Midpoints and exact points of some algebraic functions in floating-point arithmetic

Claude-Pierre Jeannerod, Nicolas Louvet, Jean-Michel Muller, *Senior member, IEEE*, Adrien Panhaleux

Abstract—When implementing a function f in floating-point arithmetic, if we wish correct rounding and good performance, it is important to know if there are input floating-point values x such that $f(x)$ is either the middle of two consecutive floating-point numbers (assuming rounded-to-nearest arithmetic), or a floating-point number (assuming rounded towards $\pm\infty$ or towards 0 arithmetic). In the first case, we say that $f(x)$ is a *midpoint*, and in the second case, we say that $f(x)$ is an *exact point*. For some usual algebraic functions, and various floating-point formats, we prove whether or not there exist midpoints or exact points. When there exist midpoints or exact points, we characterize them or list all of them (if there are not too many). The results and the techniques presented in this paper can be used in particular to deal with both the binary and the decimal formats defined in the IEEE 754-2008 standard for floating-point arithmetic.

Index Terms—floating-point arithmetic, correct rounding, algebraic function.

1 Introduction

In a floating-point system that follows the IEEE 754-1985 standard for radix-2 floating-point arithmetic [1], the user can choose an *active rounding mode*, also called *rounding-direction attribute* in the newly revised IEEE 754-2008 standard [5]: rounding toward $-\infty$, rounding toward $+\infty$, rounding toward 0, and rounding to nearest, which is the default rounding mode. Given a real number x, we denote respectively by $\mathrm{RD}(x)$, $\mathrm{RU}(x)$, $\mathrm{RZ}(x)$, and $\mathrm{RN}(x)$ these rounding modes. Let us also recall that *correct rounding* is required by the above cited IEEE standards for the four elementary arithmetic operations ($+$, $-$, $\times$, $\div$) as well as for the square root: the result of an operation is said to be correctly-rounded if for any inputs its result is the infinitely precise result rounded according to the active rounding mode. We are interested here in facilitating the delivery of correctly-rounded results for various simple algebraic functions that are frequently used in numerical analysis or signal processing.

Given a function $f : \mathbb{R}^d \to \mathbb{R}$ and a floating-point number x, the problem of computing $\mathrm{RN}(f(x))$ is closely related to the knowledge of the midpoints of the function f. We say that $f(x)$ *is a midpoint* of the function f if the exact value $f(x)$ is halfway between two consecutive floating-point numbers. Then a common strategy (see [15] and [10, chap. 10]) for returning $\mathrm{RN}(f(x))$ is as follows.

Let us first compute an approximation f_1, of accuracy ϵ_1, to $f(x)$. If there are no midpoints within distance ϵ_1 from f_1, then necessarily $\mathrm{RN}(f(x)) = \mathrm{RN}(f_1)$. If on the contrary there are such midpoints, we can progressively increase the quality of the approximations (that is, computing an approximation f_2 of accuracy $\epsilon_2 < \epsilon_1$, and so on) until we are able to provide a correctly-rounded result. The point is that this strategy may not terminate if there exist floating-point numbers x such that $f(x)$ is a midpoint. As a consequence, a correctly-rounded implementation of a given function f can be made more efficient if we know in advance that f admits no midpoints. If f admits midpoints, it is also very useful to know how to characterize them.

If now we consider one of the *directed rounding modes* (RD, RU, or RZ), the strategy that consists in progressively refining the approximations will not terminate if $f(x)$ is a floating-point number. In this case we say that $f(x)$ *is an exact point* of the function f, and it is also very useful to know a characterization of these exact points when implementing f. Moreover, a characterization of the exact points of f can be used to set the "inexact" flag required by the IEEE standards [1], [5]. For example, for $x/\sqrt{x^2+y^2}$ in radix 2, our study shows that this flag must always be raised except when x or y is zero, which can be detected easily.

In this paper, we present results on the existence of midpoints and exact points for some algebraic functions: beyond division, inversion, and square root, we study functions like the reciprocal square root $1/\sqrt{y}$, the 2D Euclidean norm $\sqrt{x^2+y^2}$ and its reciprocal $1/\sqrt{x^2+y^2}$, and the 2D-normalization function $x/\sqrt{x^2+y^2}$. A part of the results presented on division and square root have been known for some time in binary arithmetic; see for instance the pioneering work by Markstein [9], as well as studies by Iordache and Matula [6] and Parks [11]. Let us also recall the work by Lauter and Lefèvre [8] on the function x^y, which thus covers integer powers. We present these results for completeness, and we extend some of them to other radices, in particular to radix 10.

Before going into further details, we introduce some definitions. A radix-β, precision-p floating-point number x is either 0 or a rational number of the form

$$x = \pm X \cdot \beta^{e_x - p + 1},$$

where X is a positive integer such that $X < \beta^p$. If in addition $\beta^{p-1} \leq X$, then $x = \pm X \cdot \beta^{e_x - p + 1}$ is called the *normalized* representation of x, and the integers X and

e_x are called, respectively, the *integral significand* and the *exponent* of x. We can in fact speak of the exponent for any nonzero real x: in radix β, it is the unique integer e_x such that $\beta^{e_x} \leq |x| < \beta^{e_x+1}$. On computing systems conforming to the IEEE 754-2008 standard [5], the radix β is 2 or 10. Radix 16 is also sometimes used [12]. The exponent e_x is bounded: $e_{\min} \leq e_x \leq e_{\max}$, where $e_{\min}$ and $e_{\max}$ are the extremal exponents of the considered floating-point format. A nonzero number without a normal representation is said *subnormal*: all subnormal numbers have absolute value less than $\beta^{e_{\min}}$ and exponent equal to $e_{\min}$.

Assuming we are working with a radix-β, precision-p floating-point arithmetic, a *midpoint* is a rational number of the form

$$z = \pm\,(Z + 1/2) \cdot \beta^{e_z - p + 1},$$

where Z is a nonnegative integer such that

$$\begin{cases} \beta^{p-1} \leq Z < \beta^p, & \text{if } e_{\min} < e_z \leq e_{\max}, \\ 0 \leq Z < \beta^p, & \text{if } e_z = e_{\min}. \end{cases}$$

Such a number is exactly halfway between two consecutive floating-point numbers. The midpoints are the values where the function $x \mapsto \mathrm{RN}(x)$ is discontinuous, as illustrated by Figure 1 on a toy floating-point format ($\beta = 2$, $p = 2$, $e_{\min} = 0$, $e_{\max} = 2$).

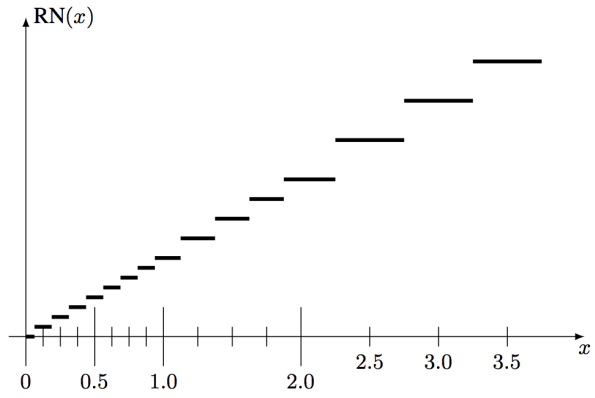

Fig. 1. The $\mathrm{RN}(x)$ function (radix $\beta = 2$, precision $p = 2$)

When using the implementation of a mathematical function in floating-point arithmetic, in most practical cases, the input and output precisions are the same. However, a user may for example wish to calculate the single-precision/binary32 number that is closest to the square-root of a double-precision/binary64 floating-point number. For the sake of simplicity, we assume in this paper that the input and output precisions are the same. Moreover, we give our results assuming an unbounded exponent range, that is, under the hypothesis that no underflow nor overflow occurs. For that purpose, we define $\mathbb{F}_{\beta,p}$ as the set of the radix-β, precision-p floating-point numbers, with an unbounded exponent range. Similarly, midpoints are restricted to the set

$$\mathbb{M}_{\beta,p} = \big\{ \pm\,(Z + 1/2) \cdot \beta^{e_z - p + 1} \;\big|\; Z \in \mathbb{N},\, \beta^{p-1} \leq Z < \beta^p,\, e_z \in \mathbb{Z} \big\},$$

where $\mathbb{Z}$ denotes the set of integers, and $\mathbb{N}$ denotes the set $\{0, 1, 2, \ldots\}$ of nonnegative integers.

The purpose of this paper is, for the floating-point number systems and the algebraic functions mentioned above, to investigate whether these functions admit midpoints or exact points, and to characterize such midpoints and exact points when they exist. The results we obtain are for $\beta = 2^q$ with q a positive integer, and for $\beta = 10$, but in some cases, we managed to weaken these assumptions on β. Moreover, most of the examples proposed are based on the basic formats defined in the IEEE 754-2008 standard [5], that are briefly recalled below:

Binary formats	p	Decimal formats	p
binary32	24	decimal32	7
binary64	53	decimal64	16
binary128	113	decimal128	34

Table 1 summarizes the results presented in the paper. In this table, "many" indicates that the techniques we used did not allow us to find a simple characterization of the midpoints or of the exact points of the function, that an exhaustive enumeration was impractical because of the too large number of cases to consider, and that we have experimental evidence that the number of midpoints and/or exact points is large. Most of the results displayed here for $\beta = 2$ are in fact obtained in a more general setting, namely for $\beta = 2^q$, q a positive integer.

Notice that since we considered an unbounded exponent range, subnormal floating-point numbers of the various IEEE 754-2008 formats can be written in normalized form. Hence, subnormal numbers are a subset of floating-point numbers with unbounded exponent range. This implies that the results presented in Table 1 remain unchanged when the inputs are subnormal numbers. If there are no exact points or midpoints for normal floating-point numbers with unbounded exponent range for a given function, then midpoints or exact points cannot occur if the inputs are subnormals. Similarly, if the exact points and midpoints are characterized by one of the theorems, assuming the inputs are subnormals will only restrict the characterization of the theorem, without creating new possible exact points or midpoints.

However, some results presented in Table 1 change when we want to know if a given function outputs midpoints in the range of subnormal floating-point numbers. In radix 2, division admits midpoints in the subnormal range, as well as the function $x/\|y\|_2$, while they have no midpoints in the normal range. The square-root function admits no midpoints, ever in the subnormal range, for the square-root of a floating-point number cannot be in the subnormal range. Although the results are not detailed in the sequel, the techniques presented in this paper can be used to deal with midpoints in the subnormal range for the other functions listed in Table 1.

Outline. We start with extensions to radices 2^q and 10 of classical, radix-2 results for square roots (Section 2), reciprocal square roots (Section 3), and positive integer powers (Section 4). In Section 5 we move to the function that maps a real x and a d-dimensional real vector $y = [y_k]_{1 \leq k \leq d}$ to $x/\|y\|_2$. Here $\|\cdot\|_2$ denotes the Euclidean norm of

TABLE 1
Summary of the results given in this paper.

	Midpoints		Exact points	
Function	Radix 2	Radix 10	Radix 2	Radix 10
$\sqrt{y}$	none	none	many	many
$1/\sqrt{y}$	none	Theorem 3	$y = 2^{2k}$	Theorem 5
x^k for $k \in \mathbb{N}_{>0}$	Theorem 6	Theorem 6	Theorem 6	Theorem 6
$x/\|y\|_2$	none	many	many	many
x/y	none	many	many	many
$1/y$	none	Theorem 8	$y = \pm 2^k$	Theorem 9
$\frac{1}{\sqrt{x^2+y^2}}$	none	Theorem 12	$\{x, y\} = \{0, \pm 2^k\}$	Theorem 13
$\frac{x}{\sqrt{x^2+y^2}}$	none	none	$x = 0$ or $y = 0$	many
$\sqrt{x^2+y^2}$	many	many	many	many

vectors: $\|y\|_2 = \sqrt{y_1^2 + \cdots + y_d^2}$. The function $x/\|y\|_2$ is interesting for it covers several important special cases, each of them being detailed in a subsequent section: for $d = 1$, division and reciprocal (Sections 6 and 7); for $d = 2$, reciprocal two-dimensional Euclidean norm $1/\sqrt{x^2+y^2}$ and normalization of two-dimensional vectors $x/\sqrt{x^2+y^2}$ (Sections 8 and 9). We comment on the two-dimensional Euclidean norm in Section 10.

Notation. Throughout the paper, the symbols $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{N}_{>0}$ denote the rational numbers, the real numbers, and the positive integers, respectively. We write i for the complex number whose square is -1, and $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ for the usual floor and ceiling functions. Also, for $x, y \in \mathbb{Z}$ such that $y \neq 0$, we use the standard notation $x \bmod y = x - y\lfloor x/y \rfloor$ (see for instance Graham, Knuth, and Patashnik [4, p. 82]).

