)^s \, 2^{p - 1023} \left(1 + \frac{m}{2^{52}}\right)$$

Since $p \geq 0$, $m \geq 0$, and $N - n - 1 \geq 0$, then $N + b^{n-1} - n - 2 \geq 0$, and

$$x = r \left(\frac{(-1)^s \, b^p \left(b^{N-n-1} + m\right)}{b^{N + b^{n-1} - n - 2}} \right)$$

is the reduced rational value of x. Therefore, every N digit floating point real number x in normal form has a reduced rational value.

Theorem 29 *Every positive reduced rational number $\frac{u}{u'}$ may be expressed as an exact $\{N, b, n\}$ floating point real number if $N < b^{n-1}$ and $u' = b^w$ for some integer w using $p = b^{n-1} + \left\lfloor \log_b \frac{u}{u'} \right\rfloor - 1 - q$ and $m = b^{N-n-1} \left(u b^{q - \left\lfloor \log_b \frac{u}{u'} \right\rfloor - w} - 1 \right)$, where $w + \left\lfloor \log_b \frac{u}{u'} \right\rfloor \leq q \leq b^{n-1} + \left\lfloor \log_b \frac{u}{u'} \right\rfloor - 1$; otherwise, for $N < b^{n-1}$ the floating point real number is an approximation using $p = b^{n-1} + \left\lfloor \log_b \frac{u}{u'} \right\rfloor - 1 - q$ and $m = b^{N-n-1} \left(u \left\lfloor \frac{b^{q - \left\lfloor \log_b \frac{u}{u'} \right\rfloor}}{u'} \right\rfloor - 1 \right)$ or $m = b^{N-n-1} \left(u \left\lceil \frac{b^{q - \left\lfloor \log_b \frac{u}{u'} \right\rfloor}}{u'} \right\rceil - 1 \right)$, where $q = \max \left\{ q \leq b^{n-1} + \left\lfloor \log_b \frac{u}{u'} \right\rfloor - 1 : u \left\lceil \frac{b^{q - \left\lfloor \log_b \frac{u}{u'} \right\rfloor}}{u'} \right\rceil < b^n + 1 \right\}$ minimizes the absolute error.*

Proof. Given $\{N, b, n\}$, let

$$s = \frac{1}{2} \left| sgn\left(\frac{u}{u'}\right) - 1 \right| = 0$$

and

$$p = b^{n-1} + \left\lfloor \log_b \frac{u}{u'} \right\rfloor - 1 - q$$

Note that

$$1 \leq u, u' \leq b^N - 1$$

means

$$b^{-N} < \frac{u}{u'} \leq b^N - 1$$

so that

$$-b^{n-1} < -N < p - b^{n-1} + 1 < N < b^{n-1}$$

Also let

$$m = b^{N-n-1}\left(\frac{u}{u'}b^{q-\left\lfloor \log_b \frac{u}{u'}\right\rfloor}-1\right)$$

Then m is an integer if and only if u' is of the form b^w, for some integer $w \leq q - \left\lfloor \log_b \frac{u}{u'}\right\rfloor$. In this case, we have

$$m = b^{N-n-1}\left(ub^{q-\left\lfloor \log_b \frac{u}{u'}\right\rfloor-w}-1\right)$$

is an integer, and we have

$$(-1)^s\, b^{p-b^{n-1}+1}\left(1+\frac{m}{b^{N-n-1}}\right) = b^{b^{n-1}+\left\lfloor \log_b \frac{u}{u'}\right\rfloor-1-q-b^{n-1}+1}\left(1+\frac{b^{N-n-1}\left(ub^{q-\left\lfloor \log_b \frac{u}{u'}\right\rfloor-w}-1\right)}{b^{N-n-1}}\right) = ub^{-w} = \frac{u}{u'}$$

