Accurate calculation of Euclidean Norms using Double-word arithmetic

Vincent Lefèvre, Nicolas Louvet, Jean-Michel Muller, Joris Picot, Laurence Rideau

Accurate calculation of Euclidean Norms using Double-word arithmetic

Vincent Lefèvre Nicolas Louvet Jean-Michel Muller
Joris Picot Laurence Rideau

December 16, 2021

Abstract

We consider the computation of the Euclidean (a.k.a. L2) norm of an n-dimensional vector in floating-point arithmetic. We review the classical solutions used to avoid spurious overflow or underflow and/or to obtain very accurate results. We modify a recently published algorithm [13] to allow for a very accurate solution, free of spurious overflows and underflows. The returned result will be within very slightly more than 0.5 ulp from the exact result, which means that we will almost always provide correct rounding.

Keywords: Floating-Point arithmetic; Euclidean norms; Double-word arithmetic; Double-double arithmetic; Overflow; Underflow; Square-root.

1 Introduction

1.1 Computation of Euclidean norms

We consider the computation of Euclidean norms in binary floating-point arithmetic. The Euclidean (a.k.a. L2) norm of a vector $(a_0, a_1, a_2, \ldots, a_{n-1})$ is the number

$$N = \sqrt{\sum_{i=0}^{n-1} a_i^2}. \tag{1}$$

Computing Euclidean norms is important in many scientific and engineering applications. A good implementation of the Euclidean norm must be fast and accurate. It must also avoid spurious underflows and overflows. A spurious underflow or overflow is an underflow or overflow that occurs during an intermediate step, resulting in an inaccurate, infinite or NaN returned result, whereas the exact result is well within the domain of normal floating-point numbers.

To illustrate how spurious underflows and overflows can jeopardize the computation of a Euclidean norm, consider the following examples, assuming IEEE 754 binary64/double-precision arithmetic and $n = 3$, with the default round-to-nearest, ties-to-even, rounding function, and suppose that we implement Formula (1) naively by first summing the squares serially and then taking the square root.

- with $a_0 = 1.5 \times 2^{511}$, $a_1 = 0$, and $a_2 = 2^{512}$, we will obtain an infinite result (because the computation of a_2^2 overflows), whereas the exact result is 5×2^{510}, which is much smaller than the overflow threshold;
- with $a_0 = a_1 = a_2 = (45/64) \times 2^{-537}$, the computed result is 0, whereas the exact result is around 1.2178×2^{-537}, which is much above the underflow threshold.

Note that from an *accuracy* point-of-view, spurious underflow is a problem only if *all* terms a_i are tiny (otherwise, the errors due to underflows that occur when squaring the "tiny" terms vanish in front of the squares of the "big" terms). However, spurious underflow can be very harmful from a *performance* point-of-view on a system on which subnormal numbers are handled in software, through a trapping underflow mechanism.

There are no catastrophic cancellations when computing a Euclidean norm: all added terms are nonnegative. Hence, even a naive use of Formula (1) will be rather accurate when no underflow or overflow occurs (roughly speaking, at most $\log_2 n$ bits of the final result can be incorrect). However, we can try to take advantage of that property to always obtain results *very near* the exact result.

The particular case $n = 2$ (the so-called "hypotenuse" function) has been studied in excellent references [1, 10]. In this paper, we assume that n is larger (more precisely, our algorithms do work in the cases $n = 1$ or 2, but it is for larger values that good performance is aimed at).

Due to the importance of the topic, several solutions have been suggested and analyzed until recently for computing norms accurately and/or without spurious underflows and overflows [2, 19, 13, 14]. We will present them in Section 2.1. Before that, let us present some definitions and properties related to floating-point arithmetic, that will be useful in the sequel of this paper.

1.2 The underlying FP arithmetic

In the following, we assume a radix-2, precision-p, floating-point (FP) arithmetic (where $p \geq 5$), with extremal exponents $e_{\min} < 0$ and $e_{\max} > 0$. We also assume $e_{\min} = 1 - e_{\max}$ (which is a requirement of the IEEE 754-2019 Standard for FP arithmetic [21]). In such a system, a floating-point number (FPN) is a number of the form [32]:

$$M \cdot 2^{e-p+1},$$

with $M \in \mathbb{Z}$, $|M| \leq 2^p - 1$, and $e \in \mathbb{Z}$, $e_{\min} \leq e \leq e_{\max}$. A FPN x is *normal* if $|x| \geq 2^{e_{\min}}$ or $x = 0$, and *subnormal* otherwise. The largest representable FP number is

$$\Omega = 2^{e_{\max}} \cdot (2 - 2^{1-p}), \tag{2}$$

the smallest positive normal FP number, also called "underflow threshold", is $2^{e_{\min}}$. In the following, we will say that an arithmetic operation *underflows* if its result is both subnormal *and* inexact.This choice may seem strange, but we want to avoid underflows because of accuracy concerns: when the result is exact, there is no need to worry about accuracy (incidentally, this is why the underflow flag is nor raised in

such a case under the default exception handling for underflow of the IEEE 754-2019 Standard).

The smallest positive FP number is

$$\alpha = 2^{e_{\min}-p+1}. \tag{3}$$

The notation $\mathrm{RN}(t)$ stands for t rounded to the nearest FP number. We do not assume a particular tie-breaking rule in our proofs,[1] and we use the default ties-to-even rule in our examples. For instance $\mathrm{RN}(c \cdot d)$ is the result of the FP multiplication $c * d$, assuming round-to-nearest rounding mode (which is the default in IEEE 754-2019). The number $\mathrm{ulp}(x)$, for $x \neq 0$ is

$$\mathrm{ulp}(x) = 2^{\max\{\lfloor \log_2 |x| \rfloor, e_{\min}\}-p+1},$$

and $u = 2^{-p} = \frac{1}{2}\mathrm{ulp}(1)$ denotes the roundoff error unit. The constraint $p \geq 5$ implies $u \leq 1/32$, which will serve many times in our proofs. The relative error due to rounding to nearest a real number x such that $|x| \in [2^{e_{\min}}, \Omega]$, namely $|(\mathrm{RN}(x) - x)/x|$, is bounded by $u/(1+u)$ [24]. When tightness is not necessary, we will use the simpler yet very slightly looser bound u. We will denote $\mathrm{succ}(t)$ the floating-point successor of t, and η the number $2^{(e_{\min}+p)/2}$ (beware: it is a FP number only when $e_{\min}+p$ is even). Barring overflow, the square of a FPN $\geq \eta$ can be expressed exactly as the sum of two FPNs [4]. We also have the following property (see for instance [32]):

Property 1.1. *If a FP number $\hat{t}$ approximates a real number t with relative error ϵ, then $\hat{t}$ is within $(\epsilon/u) \cdot \mathrm{ulp}(t)$ from t.*

The FP numbers between 2^k and 2^{k+1} are multiples of $2^{k+1}u$: for instance, the FP numbers between 1 and 2 are $1, 1+2u, 1+4u, 1+6u, \ldots, 2-2u, 2$. We call *binade* an interval of the form $[2^k, 2^{k+1}),\ k \in \mathbb{Z}$. Table 1 reminds the values of p, $e_{\min}$ and $e_{\max}$ for the binary interchange formats of IEEE 754-2019 up to 128 bits [21] and the more recent bfloat16 format [16], and Table 2 summarizes our notation for the important FP parameters.

Table 1: Main parameters of the binary interchange formats of size up to 128 bits specified by the 754-2019 standard [21], and the bfloat16 format [16].

Name	binary16	binary32 (basic)	binary64 (basic)	binary128 (basic)	bfloat16
Former name	N/A	single precision	double precision	N/A	N/A
p	11	24	53	113	8
$e_{\max}$	$+15$	$+127$	$+1023$	$+16383$	$+127$
$e_{\min}$	-14	-126	-1022	-16382	-126

[1] Our proofs, however, are based on the assumptions that $\mathrm{RN}(-x) = -\mathrm{RN}(x)$, and that if $k \in \mathbb{Z}$ and if both x and $2^k x$ are in the normal domain, then $\mathrm{RN}(2^k x) = 2^k \mathrm{RN}(x)$.