2 Square root

2.1 Midpoints for square root

The following theorem can be viewed as a consequence of a result of Markstein [9, Theorem 9.4]. It says that the square root function has no midpoints, whatever the radix β is. A detailed proof is given here for completeness.

Theorem 1 (Markstein [9]): Let $y \in \mathbb{F}_{\beta,p}$ be positive. Then $\sqrt{y} \notin \mathbb{M}_{\beta,p}$.

Proof: Let $z = \sqrt{y}$ and assume that z is in $\mathbb{M}_{\beta,p}$. Then there exist some integers Z and e_z such that $z = (Z + 1/2) \cdot \beta^{e_z - p + 1}$ and $\beta^{p-1} \leq Z < \beta^p$. Using $y = z^2$ and $y = Y \cdot \beta^{e_y - p + 1}$, we deduce that

$$4Y \cdot \beta^{e_y - 2e_z + p - 1} = (2Z + 1)^2. \tag{1}$$

Now, one may check that $e_z = \lfloor e_y/2 \rfloor$, so that

$$e_y - 2e_z = e_y \bmod 2, \tag{2}$$

which is non-negative. Thus, for $p \geq 1$, the left-hand side of (1) is an even integer. This contradicts the fact that the right-hand side is an odd integer. □

2.2 Exact points for square root

We saw in the previous section that the square root function has no midpoints. The situation for exact points is just opposite: for a given input exponent, the number N of floating-point numbers having this exponent and whose square root is also a floating-point number grows essentially like $\beta^{p/2}$. In this section, we make this claim precise for $\beta = 2^q$ ($q \in \mathbb{N}_{>0}$) and $\beta = 10$ by giving an explicit expression for N in Theorem 2. To establish this counting formula, we need the following two lemmata.

Lemma 1: For $a, b \in \mathbb{R}$ such that $0 \leq a \leq b$, and $c \in \mathbb{N}_{>0}$, the number of integer multiples of c that lie in $[a, b)$ is $\lceil b/c \rceil - \lceil a/c \rceil$.

Proof: Let us write $N_{a,b}^{(c)}$ for the number of integer multiples of c lying in $[a, b)$. Since $0 \leq a \leq b$, the set $[0, b)$ is the union of the disjoint sets $[0, a)$ and $[a, b)$. Hence $N_{a,b}^{(c)} = N_{0,b}^{(c)} - N_{0,a}^{(c)}$ and it remains to check that $N_{0,a}^{(c)} = \lceil a/c \rceil$. If $a \notin \mathbb{N}$, it follows from $c \in \mathbb{N}_{>0}$ that $N_{0,a}^{(c)} = 1 + \lfloor a/c \rfloor$. If $a \in \mathbb{N}$, either c divides a in which case $N_{0,a}^{(c)} = a/c$, otherwise $N_{0,a}^{(c)} = 1 + \lfloor a/c \rfloor$. □

Lemma 2: Let y be a positive number in $\mathbb{F}_{\beta,p}$. The real number $\sqrt{y}$ is also in $\mathbb{F}_{\beta,p}$ if and only if the integral significand Y of y satisfies $\beta^{p-1} \leq Y < \beta^p$ and $Y = Z^2 \cdot \beta^{1-p-(e_y \bmod 2)}$ for some integer Z such that $\beta^{p-1} \leq Z < \beta^p$.

Proof: Let $z = \sqrt{y}$. Assume first that $z \in \mathbb{F}_{\beta,p}$. Then there exists an integer Z such that $z = Z \cdot \beta^{e_z - p + 1}$ and $\beta^{p-1} \leq Z < \beta^p$. Using $y = z^2$ and $y = Y \cdot \beta^{e_y - p + 1}$, we deduce that

$$Y = Z^2 \cdot \beta^{1-p-(e_y - 2e_z)}. \tag{3}$$

The "only if" statement then follows from (2). Conversely, using $y = Y \cdot \beta^{e_y - p + 1}$, we may rewrite the equality $Y = Z^2 \cdot \beta^{1-p-(e_y \bmod 2)}$ as

$$\sqrt{y} = Z \cdot \beta^{e_z - p + 1}, \quad \text{where } e_z = (e_y - (e_y \bmod 2))/2.$$

By definition, e_z is an integer and, by assumption, Z is an integer lying in $[\beta^{p-1}, \beta^p)$. Therefore, $\sqrt{y}$ belongs to $\mathbb{F}_{\beta,p}$. □

Theorem 2: For a given exponent e_y, let N denote the number of positive values $y \in \mathbb{F}_{\beta,p}$ such that $\sqrt{y} \in \mathbb{F}_{\beta,p}$, and let $\epsilon_y = (e_y + p - 1) \bmod 2$.

- If $\beta = 2^q$, $q \in \mathbb{N}_{>0}$, then

$$N = \left\lceil 2^{(qp - \epsilon_y (q \bmod 2))/2} \right\rceil - \left\lceil 2^{(q(p-1) - \epsilon_y (q \bmod 2))/2} \right\rceil.$$

- If $\beta = 10$, then $N = \lceil 10^{(p - \epsilon_y)/2} \rceil - \lceil 10^{(p-1-\epsilon_y)/2} \rceil$.

Proof: Let $\gamma = p - 1 + (e_y \bmod 2)$. From Lemma 2, N is the number of integers Y in $[\beta^{p-1}, \beta^p)$ and of the form $Z^2 \cdot \beta^{-\gamma}$ for some integer Z such that $\beta^{p-1} \leq Z < \beta^p$.

Rewriting $Y = Z^2 \cdot \beta^{-\gamma}$ as $Y \cdot \beta^{\epsilon_y} \cdot \beta^{\gamma - \epsilon_y} = Z^2$, we see that $\beta^{\gamma - \epsilon_y}$ divides Z^2. Since $\epsilon_y = \gamma \bmod 2$, we know that $\gamma - \epsilon_y$ is even and, for $p \geq 1$, nonnegative. Using for instance the factorizations of $\beta^{(\gamma - \epsilon_y)/2}$ and Z into primes, we deduce that $\beta^{(\gamma - \epsilon_y)/2}$ divides Z. Consequently, there exists an integer X such that

$$Y \cdot \beta^{\epsilon_y} = X^2 \qquad \text{and} \qquad Z = X \cdot \beta^{(\gamma - \epsilon_y)/2}.$$

Now, the assumption $\beta^{p-1} \le Y < \beta^p$ is equivalent to

$$\beta^{(p-1+\epsilon_y)/2} \le X < \beta^{(p+\epsilon_y)/2}, \tag{4}$$

while the same assumption on Z is equivalent to $\beta^{p-1-(\gamma-\epsilon_y)/2} \le X < \beta^{p-(\gamma-\epsilon_y)/2}$. The latter interval contains the former because $p-1 \le \delta \le p$. Hence, N is the number of integers X satisfying (4) and whose square is an integer multiple of β^{ϵ_y}. We distinguish between the two cases $\epsilon_y = 0$ and $\epsilon_y = 1$.

If $\epsilon_y = 0$ then N is the number of integers X satisfying (4). Consequently, $N = \lceil \beta^{p/2} \rceil - \lceil \beta^{(p-1)/2} \rceil$ (using either Lemma 1 with $c = 1$, or [4, (3.12)]).

If $\epsilon_y = 1$ then X^2 is a multiple of β: When β has linear factors only (like $\beta = 2$ or $\beta = 10 = 2 \cdot 5$), this implies that X is a multiple of β. In this case, N is the number of integers X that are multiples of β and satisfy $\beta^{p/2} \le X < \beta^{(p+1)/2}$. Hence, using Lemma 1, $N = \lceil \beta^{(p-1)/2} \rceil - \lceil \beta^{(p-2)/2} \rceil$. Assume now that $\beta = 2^q$ for some positive integer q. If q is even then 2^q divides X^2 implies $2^{q/2}$ divides X, so that we take the number of X's being an integer multiple of $2^{q/2}$. Lemma 1 thus gives $N = \lceil 2^{qp/2} \rceil - \lceil 2^{q(p-1)/2} \rceil$. If q is odd then $Y \cdot 2 = (X \cdot 2^{-\lfloor q/2 \rfloor})^2$, which means that $X \cdot 2^{-\lfloor q/2 \rfloor}$ is even. Hence we keep all the X's that are an integer multiple of $2^{1+\lfloor q/2 \rfloor}$. Using Lemma 1, this gives $N = \lceil 2^{(qp-1)/2} \rceil - \lceil 2^{(q(p-1)-1)/2} \rceil$. □

For a fixed e_y, using Theorem 2, one can count the number of input floating-point numbers y whose square root is an exact point. We give below the number N of exact points for the basic formats of the IEEE 754-2008 standard.

Format	binary16	binary32	binary64	binary128
p	11	24	53	113
$\epsilon_y = 0$	14	1199	$\approx 2.78 \cdot 10^7$	$\approx 2.98 \cdot 10^{16}$
$\epsilon_y = 1$	9	849	$\approx 1.97 \cdot 10^7$	$\approx 2.11 \cdot 10^{16}$

Format	decimal32	decimal64	decimal128
p	7	16	34
$\epsilon_y = 0$	2163	$\approx 6.84 \cdot 10^7$	$\approx 6.84 \cdot 10^{16}$
$\epsilon_y = 1$	683	$\approx 2.16 \cdot 10^7$	$\approx 2.16 \cdot 10^{16}$

Also, for a fixed exponent e_y, one can see from Theorem 2 that the number of exact points for the square root function is $\Theta(2^{qp/2})$ when $\beta = 2^q$, and $\Theta(10^{p/2})$ when $\beta = 10$ (see for instance Graham, Knuth, and Patashnik [4, p. 448] for a precise definition of the Θ notation). Except for small precisions, Theorem 2 implies therefore that it can be regarded as impractical to enumerate the exact points for the square root. It also shows that when computing the square root of a floating-point number, the probability of that square root being an exact point is very small (it vanishes as p increases). This property may be taken into account when tuning a square-root algorithm.

3 Reciprocal Square Root

3.1 Midpoints for reciprocal square root

Theorem 3: Let $y \in \mathbb{F}_{\beta,p}$ be positive and let δ_y denote $e_y \bmod 2$. If $\beta = 2^q$ ($q \in \mathbb{N}_{>0}$) then $1/\sqrt{y} \notin \mathbb{M}_{\beta,p}$. If $\beta = 10$, one has $1/\sqrt{y} \in \mathbb{M}_{\beta,p}$ if and only if the integral significand Y of y has the form

$$Y = 2^{3p-\delta_y+1} \cdot 5^{3p-2\ell-\delta_y-1},$$

with $\ell \in \mathbb{N}$ such that $\ell \le (3p-\delta_y-1)/2$ and

$$\begin{cases} 2 \cdot 10^{p-1} < 5^\ell < 2 \cdot 10^{p-1/2}, & \text{if } e_y \text{ is odd,} \\ 2 \cdot 10^{p-1/2} < 5^\ell < 2 \cdot 10^p, & \text{if } e_y \text{ is even.} \end{cases} \tag{5}$$

Proof: Let $z = 1/\sqrt{y}$ and assume $z \in \mathbb{M}_{\beta,p}$. Let $y = Y \cdot \beta^{e_y-p+1}$ and $z = (Z+1/2) \cdot \beta^{e_z-p+1}$ be the normalized representations of y and z. From $yz^2 = 1$ we deduce that

$$Y(2Z+1)^2 = 4 \cdot \beta^{-e_y-2e_z+3p-3}. \tag{6}$$

Since z is a midpoint, one has $\beta^{e_z} < z < \beta^{e_z+1}$ and so $\beta^{-2e_z-2} < y < \beta^{-2e_z}$. From this, one may check that

$$-e_y - 2e_z = 2 - \delta_y, \qquad \delta_y = e_y \bmod 2. \tag{7}$$

Hence we obtain from Equations (6) and (7)

$$Y(2Z+1)^2 = 4 \cdot \beta^{3p-\delta_y-1}. \tag{8}$$

When $\beta = 2^q$, Equation (8) has no solution, since the right-hand side of the equality is a power of two while the left-hand side has an odd factor $(2Z+1)^2$.