Otherwise, using

$$m = b^{N-n-1}\left(u\left\lfloor\frac{b^{q-\left\lfloor \log_b \frac{u}{u'}\right\rfloor}}{u'}\right\rfloor-1\right) \quad \text{or} \quad m = b^{N-n-1}\left(u\left\lceil\frac{b^{q-\left\lfloor \log_b \frac{u}{u'}\right\rfloor}}{u'}\right\rceil-1\right)$$

then the absolute difference between

$$b^{\left\lfloor \log_b \frac{u}{u'}\right\rfloor-q}\left(1+\frac{b^{N-n-1}\left(u\left\lfloor\frac{b^{q-\left\lfloor \log_b \frac{u}{u'}\right\rfloor}}{u'}\right\rfloor-1\right)}{b^{N-n-1}}\right) \quad \text{and} \quad b^{\left\lfloor \log_b \frac{u}{u'}\right\rfloor-q}\left(1+\frac{b^{N-n-1}\left(u\left\lceil\frac{b^{q-\left\lfloor \log_b \frac{u}{u'}\right\rfloor}}{u'}\right\rceil-1\right)}{b^{N-n-1}}\right)$$

is given by

$$ub^{\left\lfloor \log_b \frac{u}{u'}\right\rfloor-q}$$

and by choosing q as large as possible, namely

$$\max\left\{q \leq b^{n-1}+\left\lfloor \log_b \frac{u}{u'}\right\rfloor-1 : u\left\lceil\frac{b^{q-\left\lfloor \log_b \frac{u}{u'}\right\rfloor}}{u'}\right\rceil < b^n+1\right\}$$

then this absolute error is minimized. ■

Note that every negative reduced rational number may be expressed as the negative of the floating point real number found in Theorem 29 for its absolute value.

For example, the (IEEE-754 64-bit) base 2 floating point representation of the (IEEE-754 64-bit) base 10 floating point real number 34.77821 may be approximated as

$$(34.77821)_{10} > (-1)^0\, 2^{1028-1023}\left(1+\frac{(1011000111001110001100010101000011011010111000111)_2}{2^{52}}\right)$$
$$(34.77821)_{10} < (-1)^0\, 2^{1028-1023}\left(1+\frac{(1011000111001110001100010101000011011010111001000)_2}{2^{52}}\right)$$

yet

$$(34.77821)_{10} = \frac{(11010100001000100111101)_2}{(11000011010100000)_2}$$

exactly (as a reduced and rectified non-zero rational number).

3. Representation Of Irrational Numbers

While it may be optimal to utilize rational arithmetic and conversion operators as documented in this memorandum, et seq., and to implement all statistical calculation applications strictly with rational arithmetic and conversion operators (with appropriate extensions), there will inevitably be analytical contexts where irrational numbers must be included in a calculation. These contexts may be as straightforward as an intermediate addition/multiplication of an irrational number to a rational number, or as specific as an infinite series of rational numbers asymptotic to the irrational number, to name only a few such possibilities. In all cases, a formal system of arithmetic operators must be defined that take as their operands a combination of rational and irrational numbers, and that provides for separate accounting of each type of term. In this manner, any error encountered by using rational terms to approximate the irrational terms at any intermediate point in a calculation may be bounded by absolute or relative amounts at the discretion of the implementing analyst.

4. Operators

In addition to the usual arithmetic binary operations of addition, subtraction, multiplication, and division, rational arithmetic also includes several unitary operators that involve negation, rectification, reduction, rationalization, and conversions between representative forms of the operand.

All arithmetic operations take place with base b calculations, even if the operands and resulting forms may be given in/converted to a different base. When $b > 10$, the uppercase letters $A - Z$ shall be used to designate the place values for $10 - 35$, and lowercase letters $a - z$ for place values for $36 - 61$. When $b > 36$, additional non-ASCII characters would be needed to uniquely name each such base b number.

4.1 Rectification

The unitary operator *Rectification* takes a non-zero rational number and returns either a positive rational number or a negative rational number with positive denominator according to the results of Table X. This operation does not affect 0. All calculations take place in context with base b arithmetic.

x	Arg1	Result
	$\frac{u}{u'}$	$\begin{cases} \frac{u}{u'}, & u' > 0 \\ \frac{-u}{-u'}, & u' < 0 \end{cases}$
	0	0

Given the non-zero rational number $\frac{u}{v}$, the following algorithm returns the rectified value of $\frac{u}{v}$, namely $x\left(\frac{u}{v}\right)$.

1. If $v > 0$, then return $\frac{u}{v}$; otherwise continue.
2. If $v < 0$, then return $\frac{-u}{-v}$; otherwise return $\frac{NaN}{NaN}$.

Note that $u \neq 0$ in this algorithm since $\frac{u}{v}$ must be a non-zero rational number.