Table 2: Notation for the important FP parameters

Notation	numerical value	explanation
Ω	$2^{e_{\max}} \cdot (2 - 2^{-p+1})$	largest finite FPN
α	$2^{e_{\min}-p+1}$	smallest positive FPN
$2^{e_{\min}}$	$2^{e_{\min}}$	smallest positive normal FPN (underflow threshold)
$\mathrm{succ}(t)$	$\mathrm{succ}(t)$	floating-point successor of t.
η	$2^{(e_{\min}+p)/2}$	the square of a FPN $\geq \eta$ (unless it overflows) is exactly representable as the sum of two FPNs
u	2^{-p}	roundoff error unit (the relative error due to rounding to nearest a number between $2^{e_{\min}}$ and Ω is $\leq u/(1+u) < u$)
$\mathrm{ulp}(x)$ $(x \in \mathbb{R}, x \neq 0)$	$2^{\max\{\lfloor \log_2 \lvert x \rvert \rfloor, e_{\min}\}-p+1}$	unit in the last place (distance between consecutive FPNs in the neighborhood of x)

A computed result is *faithfully rounded* if $i)$ it is equal to the exact result if this one is a FP number, and $ii)$ it is one of the two FP numbers that surround the exact result otherwise. This implies (barring overflow) that the returned result is within one ulp of the exact result from the exact result.

Figure 1 illustrates the notions presented in this section.

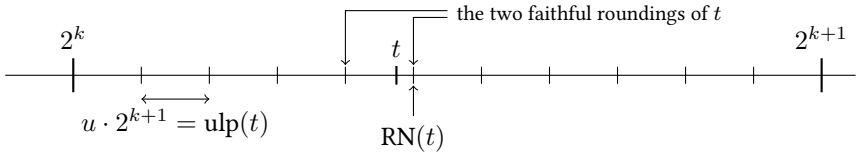

Figure 1: The floating-point numbers between 2^k and 2^{k+1} (assuming $e_{\min} \leq k < e_{\max}$)

1.3 Double-word and pair arithmetics

Evaluating norms with an accuracy significantly better than that of the naive algorithm may require representing intermediate results with a precision higher than the working FP precision. This can be done by representing these intermediate results by a pair of FP numbers. For instance, Graillat et al. [13] recently published an algorithm that computes *faithfully rounded* norms. To achieve that goal, they use double-word (a.k.a. "double-double") arithmetic in their intermediate calculations. Lange and

Rump [38] recently defined a "pair arithmetic" (which is a somehow "relaxed" version of double-word arithmetic), and showed how it can be used, under some conditions, to obtain faithfully rounded results in floating-point arithmetic. The algorithms used to perform operations with these arithmetics are usually based on the three basic "building blocks" presented in Section 1.3.1: Fast2Sum, 2Sum, and Fast2Mult. However, it is possible that new operations recently introduced in the IEEE 754 Standard for FP arithmetic [21] and briefly presented in Section 1.3.2 replace these building blocks in a near future.

1.3.1 The basic building blocks: Fast2Sum, 2Sum, and Fast2Mult

Algorithm 1 – Fast2Sum(a, b). The Fast2Sum algorithm [11]. It takes 3 FP operations.

$s \leftarrow \mathrm{RN}(a + b)$
$z \leftarrow \mathrm{RN}(s - a)$
$t \leftarrow \mathrm{RN}(b - z)$

If $|a| \geq |b|$, unless overflow occurs, the two FP numbers s and t returned by Algorithm 1 satisfy $s + t = a + b$. Since s is the result of the conventional floating-point addition of a and b, t is the error of that addition. Also, if the first operation does not overflow, the other operations cannot overflow [6]. For that algorithm, underflow is harmless (this is an immediate consequence of Lemma 1.3).

Algorithm 2 – 2Sum(a, b). The 2Sum algorithm [30, 28]. It takes 6 FP operations.

$s \leftarrow \mathrm{RN}(a + b)$
$a' \leftarrow \mathrm{RN}(s - b)$
$b' \leftarrow \mathrm{RN}(s - a')$
$\delta_a \leftarrow \mathrm{RN}(a - a')$
$\delta_b \leftarrow \mathrm{RN}(b - b')$
$t \leftarrow \mathrm{RN}(\delta_a + \delta_b)$

Unless overflow occurs, the two FP numbers s and t returned by Algorithm 2 satisfy $s + t = a + b$: this algorithm returns the same result as Algorithm 1 without any condition on a and b. On the other hand, it is slightly less overflow-proof: If the first operation does not overflow and if $|a| < \Omega$ then the other operations cannot overflow [6]. Underflow is harmless.

Algorithm 3 – Fast2Mult(a, b)**.** The Fast2Mult algorithm (see for instance [27, 35, 31]). It requires the availability of a fused multiply-add (FMA) instruction for computing $\mathrm{RN}(ab - \pi_h)$.

$\pi_h \leftarrow \mathrm{RN}(a \cdot b)$
$\pi_\ell \leftarrow \mathrm{RN}(a \cdot b - \pi_h)$

In this paper, Algorithm Fast2Mult will be used for expressing the square of a FP number a as a double-word number. One should keep in mind that, barring overflow, the condition for that algorithm to guarantee that $\pi_h + \pi_\ell = a \cdot b$ is stronger than just requiring the absence of underflow in the first multiplication. Several slightly different conditions appear in the literature (see [4] for a necessary and sufficient condition). One can show (see for instance [5]) that if $2^{e_{\min}+p} \leq |a \cdot b|$, then $\pi_h + \pi_\ell = a \cdot b$. In the case of the computation of the square of a, this condition becomes

$$|a| \geq \eta = 2^{(e_{\min}+p)/2}. \tag{4}$$

Algorithm Fast2Mult requires the availability of an FMA instruction. Without an FMA instruction, the calculation of (π_h, π_ℓ) remains possible, but at a significantly higher cost (17 floating-point operations instead of 2 [11]).

1.3.2 An alternative: the new "augmented" arithmetic operations

The latest release of the IEEE Standard for Floating-Point Arithmetic, published in 2019 [21], specifies new "augmented" operations, called *augmentedAddition*, *augmentedSubtraction*, and *augmentedMultiplication* (history and motivation are presented in [36]). These operations use a new "rounding direction", round-to-nearest *ties-to-zero*, denoted RN_0 in this paper, that satisfies [21]:

> $\mathrm{RN}_0(t)$ (where t is a real number) is the FP number nearest t. If the two nearest FP numbers bracketing t are equally near, $\mathrm{RN}_0(t)$ is the one with smaller magnitude. If $|t| > \Omega + 2^{e_{\max}-p}$ then $\mathrm{RN}_0(t) = \infty$, with the same sign as t.

The augmented operations are defined as follows [21, 36]:

- *augmentedAddition*(x, y) delivers (a_0, b_0) such that $a_0 = \mathrm{RN}_0(x+y)$ and, when $a_0 \notin \{\pm\infty, \mathrm{NaN}\}$, $b_0 = (x + y) - a_0$. When $b_0 = 0$, it is required to have the same sign as a_0;
- *augmentedSubtraction*(x, y) is augmentedAddition$(x, -y)$;
- *augmentedMultiplication*(x, y) delivers (a_0, b_0) such that $a_0 = \mathrm{RN}_0(x \cdot y)$ and, where $a_0 \notin \{\pm\infty, \mathrm{NaN}\}$, $b_0 = \mathrm{RN}_0((x \cdot y) - a_0)$. When $(x \cdot y) - a_0 = 0$, the floating-point number b_0 (equal to 0) is required to have the same sign as a_0.