Let us now consider the case where $\beta = 10$. Equation (8) then becomes

$$Y(2Z+1)^2 = 2^{3p-\delta_y+1} \cdot 5^{3p-\delta_y-1}. \tag{9}$$

Since $2Z+1$ is odd, we deduce from (9) that $2Z+1 = 5^\ell$ for some $\ell \in \mathbb{N}$. Hence

$$Y = 2^{3p-\delta_y+1} \cdot 5^{3p-2\ell-\delta_y-1}$$

and it remains to prove the bounds on ℓ. Since Y is an integer, we have $3p-2\ell-\delta_y-1 \ge 0$, and the first bound $\ell \le (3p-\delta_y-1)/2$ follows. To prove the bounds in (5), note first that $10^{e_y} \le y < 10^{e_y+1}$ and (7) give $10^{e_z+(1-\delta_y)/2} < z = 1/\sqrt{y} \le 10^{e_z+1-\delta_y/2}$. Then, using $z = (Z+1/2) \cdot 10^{e_z-p+1}$, we obtain

$$2 \cdot 10^{p-(\delta_y+1)/2} < 2Z+1 = 5^\ell \le 2 \cdot 10^{p-\delta_y/2}.$$

In fact, the upper bound is strict, for 5^ℓ is an odd integer while $2 \cdot 10^{p-\delta_y/2}$ is either an even integer ($\delta_y = 0$) or an irrational number ($\delta_y = 1$). Conversely, let $Y = 2^{3p-\delta_y+1} \cdot 5^{3p-2\ell-\delta_y-1}$, with ℓ as in (5), and let $z = 1/\sqrt{y}$. From (8) we deduce that $y = 2^{2p-2e_z} \cdot 5^{2p-2\ell-2e_z-2}$ and $z = \big((5^\ell - 1)/2 + 1/2\big) \cdot 10^{1-p+e_z}$. Now $2 \cdot 10^{p-1} < 5^\ell < 2 \cdot 10^p$ implies $10^{p-1} \le (5^\ell - 1)/2 < 10^p$ and thus $z \in \mathbb{M}_{10,p}$. □

To find in radix 10 the significands Y of all the inputs y such that $1/\sqrt{y}$ is a midpoint, it suffices to find the at most two $\ell \in \mathbb{N}$ such that $2 \cdot 10^{p-1} < 5^\ell < 2 \cdot 10^p$, and to determine from the bounds (5) whether e_y is even or odd. Table 2 gives the integral significands Y and the parity of the exponent e_y such that $z = 1/\sqrt{y}$ is a midpoint in the basic decimal formats of IEEE 754-2008.

Notice that for radices different from 10 or a power of 2, we do not have general results (which is in contrast with square root; see Section 2.1). Equation (8) may have solutions; for example, in radix 3 with $p = 6$, one may check that $(Y, Z, \delta_y) = (324, 364, 1)$ satisfies Equation (8), and gives a midpoint for the reciprocal square root.

TABLE 2
Integral significands Y of $y \in \mathbb{F}_{10,p}$ such that $1/\sqrt{y} \in \mathbb{M}_{10,p}$, for the decimal formats of the IEEE 754-2008 standard [5].

Format	Integral significand Y	e_y
decimal32 ($p = 7$)	$2^{22} \cdot 5^0 = 4194304$	even
decimal64 ($p = 16$)	$2^{48} \cdot 5^2 = 7036874417766400$	odd
	$2^{49} \cdot 5^1 = 2814749767106560$	even
decimal128 ($p = 34$)	$2^{102} \cdot 5^4 = 3169126500570573503741758013440000$	odd
	$2^{103} \cdot 5^3 = 1267650600228229401496703205376000$	even

3.2 Exact points for reciprocal square root

The following theorem gives a characterization of the exact points of the square-root reciprocal when the radix is a prime number (which includes the most frequent case $\beta = 2$) and also when the radix is a positive integer power of 2. The case $\beta = 10$ is treated separately in Theorem 5.

Theorem 4: Let $y \in \mathbb{F}_{\beta,p}$ be positive. Then
• for β a prime number, one has $1/\sqrt{y} \in \mathbb{F}_{\beta,p}$ if and only if $y = \beta^{2k}$ with $k \in \mathbb{Z}$;
• for $\beta = 2^q$ ($q \in \mathbb{N}_{>0}$), one has $1/\sqrt{y} \in \mathbb{F}_{\beta,p}$ if and only if $y = 2^{2k}$ with $k \in \mathbb{Z}$.

Proof: Taking $z = 1/\sqrt{y}$, note first that (7) still holds. Now assume that $z \in \mathbb{F}_{\beta,p}$ and let Y and Z be the integral significands of y and z. From $yz^2 = 1$ and (7), we deduce

$$YZ^2 = \beta^{3p-\delta_y-1}. \tag{10}$$

If β is prime, we deduce from (10) that $Z = \beta^\ell$ for some $\ell \in \mathbb{N}$. Hence $Y = \beta^{3p-\delta_y-1-2\ell}$ and, using (7), $y = \beta^{2(p-1-e_z-\ell)}$ is indeed an even power of β. Conversely, if $y = \beta^{2k}$, then $z = \beta^{-k}$ is in $\mathbb{F}_{\beta,p}$.

If $\beta = 2^q$ with $q \in \mathbb{N}_{>0}$, we deduce from (10) that $Z = 2^\ell$ for some $\ell \in \mathbb{Z}$ and, similarly to the previous case, we find $y = 2^{2(q(p-1-e_z)-\ell)}$, which is an even power of two. Conversely, if $y = 2^{2k}$, then $z = 2^{-k}$. Since any integral power of 2 is representable in $\mathbb{F}_{2^q,p}$, we conclude that z is an exact point. □

All the floating-point numbers y such that $1/\sqrt{y}$ is an exact point can be deduced from the ones lying in the interval $[1, \beta^2)$. In radix 2^q, Theorem 4 implies that at most q values of y in $[1, 2^{2q})$ suffice to characterize the exact points for the reciprocal square root. In radix $16 = 2^4$ for instance, the only exact points for input values $y \in [1, 256)$ are:

y	1	4	16	64
$1/\sqrt{y}$	1	$1/2 = 0.8_{16}$	$1/4 = 0.4_{16}$	$1/8 = 0.2_{16}$

Theorem 5: Let $y \in \mathbb{F}_{10,p}$ be positive and let δ_y denote $e_y \bmod 2$. One has $1/\sqrt{y} \in \mathbb{F}_{10,p}$ if and only if either $y = 10^{-2e_z}$ or the integral significand Y of y differs from 10^{p-1} and has the form

$$Y = 2^{3p-1-\delta_y-2k} \cdot 5^{3p-1-\delta_y-2\ell},$$

with $k, \ell \in \mathbb{N}$ such that $0 \le k, \ell \le (3p - 1 - \delta_y)/2$.

Proof: Let $z = 1/\sqrt{y}$ and assume $z \in \mathbb{F}_{10,p}$. If $z = 10^{e_z}$ then obviously $y = 10^{-2e_z}$. On the other hand, z must differ from the irrational number $10^{e_z+1/2}$. Hence we now assume $z \in (10^{e_z}, 10^{e_z+1/2}) \cup (10^{e_z+1/2}, 10^{e_z+1})$. This implies $y \in (10^{-2e_z-2}, 10^{-2e_z-1}) \cup (10^{-2e_z-1}, 10^{-2e_z})$. Therefore, y is not a power of 10 and its normalized representation $y = Y \cdot 10^{e_y-p+1}$ is such that $Y \neq 10^{p-1}$. Note now that (7) and (10) still hold here, so that $yz^2 = 1$ implies $YZ^2 = 10^{3p-1-\delta_y}$. In particular, Z must have the form $Z = 2^k \cdot 5^\ell$ for some k, ℓ in $\mathbb{N}$. Thus

$$Y = 2^{3p-1-\delta_y-2k} \cdot 5^{3p-1-\delta_y-2\ell},$$

where, since Y is an integer, $0 \le k, \ell \le (3p - 1 - \delta_y)/2$.

Conversely, the case $y = 10^{-2e_z}$ being straightforward, let $Y = 2^{3p-1-\delta_y-2k} \cdot 5^{3p-1-\delta_y-2\ell}$ be the integral significand of y such that $10^{p-1} < Y < 10^p$, and let $z = 1/\sqrt{y}$. Using (7) further leads to $z = 2^k \cdot 5^\ell \cdot 10^{e_z-p+1}$. One has $2^k \cdot 5^\ell \in \mathbb{N}$ and, from $10^{p-1} < Y < 10^p$, we get $10^{p-(1+\delta_y)/2} < 2^k \cdot 5^\ell < 10^{p-\delta_y/2}$. Hence $z \in \mathbb{F}_{10,p}$. □

Enumerating the integral significands $Y = 2^{3p-1-\delta_y-2k} \cdot 5^{3p-1-\delta_y-2\ell}$ with $k, \ell \in \mathbb{N}$ such that $0 \le k, \ell \le (3p - 1 - \delta_y)/2$ and $10^{p-1} < Y < 10^p$ is easily done by a simple program. Table 3 gives all the integral significands Y of y, and the parity of the exponent e_y, such that $1/\sqrt{y}$ is a floating-point number too, in the decimal32 format (see also Table 8 in the appendix for the decimal64 format).

For the basic decimal formats of the IEEE 754-2008, the table below gives the number of significands Y such that $1/\sqrt{y}$ is an exact point, with respect to the parity δ_y of the exponent of y.

Format p	decimal32 7	decimal64 16	decimal128 34
$\delta_y = 0$	9	17	37
$\delta_y = 1$	7	17	36

TABLE 3
Integral significands Y of $y \in \mathbb{F}_{10,7}$, such that $1/\sqrt{y} \in \mathbb{F}_{10,7}$.

Y	$1/\sqrt{Y \cdot 10^{\delta_y-p+1}}$	e_y
$2^6 \cdot 5^6 = 1000000$	$1.000000 \cdot 10^0$	even
$2^{20} \cdot 5^0 = 1048576$	$9.765625 \cdot 10^{-1}$	even
$2^{18} \cdot 5^2 = 6553600$	$3.906250 \cdot 10^{-1}$	even
$2^{16} \cdot 5^2 = 1638400$	$7.812500 \cdot 10^{-1}$	even
$2^{12} \cdot 5^4 = 2560000$	$6.250000 \cdot 10^{-1}$	even
$2^8 \cdot 5^6 = 4000000$	$5.000000 \cdot 10^{-1}$	even
$2^4 \cdot 5^8 = 6250000$	$4.000000 \cdot 10^{-1}$	even
$2^2 \cdot 5^8 = 1562500$	$8.000000 \cdot 10^{-1}$	even
$2^0 \cdot 5^{10} = 9765625$	$3.200000 \cdot 10^{-1}$	even
$2^{19} \cdot 5^1 = 2621440$	$1.953125 \cdot 10^{-1}$	odd
$2^{15} \cdot 5^3 = 4096000$	$1.562500 \cdot 10^{-1}$	odd
$2^{13} \cdot 5^3 = 1024000$	$3.125000 \cdot 10^{-1}$	odd
$2^{11} \cdot 5^5 = 6400000$	$1.250000 \cdot 10^{-1}$	odd
$2^9 \cdot 5^5 = 1600000$	$2.500000 \cdot 10^{-1}$	odd
$2^5 \cdot 5^7 = 2500000$	$2.000000 \cdot 10^{-1}$	odd
$2^1 \cdot 5^9 = 3906250$	$1.600000 \cdot 10^{-1}$	odd

4 Positive integer powers

We consider here the function $(x, k) \mapsto x^k$ with $x \in \mathbb{R}$ and $k \in \mathbb{N}_{>0}$, assuming that each prime factor appears only

once in the prime decomposition of β, which is the case for $\beta = 2$ and $\beta = 10$. We provide a sufficient condition for the nonexistence of midpoints in such radices. In the particular case $\beta = 2$, the results given in this section can be deduced easily from Lauter and Lefèvre's study of the power function $(x, y) \mapsto x^y$ [8], which shows how to check quickly if x^y is a midpoint or an exact point, in double precision (binary64 format).

Definition 1: A number *fits in n digits exactly* in radix β if it is a precision-n floating-point number that cannot be exactly represented in precision $n-1$. More precisely, it is a number of the form $x = X \cdot \beta^{e_x}$, where $e_x, X \in \mathbb{Z}$, $\beta^{n-1} < |X| < \beta^n$, and X is not a multiple of β.

Lemma 3: Let $k \in \mathbb{N}_{>0}$ be given. If each factor of β appears only once in its prime number decomposition (which is true for β equal to 2 or 10), and if x fits in n digits exactly then x^k fits in m digits exactly, with $m \in \mathbb{N}$ such that $k(n-1) < m \leq kn$.

Proof: Let $x = X \cdot \beta^{e_x}$ be a number that fits in n digits exactly. From $\beta^{n-1} < |X| < \beta^n$ it follows that $\beta^{k(n-1)} < |X^k| < \beta^{kn}$. Consequently, there exists $m \in \mathbb{N}$ such that $k(n-1) < m \leq kn$ and $\beta^{m-1} < |X^k| < \beta^m$. Moreover, the assumption on the prime factor decomposition of β and the fact that β does not divide X imply that X^k is not a multiple of β. □

An immediate consequence of the previous lemma is the following result.