4.2 Reduction

The unitary operator *Reduction* takes a non-zero base b rational number $\frac{u}{u'}$ and returns the reduced rational number $r\left(\frac{u}{u'}\right)$ which is equal to the operand as a real number according to the definition of a reduced rational number. This operation does not affect 0. All calculations take place in context with base b arithmetic.

Given a non-zero base b rational number $\frac{u}{u'}$ and a precision p, the following algorithm calculated the reduced rational number of $\frac{u}{u'}$, namely $r\left(\frac{u}{u'}\right)$.

1. Determine the signs[2] of u and v; call these values $sgnu$ and $sgnv$.
2. Convert $|u|$ from base b to decimal; call this value ux.
3. Convert $|v|$ from base b to decimal; call this value vx.
4. Calculate the greatest common divisor k of ux and vx.
5. Convert $\frac{ux}{k}$ from decimal to base b; call this value uu.
6. Convert $\frac{vx}{k}$ from decimal to base b; call this value vv.
7. Return $\frac{sgnu \times uu}{sgnv \times vv}$.

4.3 Negation

The unitary operator *Negation* takes a non-zero rational number $\frac{u}{u'}$ and returns the rectified form of the negative of the rational number $\frac{u}{u'}$, namely $n\left(\frac{u}{u'}\right)$. This operation does not affect 0. Note that integer negation is performed with respect to 0, and *not* with respect to b. All calculations take place in context with base b arithmetic.

x	Arg1	Result
	$\frac{u}{u'}$	$\begin{cases} \frac{-u}{u'}, & u' > 0 \\ \frac{u}{-u'}, & u' < 0 \end{cases}$
	0	0

The following algorithm implements this operator.

1. If $u' > 0$, then return $\frac{-u}{u'}$; otherwise continue.
2. If $u' < 0$, then return $\frac{u}{-u'}$; otherwise continue.
3. Otherwise return 0.

4.4 Addition/Subtraction

Given two non-zero rational numbers $\frac{u}{u'}$ and $\frac{v}{v'}$, base $b > 1$, and precision p, the following algorithm calculates the reduced and rectified arithmetic sum of $\frac{u}{u'}$ and $\frac{v}{v'}$, namely $r\left(x\left(\frac{u}{u'} + \frac{v}{v'}\right)\right)$. All calculations take place in context with base b arithmetic.

[2]Note that the operand does not need to be rectified.

$\pm$	Arg1	Arg2	Result
Addition	$\frac{u}{u'}$	0	$\frac{u}{u'}$
	$\frac{u}{u'}$	$\frac{v}{v'}$	$r\left(\frac{u}{u'}+\frac{v}{v'}\right)$
	0	$\frac{u}{u'}$	$\frac{u}{u'}$
	0	0	0
Subtraction	$\frac{u}{u'}$	0	$\frac{u}{u'}$
	$\frac{u}{u'}$	$\frac{v}{v'}$	$r\left(\frac{u}{u'}+n\left(\frac{v}{v'}\right)\right)$
	0	$\frac{u}{u'}$	$n\left(\frac{u}{u'}\right)$
	0	0	0

The following algorithm implements this operator.

1. Rectify $\frac{u}{u'}$; call this rational number R.
2. Convert the numerator of R from base b to decimal; call this value ux.
3. Convert the denominator of R from base b to decimal; call this value upx.
4. Rectify $\frac{v}{v'}$; call this rational number (the new value of) R.
5. Convert the numerator of R from base b to decimal; call this value vx.
6. Convert the denominator of R from base b to decimal; call this value vpx.
7. Calculate the greatest common divisor of $|upx|$ and $|vpx|$; call this value k.
8. Reduce $\frac{\frac{ux\times vpx}{k}+\frac{vx\times upx}{k}}{\frac{upx\times vpx}{k}}$ base 10; call this rational number (the new value of) R.
9. Convert the numerator of R from decimal to base b; call this value uxx.
10. Convert the denominator of R from decimal to base b; call this value vxx.
11. Rectify $\frac{uxx}{vxx}$, and return the results.

Note that subtraction is simply addition of the first rational number with the negation of the second rational number.

4.5 Multiplication

Given two non-zero rational numbers $\frac{u}{u'}$ and $\frac{v}{v'}$, base $b > 1$, and precision p, the following algorithm calculates the reduced and rectified arithmetic product of $\frac{u}{u'}$ and $\frac{v}{v'}$, namely $r\left(x\left(\frac{u}{u'}\times\frac{v}{v'}\right)\right)$. All calculations take place in context with base b arithmetic.