As we are writing these lines, no fast hardware implementation of these operations is offered on widely available platforms. When this happens, in the algorithms presented in this paper, it can be worth replacing 2Sum and Fast2Sum by augmentedAddition, and replacing Fast2Mult by augmentedMultiplication.

1.3.3 Double-word arithmetic

We define a double-word number as follows

Definition 1.2. *A* double-word *(DW) number x is the unevaluated sum $x_h + x_\ell$ of two floating-point numbers x_h and x_ℓ such that*

$$x_h = \mathrm{RN}(x).$$

Double-word arithmetic goes back to the seminal work of Dekker [11]. Algorithms for manipulating double-word numbers have been published and analyzed by Li et al. [29], Hida, Li and Bailey [18, 17], Joldes et al. [26], Muller and Rideau [33]. Let us now give a two classical DW algorithms. Some new results on DW arithmetic necessary for this study will be given later on, in Section 3.

Let us first consider the addition of a DW number and a FP number. Consider Algorithm 4 below. It was implemented in the QD library [18].

Algorithm 4 – DWPlusFP(x_h, x_ℓ, y). Algorithm for computing $(x_h, x_\ell) + y$ in binary, precision-p, floating-point arithmetic, implemented in the QD library. The number $x = (x_h, x_\ell)$ is a DW number (i.e., it satisfies Definition 1.2). It takes 10 FP operations.

1: $(s_h, s_\ell) \leftarrow \mathrm{2Sum}(x_h, y)$
2: $v \leftarrow \mathrm{RN}(x_\ell + s_\ell)$
3: $(z_h, z_\ell) \leftarrow \mathrm{Fast2Sum}(s_h, v)$
4: **return** (z_h, z_ℓ)

That algorithm was analyzed by Joldes et al. [26]. They found that its relative error

$$\left| \frac{(z_h + z_\ell) - (x + y)}{x + y} \right|$$

is bounded by

$$\frac{2 \cdot u^2}{1 - 2u} = 2u^2 + 4u^3 + 8u^4 + \cdots \tag{5}$$

They also showed that the bound (5) is asymptotically optimal, by exhibiting "generic" (i.e., parameterized by the precision p) input values for which the ratio between the attained relative error and the bound goes to 1 as p goes to infinity.

Now, let us turn to the addition of two DW numbers. Algorithm 5 below was first given by Dekker [11], under the name of *add2*. It was implemented by Hida, Li, and Bailey in the QD library [18] under the name of "sloppy addition". The reason for that name is that if the input operands have different signs, the relative error can be arbitrarily large. We will *not* use that algorithm, but since it is the algorithm used by Graillat et al to perform their summations [13], we briefly present it and some of its properties for the sake of completeness and for helping to compare our solutions.

Algorithm 5 – SloppyDWPlusDW$(x_h, x_\ell, y_h, y_\ell)$. "Sloppy" calculation of $(x_h, x_\ell) + (y_h, y_\ell)$ in binary, precision-p, floating-point arithmetic. It takes 11 FP operations.

1: $(s_h, s_\ell) \leftarrow \mathsf{2Sum}(x_h, y_h)$
2: $v \leftarrow \mathsf{RN}(x_\ell + y_\ell)$
3: $w \leftarrow \mathsf{RN}(s_\ell + v)$
4: $(z_h, z_\ell) \leftarrow \mathsf{Fast2Sum}(s_h, w)$
5: **return** (z_h, z_ℓ)

If the inputs operands x_h and y_h have the same sign (which is of course the case when summing squares), the relative error of Algorithm 5 is bounded by $3u^2$ [13]. This bound is asymptotically optimal. Consider:

$$\begin{cases} x_h &= 1 + 2u, \\ x_\ell &= -u + u^2, \\ y_h &= 9u, \\ y_\ell &= -6u^2 - 8u^3, \end{cases}$$

for which the double-word number returned by Algorithm 5 is equal to $1 + 10u - 8u^2$ and the exact sum is equal to $1 + 10u - 5u^2 - 8u^3$, resulting in a relative error

$$u^2 \cdot \frac{3 - 8u}{1 + 10u - 5u^2 - 8u^3} = 3u^2 - 38u^3 + \mathcal{O}(u^4).$$

1.3.4 Lange and Rump's pair arithmetic

Lange and Rump [38] recently defined a "pair arithmetic" (which is a somehow "relaxed" version of double-word arithmetic), and showed how it can be used, under some conditions, to obtain faithfully rounded results in floating-point arithmetic.

Rewritten with our notation, the pair algorithms used by Lang and Rump [38] for addition and square root are the following.

Algorithm 6 – Pair_addition$(x_h, x_\ell, y_h, y_\ell)$. Lange and Rump's calculation of $(x_h, x_\ell) + (y_h, y_\ell)$ in binary, precision-p, floating-point arithmetic. It is Algorithm 5 without the last "renormalization". It takes 8 FP operations.

1: $(s_h, s_\ell) \leftarrow \mathsf{2Sum}(x_h, y_h)$
2: $v \leftarrow \mathsf{RN}(x_\ell + y_\ell)$
3: $w \leftarrow \mathsf{RN}(s_\ell + v)$
4: **return** (s_h, w)

Algorithm 7 – Pair_sqrt(x_h, x_ℓ). Lange and Rump's calculation of the square-root of (x_h, x_ℓ) in binary, precision-p, floating-point arithmetic. It is Algorithm 8 without the last "renormalization". It takes 5 FP operations (counting the square root as one).

1: $s_h \leftarrow \text{RN}(\sqrt{x_h})$
2: $\rho_1 \leftarrow \text{RN}(x_h - s_h^2)$ (with an FMA instruction)
3: $\rho_2 \leftarrow \text{RN}(x_\ell + \rho_1)$
4: $s_\ell \leftarrow \text{RN}(\rho_2/(2 \cdot s_h))$
5: **return** (s_h, s_ℓ)

These algorithms are similar to the DW algorithms presented in this paper, with the difference that they avoid the last "renormalizing" Fast2Sum operation. This makes them significantly faster, but this may sometimes make them less accurate, especially when cancellations occur. When adding squares, however, there are no cancellations: this makes Lange and Rump's pair arithmetic a very good candidate.

1.4 Some results useful later on

The following Lemma is frequently used to show that some calculations remain valid even when operands are below the underflow threshold (the proof is straightforward).

Lemma 1.3 (Hauser Lemma [15])**.** *If x and y are floating-point numbers, and if the number* $\text{RN}(x + y)$ *is subnormal, then $x + y$ is a floating-point number, which implies* $\text{RN}(x + y) = x + y$.

For bounding the error committed during the evaluation of a sum of squares, we will use the following lemma, which is a direct consequence of Lemma 2.1 in [37], due to Lange and Rump.

Lemma 1.4 (Lange and Rump [37])**.** *Let $\mathbb{F}$ be an arbitrary subset of $\mathbb{R}$ and let $\tilde{+}$ be an operation in $\mathbb{F}$ with the only assumption that*

$$\forall a, b \in \mathbb{F}, |(a\tilde{+}b) - (a + b)| \leq \min\{|a|, |b|\}.$$

Let $x_1, x_2, \cdots, x_n$ be elements of $\mathbb{F}$ and define numbers s_i and ϵ_i as follows:

$$\begin{array}{rcl} s_1 & = & x_1, \\ s_i & = & x_i\tilde{+}s_{i-1} = (x_i + s_{i-1})(1 + \epsilon_i) \text{ for } i = 2, \ldots, n. \end{array}$$

We have

$$\left| s_n - \sum_{i=1}^{n} x_i \right| \leq \sum_{i=2}^{n} |\epsilon_i| \cdot \sum_{i=1}^{n} |x_i|.$$

For computing square roots in double-word arithmetic, we will need the following result, due to Boldo and Daumas [4].