Theorem 6: Assume the radix β is such that each factor appears only once in its prime number decomposition, and let p be the precision. If x fits in n digits exactly then x^k cannot be a midpoint as soon as $k(n-1) > p$, and it cannot be an exact point as soon as $k(n-1)+1 > p$.

Theorem 6 is not helpful when k is small. For large values of k, however, it allows to quickly determine the possible midpoints and exact points. For instance, in the binary64 format ($\beta = 2$ and $p = 53$), the only floating-point numbers x such that x^{10} can be an exact point are those that fit in n bits exactly, where $n \leq 6$. For a given value of the exponent, there are at most $2^6 = 64$ such points: it therefore suffices to check these 64 values to know all the exact points. By accurately computing x^{10} for these 64 points, we easily find that the exact points for function x^{10} in the binary64 format correspond to the input values of the form $x = X \cdot 2^{e_x}$, where X is an integer between 0 and 40.

5 The function $(x, y) \mapsto x \,/\, \|y\|_2$

Given $d \in \mathbb{N}_{>0}$, the number of exact points of the function that maps $(x, y) \in \mathbb{R} \times (\mathbb{R}^d \backslash \{0\})$ to $x/\|y\|_2 = x/\sqrt{\sum_{1 \leq k \leq d} y_k^2}$ is huge. Indeed, all the exact points for the division operation, whose number is huge as we will see later in Section 6.2, are exact points for the function $x/\|y\|_2$ as well. Therefore, we shall focus here exclusively on midpoints: our aim is to decide whether there exist floating-point inputs $x, y_1, \ldots, y_d \in \mathbb{F}_{\beta,p}$ such that $x/\|y\|_2 \in \mathbb{M}_{\beta,p}$. We start with the following theorem, which says that midpoints cannot exist in radix 2.

Theorem 7: Let $x \in \mathbb{F}_{\beta,p}$ and, for $d \in \mathbb{N}_{>0}$, let y be a nonzero, d-dimensional vector of elements of $\mathbb{F}_{\beta,p}$. If $\beta = 2$ then $x/\|y\|_2 \notin \mathbb{M}_{\beta,p}$.

Proof: Because of the symmetries of the function that maps (x, y) to $x/\|y\|_2$, we can restrict to the case where x and all the entries of $y = [y_k]$ are positive. Hence $x = X \cdot \beta^{e_x - p + 1}$ and $y_k = Y_k \cdot \beta^{e_{y_k} - p + 1}$ for some integers X and Y_k such that $\beta^{p-1} \leq X, Y_k < \beta^p$. Let $z = x/\|y\|_2$ and assume z is a midpoint, that is, $z = (Z + 1/2) \cdot \beta^{e_z - p + 1}$ for some integer Z in the same range as X and the Y_k above. The identity $x^2 = \|y\|_2^2 \, z^2$ thus becomes

$$4X^2 \cdot \beta^{2(e_x - e_z + p - 1)} = \Big(\sum_k Y_k^2 \cdot \beta^{2e_{y_k}} \Big) (2Z+1)^2. \quad (11)$$

In order to have integers on both sides, it suffices to multiply (11) by β^{-2e^*}, where $e^* = \min_k e_{y_k}$. This gives

$$4X^2 \cdot \beta^{2(e_x - e_z - e^* + p - 1)} = \Big(\sum_k Y_k^2 \cdot \beta^{2(e_{y_k} - e^*)} \Big) (2Z+1)^2. \quad (12)$$

Now, the power of β involved in the left-hand side of (12) is itself an integer. This is is due to the fact that the integer $e_x - e_z - e^*$ is non-negative, which can be seen as follows. Since $d \geq 1$ and $y_k \geq \beta^{e^*}$ for $k = 1, \ldots, d$, one has $z \leq x/\beta^{e^*}$. Using $x < \beta^{e_x + 1}$ and $\beta^{e_z} \leq z$ (in fact this lower bound is strict, for z is a midpoint), we deduce that $\beta^{e_z} < \beta^{e_x - e^* + 1}$. The exponents on both sides of the latter inequality being integers, we conclude that $e_z \leq e_x - e^*$. When $\beta = 2$, Equation (12) becomes

$$X^2 \cdot 2^{2(e_x - e_z - e^* + p)} = \Big(\sum_k Y_k^2 \cdot 2^{2(e_{y_k} - e^*)} \Big) (2Z+1)^2. \quad (13)$$

The left-hand side of (13) is a multiple of the odd integer $(2Z+1)^2$. Since $e_x - e_z - e^*$ is non-negative, this implies that X is a multiple of $2Z+1$ and thus $X \geq 2Z+1$. However, recalling that $2^{p-1} \leq X, Z < 2^p$, we have

$$X < 2Z + 1. \quad (14)$$

Hence a contradiction, which concludes the proof. □

Theorem 7 implies the non-existence of midpoints in radix $\beta = 2$ for a number of important special cases: division x/y (see Corollary 1) and thus reciprocal $1/y$ as well; reciprocal 2D Euclidean norm $1/\sqrt{x^2 + y^2}$ and 2D-vector normalization $x/\sqrt{x^2 + y^2}$.

However, when $\beta > 2$, the function $x/\|y\|_2$ does have midpoints and some examples will be given in Section 6.1 for $\beta \in \{3, 4, 10\}$. Thus, rather than trying to characterize all the midpoints of that general function, we focus from Section 6 to Section 9 on the four special cases just mentioned.

6 Division

6.1 Midpoints for division

Concerning midpoints for division, Theorem 7 gives an answer for the far most frequent case in practice: the radix is 2, the input precision equals the output precision, and the results are above the underflow threshold. Indeed, choosing $d = 1$ in Theorem 7, we obtain the following corollary.

Corollary 1: In binary arithmetic, the quotient of two floating-point numbers cannot be a midpoint in the same precision.

In radix-2 floating-point arithmetic, Corollary 1 can be seen as a consequence of a result presented by Markstein in [9, Theorem 8.4, p. 114]. Note that this result only holds when $\beta = 2$ and when the input precision is less than or equal to the output precision. Nevertheless, it is sometimes believed that it holds in prime radices: the first example given below shows that this is not the case. The following examples also illustrate the existence of midpoints when $\beta > 2$.

- In radix 3, with precision $p = 4$,
$$\frac{28_{10}}{56_{10}} = \frac{1001_3}{2002_3} = 0.1111_3 + \frac{1}{2} \cdot 3^{-4}.$$
- In radix 4, with $p = 4$
$$\frac{129_{10}}{128_{10}} = \frac{2001_4}{2000_4} = 1.000_4 + \frac{1}{2} \cdot 4^{-3}.$$
- In radix 10, midpoint quotients are quite frequent. For instance, with $p = 2$ we have 181 midpoints for X/Y with $10 \leq X, Y \leq 99$ (e.g., $10/16 = 0.625$), and with $p = 3$, we have 2633 cases with $100 \leq X, Y \leq 999$.

We now briefly discuss the case of different input (p_i) and output (p_o) precisions. If $p_i > p_o$, many quotients can be midpoints, even in radix-2 arithmetic. For example, we can compute the quotient $x/1$ in precision p_o. Since x is in precision $p_i > p_o$, x can be a midpoint in precision p_o. It is also possible to find less trivial cases. For example, if x and y are binary64 numbers ($p_i = 53$) with

$$\begin{aligned} x &= 1.00000000000000000000000 \\ &\quad 11111111111111111111110100000, \\ y &= 1.11111111111111111111 \\ &\quad 00000000000000000000000000000, \end{aligned}$$

then one has

$$x/y = 0.\underbrace{100000000000000000000001}_{p_o=24}1,$$

which is a midpoint in the binary32 floating-point format ($p_o = 24$).

6.2 Exact points for division

Let x and y be two numbers in $\mathbb{F}_{\beta,p}$, and assume that the quotient $z = x/y$ is also in $\mathbb{F}_{\beta,p}$. Using the normalized representations $x = X \cdot \beta^{e_x-p+1}$, $y = Y \cdot \beta^{e_y-p+1}$ then z can be written $z = Z \cdot \beta^{e_x-e_y+\delta-p}$, with $\delta \in \{0, 1\}$. Hence from $x = yz$ it follows that

$$\beta^{p-\delta} X = YZ, \tag{15}$$

with $\delta \in \{0, 1\}$. In other words, if z is an exact point then Equation (15) must be satisfied. For any radix β, Equation (15) has many solutions: for each value of X there is at least the straightforward solution $(X, Y) = (Z, \beta^{p-1})$, which corresponds to x/β^{e_y}. As a consequence, the number of exact points of the function $(x, y) \mapsto x/y$ grows at least like $\beta^{p-1}(\beta - 1)$ for any given exponents e_x, e_y. This is too large to enumerate all the exact points of division in practice.

7 Reciprocal

As we have seen above, except in radix 2, division admits many midpoints. Moreover, whatever the radix is, division also admits a lot of exact points. Consequently, we now focus on a special case, the reciprocal function $y \mapsto 1/y$, for which more useful results can be obtained.

7.1 Midpoints for reciprocal

Theorem 8: Let $y \in \mathbb{F}_{\beta,p}$ be nonzero. If $\beta = 2^q$ ($q \in \mathbb{N}_{>0}$) then $1/y \notin \mathbb{M}_{\beta,p}$. If $\beta = 10$, one has $1/y \in \mathbb{M}_{\beta,p}$ if and only if the integral significand Y of y has the form

$$Y = 2^{2p} \cdot 5^{2p-1-\ell}, \tag{16}$$

with $\ell \in \mathbb{N}$ such that $2 \cdot 10^{p-1} < 5^\ell < 2 \cdot 10^p$.

Proof: Without loss of generality, we assume $y > 0$. Let $z = 1/y$. First, one may check that

$$e_z = -e_y - 1. \tag{17}$$

Now, if $z \in \mathbb{M}_{\beta,p}$ then $z = (Z + 1/2) \cdot \beta^{e_z-p+1}$ for some integer Z such that $\beta^{p-1} \leq Z < \beta^p$. Using $yz = 1$ thus gives

$$Y(2Z + 1) = 2 \cdot \beta^{2p-1}. \tag{18}$$

When $\beta = 2^q$, Equation (18) has no solution, since the right-hand side of the equality is a power of two while the left-hand side has an odd factor $2Z + 1$. When $\beta = 10$, (18) becomes

$$Y(2Z + 1) = 2^{2p} \cdot 5^{2p-1}. \tag{19}$$

As $2Z + 1$ is odd, we deduce from (19) that $2Z + 1$ is a power of 5. Also, since $2 \cdot 10^{p-1} < 2Z + 1 < 2 \cdot 10^p$, there are at most two such powers of 5. Hence y is necessarily as in (16). Conversely, if $y = Y \cdot 10^{e_y-p+1}$ with Y as in (16) then, using (17), $y = 5^{-\ell-1} \cdot 10^{-e_z+p}$. It follows that z can be written $z = ((5^\ell - 1)/2 + 1/2) \cdot 10^{e_z-p+1}$. Since $(5^\ell - 1)/2$ is an integer, and by hypothesis $10^{p-1} \leq (5^\ell - 1)/2 < 10^p$, we deduce that $z \in \mathbb{M}_{10,p}$, which concludes the proof. □

In radix 10, there are at most two values of $\ell \in \mathbb{N}$ such that $2 \cdot 10^{p-1} < 5^\ell < 2 \cdot 10^p$. Therefore, to determine all inputs y that give a midpoint $1/y$ for a fixed exponent e_y, it suffices to find the at most two ℓ such that $2 \cdot 10^{p-1} < 5^\ell < 2 \cdot 10^p$. This is easily done, even when the precision p is large. Table 4 gives the integral significands Y of the floating-point numbers y such that $1/y$ is a midpoint, for the decimal formats of the IEEE 754-2008 standard [5].

7.2 Exact points for reciprocal

For radices either 10 or a positive power of 2, the exact points of the reciprocal function can all be enumerated according to the following theorem.

TABLE 4
Integral significands Y of $y \in \mathbb{F}_{10,p}$ such that $1/y \in \mathbb{M}_{10,p}$, for the decimal formats of the IEEE 754-2008 standard [5].