$*$	Arg1	Arg2	Result
Multiplication	$\frac{u}{u'}$	0	0
	$\frac{u}{u'}$	$\frac{v}{v'}$	$r\left(\left(\frac{u}{u'}\right)\left(\frac{v}{v'}\right)\right)$
	0	$\frac{u}{u'}$	0
	0	0	0

The following algorithm implements this operator.

1. Convert the u from base b to decimal; call this value ux.
2. Convert the u' from base b to decimal; call this value upx.
3. Convert the v from base b to decimal; call this value vx.
4. Convert the v' from base b to decimal; call this value vpx.
5. Calculate the greatest common divisor of $|ux|$ and $|vpx|$; call this value k.
6. Calculate the greatest common divisor of $|upx|$ and $|vx|$; call this value r.
7. Reduce $\frac{\frac{ux}{k} \times \frac{vx}{r}}{\frac{upx}{r} \times \frac{vpx}{k}}$ base 10; call this rational number (the new value of) R.
8. Convert the numerator of R from decimal to base b; call this value uxx.
9. Convert the denominator of R from decimal to base b; call this value vxx.
10. Rectify $\frac{uxx}{vxx}$, and return the results.

4.6 Division

Given two non-zero rational numbers $\frac{u}{u'}$ and $\frac{v}{v'}$, base $b > 1$, and precision $[$, the following algorithm calculates the reduced and rectified arithmetic quotient (ratio) of $\frac{u}{u'}$ and $\frac{v}{v'}$, namely $r\left(x\left(\frac{u}{u'} \times \frac{v}{v'}\right)\right)$. All calculations take place in context with base b arithmetic.

$\div$	Arg1	Arg2	Result
Division	$\frac{u}{u'}$	0	NaN
	$\frac{u}{u'}$	$\frac{v}{v'}$	$r\left(\frac{\left(\frac{u}{u'}\right)}{\left(\frac{v}{v'}\right)}\right)$
	0	$\frac{u}{u'}$	0
	0	0	NaN

The following algorithm implements this operator.

1. Convert the u from base b to decimal; call this value ux.
2. Convert the u' from base b to decimal; call this value upx.
3. Convert the v from base b to decimal; call this value vx.
4. Convert the v' from base b to decimal; call this value vpx.
5. Calculate the greatest common divisor of $|ux|$ and $|vpx|$; call this value k.
6. Calculate the greatest common divisor of $|upx|$ and $|vx|$; call this value r.
7. Reduce $\frac{\frac{ux}{k} \times \frac{vpx}{r}}{\frac{upx}{r} \times \frac{vx}{k}}$ base 10; call this rational number (the new value of) R.
8. Convert the numerator of R from decimal to base b; call this value uxx.
9. Convert the denominator of R from decimal to base b; call this value vxx.
10. Rectify $\frac{uxx}{vxx}$, and return the results.

4.7 b-Float

Given two base b integers $u = (\pm u_{n-1}u_{n-2}\cdots u_1u_0)_b$ and $v = (v_{m-1}v_{m-2}\cdots v_1v_0)_b$, the following algorithm calculates the base b positive floating point representation of $\frac{u}{v}$ to p digits. It is understood that $u_{k<0} = 0$, and all calculations take place in context with base b arithmetic.[3]

1. Determine the sign[4] of u; call this value sgn.
2. Calculate the decimal value of $|u|$; call this value ux.
3. Calculate the decimal value of v; call this value vx.
4. Partition ux into an array of values uu.
5. Set $x = 0$, $w = 0$, and $k = 1$.
6. Set $N = uu\,[n - k]$.
7. Calculate $q = \left\lfloor \frac{N}{v} \right\rfloor$ and $r = N - qv$.
8. If $q = 0$, then increment k by 1 and set $10N + u_{n-k}$ to be the new value of N, then skip to Step 3; otherwise, continue.
9. Set $x + q10^{n-k}$ to be the new value of x and increment w by 1.
10. If $w = p$, then skip to Step 12; otherwise, continue.
11. Increment k by 1, set $N = 10r + u_{n-k}$, and skip to Step 3.
12. Convert $sgn \times x$ from decimal to base b, and return this result.