Lemma 1.5 (Exact representation of the square root remainder. This is Theorem 5 of [4], restricted to binary arithmetic and rewritten with our notation)**.** *In binary,*

precision-p, FP arithmetic, let $s = \mathrm{RN}(\sqrt{x})$, where x is a FP number. The correcting term

$$x - s^2$$

is a FPN if and only if there exists a pair of integers (m, e) (with $|m| \leq 2^p - 1$) such that $s = m \cdot 2^{e-p+1}$ and $2e \geq e_{\min} + p - 1$.

1.5 Aim and organization of this paper

Ideally, one would like to always return *correctly rounded* results (i.e., the computed result is the floating-point nearest to the exact result, which implies that the error is less than or equal to 0.5 ulp of the exact result). This seems difficult to guarantee without making the calculation much slower. However, we are going to show that a modification of Graillat et al.'s algorithm [13] can be used to always obtain a maximum error *very slightly* above 0.5 ulp. This means that we will *almost always* obtain the correctly rounded result, except in rare cases when the exact norm is very near the exact middle of two consecutive floating-point numbers. Beyond the gain in terms of accuracy, this will enhance the reproducibility of the calculations, which is becoming an important issue [12].

The sequel of the paper is organized as follows. Section 2 presents the algorithms one can find in the literature. More precisely, in Subsection 2.1 we quickly review the classical solutions suggested for avoiding spurious overflows and underflows, Subsection 2.2 briefly presents the use by Graillat et al. of double-word arithmetic for obtaining more accurate results, and in Subsection 2.3 we consider applying a recent result by Lange and Rump [38] to obtain faithfully rounded norms in pair arithmetic. In Section 3 we give some new results on double-word arithmetic that will be helpful for our study. In particular we give a new bound (that takes into account the fact that we manipulate positive numbers) for and existing addition algorithm, and we present and analyze an algorithm that computes the square-root of a double-word number. Since the proof of that square-root algorithm is long and rather complex, and since the error bound of our euclidean norm algorithm derives from the error bound of the square-root algorithm, to give more confidence on our result, we have formally proven the square-root algorithm, using the Coq proof assistant (see for instance [7]). This part of the paper continues the work undertaken by two of us on the formal proof of double-word algorithms [34]. Section 4 presents our algorithms for computing Euclidean norms. We first assume that no underflow or overflow occurs in Subsection 4.1, then we deal with the general case in Subsection 4.2.

Our solution builds on Graillat et al.'s solution [13], with the following differences:

- we introduce a more accurate algorithm for summing the squares of the terms a_i in DW arithmetic;
- once we have obtained an approximation to the sum of squares as a DW number, we directly take its square root using a specific DW to FP square-root algorithm, whereas Graillat et al "convert" the sum of squares to floating-point (by just retaining the most significant part) and take its square root using the conventional FP square root;

- we use different comparison constants for preventing underflows and overflows.

2 Conventional solutions for computing Euclidean norms

2.1 Avoiding spurious overflows/underflows

Several solutions have been suggested for dealing with spurious underflows and overflows when computing Euclidean norms. A first solution [20] would be to use the exception-handling mechanism provided by the IEEE 754 Standard for FP arithmetic: one could first use the naive method (i.e., straightforward implementation of (1)), check if an underflow or overflow exception occurred, and use a more sophisticated method only in that case. It is unlikely that this approach will allow good performance on modern highly pipelined processors. All other approaches consist in *scaling* the terms a_i, i.e. we multiply or divide them by one (or several) constant(s) such that computing sums of squares of the scaled values is overflow-free, and that underflow is either impossible or harmless (a good presentation, along with comparisons of existing Fortran codes can be found in [14]). A straightforward choice is to scale all values by the factor $\max|a_i|$, i.e., to evaluate the norm as

$$\max|a_i| \times \sqrt{\sum_{k=0}^{n-1}\left(\frac{a_k}{\max|a_i|}\right)^2}.$$

That approach has several drawbacks:

- it requires two passes over the data (finding the maximum of the $|a_i|$ takes time and no computation can start before that max is found);
- it requires divisions, and FP divisions are in general significantly slower than FP additions and multiplications;
- multiplying and dividing by $\max|a_i|$ are, in general, non-exact operations, which leads to a slightly larger final error than the error of directly using (1) when no underflow/overflow occurs.

An already better approach (at least in terms of accuracy, delay may be another matter) consists in choosing a scale factor equal to a power of 2 close to $\max|a_i|$, obtained for instance by the means of the **scaleB** and **logB** functions[2] specified by the IEEE-754 Standard [21] (when an efficient implementation of these functions is available. If this is not the case, a possible turnaround is suggested in [23, Theorem IV.2]).

[2] **scaleB**(x, k) returns (in a binary format, which is the case considered in this paper) $x \cdot 2^k$ (where x is a FP number and k is an integer), and **logB**(x) returns (in a binary format) $\lfloor \log_2 |x| \rfloor$ (where x is a FP number). In the C programming language, these functions are called **scalebn** and **logb** (resp. **scalebnf** and **logbf**) for binary64/double precision operands (resp. binary32/single precision) operands.

Higham [19, Pages 500 and 507] attributes to Hammarling a smart algorithm that consists in *dynamically* scaling the data. We start from $s_0 = 1$ and $t_0 = |x_0|$. At step i of the algorithm, we have already computed

$$s_{i-1} = \sum_{k=0}^{i-1} (x_k/t_{i-1})^2,$$

where s_{i-1} is the current scaled sum and t_{i-1} is the current value of the scale factor. If $|x_i| \leq t_{i-1}$ then $s_i = s_{i-1} + (x_i/t_{i-1})^2$ and the scale factor does not change: $t_i = t_{i-1}$. If $|x_i| > t_{i-1}$ then we need to update the scale factor. We compute

$$s_i = 1 + s_{i-1} \cdot (t_{i-1}/x_i)^2,$$

and we replace the scale factor by $|x_i|$: $t_i = |x_i|$. After this, one easily checks that $s_i = \sum_{k=0}^{i} (x_k/t_i)^2$. The final result is $t_{n-1}\sqrt{s_{n-1}}$. With this method, a single pass over the data suffices. That algorithm was implemented in the Lapack [14] package released by `netlib`[3]. In Section 5.2 it will be called the "Netlib algorithm". However, the number of scale factor updates may be large: up to $n-1$ updates if the $|x_i|$'s are in increasing order (although its average value is around $\log(n)$), which may result in delays and additional rounding errors due to (in general, nonexact) multiplications and divisions. Again, a slight improvement in terms of accuracy can consist in choosing, when $|x_i| > t_{i-1}$, a value t_i equal to a power of two close to (and preferably above) $|x_i|$, and then taking $s_i = (x_i/t_i)^2 + s_{i-1} \cdot (t_{i-1}/t_i)^2$.

With the methods examined so far, a scaling is applied even when not needed.

Blue [2] takes a decisive step by suggesting to split the input numbers into 3 classes (that we will call TINY, MED, and BIG), depending on their order of magnitude:

- numbers of the MED Class can be squared, and their squares can be accumulated, without underflows or overflows. A FP number a_i is in the MED Class if[4]

  $$a_i = 0 \text{ or } \mathrm{minmed} \leq |a_i| \leq \mathrm{maxmed},$$

  where the choice of minmed and maxmed depends on the parameters (p, $e_{\min}$ and $e_{\max}$) of the FP arithmetic, and on the largest value of n, say $n_{\max}$, for which a correct behavior is to be guaranteed. We compute

  $$S_{\mathrm{med}} = \sum_{a_i \in \mathrm{MED}} a_i^2;$$

- numbers of the BIG Class must be "scaled down" to make sure that we can accumulate their squares safely (i.e., without overflow). A FP number a_i is in the BIG Class if

  $$\mathrm{maxmed} < |a_i|.$$

[3] `www.netlib.org`

[4] As a matter of fact, 0 could be as well in the TINY class instead of the MED class if this simplifies the programming. In any case, accumulating 0 in one part or another one will of course not change the result. However, depending on the underlying computer architecture, the choice may have a significant impact on performance, due to branch prediction issues.