Format	Integral significand Y
decimal32 ($p = 7$)	$2^{14} \cdot 5^3 = 2048000$
decimal64 ($p = 16$)	$2^{32} \cdot 5^8 = 1677721600000000$ $2^{32} \cdot 5^9 = 8388608000000000$
decimal128 ($p = 34$)	$2^{68} \cdot 5^{18} = 1125899906842624000000000000000000$ $2^{68} \cdot 5^{19} = 5629499534213120000000000000000000$

Theorem 9: Let $y \in \mathbb{F}_{\beta,p}$ be nonzero. One has $1/y \in \mathbb{F}_{\beta,p}$ if and only if the integral significand Y of y satisfies $\beta^{p-1} \leq Y < \beta^p$ and

$$Y = \begin{cases} 2^k, & 0 \leq k \leq q(2p-1), \quad \text{if } \beta = 2^q,\ q \in \mathbb{N}_{>0}; \\ 2^k \cdot 5^\ell, & 0 \leq k, \ell \leq 2p-1, \quad \text{if } \beta = 10. \end{cases}$$

Proof: For the "only if" statement, let $y > 0$ in $\mathbb{F}_{\beta,p}$ be given, let $z = 1/y$, and assume that $z \in \mathbb{F}_{\beta,p}$. First, one may check that the exponent of z satisfies $e_z = -e_y - \delta$ with $\delta \in \{0, 1\}$. Then, using the identity $yz = 1$ together with the normalized representations $y = Y \cdot \beta^{e_y - p + 1}$ and $z = Z \cdot \beta^{e_z - p + 1}$, we get

$$YZ = \beta^{2p-2+\delta}, \qquad \beta^{p-1} \leq Y, Z < \beta^p. \tag{20}$$

If $\beta = 2^q$ for some integer $q \geq 1$ then (20) implies that $Y = 2^k$ for some integer k such that $0 \leq k \leq q(2p-1)$. If $\beta = 10$ then (20) implies that $Y = 2^k \cdot 5^\ell$ for some integers k and ℓ such that $0 \leq k, \ell \leq 2p-1$.

Let us now prove the "if" statement. If $Y = \beta^{p-1}$ then y is a power of the radix and thus $1/y$ belongs to $\mathbb{F}_{\beta,p}$. If $\beta^{p-1} < Y < \beta^p$ then, defining $Z = Y^{-1} \cdot \beta^{2p-1}$, we obtain

$$1/y = Z \cdot \beta^{-e_y - p}, \qquad \beta^{p-1} < Z < \beta^p. \tag{21}$$

To conclude that $1/y$ belongs to $\mathbb{F}_{\beta,p}$ it remains to show that Z is an integer: If $\beta = 2^q$ and $Y = 2^k$, one has $Z = 2^{q(2p-1)-k}$, which is an integer for $k \leq q(2p-1)$; If $\beta = 10$ and $Y = 2^k \cdot 5^\ell$ then $Z = 2^{2p-1-k} \cdot 5^{2p-1-\ell}$, which is an integer for $k, \ell \leq 2p-1$. Hence Z is an integer in both cases, showing that $1/y$ is indeed an exact point. This concludes the proof. □

In radix $16 = 2^4$ for instance, the exact points $1/y$ with y in the interval $[1, 16)$ are listed below:

y	1	2	4	8
$1/y$	1	$1/2 = 0.8_{16}$	$1/4 = 0.4_{16}$	$1/8 = 0.2_{16}$

In radix 10, all the integers $Y = 2^k \cdot 5^\ell$ with $0 \leq k, \ell \leq 2p-1$ and $10^{p-1} \leq Y < 10^p$ can be enumerated by a simple program, and each one of them gives an exact point. Table 5 gives the 21 integral significands Y such that $1/y$ is an exact point, in the case of the decimal32 format (see also Table 7 in the appendix for the decimal64 format).

Furthermore, given an input exponent, the result below provides an explicit formula for the number N of floating-point inputs having this exponent and whose reciprocal is a floating-point number.

TABLE 5
Integral significands Y of $y \in \mathbb{F}_{10,7}$ such that $1/y \in \mathbb{F}_{10,7}$.

Y	$1/Y$
$2^0 \cdot 5^9 = 1953125$	$5.120000 \cdot 10^{-7}$
$2^0 \cdot 5^{10} = 9765625$	$1.024000 \cdot 10^{-7}$
$2^1 \cdot 5^9 = 3906250$	$2.560000 \cdot 10^{-7}$
$2^2 \cdot 5^8 = 1562500$	$6.400000 \cdot 10^{-7}$
$2^2 \cdot 5^9 = 7812500$	$1.280000 \cdot 10^{-7}$
$2^3 \cdot 5^8 = 3125000$	$3.200000 \cdot 10^{-7}$
$2^4 \cdot 5^7 = 1250000$	$8.000000 \cdot 10^{-7}$
$2^4 \cdot 5^8 = 6250000$	$1.600000 \cdot 10^{-7}$
$2^5 \cdot 5^7 = 2500000$	$4.000000 \cdot 10^{-7}$
$2^6 \cdot 5^6 = 1000000$	$1.000000 \cdot 10^{-6}$
$2^6 \cdot 5^7 = 5000000$	$2.000000 \cdot 10^{-7}$
$2^7 \cdot 5^6 = 2000000$	$5.000000 \cdot 10^{-7}$
$2^8 \cdot 5^6 = 4000000$	$2.500000 \cdot 10^{-7}$
$2^9 \cdot 5^5 = 1600000$	$6.250000 \cdot 10^{-7}$
$2^9 \cdot 5^6 = 8000000$	$1.250000 \cdot 10^{-7}$
$2^{10} \cdot 5^5 = 3200000$	$3.125000 \cdot 10^{-7}$
$2^{11} \cdot 5^4 = 1280000$	$7.812500 \cdot 10^{-7}$
$2^{11} \cdot 5^5 = 6400000$	$1.562500 \cdot 10^{-7}$
$2^{12} \cdot 5^4 = 2560000$	$3.906250 \cdot 10^{-7}$
$2^{13} \cdot 5^3 = 1024000$	$9.765625 \cdot 10^{-7}$
$2^{13} \cdot 5^4 = 5120000$	$1.953125 \cdot 10^{-7}$

Theorem 10: For a given exponent e_y, the number N of positive values $y \in \mathbb{F}_{\beta,p}$ such that $1/y \in \mathbb{F}_{\beta,p}$ is

$$N = \begin{cases} q, & \text{if } \beta = 2^q,\ q \in \mathbb{N}_{>0}; \\ 2\lfloor p \log_5(10) \rfloor + 1, & \text{if } \beta = 10. \end{cases}$$

Proof: When $\beta = 2^q$, Theorem 9 says that each exact point corresponds to an integer k such that $2^{q(p-1)} \leq 2^k < 2^{qp}$ and $0 \leq k \leq q(2p-1)$. The former condition is equivalent to $q(p-1) \leq k < qp$ and thus implies the latter. From this we deduce that the number of possible values of k is q when $\beta = 2^q$.

When $\beta = 10$, Theorem 9 says that each exact point corresponds to a pair of integers (k, ℓ) such that

$$10^{p-1} \leq 2^k \cdot 5^\ell < 10^p \qquad \text{and} \qquad 0 \leq k, \ell \leq 2p-1.$$

The value of N is the number of points $(k, \ell) \in \mathbb{Z}^2$ that satisfy those two sets of constraints. Let $\sigma = \log_5(2) = 0.4306765581\ldots$. The first set of constraints is equivalent to

$$(p-1)(1+\sigma) \leq \sigma k + \ell < p(1+\sigma). \tag{22}$$

It implies in particular that $(p-1)(1+\sigma) \leq \ell < p(1+\sigma)$, which is stronger than $0 \leq \ell \leq 2p-1$ for $p \geq 2$, since $1 + \sigma \approx 1.43$. Hence $N = \sum_{0 \leq k < 2p} N_k$, where N_k is the number of integers ℓ satisfying (22) for a given k.

Recalling that half-open real intervals $[a, b)$ such that $a \leq b$ contain exactly $\lceil b \rceil - \lceil a \rceil$ integers [4, p. 74], we deduce that, for $0 \leq k < 2p$,

$$\begin{aligned} N_k &= \big\lceil p(1+\sigma) - \sigma k \big\rceil - \big\lceil (p-1)(1+\sigma) - \sigma k \big\rceil \\ &= \big\lceil (p-k)\sigma \big\rceil - \big\lceil (p-k-1)\sigma \big\rceil + 1. \end{aligned}$$

Consequently, the sum $\sum_{0 \leq k < 2p} N_k$ telescopes to $2p + \lfloor p\sigma \rfloor + \lceil p\sigma \rceil$. Since the integer p is nonzero and σ is

irrational, $p\sigma$ cannot be an integer. Hence $\lceil p\sigma \rceil = \lfloor p\sigma \rfloor + 1$, which leads to $N = 2\big\lfloor p(1+\sigma) \big\rfloor + 1$. □

According to Theorem 10, when $\beta = 2^q$, the number N of different integral significands leading to an exact point is q. In radix 10, we have $N = \Theta(p)$, which confirms the fact that the midpoints for the reciprocal can be easily enumerated, even when the precision p is large. This is in contrast with the exact points of square root in radix 10 or 2^q, whose number is exponential in p (see Section 2.2). For the decimal formats of IEEE 754-2008, the corresponding values of N are listed below:

Format	decimal32	decimal64	decimal128
p	7	16	34
N	21	45	97

8 Reciprocal 2D Euclidean norm

Given a d-dimensional vector y with entries in $\mathbb{F}_{2,p}$, we know from Theorem 7 that $z = 1/\|y\|_2$ cannot be a midpoint in radix 2. In this section, we focus on the two-dimensional case, studying the midpoints and the exact points of the reciprocal 2D Euclidean norm, in radices 2^q and 10. In radix 10, our study relies on the representation of products of the form $2^r \cdot 5^s$ as sums of two squares $a^2 + b^2$, where $a, b \in \mathbb{N}$. Thus, we first explain in Section 8.1 the method we used for enumerating all the representations of such a product as the sum of two integer squares. Then midpoints and exact points are studied in Sections 8.2 and 8.3, respectively.

8.1 Decomposing $2^r \cdot 5^s$ into sums of two squares

Decomposing an integer into sums of two squares is a well studied problem in the mathematical literature (see for instance Wagon [13] and the references therein). In our particular case of interest, we deduce the following theorem that allows to compute all decompositions of $2^r \cdot 5^s$ as sums of two squares. The proof of Theorem 11 relies on the uniqueness of the decomposition of a number into prime factors in the ring of Gaussian integers $\mathbb{Z}[i]$ (see for instance Everest and Ward [3, chap. 2] for more details on this topic).

Theorem 11: Let $r, s \in \mathbb{N}$ be given, and assume $k \in \mathbb{N}$. All the unordered pairs $\{a, b\}$ with $a, b \in \mathbb{N}$ and $a^2 + b^2 = 2^r \cdot 5^s$ are given by $a = |\Re(c)|$ and $b = |\Im(c)|$ with

$$c = 2^{\lfloor r/2 \rfloor}(1+i)^{r \bmod 2}(2+i)^k(2-i)^{s-k}, \qquad 0 \le k < \lceil (s+1)/2 \rceil.$$

In particular, there are $\lceil (s+1)/2 \rceil$ different decompositions of $2^r \cdot 5^s$ as the sum of two squares.

Proof: Let us assume $2^r \cdot 5^s = a^2 + b^2$. Since the decomposition of $2^r \cdot 5^s$ into prime factors in $\mathbb{Z}[i]$ is unique apart from multiplications by ± 1 or $\pm i$, one has $2^r \cdot 5^s = \delta_0(1+i)^r(1-i)^r(2+i)^s(2-i)^s$ with $\delta_0 \in \{\pm 1, \pm i\}$. On the other hand one has $a^2 + b^2 = (a+ib)(a-ib)$, hence by uniqueness of the decomposition into prime factors it follows that $a + ib = \delta_1(1+i)^{k_1}(1-i)^{k_2}(2+i)^{k_3}(2-i)^{k_4}$ for some $k_1, k_2, k_3, k_4 \in \mathbb{N}$ and $\delta_1 \in \{\pm 1, \pm i\}$. Then one has $a^2 + b^2 = \delta_1\overline{\delta_1}(1+i)^{k_1+k_2}(1-i)^{k_1+k_2}(2+i)^{k_3+k_4}(2-i)^{k_3+k_4}$, and from $2^r \cdot 5^s = a^2 + b^2$ we deduce that $k_1 + k_2 = r$ and $k_3 + k_4 = s$. Moreover, distinguishing two cases corresponding to the parity of r, it can be checked that

$$(1+i)^{k_1}(1-i)^{k_2} = \delta_2 \cdot 2^{\lfloor r/2 \rfloor}(1+i)^{r \bmod 2},$$

with $\delta_2 \in \{\pm 1, \pm i\}$. Hence we obtain

$$a + ib = \delta \cdot 2^{\lfloor r/2 \rfloor}(1+i)^{r \bmod 2}(2+i)^k(2-i)^{s-k},$$

for some $\delta \in \{\pm 1, \pm i\}$ and $k \in \mathbb{N}$ such that $0 \le k \le s$. Since $a, b \ge 0$, we deduce that necessarily $a = |\Re(c)|$ and $b = |\Im(c)|$ with $c = 2^{\lfloor r/2 \rfloor}(1+i)^{r \bmod 2}(2+i)^k(2-i)^{s-k}$. However, since both c and $\overline{c} = 2^{\lfloor r/2 \rfloor}(1-i)^{r \bmod 2}(2-i)^k(2+i)^{s-k}$ lead to the same unordered pair $\{a, b\}$, there are at most $\lceil (s+1)/2 \rceil$ such unordered pairs $\{a, b\}$. This implies that we only need the assumption $0 \le k < \lceil (s+1)/2 \rceil$ for k.