4.8 b-Floor

Use the quotient and remainder process to find the floor function (in all cases where the operand may be positive or negative). All calculations take place in context with base b arithmetic.

Given the non-zero rational number $\frac{u}{v}$, base $b > 1$, and precision p, the following algorithm calculates the floor function of $\frac{u}{v}$, namely $f\left(\frac{u}{v}\right)$.

1. Determine the sign[5] of u; call this value sgn (which may be positive, negative, or zero).
2. Calculate the quotient q and remainder r of $\frac{u}{v}$.
3. If $sgn < 0$, then ...
   (a) If the quotient q is 0, then return $-q$; otherwise continue.
   (b) If the quotient q is non-zero, then return $-q - 1$.
4. If $sgn = 0$, then return 0.
5. If $sgn > 0$, then return q.

[3]The floor function used in the b-Float operator $\lfloor\cdots\rfloor$ acts the same way regardless of base $b > 1$. In particular, if $u = \left(u_{n-1}u_{n-2}\cdots u_{n-k}.u_{n-(k+1)}\cdots u_1u_0\right)_b$ is a base b floating point real number of length n, where $0 \le u_j < b$, and where the fractional part begins between u_{n-k} and u_{n-k-1} (thereby having least significant digit value b^{n-k}), then $\lfloor u \rfloor = (u_{n-1}u_{n-2}\cdots u_{n-k})_b$ regardless of b.

[4]Note that the operand does need to be rectified.

[5]Note that the operand does need to be rectified; the subtraction in Step 3b is made in base b arithmetic.

4.9 b-Ceiling

Use the quotient and remainder process to find the ceiling function (in all cases where the operand may be positive or negative). All calculations take place in context with base b arithmetic.

Given the non-zero rational number $\frac{u}{v}$, base $b > 1$, and precision p, the following algorithm calculates the ceiling function of $\frac{u}{v}$, namely $c\left(\frac{u}{v}\right)$.

1. Determine the sign[6] of u; call this value sgn (which may be positive, negative, or zero).
2. Calculate the quotient q and remainder r of $\frac{u}{v}$.
3. If $sgn > 0$, then ...
   (a) If the quotient q is 0, then return q; otherwise continue.
   (b) If the quotient q is non-zero, then return $q + 1$.
4. If $sgn = 0$, then return 0.
5. If $sgn > 0$, then return $-q$.

4.10 (b, v)-Quantization

The following claim justifies the analytical basis for the quantization operator.

Claim 30 *Given the non-zero rectified rational number $\frac{u}{u'}$ and a non-zero integer v', such that $2uv' < -u'$ or $2uv' \geq u'$, then $\frac{\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor}{v'}$ is the closest rational number with denominator v' to $\frac{u}{u'}$ in absolute value, and the difference is $\left| \frac{u'\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor - uv'}{u'v'} \right|$. For $-u' \leq 2uv' < u'$, then 0 is the closest number to $\frac{u}{u'}$ in absolute value.*

Proof. Given $u', v' \neq 0$, for every $u \neq 0$ we have

$$-1 < \left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor - \left(\frac{u}{u'}v' + \frac{1}{2} \right) \leq 0$$

which means

$$-\frac{1}{2v'} < \frac{1}{v'}\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor - \frac{u}{u'} \leq \frac{1}{2v'}$$

Therefore, $\frac{\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor}{v'}$ is within $\frac{1}{2v'}$ of $\frac{u}{u'}$. However, for integer $q \neq 0$, this means

$$\frac{2q-1}{2v'} < \frac{1}{v'}\left(\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor + q \right) - \frac{u}{u'} \leq \frac{2q+1}{2v'}$$

For $\frac{1}{v'}\left(\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor + q \right)$ to be closer to $\frac{u}{u'}$ than is $\frac{\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor}{v'}$, we have that $2q - 1 > -1$ or $q > 0$, and $2q + 1 < 1$ or $q < 0$, must occur at the same time; this is a contradiction.

Furthermore, $\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor = 0$ if and only if $-\frac{1}{2} \leq \frac{u}{u'}v' < \frac{1}{2}$, which means $-u' \leq 2uv' < u'$ since $u' > 0$, i.e., since $\frac{u}{u'}$ is rectified.

[6]Note that the operand does need to be rectified; the addition in Step 3b is made in base b arithmetic.