All numbers of the BIG Class are multiplied by *the same* predefined constant t_{big}, chosen equal to a power of 2 (to make the multiplication errorless), and such that for $a_i \in \mathrm{BIG}$, $t_{\mathrm{big}} \cdot a_i \in \mathrm{MED}$. We compute

$$S_{\mathrm{big}} = \sum_{a_i \in \mathrm{BIG}} (t_{\mathrm{big}} \cdot a_i)^2 .$$

(usual presentation of the method is with divisions by constants, of course when actually implementing it, multiplication is preferable for performance reasons)

- numbers of the TINY Class must be "scaled up" to make sure that we can compute their squares safely (i.e., without underflow: each square must be larger than the subnormal threshold[5] $2^{e_{\min}}$). A FP number a_i is in the TINY Class if

$$|a_i| < \mathrm{minmed} \text{ and } a_i \neq 0.$$

All numbers of the TINY Class are multiplied by *the same* constant t_{tiny}, chosen equal to a power of 2, and such that for $a_i \in \mathrm{TINY}$, $t_{\mathrm{tiny}} \cdot a_i \in \mathrm{MED}$. We compute

$$S_{\mathrm{tiny}} = \sum_{a_i \in \mathrm{TINY}} (t_{\mathrm{tiny}} \cdot a_i)^2 .$$

Let us summarize the various constraints that the parameters minmed, maxmed, t_{big}, and t_{tiny} must satisfy:

$$\mathrm{minmed}^2 \geq 2^{e_{\min}}, \tag{6a}$$

$$n_{\max} \cdot \mathrm{maxmed}^2 \cdot (1+\rho) < \Omega + \frac{1}{2}\mathrm{ulp}(\Omega) = 2^{e_{\max}+1} - 2^{e_{\max}-p}, \tag{6b}$$

$$\mathrm{maxmed} \cdot t_{\mathrm{big}} \geq \mathrm{minmed}, \tag{6c}$$

$$\Omega \cdot t_{\mathrm{big}} \leq \mathrm{maxmed}, \tag{6d}$$

$$\mathrm{minmed} \cdot t_{\mathrm{tiny}} \leq \mathrm{maxmed}, \tag{6e}$$

$$\alpha \cdot t_{\mathrm{tiny}} \geq \mathrm{minmed}, \tag{6f}$$

where Ω and α are defined in (2) and (3) respectively, and ρ is a bound on the relative error of the algorithm used for computing the sum of squares in MED. Assuming that maxmed and $n_{\max}$ are powers of 2, Eq. (6b) can be replaced by

$$n_{\max} \cdot \mathrm{maxmed}^2 < 2^{e_{\max}+1} - 2^{e_{\max}-p}.$$

Later on, when we use double-word arithmetic, (6a) will need to be replaced by the stronger condition $\mathrm{minmed} \geq \eta$.

In Blue's original algorithm [2], the three terms S_{tiny}, S_{med}, and S_{big} are all computed, in three accumulators. However, if BIG is nonempty, provided that the ratio $\mathrm{maxmed}/\mathrm{minmed}$ is large enough, the value of S_{tiny} has negligible influence on the final result. Graillat et al's algorithm [13] and the variant of Blue's algorithm presented by Hanson and Hopkins in [14] take this into account and use two accumulators only: as soon as an element of BIG is met, we no longer need to accumulate elements of TINY.

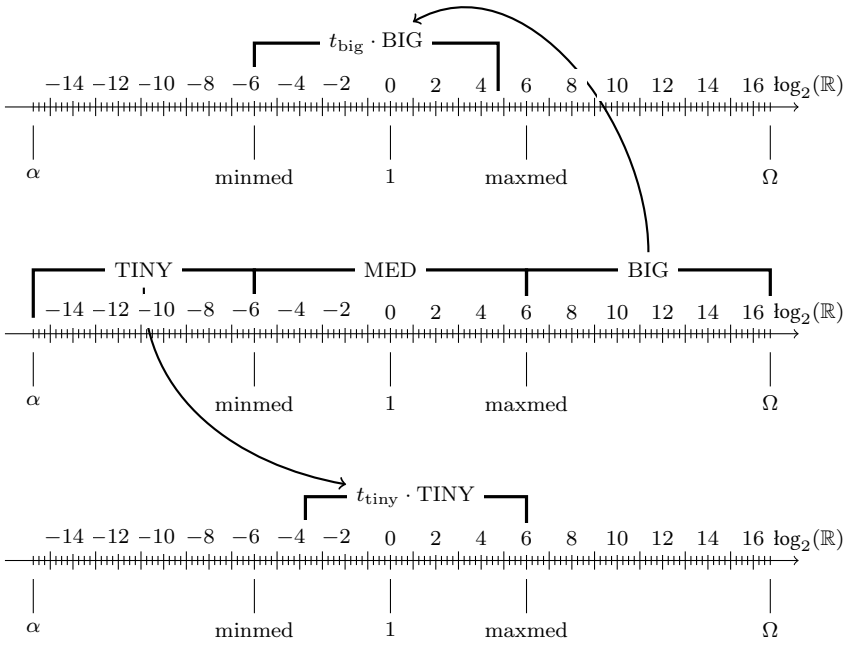

Figure 2: The splitting and the scaling of the FP numbers into 3 classes TINY, MED, and BIG.

Figure 2 illustrates this splitting of the FP numbers into three classes.
The norm

$$N = \sqrt{\sum_{i=0}^{n-1} a_i^2}$$

is equal to

$$\sqrt{S_{\rm tiny}/t_{\rm tiny}^2 + S_{\rm med} + S_{\rm big}/t_{\rm big}^2},$$

but obviously that formula cannot be employed since $S_{\rm big}/t_{\rm big}^2$—and, more rarely, the sum—could overflow, and $S_{\rm tiny}/t_{\rm tiny}^2$ could underflow. Blue suggests obtaining N as follows:

1. if BIG and TINY are empty (i.e., if $S_{\rm big} = S_{\rm tiny} = 0$), then $\sqrt{S_{\rm med}}$ is returned;
2. otherwise, if BIG is nonempty then if $\sqrt{S_{\rm big}}$ is larger than the precomputed constant $\Omega \cdot t_{\rm big}$, $+\infty$ is returned[6], otherwise we define

$$\begin{aligned} y_{\rm min} &= \min\left\{\sqrt{S_{\rm med}}, \tfrac{1}{t_{\rm big}}\sqrt{S_{\rm big}}\right\}, \\ y_{\rm max} &= \max\left\{\sqrt{S_{\rm med}}, \tfrac{1}{t_{\rm big}}\sqrt{S_{\rm big}}\right\}. \end{aligned}$$

   and we go to step 4.
3. if BIG is empty and TINY is nonempty, we define

$$\begin{aligned} y_{\rm min} &= \min\left\{\sqrt{S_{\rm med}}, \tfrac{1}{t_{\rm tiny}}\sqrt{S_{\rm tiny}}\right\}, \\ y_{\rm max} &= \max\left\{\sqrt{S_{\rm med}}, \tfrac{1}{t_{\rm tiny}}\sqrt{S_{\rm tiny}}\right\}. \end{aligned}$$

   and we go to step 4
4. if $y_{\rm min} < \sqrt{u} \cdot y_{\rm max}$ we return $y_{\rm max}$, otherwise we return

$$y_{\rm max} \cdot \left(1 + \left(\frac{y_{\rm min}}{y_{\rm max}}\right)^2\right)^{1/2}. \tag{7}$$