Conversely, if $a = |\Re(c)|$ and $b = |\Im(c)|$ with $c = 2^{\lfloor r/2 \rfloor}(1+i)^{r \bmod 2}(2+i)^k(2-i)^{s-k}$, then $a+ib = \delta 2^{\lfloor r/2 \rfloor}(1+i)^{r \bmod 2}(2+i)^k(2-i)^{s-k}$ with $\delta \in \{\pm 1, \pm i\}$. Then one can easily check that $a^2 + b^2 = (a+ib)(a-ib) = 2^r \cdot 5^s$.

By uniqueness of the factorization into primes in $\mathbb{Z}[i]$, it can be shown that if we take $k_1 \ne k_2$ with $0 \le k_1, k_2 < \lceil (s+1)/2 \rceil$, then the corresponding unordered pairs $\{a_1, b_1\}$ and $\{a_2, b_2\}$ are necessarily different. It means that there are exactly $\lceil (s+1)/2 \rceil$ unordered pairs $\{a, b\}$. □

For later use, we also state the following corollary of Theorem 11.

Corollary 2: Given $r \in \mathbb{N}$, the unique decomposition of 2^r as a sum of two integer squares is

$$2^r = \begin{cases} 0^2 + (2^{r/2})^2, & \text{if } r \text{ is even,} \\ (2^{(r-1)/2})^2 + (2^{(r-1)/2})^2, & \text{if r is odd.} \end{cases}$$

8.2 Midpoints for reciprocal 2D norm

Theorem 12 below can be used to determine all the the midpoints of the reciprocal 2D-norm function with exponent e_z.

Theorem 12: Let $x, y \in \mathbb{F}_{\beta,p}$ be such that $(x, y) \ne (0, 0)$, and let $z = 1/\sqrt{x^2 + y^2}$. If $\beta = 2^q$ ($q \in \mathbb{N}_{>0}$) then $z \notin \mathbb{M}_{\beta,p}$. If $\beta = 10$, one has $z \in \mathbb{M}_{\beta,p}$ if and only if z has the form

$$\left(\frac{5^\ell - 1}{2} + \frac{1}{2}\right) \cdot 10^{e_z - p + 1},$$

with $e_z \in \mathbb{Z}$, and $\ell \in \mathbb{N}$ such that $2 \cdot 10^{p-1} < 5^\ell < 2 \cdot 10^p$.

Proof: Let $z = 1/\sqrt{x^2 + y^2}$ be a midpoint, with $x, y \in \mathbb{F}_{\beta,p}$. Without loss of generality, we assume that z is in $[1, \beta)$, and since z is a midpoint then one has $1 < z < \beta$. Let us also assume that $x \ge y \ge 0$, which implies

$$\frac{1}{\sqrt{2}x} \le \frac{1}{\sqrt{x^2 + y^2}} \le \frac{1}{x}. \tag{23}$$

Denoting by e_x and e_y the exponents of x and y respectively, from (23) it follows that $\beta^{-e_x - 2} < z \le \beta^{-e_x}$, and since $1 < z < \beta$, necessarily $e_x \in \{-1, -2\}$. Writing $z = (Z + 1/2) \cdot \beta^{-p+1}$, with $Z \in \mathbb{N}$ such that $\beta^{p-1} \le Z < \beta^p$, from $(x^2 + y^2)z^2 = 1$ we deduce

$$\left(X^2 \cdot \beta^{2e_x - 2e_y} + Y^2\right)(2Z+1)^2 = 4 \cdot \beta^{4p - 2e_y - 4}. \tag{24}$$

Note that $x \geq y$ implies $e_x \geq e_y$, so that the left-hand side of Equation (24) is indeed in $\mathbb{N}$. When $\beta = 2^q$, Equation (24) has no solution, since the right-hand side of the equality is a power of two while the left-hand side has an odd factor. When $\beta = 10$, (24) becomes

$$\left(X^2 \cdot 10^{2e_x-2e_y} + Y^2\right)(2Z+1)^2 = 2^{4p-2e_y-2} \cdot 5^{4p-2e_y-4}. \quad (25)$$

Then one has necessarily $2Z + 1 = 5^\ell$ with $\ell \in \mathbb{N}$. The bounds on 5^ℓ follow from $10^{p-1} \leq Z \leq 10^p - 1$. Conversely, if z has the form given in Theorem 12 it is clearly a midpoint. □

For instance, in the decimal32 format of IEEE 754-2008 ($p = 7$), function $1/\sqrt{x^2+y^2}$ has only one midpoint in $[1, 10)$, namely $z = 4.8828125$. This midpoint corresponds to $5^{10} = 9765625$, which is the only power of 5 in the interval $(2 \cdot 10^6, 2 \cdot 10^7)$. All the other midpoints of the function are obtained by multiplying 4.8828125 by an integral power of 10.

Theorem 12 can only be used to determine the midpoints of the reciprocal norm function. Given such a midpoint z, let us now show how to find x and y in $\mathbb{F}_{10,p}$ such that $z = 1/\sqrt{x^2+y^2}$. For this, we shall use the following trivial lemma.

Lemma 4: Let a be in $\mathbb{Q}$. One has $a^2 \in \mathbb{N}$ if and only if $a \in \mathbb{Z}$.

As in the proof of Theorem 12, let us assume that $1 < z < 10$ and $x \geq y \geq 0$, which implies $e_x \in \{-1, -2\}$. We denote by X and Y the integral significands of x and y respectively. From Equation (25) we can deduce that X and Y must satisfy

$$2^{4p+2} \cdot 5^{4p-2\ell} = \left(X \cdot 10^{e_x+2}\right)^2 + \left(Y \cdot 10^{e_y+2}\right)^2. \quad (26)$$

From $2 \cdot 10^{p-1} < 5^\ell < 2 \cdot 10^p$, one has $5^{4p-2\ell} \in \mathbb{N}$. Since moreover $e_x \in \{-1, -2\}$, necessarily $X \cdot 10^{e_x+2} \in \mathbb{N}$, and $Y^2 \cdot 10^{2(e_y+2)}$ is also in $\mathbb{N}$. Since $Y \cdot 10^{e_y+2}$ is a nonnegative rational number whose square is a natural integer, it follows from Lemma 4 that $Y \cdot 10^{e_y+2} \in \mathbb{N}$. Hence we know that $X \cdot 10^{e_x+2}$ and $Y \cdot 10^{e_y+2}$ both necessarily belong to $\mathbb{N}$.

As a consequence, to find all inputs (X, Y) that give a midpoint for the function $1/\sqrt{x^2+y^2}$, we know from Equation (26) that we need to find all the decompositions of the at most two integers $2^{4p+2} \cdot 5^{4p-2\ell}$ as the sum of two squares. We used Theorem 11 to build all values x and y, $x \geq y$, such that $1/\sqrt{x^2+y^2}$ is a midpoint, for the decimal formats of the IEEE 754-2008 standard. For the decimal32 format, all the pairs of floating-point numbers (x, y) for which $1/\sqrt{x^2+y^2}$ is a midpoint can be deduced from the pairs listed in Table 6 by either exchanging x and y or by multiplying them by the same power of 10 (see also Table 9 in the appendix for the decimal64 format).

The following table gives the number N_z of midpoints z in a decade (i.e., with a fixed exponent e_z), with respect to the decimal format considered. The table also gives the number N of pairs of integral significand (X, Y) with $X \geq Y$ that give these midpoints. In decimal64 arithmetic for instance, the function $(x, y) \mapsto 1/\sqrt{x^2+y^2}$ has 2 midpoints $z_1 < z_2$ in the decade $[1, 10)$: the number of pairs (X, Y) that give z_1 is 10, and 9 pairs give z_2.

TABLE 6
Floating-point numbers $x, y \in \mathbb{F}_{10,7}$ with $X \geq Y$ such that $z = 1/\sqrt{x^2+y^2}$ is a midpoint, with $10^{-8} \leq z < 10^{-7}$.

x	y	$z = 1/\sqrt{x^2+y^2}$
1966080	573440.0	4.8828125×10^{-7}
1638400	1228800	
1916928	720896.0	
2048000	0	

Format	decimal32	decimal64	decimal128
p	7	16	34
N_z	1	2	2
N	4	$10 + 9$	$20 + 19$

8.3 Exact points for reciprocal 2D norm

Theorem 13: Let $x, y \in \mathbb{F}_{\beta,p}$ be such that $(x, y) \neq (0, 0)$. Let X, Y denote the integral significands of x, y, and let also z denote $1/\sqrt{x^2+y^2}$.

• For $\beta = 2^q$ ($q \in \mathbb{N}_{>0}$), the real z is also in $\mathbb{F}_{2^q,p}$ if and only if $\{x, y\} = \{0, \pm 2^k\}$ for some $k \in \mathbb{Z}$.

• For $\beta = 10$, the number z is in $\mathbb{F}_{10,p}$ if and only if its integral significand Z satisfies $Z = 2^k \cdot 5^\ell$, with $10^{p-1} \leq 2^k \cdot 5^\ell < 10^p$ and $k, \ell \in \mathbb{N}$. In this case one has $2^{8p-2k} \cdot 5^{8p-2\ell} \in \mathbb{N}$, and (X, Y) must satisfy

$$(X \cdot 10^m)^2 + (Y \cdot 10^n)^2 = 2^{8p-2k} \cdot 5^{8p-2\ell},$$

where $m, n \in \mathbb{Z}$ such that $X \cdot 10^m$ and $Y \cdot 10^n$ are in $\mathbb{N}$.

Proof: Without loss of generality, we assume that $1 \leq z < \beta$ and that $0 \leq y \leq x$. Reasoning as in the proof of Theorem 12, one may check that necessarily $e_x \in \{-2, -1, 0\}$. Using as usual the normalized representations of x, y and z, from $(x^2+y^2)z = 1$ we deduce

$$Z^2(X^2 \cdot \beta^{2e_x-2e_y} + Y^2) = \beta^{4p-4-2e_y}. \quad (27)$$

If $\beta = 2^q$ for some $q \in \mathbb{N}_{>0}$, then Equation (27) implies that $Z = 2^\ell$ for some $\ell \in \mathbb{Z}$. From Equation (27), we then deduce

$$\left(X \cdot 2^{q(e_x+2)}\right)^2 + \left(Y \cdot 2^{q(e_y+2)}\right)^2 = 2^{4qp-2\ell}. \quad (28)$$

Since $2^{q(p-1)} \leq Z < 2^{qp}$, we deduce that $q(p-1) \leq \ell < qp$, hence $2^{4qp-2\ell}$ is in $\mathbb{N}$. Since both $2^{4qp-2\ell}$ and $X \cdot 2^{q(e_x+2)}$ are in $\mathbb{N}$, it follows that $(Y \cdot 2^{q(e_y+2)})^2$ is also in $\mathbb{N}$, and from Lemma 4 we deduce that $Y \cdot 2^{q(e_y+2)} \in \mathbb{N}$. Then Corollary 2 implies that the only possible decomposition of $2^{4qp-2\ell}$ as the sum of two squares is $2^{4qp-2\ell} = 0^2 + (2^{2qp-\ell})^2$, so that $\{X, Y\} = \{0, 2^{2qp-\ell}\}$. Conversely, if $\{x, y\} = \{0, \pm 2^k\}$, then $1/\sqrt{x^2+y^2} = 2^{-k}$ belongs to $\mathbb{F}_{2^q,p}$.

Now let us assume that $\beta = 10$. Then Equation (27) becomes

$$Z^2\left(X^2 \cdot 10^{2e_x-2e_y} + Y^2\right) = 10^{4p-4-2e_y}. \quad (29)$$

Since $10^{4p-4-2e_y}$ is a multiple of Z, necessarily $Z = 2^k \cdot 5^\ell$ with $k, \ell \in \mathbb{N}$ such that $10^{p-1} \leq 2^k \cdot 5^\ell < 10^p$, which implies

$\ell \le 2p$ and $k \le 4p$. Moreover, from (29) with $Z = 2^k \cdot 5^\ell$ we have

$$\left(X \cdot 10^{2p+e_x+2}\right)^2 + \left(Y \cdot 10^{2p+e_y+2}\right)^2 = 2^{8p-2k} \cdot 5^{8p-2\ell}. \tag{30}$$

Since $(X \cdot 10^{2p+e_x+2})^2$ and $2^{8p-2k} \cdot 5^{8p-2\ell}$ are both in $\mathbb{N}$, then $Y \cdot 10^{2p+e_y+2}$ also belongs to $\mathbb{N}$, which concludes the proof. □

In radix 2^q, the pairs (x, y) such that $1/\sqrt{x^2+y^2}$ is a midpoint are characterized by Theorem 13. In radix 10, for each $Z = 2^k \cdot 5^\ell$ with $k, \ell \in \mathbb{N}$ such that $10^{p-1} \le 2^k \cdot 5^\ell < 10^p$, we are reduced to find all decompositions of $2^{8p-2k} \cdot 5^{8p-2\ell}$ as sums of two squares. This is done as explained in Subsection 8.1. For each basic decimal format of the IEEE 754-2008 standard, the following table gives the number N_z of midpoints with a fixed exponent e_z, together with the number N of pairs of significands (X, Y) with $X \ge Y$ such that $1/\sqrt{x^2+y^2}$ is in $\mathbb{F}_{10,p}$.