Hence, for $2uv' < -u'$ or $2uv' \geq u'$,

$$v_0 = \left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor$$

provides the closest rational number $\frac{v_0}{v'}$ with denominator v' to $\frac{u}{u'}$ in absolute value; otherwise, 0 provides the closest number to $\frac{u}{u'}$ in absolute value. ■

Corollary 31 *If $v' = u'$ in Claim 30, then $\frac{u}{v'}$ is the closest rational number with denominator v' to $\frac{u}{u'}$ in absolute value.*

Proof. We have $v' = u'$ means either $2uv' > v' = u'$ or $2uv' < v' = u'$ (since $|u| \geq 1$ and $v' = u' > 0$, i.e., $\frac{u}{u'}$ is rectified). Therefore, the numerator of the closest rational number with denominator v' to $\frac{u}{u'}$ in absolute value is non-zero.

In particular, we have

$$\left\lfloor \frac{u}{u'}v' + \frac{1}{2} \right\rfloor = \left\lfloor u + \frac{1}{2} \right\rfloor = u$$

regardless of the sign of u. ■

Definition 32 *0 is the closest number to 0 in absolute value, and there is no rational number closest to 0 in absolute value.*

Remark 33 *These definitions are consistent with Claim 30 in the sense that $\left\lfloor \frac{0}{u'}v' + \frac{1}{2} \right\rfloor = 0$ for all $u', v' \neq 0$, and if there were a rational number $\frac{v}{v'}$ closest to 0 in absolute value, then $\frac{v}{2v'}$ would be closer, i.e., $\left|\frac{v}{2v'} - 0\right| < \left|\frac{v}{v'} - 0\right|$, which is a contradiction.*

For example, the closest rational number with denominator 18 to $\frac{362}{9201}$ is $\left\lfloor \frac{362}{9201}(18) + \frac{1}{2} \right\rfloor = 1$. This is confirmed by

$$0.039344 \approx \left|\frac{0}{18} - \frac{362}{9201}\right| > \left|\frac{1}{18} - \frac{362}{9201}\right| \approx 0.016212$$
$$0.071768 \approx \left|\frac{2}{18} - \frac{362}{9201}\right| > \left|\frac{1}{18} - \frac{362}{9201}\right| \approx 0.016212$$

The rectified and reduced form of $\frac{1}{18}$ is $\frac{1}{18}$, and the difference is $\left|\frac{(9201)\left\lfloor \frac{362}{9201}(18) + \frac{1}{2} \right\rfloor - (362)(18)}{(9201)(18)}\right| \approx 0.016212$.

As another example, the closest rational number with denominator -75 to $\frac{177}{381}$ is $\left\lfloor \frac{177}{381}(-75) + \frac{1}{2} \right\rfloor = -35$. This is confirmed by

$$0.011234 \approx \left|\frac{-34}{-75} - \frac{177}{381}\right| > \left|\frac{-35}{-75} - \frac{177}{381}\right| \approx 0.0020997$$
$$0.015433 \approx \left|\frac{-36}{-75} - \frac{177}{381}\right| > \left|\frac{-35}{-75} - \frac{177}{381}\right| \approx 0.0020997$$

The rectified and reduced form of $\frac{-35}{-75}$ is $\frac{7}{15}$, and the difference is $\left|\frac{(381)\left\lfloor \frac{177}{381}(-75) + \frac{1}{2} \right\rfloor - (177)(-75)}{(381)(-75)}\right| \approx 0.0020997$.

Given a reduced rational number $\frac{u}{u'}$ and a non-zero integer v', the ternary operator (b, v) *-Quantization* q calculates the non-zero base b integer v and returns the base b rational number $\frac{v}{v'}$ that is closest to $\frac{u}{u'}$, or returns 0 when the operand is 0. Note that $r\left(\frac{v}{v'}\right)$ is also just as close to $\frac{u}{u'}$ as is $\frac{v}{v'}$, even though the denominator of $r\left(\frac{v}{v'}\right)$ may not be equal to v'.

In the following table, all additions and multiplications are performed relative to base b arithmetic.

q	Arg1	Arg2	Result
	$\frac{u}{u'}$	$v' \neq u'$	$\begin{cases} \frac{v}{v'}, & \lvert u'v - uv' \rvert \leq \lvert u'(v+1) - uv' \rvert \\ \frac{v+1}{v'}, & \lvert u'v - uv' \rvert > \lvert u'(v+1) - uv' \rvert \end{cases}$
	$\frac{u}{u'}$	$v' = u'$	$\frac{u}{u'}$
	0	v'	0
	0	0	NaN

The following algorithm implements this operator.