The additional division and square root that appear in (7) were perhaps unavoidable in the pre-IEEE-754 era. However, they involve additional delay and rounding error in the calculation. Graillat et al [13] also split the input values into 3 classes, and give a simpler solution for the final reconstruction of the norm N from $S_{\rm big}$, $S_{\rm med}$, and $S_{\rm tiny}$. Their work uses double-word arithmetic (see Section 1.2) for accumulating the sums $S_{\rm big}$, $S_{\rm med}$, and $S_{\rm tiny}$, so the context is slightly different (we will come to that later on in this paper), but let us momentarily present their solution for avoiding under/overflow in the context of simple FP numbers. They choose:

$$t_{\rm big} = 2^{-E}, \tag{8a}$$

[5]We will need a stronger condition when we represent numbers in double-word arithmetic.
[6]One can check that with an IEEE-754 compliant system, that test is not needed.

$$t_{\text{tiny}} = 2^E, \text{ with} \tag{8b}$$

$$E = 2 \times \left\lceil \frac{1}{2} \cdot \left\lceil \frac{e_{\max} - e_{\min} + p}{3} \right\rceil \right\rceil, \tag{8c}$$

$$\text{minmed} = 2^{e_{\max}+1-2E}, \tag{8d}$$

$$\text{maxmed} = 2^{e_{\max}+1-E}, \tag{8e}$$

so that $t_{\text{big}} = 1/t_{\text{tiny}}$ and $\text{minmed} = t_{\text{big}} \cdot \text{maxmed}$. The choice (8c) indicates that they obtain TINY, MED, and BIG by splitting the exponent range of the FP format into three parts of approximately the same size. Assume $n < 1/u$. Graillat et al. show that:

- If BIG is nonempty, we can neglect the elements of TINY, so that we need to compute
$$\sqrt{S_{\text{med}} + \frac{S_{\text{big}}}{t_{\text{big}}^2}}.$$
This can be done without under/overflows as follows.
  - If $S_{\text{big}} \geq \text{minmed}^2/u^3$ (i.e., $\frac{S_{\text{big}}}{t_{\text{big}}^2} \geq \frac{\text{maxmed}^2}{u^3} \geq n \cdot \frac{\text{maxmed}^2}{u^2}$) or $S_{\text{med}} \leq \text{maxmed}^2 u^2$ then S_{med} is negligible in front of $\frac{S_{\text{big}}}{t_{\text{big}}^2}$ and we can return $\frac{1}{t_{\text{big}}}\sqrt{S_{\text{big}}} = t_{\text{tiny}}\sqrt{S_{\text{big}}}$.
  - if $S_{\text{med}} > \text{maxmed}^2 u^2$ and $S_{\text{big}} < \text{minmed}^2/u^3$ then we can compute
$$\frac{S_{\text{big}}}{t_{\text{big}}} + t_{\text{big}} S_{\text{med}} = t_{\text{tiny}} S_{\text{big}} + t_{\text{big}} S_{\text{med}}$$
without overflow or underflow, so that we can return
$$\frac{1}{\sqrt{t_{\text{big}}}} \cdot \sqrt{t_{\text{tiny}} S_{\text{big}} + t_{\text{big}} S_{\text{med}}}, \tag{9}$$
and the (precomputed) constant $\frac{1}{\sqrt{t_{\text{big}}}} = \sqrt{t_{\text{tiny}}}$ is a power of 2 since (8a) and (8c) imply that t_{big} is an *even* power of 2, so all multiplications in (9) are exact.
- If BIG is empty, we need to compute
$$\sqrt{S_{\text{med}} + \frac{S_{\text{tiny}}}{t_{\text{tiny}}^2}} = \frac{1}{t_{\text{tiny}}}\sqrt{S_{\text{tiny}} + \frac{S_{\text{med}}}{t_{\text{big}}^2}}.$$
This can be done as follows.
  - If $S_{\text{med}} \geq \text{minmed}/u^3$ or $S_{\text{tiny}} \leq \text{maxmed}^2 u^2$ we can return $\sqrt{S_{\text{med}}}$;
  - otherwise, we can safely compute the desired result as
$$\frac{1}{\sqrt{t_{\text{tiny}}}} \cdot \sqrt{t_{\text{tiny}} S_{\text{med}} + t_{\text{big}} S_{\text{tiny}}}$$
and the term $\frac{1}{\sqrt{t_{\text{tiny}}}} = \sqrt{t_{\text{big}}}$ is a power of 2.

2.2 Using double-word numbers to improve accuracy: Graillat et al.'s solution

Let us now temporarily put aside the problem of avoiding spurious under/overflow, and let us focus on the need for accurately computing the sum of squares and the square root involved in the computation of a Euclidean norm. The goal of Graillat et al [13] was to guarantee *faithful rounding* of the Euclidean norm N (1). For that purpose, to save accuracy as much as possible, they compute the sum of squares $\sum_{i=0}^{n-1} a_i^2$ in double-word arithmetic. They first express the squares of the FP numbers a_i as DW numbers using Algorithm 3 (Fast2Mult). Then they sum the obtained DW numbers using a double-word addition algorithm, Algorithm 5 (SloppyDWPlusDW). As they mention, that summation is easily parallelizable. The obtained result is a double-word approximation (S_h, S_ℓ) to the sum of squares.

After this, they take the square root of S_h, using the correctly rounded square root instruction that is available on all IEEE 754 compliant systems. They show that, under reasonable conditions, that square root is a faithful rounding of the norm. More precisely, the condition they give on n for their algorithm to return a faithful result is

$$n < \frac{1}{24u + u^2}, \tag{10}$$

i.e., $n \leq 699050$ in binary32 arithmetic, and $n \leq 3.752 \times 10^{14}$ in binary64 arithmetic.

Assuming sequential addition of the squares of the a_is, and assuming that all numbers are in the MED class (i.e., no scaling is needed), Graillat et al's algorithm uses $13n - 10$ floating-point operations.

Let us mention, however, that the choice of "dropping" S_ℓ is tantamount to losing a non-negligible information on the sum of squares.

Incidentally, Graillat et al make a little and reasonably harmless mistake: they did not realize that the value of minmed they choose (called β_0 in their paper), given in Eq. (8d), is less than η in binary32 arithmetic (it, is however, larger than η in binary64 and binary128 arithmetics). Hence, in very rare cases (computations of norms in binary32 arithmetic with all scaled operands slightly over minmed), some squares will not be expressed exactly as DW numbers. Whether this can lead to errors slightly larger than the claimed bound remains an open question.

They target faithful rounding (i.e., error less than 1 ulp). We have a different goal in mind: we wish to achieve a final error extremely close to 0.5 ulp of the exact result, i.e., we wish to almost always provide a correctly rounded result. This will be done by keeping the sum (S_h, S_ℓ) of the squares in DW arithmetic, and using an algorithm that computes the square root of a DW number (Algorithm 9). We will also compute (S_h, S_ℓ) more accurately, by using a different summation scheme, based on Algorithm 4.

2.3 An alternative: computing norms with pair arithmetic

In [38], Lange and Rump give conditions for their pair arithmetic to return *faithfully rounded* results. We have applied Theorems 4.2 and 5.4 of [38] to two cases: the computation of N when the squares are added sequentially using the Pair_addition

algorithm, and the same computation where the a_i^2 are added *blockwise*: we divide them in k blocks of m terms, with $km = n$, we first add all the terms of each block together, and then we add the k obtained sums. We obtain the following results:

- with the sequential summation, we will obtain a faithfully rounded result if

$$\left\lceil \frac{4}{5}n + \frac{5}{4} \right\rceil \leq \frac{1}{\sqrt{2u - u^2}} - 2;$$

- with the blockwise summation , we will obtain a faithfully rounded result if

$$\left\lceil \frac{4}{5}(m + k - 1) + \frac{5}{4} \right\rceil \leq \frac{1}{\sqrt{2u - u^2}} - 2.$$

Table 3 presents the maximum possible values of n allowed by these conditions, for binary32 and binary64 arithmetics. For the blockwise algorithm we have chosen the "optimal" choice $k = m = \lceil \sqrt{n} \rceil$.