Format	decimal32	decimal64	decimal128
p	7	16	34
N_z	42	93	196
N	160	764	3373

9 Normalization of 2D-vectors

Theorem 7 shows that $x/\sqrt{x^2+y^2}$, cannot be a midpoint in radix 2. Here we first extend this result to radices 2^q and 10. Then we characterize the exact points of the 2D-normalization function in radix 2^q.

9.1 Midpoints for 2D normalization

Theorem 14: Let $x, y \in \mathbb{F}_{\beta,p}$ such that $(x, y) \neq (0, 0)$. If $\beta = 2^q$ ($q \in \mathbb{N}_{>0}$) or $\beta = 10$ then $x/\sqrt{x^2+y^2} \notin \mathbb{M}_{\beta,p}$.

Proof: Without loss of generality, let us assume $x, y > 0$, and assume that $z = x/\sqrt{x^2+y^2}$ is a midpoint. Hence we write as usual $z = (Z + 1/2) \cdot 10^{e_z-p+1}$ with $e_z \in \mathbb{Z}$ and $Z \in \mathbb{N}$ such that $\beta^{p-1} \le Z < \beta^p$. From $x/\sqrt{x^2+y^2} \le 1$ we deduce that $z \le 1$, hence $e_z \le 0$. Using $x^2(1-z^2) = y^2 z^2$ and the normalized representations of x and y gives

$$X^2\left(4 \cdot \beta^{2p-2-2e_z} - (2Z+1)^2\right) = Y^2(2Z+1)^2 \cdot \beta^{2e_y-2e_x}. \tag{31}$$

From $e_z \le 0$, the left-hand side of (31) is in $\mathbb{N}$ and thus, using Lemma 4, $Y(2Z+1) \cdot \beta^{e_y-e_x} \in \mathbb{N}$. Since $Y^2(2Z+1)^2 \cdot \beta^{2e_y-2e_z}$ is a multiple of X^2, it follows that $Y(2Z+1) \cdot \beta^{e_y-e_x} = JX$ for some J in $\mathbb{N}_{>0}$. Equation (31) then becomes

$$(2 \cdot \beta^{p-1-e_z})^2 = J^2 + (2Z+1)^2, \tag{32}$$

which expresses $(2 \cdot \beta^{p-1-e_z})^2$ as a sum of two integer squares.

If $\beta = 2^q$ then $(2 \cdot \beta^{p-1-e_z})^2$ is an even power of two and Corollary 2 then implies that it has only one possible decomposition, which is $0^2 + (2^{q(p-1-e_z)+1})^2$. However, this contradicts the fact that both J and $2Z+1$ are positive integers.

With $\beta = 10$, Equation (32) becomes

$$2^{2p-2e_z} \cdot 5^{2p-2-2e_z} = J^2 + (2Z+1)^2. \tag{33}$$

Since $2p - 2e_z$ is even, according to Theorem 11, one has

$$2^{2p-2e_z} \cdot 5^{2p-2-2e_z} = |\Re(c)|^2 + |\Im(c)|^2,$$

with $c = 2^{p-e_z}(2+i)^k(2-i)^{2p-2-2e_z-k}$ for some $k \in \mathbb{N}$, and one may check that both $|\Re(c)|$ and $|\Im(c)|$ are even. Hence the two squares in the right-hand side of Equation (33) must be even, which is a contradiction since $2Z+1$ is odd. □

9.2 Exact points for 2D normalization

The next theorem provides a characterization of the exact points of the 2D-normalization function in radix 2^q.

Theorem 15: Let $q \in \mathbb{N}_{>0}$ and let $x, y \in \mathbb{F}_{2^q,p}$ be such that $(x, y) \neq (0, 0)$. One has $z = x/\sqrt{x^2+y^2} \in \mathbb{F}_{2^q,p}$ if and only if $x = 0$ or $y = 0$.

Proof: The "if" statement is obvious. Conversely, assume that $z \in \mathbb{F}_{2^q,p}$ and that both x and y are nonzero. We can restrict to $x, y > 0$ with no loss of generality. Let $z = x/\sqrt{x^2+y^2}$. Since $z \le 1$, necessarily $e_z \le 0$. Then, using $x^2(1-z^2) = y^2 z^2$ and the normalized representations of x and y,

$$X^2(\beta^{2p-2e_z-2} - Z^2) = Y^2 Z^2 \cdot \beta^{2e_y-2e_x}, \tag{34}$$

From $e_z \le 0$ it follows that the left-hand side of (34) is in $\mathbb{N}$ and, due to Lemma 4, so is $YZ \cdot \beta^{e_y-e_y}$. Now, since $Z^2Y^2 \cdot \beta^{2e_y-2e_x}$ is a multiple of X^2, we have $ZY \cdot \beta^{e_y-e_x} = JX$ for some $J \in \mathbb{N}_{>0}$. Then we obtain from Equation (34)

$$\left(\beta^{e_z-p+1}\right)^2 = J^2 + Z^2. \tag{35}$$

When $\beta = 2^q$, Corollary 2 implies that either J or Z is zero, a contradiction. □

In radix 10, we do not have simple results to characterize the exact points of the 2D-normalization function. But they can of course be enumerated using Equation (35), at least for some small precisions. Using Theorem 11, we enumerate all the pairs (Z, J) for a fixed e_z such that (35) holds. Without loss of generality, we fix $e_x = 0$. The inputs x and y can then be found by searching the points $(X, Y \cdot 10^{e_y})$ on the line $YZ \cdot 10^{e_y} = JX$, with $10^{p-1} \le X < 10^p$ and $X \in \mathbb{N}$. For some small precisions, the following table gives the number of pairs of inputs (X, Y) such that $x/\sqrt{x^2+y^2}$ is an exact point:

p	1	2	3	4	5	6	7
$e_z = -1$	4	54	558	5622	56254	562696	5630268
$e_z = -2$	0	0	0	6	60	597	2889

This experiment suggests that the number of (x, y) such that $x/\sqrt{x^2+y^2}$ is an exact point grows very rapidly with p, and that no useful enumeration can be performed.

10 2D Euclidean norm

Let x and y be two numbers in $F_{\beta,p}$, and assume that $z = \sqrt{x^2+y^2}$ is a midpoint. We use the normalized representations of x, y and we write as usual $z = (Z+1/2) \cdot \beta^{e_z-p+1}$. Without loss of generality, we assume that $x \ge y$, which implies $e_z \ge e_x \ge e_y$. Then from $x^2 + y^2 = z^2$ it follows that

$$4(Y^2 + X^2 \cdot \beta^{2e_x-2e_y}) = (2Z+1)^2 \cdot \beta^{2e_z-2e_y}. \tag{36}$$

When β is odd, the right-hand side of Equation (36) is odd, while the left-hand side is always even. Hence, if the radix β is odd, $\sqrt{x^2+y^2}$ cannot be a midpoint, and this observation can be generalized to the Euclidean norm in higher dimensions. Nevertheless, this not a very useful result since it does not hold for binary, decimal nor hexadecimal arithmetic.

For even radices, we do not have general results. Equation (36) has solutions, and exhaustive enumeration can be performed at least for small precisions. In radices 2 and 10, and for some small precisions p, the following tables display the number N of input pairs (x, y), with $x \geq y$, such that $z = \sqrt{x^2+y^2}$ is a midpoint in the interval $[1, \beta)$.

Radix 2

p	1	2	3	4	5	6	7	8	9	10
N	0	1	1	3	5	18	30	76	155	348

Radix 10

p	1	2	3	4
N	0	11	177	2428

These experiments suggest that the number of midpoints for the function $(x, y) \mapsto \sqrt{x^2+y^2}$ grows very rapidly with p.

On the other hand, in one-dimension the Euclidean norm reduces to the absolute value, which suffices to see that it admits only exact points, whatever the parity of β.

11 Conclusion

We have shown that for several simple algebraic functions ($\sqrt{y}$, $1/\sqrt{y}$, x^k for $k \in \mathbb{N}_{>0}$, $x/\|y\|_2$, x/y, $1/y$, $1/\sqrt{x^2+y^2}$, $x/\sqrt{x^2+y^2}$, ...), we can obtain useful information on the existence of midpoints and exact points. This information can be used for simplifying or improving the performance of programs that evaluate these functions.

Finding midpoints and exact points would also be of interest for the most common transcendental functions (sine, cosine, exponential, logarithm,...). Providing these functions with correct rounding is a difficult problem, known as the *Table-Maker's Dilemma* [7], [10]. For the most simple transcendental functions (that is, those built from the complex exponential and logarithm), one can deduce the nonexistence of midpoints from the following corollary of Lindemann's theorem (see for example [2, p. 6]):

Theorem 16 (Lindemann): e^z is transcendental for every non-zero algebraic complex number z.

Since floating-point numbers as well as midpoints are algebraic numbers, Theorem 16 allows us to deduce that for any radix and precision, if x is a floating-point number then $\ln(x)$, $\exp(x)$, $\sin(x)$, $\cos(x)$, $\tan(x)$, $\arctan(x)$, $\arcsin(x)$ and $\arccos(x)$ cannot be midpoints. Furthermore, the only exact points are $\ln(1) = 0$, $\exp(0) = 1$, $\sin(0) = 0$, $\cos(0) = 1$, $\tan(0) = 0$, $\arctan(0) = 0$, $\arcsin(0) = 0$, and $\arccos(1) = 0$.

The case of radix-2 and radix-10 exponentials and logarithms have to be treated more carefully. But one can prove that the radix-2 or 10 logarithm of a rational number is either an integer or an irrational number. This gives the following result. Assume that the exponent size is less than the precision (which is true in any reasonable floating-point system), and that x is a floating-point number. Then we have the following:

- $\log_2(x)$ cannot be a midpoint. It can be an exact point only when $x = 2^k$, where k is an integer;
- $\log_{10}(x)$ cannot be a midpoint. It can be an exact point only when $x = 10^k$, where k is an integer;

It is always possible to build ad-hoc transcendental functions for which something can be said about midpoints or exact points. Unfortunately, for the many common non-elementary transcendental functions useful in scientific applications (physics, statistics, etc.), almost nothing is known about their midpoints or exact points in floating point arithmetic.

Consider for instance the Gamma function. We know that if n is a nonnegative integer then $\Gamma(n) = (n-1)!$ is an integer too (which implies the existence of midpoints in some cases, e.g., in radix-2 arithmetic with $p = 3$, the number $6_{10} = 110_2$ is a floating-point number, and $\Gamma(6) = 5! = 120_{10} = 1111000_2$ is a midpoint). Although we have no proof of that, it is extremely unlikely that Gamma of a non-integer floating-point number could be a midpoint or an exact point. To our knowledge (see for example [14]), the only result that can be used to deal with a very few cases is that $\Gamma(x)$ is shown to be irrational if x modulo 1 belongs to $\{1/6, 1/4, 1/3, 1/2, 2/3, 3/4, 5/6\}$.

References

[1] American National Standards Institute and Institute of Electrical and Electronic Engineers. *IEEE Standard for Binary Floating-Point Arithmetic*. ANSI/IEEE Standard 754-1985, 1985.

[2] A. Baker. *Transcendental Number Theory*. Cambridge University Press, 1975.

[3] G. Everest and T. Ward. *An Introduction to Number Theory*. Graduate Texts in Mathematics. Springer-Verlag, London, 2005.

[4] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, second edition, 1994.

[5] IEEE Computer Society. *IEEE Standard for Floating-Point Arithmetic*. IEEE Standard 754-2008, August 2008. available at http://ieeexplore.ieee.org/servlet/opac?punumber=4610933.

[6] C. Iordache and D. W. Matula. On infinitely precise rounding for division, square root, reciprocal and square root reciprocal. In Koren and Kornerup, editors, *Proceedings of the 14th IEEE Symposium on Computer Arithmetic (Adelaide, Australia)*, pages 233–240. IEEE Computer Society Press, Los Alamitos, CA, April 1999.