1. Reduce $\frac{v}{u'}$; call this value R.
2. Reduce the product of R with u; call this (the new value of) R.
3. Reduce the sum of r with $\frac{1}{2}$; call this (the new value of) R.
4. Calculate the floor function of R; call this value k.
5. Return $\frac{k}{v}$.

4.11 b-Rationalization

Given a b-Float number x, the unitary operator *b-Rationalization* returns the reduced and rectified base b rational number $\frac{u}{u'}$ that has the same value as x. The following algorithm calculates the values of u and u'. All calculations take place in context with base b arithmetic.

The following algorithm implements this operator.

1. Let n be the smallest positive integer such that xb^n is an integer.[7]
2. Return the reduced value of $\frac{x \times b^n}{b^n}$.

Note that the return value is automatically rectified by the algorithm.

4.12 (b, q)-Fix

Given two base b positive integers u and v, and a precision level q, the binary operator *(b, q)-Fix* returns the base b rational number $\frac{w}{b^q}$ whose value is closest (in absolute value to q digits precision) to $\frac{u}{v}$. All calculations take place in context with base b arithmetic.

The following algorithm implements this operator.

1. Calculate the base b floating point value x of $\frac{u}{v}$ to precision q.
2. Determine the number of base b digits needed to represent the result in Step 1; call this value n.
3. Return the reduced value of $\frac{x \times b^n}{b^n}$.

Note that the return value is automatically rectified by the algorithm.

[7] Such an integer always exists: (a) if x is an integer, then $n = 0$, and (b) if x is not an integer, then xb^n is never an integer for any $n < 0$; hence, n is bounded below by 0.

5. Greatest Common Divisor Calculation

The following algorithm calculates the greatest common divisor, and by inference, the least common multiple (see Claim 22).

Algorithm 34 *(Stein Binary Method[**1**]) Given positive integers u and v, the following steps calculate* $\gcd(u, v)$.

1. *Set* $C = 0$.
2. *Continue to divide* u *and* v *by* 2 *until one of them is odd. Record the number of such divisions as* C.
   (a) *By Claim 19, we have* $\gcd(u, v) = 2^C \gcd\left(\frac{u}{2^C}, \frac{v}{2^C}\right)$, *for integer* $C \geq 0$.
3. *Set* $u \longleftarrow \frac{u}{2^C}$ *and* $v \longleftarrow \frac{v}{2^C}$.
4. *If* u *is odd, then set* $t = -v$ *and* $s = u$; *otherwise set* $t = u$ *and* $s = v$.
5. *Continue to divide* t *by* 2 *until it is odd.*
   (a) *By Claim 20, we have* $\gcd(u, v) = \gcd\left(\frac{u}{2^R}, v\right)$ *for any integer* $R \geq 0$.
   (b) *At this point both* t *and* s *are odd.*
6. *If* $t > 0$, *then set* $u \longleftarrow t$; *otherwise set* $v \longleftarrow -t$.
   (a) *If* $t > 0$, *then* $u > v$ *(when repeated from Step 8), so that* $|t|$ *is set in place of* $\max(u, v)$.
   (b) *Before a repeat from Step 8,* $|t|$ *may be set in place of either* $\max(u, v)$ *or* $\min(u, v)$.
7. *Set* $t \longleftarrow u - v$.
   (a) *By Claim 18,* $\gcd\left(\frac{u}{2^R}, v\right) = \gcd\left(v, \frac{u}{2^R} - v\right)$ *(where* $q = 1$ *in this case).*
   (b) *Since* $|u - v| < \max(u, v)$, *when repeated from Step 8, the new value of* t *is necessarily less than either* u *or* v, *even if no divisions by* 2 *occur in Step 5, so that the next application of Steps 6-7 will strictly reduce the value of* t *(and therefore either* u *or* v*). In this respect,* t *converges to* 0 *in finitely many steps (see Step 8).*
8. *If* $t \neq 0$, *repeat from Step 5; otherwise return* $2^C u$.

REFERENCES

1 Stein, J., *Computational problems associated with Racah algebra*, Journal of Computational Physics, **1**:3 (February 1967), pp. 397-405.