Table 3: Maximum values of n for which a pair-arithmetic implementation is guaranteed to return a faithful result. For the blockwise algorithm we have chosen $k = m = \lceil \sqrt{n} \rceil$.

format	sequential summation	blockwise summation
binary32	3615	$3,268,864$
binary64	$83,886,075$	$\approx 8.796 \times 10^{14}$

Assuming sequential addition of the squares of the a_is, and assuming that all numbers are in the MED class (i.e., no scaling is needed), computing a norm in pair arithmetic uses $10n-3$ floating-point operations. Of course, exactly as for conventional or double word arithmetics, one may need scalings to avoid spurious under/overflows.

3 Some results on double-word arithmetic

In this section, let us give a few new results on double-word arithmetic that can be useful for accurately computing norms. All the results of this section have been formally certified using the Coq proof assistant [7] and the Flocq [8, 9] library.

3.1 Properties of DWPlusFP

First, the relative error bound (5) on Algorithm DWPlusFP (Alg. 4) was given in [26] assuming input numbers of *arbitrary sign*. One may wonder if, when the operands have the same sign (i.e., when x_h and y have the same sign), we can obtain a better error bound. This would for instance be useful for summing squares, which is the main step of the computation of Euclidean norms. Indeed, we have,

[7] `http://coq.inria.fr/`

Theorem 3.1. *If $x = (x_h, x_\ell)$ is a nonnegative double-word number and y is a nonnegative FP number, then the relative error of Algorithm 4 is bounded by u^2. That bound is asymptotically optimal.*

Proof. We will show a result very slightly more general than the theorem, namely: if x_h, x_ℓ and y are FP numbers satisfying $x_h \geq 0$, $y \geq 0$, and $|x_\ell| \leq \frac{1}{2}\mathrm{ulp}\left(\max\{x_h, y\}\right)$ then (z_h, z_ℓ) is a DW number and

$$\left|\frac{(z_h + z_\ell) - (x + y)}{x + y}\right| \leq u^2.$$

The case $x_h = 0$ or $y = 0$ is straightforward, so we do not consider it in the following. Without loss of generality, we can assume $x_h \geq y$ (otherwise, it suffices to exchange these variables, since they have a symmetric role in the algorithm). Still without loss of generality, we assume $1 \leq x_h \leq 2 - 2u$, which implies $1 - u \leq x \leq 2 - u$, so that $1 - u < x + y \leq 4 - 3u$.

Define $\epsilon = v - (x_\ell + s_\ell)$.

- If $x + y < 1$ then $x_h = 1$ (otherwise $x + y = x_h + x_\ell + y \geq (1 + 2u) - u + 0 = 1 + u$). We also have $-u \leq x_\ell < 0$, and $0 < y < |x_\ell|$. This implies $s_h = 1$ and $s_\ell = y$. Therefore $-u \leq x_\ell + s_\ell \leq u$, which implies $|v| \leq u$ and $|\epsilon| \leq u^2/2$. Since $|v|$ is less than s_h, we can use Fast2Sum at Line 3 of the algorithm. This leads to a relative error bounded by $u^2/(2 \cdot (1 - u))$, which is always less than u^2;

- If $1 \leq x + y \leq 2$ then $x_h + y \leq 2 + u$, so that $1 \leq s_h \leq 2$ and $|s_\ell| \leq u$. Therefore $|x_\ell + s_\ell| \leq 2u$. As a consequence $|v| \leq 2u$ and $|\epsilon| \leq u^2$. Since $|v| \leq 2u$ and $s_h \geq 1$ we can use Fast2Sum at Line 3 of the algorithm, and the relative error $\epsilon/(x + y)$ is bounded by u^2.

- If $2 < x + y$ then $2 - u < x_h + y \leq 4 - 4u$, so that $2 \leq s_h \leq 4 - 4u$ and $|s_\ell| \leq 2u$. Therefore $|x_\ell + s_\ell| \leq 3u$. As a consequence $|v| \leq 3u$ and $|\epsilon| \leq 2u^2$. Since $|v| \leq 3u$ and $s_h \geq 2$ we can use Fast2Sum at Line 3 of the algorithm, and the relative error $\epsilon/(x + y)$ is bounded by u^2.

The asymptotic optimality of the bound is shown by considering the following input values

$$\begin{cases} x_h &= 1, \\ x_\ell &= u - u^2, \\ y &= u, \end{cases}$$

for which the double-word number returned by Algorithm 4 is equal to $1 + 2u$ (with 2Sum with round-to-nearest ties-to-even at Line 1 of the Algorithm. If we replace 2Sum by augmentedAddition, we obtain $1 + 2u - 2u^2$: the absolute value of the error is the same), resulting in a relative error

$$\frac{u^2}{1 + 2u - u^2} = u^2 - 2u^3 + \mathcal{O}(u^4).$$

□

Remark 3.2. *When the operands $x = x_h + x_\ell$ and y are positive, Algorithm 4 satisfies the condition of Lemma 1.4 (with $\mathbb{F}$ being the set of the DW numbers), namely:*

$$\left|\mathrm{DWPlusFP}(x_h, x_\ell, y) - (x_h + x_\ell + y)\right| \leq \min\left\{(x_h + x_\ell), y\right\}.$$

Proof. We have

$$\left|\mathrm{DWPlusFP}(x_h, x_\ell, y) - (x_h + x_\ell + y)\right| = |v - (x_\ell + s_\ell)|,$$

and

$$|v - (x_\ell + s_\ell)| \leq |x_\ell| \leq u \cdot (x_h + x_\ell) < x_h + x_\ell,$$

and

$$|v - (x_\ell + s_\ell)| \leq |s_\ell| \leq y.$$

Therefore,

$$|v| \leq \min\{x_h + x_\ell, y\}.$$

□

Later on, we will compute sums of squares using Algorithm DWPlusFP (Algorithm 4). We will need to bound the computed sum of $n \leq 1/u$ positive numbers less than some power of 2, say 2^k, by $n \cdot 2^k$ (this is of course a straightforward property of the exact sum, but this is far from obvious for the computed sum). This will be ensured by the following lemma.

Lemma 3.3. *If (x_h, x_ℓ) is a DW number and y is a FP number such that $x_h, y \geq 2^{e_{\min}}$, $x_h + x_\ell \leq m_1 \cdot 2^k$ and $y \leq m_2 \cdot 2^k$ where m_1 and m_2 are positive integers satisfying $m_1 + m_2 \leq 2^p$ then, barring overflow, the double-word number (z_h, z_ℓ) returned by Algorithm DWPlusFP satisfies*

$$z_h + z_\ell \leq (m_1 + m_2) \cdot 2^k.$$

Proof. Since $z_h + z_\ell = s_h + v$, we need to prove that $s_h + v \leq (m_1 + m_2) \cdot 2^k$. First note that $x_h = \mathrm{RN}(x_h + x_\ell) \leq m_1 \cdot 2^k$ and $s_h = \mathrm{RN}(x_h + y) \leq \mathrm{RN}\left((m_1 + m_2) \cdot 2^k\right) = (m_1 + m_2) \cdot 2^k$. We have,

$$x + y = x_h + x_\ell + y \leq (m_1 + m_2) \cdot 2^k,$$

so that (since $s_h + s_\ell = x_h + y$)

$$s_h + s_\ell + x_\ell \leq (m_1 + m_2) \cdot 2^k. \tag{11}$$

- if $s_h < (m_1 + m_2) \cdot 2^k$, then, since s_h and $(m_1 + m_2) \cdot 2^k$ are FP numbers, $\mathrm{succ}(s_h) \leq (m_1 + m_2) \cdot 2^k$. Since $|s_\ell| \leq \frac{1}{2}\mathrm{ulp}(s_h)$ and $|x_\ell| \leq \frac{1}{2}\mathrm{ulp}(x_h) \leq \frac{1}{2}\mathrm{ulp}(s_h)$, we have $|s_\ell + x_\ell| \leq \mathrm{ulp}(s_h)$, so that

  $$|v| = |\mathrm{RN}(s_\ell + x_\ell)| \leq \mathrm{RN}(\mathrm{ulp}(s_h)) = \mathrm{ulp}(s_h).$$

  Hence,

  $$s_h + v \leq s_h + \mathrm{ulp}(s_h) = \mathrm{succ}(s_h) \leq (m_1 + m_2) \cdot 2^k.$$

- if $s_h = (m_1 + m_2) \cdot 2^k$ then (11) implies $s_\ell + x_\ell \leq 0$, therefore $v = \mathrm{RN}(s_\ell + x_\ell) \leq 0$, which implies $s_h + v \leq (m_1 + m_2) \cdot 2^k$.