[7] W. Kahan. A logarithm too clever by half. Available at http://http.cs.berkeley.edu/~wkahan/LOG10HAF.TXT, 2004.

[8] C. Q. Lauter and V. Lefèvre. An efficient rounding boundary test for pow(x, y) in double precision. *IEEE Transactions on Computers*, 58(2):197–207, February 2009.

[9] P. Markstein. *IA-64 and Elementary Functions: Speed and Precision*. Hewlett-Packard Professional Books. Prentice-Hall, Englewood Cliffs, NJ, 2000.

[10] J.-M. Muller. *Elementary Functions, Algorithms and Implementation*. Birkhäuser Boston, MA, 2nd edition, 2006.

[11] M. Parks. Inexact quotients and square roots. http://www.geocities.com/ieee754/papers/parks-inexact.ps.

[12] E.M. Schwarz, R.M. Smith, and C.A. Krygowski. The S/390 G5 floating-point unit supporting hex and binary architectures. In *Proceedings of the 14th IEEE Symposium on Computer Arithmetic (ARITH-14)*, pages 258–265, April 1999.

[13] S. Wagon. The Euclidean algorithm strikes again. *The American Mathematical Monthly*, 97(2):125–129, February 1990.

[14] M. Waldschmidt. Transcendence of periods: The state of the art. *Pure and Applied Mathematics Quarterly*, 2(2):435–463, 2006.

[15] A. Ziv. Fast evaluation of elementary mathematical functions with correctly rounded last bit. *ACM Transactions on Mathematical Software*, 17(3):410–423, September 1991.

APPENDIX: EXACT POINTS FOR RECIPROCAL, SQUARE ROOT RECIPROCAL AND 2D-NORM RECIPROCAL

TABLE 7
In this table, we list the 45 integral significands Y such that $1/y$ is an exact point, in the case of the decimal64 format ($p = 16$).

Y	$1/Y$
2384185791015625	$4.194304000000000 \cdot 10^{-16}$
4768371582031250	$2.097152000000000 \cdot 10^{-16}$
1907348632812500	$5.242880000000000 \cdot 10^{-16}$
9536743164062500	$1.048576000000000 \cdot 10^{-16}$
3814697265625000	$2.621440000000000 \cdot 10^{-16}$
1525878906250000	$6.553600000000000 \cdot 10^{-16}$
7629394531250000	$1.310720000000000 \cdot 10^{-16}$
3051757812500000	$3.276800000000000 \cdot 10^{-16}$
1220703125000000	$8.192000000000000 \cdot 10^{-16}$
6103515625000000	$1.638400000000000 \cdot 10^{-16}$
2441406250000000	$4.096000000000000 \cdot 10^{-16}$
4882812500000000	$2.048000000000000 \cdot 10^{-16}$
1953125000000000	$5.120000000000000 \cdot 10^{-16}$
9765625000000000	$1.024000000000000 \cdot 10^{-16}$
3906250000000000	$2.560000000000000 \cdot 10^{-16}$
1562500000000000	$6.400000000000000 \cdot 10^{-16}$
7812500000000000	$1.280000000000000 \cdot 10^{-16}$
3125000000000000	$3.200000000000000 \cdot 10^{-16}$
1250000000000000	$8.000000000000000 \cdot 10^{-16}$
6250000000000000	$1.600000000000000 \cdot 10^{-16}$
2500000000000000	$4.000000000000000 \cdot 10^{-16}$
1000000000000000	$1.000000000000000 \cdot 10^{-15}$
5000000000000000	$2.000000000000000 \cdot 10^{-16}$
2000000000000000	$5.000000000000000 \cdot 10^{-16}$
4000000000000000	$2.500000000000000 \cdot 10^{-16}$
1600000000000000	$6.250000000000000 \cdot 10^{-16}$
8000000000000000	$1.250000000000000 \cdot 10^{-16}$
3200000000000000	$3.125000000000000 \cdot 10^{-16}$
1280000000000000	$7.812500000000000 \cdot 10^{-16}$
6400000000000000	$1.562500000000000 \cdot 10^{-16}$
2560000000000000	$3.906250000000000 \cdot 10^{-16}$
1024000000000000	$9.765625000000000 \cdot 10^{-16}$
5120000000000000	$1.953125000000000 \cdot 10^{-16}$
2048000000000000	$4.882812500000000 \cdot 10^{-16}$
4096000000000000	$2.441406250000000 \cdot 10^{-16}$
1638400000000000	$6.103515625000000 \cdot 10^{-16}$
8192000000000000	$1.220703125000000 \cdot 10^{-16}$
3276800000000000	$3.051757812500000 \cdot 10^{-16}$
1310720000000000	$7.629394531250000 \cdot 10^{-16}$
6553600000000000	$1.525878906250000 \cdot 10^{-16}$
2621440000000000	$3.814697265625000 \cdot 10^{-16}$
1048576000000000	$9.536743164062500 \cdot 10^{-16}$
5242880000000000	$1.907348632812500 \cdot 10^{-16}$
2097152000000000	$4.768371582031250 \cdot 10^{-16}$
4194304000000000	$2.384185791015625 \cdot 10^{-16}$

TABLE 8
This table gives all integral significands Y of y, and the parity of the exponent e_y, such that $z = 1/\sqrt{y}$ is a floating point number too, in decimal64 format.

Y			$1/\sqrt{Y \cdot 10^{\delta_y - p + 1}}$	δ_y
$2^{15}.5^{15}$	=	1000000000000000	$1.000000000000000 \cdot 10^{0}$	0
$2^{45}.5^{3}$	=	4398046511104000	$4.768371582031250 \cdot 10^{-1}$	0
$2^{43}.5^{3}$	=	1099511627776000	$9.536743164062500 \cdot 10^{-1}$	0
$2^{41}.5^{5}$	=	6871947673600000	$3.814697265625000 \cdot 10^{-1}$	0
$2^{39}.5^{5}$	=	1717986918400000	$7.629394531250000 \cdot 10^{-1}$	0
$2^{35}.5^{7}$	=	2684354560000000	$6.103515625000000 \cdot 10^{-1}$	0
$2^{31}.5^{9}$	=	4194304000000000	$4.882812500000000 \cdot 10^{-1}$	0
$2^{29}.5^{9}$	=	1048576000000000	$9.765625000000000 \cdot 10^{-1}$	0
$2^{27}.5^{11}$	=	6553600000000000	$3.906250000000000 \cdot 10^{-1}$	0
$2^{25}.5^{11}$	=	1638400000000000	$7.812500000000000 \cdot 10^{-1}$	0
$2^{21}.5^{13}$	=	2560000000000000	$6.250000000000000 \cdot 10^{-1}$	0
$2^{17}.5^{15}$	=	4000000000000000	$5.000000000000000 \cdot 10^{-1}$	0
$2^{13}.5^{17}$	=	6250000000000000	$4.000000000000000 \cdot 10^{-1}$	0
$2^{11}.5^{17}$	=	1562500000000000	$8.000000000000000 \cdot 10^{-1}$	0
$2^{9}.5^{19}$	=	9765625000000000	$3.200000000000000 \cdot 10^{-1}$	0
$2^{7}.5^{19}$	=	2441406250000000	$6.400000000000000 \cdot 10^{-1}$	0
$2^{3}.5^{21}$	=	3814697265625000	$5.120000000000000 \cdot 10^{-1}$	0
$2^{46}.5^{2}$	=	1759218604441600	$2.384185791015625 \cdot 10^{-1}$	1
$2^{42}.5^{4}$	=	2748779069440000	$1.907348632812500 \cdot 10^{-1}$	1
$2^{38}.5^{6}$	=	4294967296000000	$1.525878906250000 \cdot 10^{-1}$	1
$2^{36}.5^{6}$	=	1073741824000000	$3.051757812500000 \cdot 10^{-1}$	1
$2^{34}.5^{8}$	=	6710886400000000	$1.220703125000000 \cdot 10^{-1}$	1
$2^{32}.5^{8}$	=	1677721600000000	$2.441406250000000 \cdot 10^{-1}$	1
$2^{28}.5^{10}$	=	2621440000000000	$1.953125000000000 \cdot 10^{-1}$	1
$2^{24}.5^{12}$	=	4096000000000000	$1.562500000000000 \cdot 10^{-1}$	1
$2^{22}.5^{12}$	=	1024000000000000	$3.125000000000000 \cdot 10^{-1}$	1
$2^{20}.5^{14}$	=	6400000000000000	$1.250000000000000 \cdot 10^{-1}$	1
$2^{18}.5^{14}$	=	1600000000000000	$2.500000000000000 \cdot 10^{-1}$	1
$2^{14}.5^{16}$	=	2500000000000000	$2.000000000000000 \cdot 10^{-1}$	1
$2^{10}.5^{18}$	=	3906250000000000	$1.600000000000000 \cdot 10^{-1}$	1
$2^{6}.5^{20}$	=	6103515625000000	$1.280000000000000 \cdot 10^{-1}$	1
$2^{4}.5^{20}$	=	1525878906250000	$2.560000000000000 \cdot 10^{-1}$	1
$2^{2}.5^{22}$	=	9536743164062500	$1.024000000000000 \cdot 10^{-1}$	1
$2^{0}.5^{22}$	=	2384185791015625	$2.048000000000000 \cdot 10^{-1}$	1

TABLE 9
In decimal64 arithmetic, the function $1/\sqrt{x^2+y^2}$ has two midpoints in the range $[10^{-16}, 10^{-15})$, denoted by z_1 and z_2. The pairs of floating-point numbers x and y such that $1/\sqrt{x^2+y^2}$ equals z_1 or z_2 are listed below.

$$z_1 = 1.1920928955078125 \cdot 10^{-16}$$
$$z_2 = 5.9604644775390625 \cdot 10^{-16}$$

x	y	z
8053063680000000	2348810240000000	z_1
6710886400000000	5033164800000000	
7851737088000000	2952790016000000	
7073274265600000	4509715660800000	
6309843828736000	5527622909952000	
8208004625203200	1731301317017600	
7605184490373120	3539761721507840	
7394920071430144	3960290559393792	
8364448808960000	636192030720000	
8388608000000000	0	
1342177280000000	1006632960000000	z_2
1261968765747200	1105524581990400	
1610612736000000	469762048000000	
1570347417600000	590558003200000	
1414654853120000	901943132160000	
1672889761792000	127238406144000	
1641600925040640	346260263403520	
1521036898074624	707952344301568	
1677721600000000	0	

Claude-Pierre Jeannerod received the PhD degree in applied mathematics from Institut National Polytechnique de Grenoble in 2000. After being a postdoctoral fellow in the Symbolic Computation Group at the University of Waterloo, he is now a researcher at INRIA Grenoble - Rhône-Alpes and a member of Laboratoire LIP (CNRS, ÉNS Lyon, INRIA, Université Claude Bernard Lyon 1). His research interests include computer algebra, linear algebra, and floating-point arithmetic. He is a member of the ACM and the IEEE.

Nicolas Louvet received the MSc degree from the Université de Picardie Jules Verne (Amiens, France), in 2004, and the PhD degree in computer science from the Université de Perpignan Via Domitia (Perpignan, France) in 2007. He is now assitant professor in the department of computer science of the Université Claude Bernard Lyon 1 (Lyon, France), and a member of the LIP laboratory (LIP is a joint laboratory of CNRS, École Normale Supérieure de Lyon, INRIA and Université Claude Bernard Lyon 1). His research interests are in computer arithmetic.

Jean-Michel Muller was born in Grenoble, France, in 1961. He received his Ph.D. degree in 1985 from the Institut National Polytechnique de Grenoble. He is Directeur de Recherches (senior researcher) at CNRS, France, and he is the former head of the LIP laboratory (LIP is a joint laboratory of CNRS, École Normale Supérieure de Lyon, INRIA and Université Claude Bernard Lyon 1). His research interests are in Computer Arithmetic. Dr. Muller was co-program chair of the 13th IEEE Symposium on Computer Arithmetic (Asilomar, USA, June 1997), general chair of SCAN'97 (Lyon, France, sept. 1997), general chair of the 14th IEEE Symposium on Computer Arithmetic (Adelaide, Australia, April 1999). He is the author of several books, including "Elementary Functions, Algorithms and Implementation" (2nd edition, Birkhäuser Boston, 2006), and he coordinated the writing of the "Handbook of Floating-Point Arithmetic" (Birkhäuser Boston, 2010). He served as associate editor of the IEEE Transactions on Computers from 1996 to 2000. He is a senior member of the IEEE.

Adrien Panhaleux was born in Roubaix, France, in 1985. He received his master degree in 2008 from the École Normale Supérieure de Lyon. He is now preparing his Ph.D. at École Normale Supérieure de Lyon, France, under the supervision of Jean-Michel Muller and Nicolas Louvet.