□

3.2 Square-root of a double-word number

Assume that $x = (x_h, x_\ell)$ is a DW number, and that $x_h \geq 2^{2k}$, where k is an integer larger than or equal to $(e_{\min}+p)/2$. The following two algorithms evaluate the square root of x. Algorithm 8 returns a DW number, and Algorithm 9 returns a FP number.

Algorithm 8 – SQRTDWtoDW(x_h, x_ℓ). Computes the square-root of the DW number (x_h, x_ℓ) in binary, precision-p, floating-point arithmetic and returns a DW number (z_h, z_ℓ). It takes 8 FP operations (counting the FP square root as one).

1: $s_h \leftarrow \text{RN}(\sqrt{x_h})$
2: $\rho_1 \leftarrow \text{RN}(x_h - s_h^2)$ (with an FMA instruction)
3: $\rho_2 \leftarrow \text{RN}(x_\ell + \rho_1)$
4: $s_\ell \leftarrow \text{RN}(\rho_2/(2 \cdot s_h))$
5: $(z_h, z_\ell) \leftarrow \text{Fast2Sum}(s_h, s_\ell)$
6: **return** (z_h, z_ℓ)

To obtain a floating-point number, one can replace the "Fast2Sum" of Line 5 of Algorithm 8 by a floating-point addition and obtain

Algorithm 9 – SQRTDWtoFP(x_h, x_ℓ). Computes the square-root of the DW number (x_h, x_ℓ) in binary, precision-p, floating-point arithmetic and returns a floating-point number z. It takes 6 FP operations (counting the FP square root as one).

1: $s_h \leftarrow \text{RN}(\sqrt{x_h})$
2: $\rho_1 \leftarrow \text{RN}(x_h - s_h^2)$ (with an FMA instruction)
3: $\rho_2 \leftarrow \text{RN}(x_\ell + \rho_1)$
4: $s_\ell \leftarrow \text{RN}(\rho_2/(2 \cdot s_h))$
5: $z \leftarrow \text{RN}(s_h + s_\ell)$
6: **return** z

Remark 3.4. *For performance purposes, if the FMA operator is fast enough, to avoid the multiplication of s_h by 2 that appears in Line 4 of both algorithms, one can somehow "delay" that operation, and replace Lines 4 and 5 of Algorithm 8 by (we inline the Fast2Sum algorithm for the sake of clarity):*

4: $t_\ell \leftarrow \text{RN}(\rho_2/s_h)$ $\quad t_\ell$ *equals* $2s_\ell$
5: $z_h \leftarrow \text{RN}(s_h + 0.5 \cdot t_\ell)$ $\quad$ *first line of Fast2Sum*(s_h, s_ℓ)
6: $delta \leftarrow \text{RN}(z_h - s_h)$ $\quad$ *second line of Fast2Sum*(s_h, s_ℓ)
7: $z_\ell \leftarrow \text{RN}(0.5 \cdot t_\ell - \delta)$ $\quad$ *third line of Fast2Sum*(s_h, s_ℓ)

Similarly, one can replace Lines 4 and 5 of Algorithm 9 by:

4: $t_\ell \leftarrow \text{RN}(\rho_2/s_h)$ $\quad t_\ell$ *equals* $2s_\ell$
5: $z \leftarrow \text{RN}(s_h + 0.5 \cdot t_\ell)$

By doing so, Algorithm 8 now takes 7 FP operations, and Algorithm 9 takes 5 FP operations. In both cases, the computed results are exactly the same.

Let us now analyze Algorithms 8 and 9. We assume that the input double-word operand $x = (x_h, x_\ell)$ is positive, i.e., $x_h > 0$. If x_h can be zero, a preliminary test must be added at the beginning of these two algorithms. We have,

Theorem 3.5. *If $x = (x_h, x_\ell)$ is a double-word number, $p \geq 5$, $x \geq 2^{2k}$, where k is an integer larger than or equal to $(e_{\min} + p)/2$ and no overflow or underflow occurs, then the relative error of Algorithm 8 is bounded by*

$$\frac{25}{8}u^2 = 3.125u^2,$$

and that bound is asymptotically optimal.

Theorem 3.6. *If $x = (x_h, x_\ell)$ is a double-word number, $p \geq 5$, $x \geq 2^{2k}$, where k is an integer larger than or equal to $(e_{\min} + p)/2$ and no overflow or underflow occurs, then the floating-point number returned by Algorithm 9 is within*

$$\left(\frac{1}{2} + \frac{7}{4} \cdot 2^{-p}\right) \cdot \mathrm{ulp}(\sqrt{x_h + x_\ell})$$

from $\sqrt{x_h + x_\ell}$, and the relative error of that algorithm is bounded by

$$u + \frac{17}{8}u^2 + \frac{33}{8}u^3.$$

For proving Theorem 3.10 below, we will also need the following result.

Remark 3.7. *If $x_h = 2^{2k}$ with $k \in \mathbb{Z}$ then Algorithm 9 returns 2^k.*

Since Algorithms 8 and 9 differ only in the last line, we give a common proof to the two theorems (and we will prove Remark 3.7 in passing).

Proof. Define $x = x_h + x_\ell$. First note that $s_h \geq 2^k$ and, since $k \geq (e_{\min}+p)/2 > e_{\min}$, s_h is in the normal domain. Hence, if e is the FP exponent of s_h, $2^e \leq s_h < 2^{e+1}$, and s_h can be written $m \cdot 2^{e-p+1}$ where $m \in \mathbb{N}$. We have $e \geq k$ and therefore $2e \geq e_{\min} + p$, hence the condition of Lemma 1.5 is satisfied. This implies

$$\rho_1 = x_h - s_h^2$$

(i.e., the operation at Line 2 of Algorithms 8 and 9 is exact). This is a well-known property of the floating-point square-root and the FMA instruction [3, 4].

Now, without loss of generality, we can assume $1 \leq x_h \leq 4 - 4u$. Note that this implies $1 \leq s_h \leq 2 - 2u$. We have

$$|s_h - \sqrt{x_h}| < u \tag{12}$$

(equality is impossible because the square root of a floating-point number cannot be a "midpoint" [22]).

Let us first bound the distance between $\sqrt{x}$ and s_h. We have

$$\left|\sqrt{x} - s_h\right| \leq \left|\sqrt{x_h} - s_h\right| + \left|\sqrt{x} - \sqrt{x_h}\right|. \tag{13}$$

Using the mean-value theorem, we have

$$\left|\sqrt{x} - \sqrt{x_h}\right| = \left|\frac{x - x_h}{2\sqrt{\xi}}\right| = \left|\frac{x_\ell}{2\sqrt{\xi}}\right|, \tag{14}$$

for some ξ between x and x_h. Hence,

- if $x_h = 1$ and $x < x_h$ (which implies $-u/2 \leq x_\ell < 0$) then

  $$\left|\sqrt{x} - \sqrt{x_h}\right| \leq \frac{u/2}{2\sqrt{1 - u/2}},$$

  which is less than $u/2$ for all $u \leq 1$;

- if $x_h > 1$ or ($x_h = 1$ and $x_\ell \geq 0$) then the term $\sqrt{\xi}$ in (14) is larger than 1, so that

  $$\left|\sqrt{x} - \sqrt{x_h}\right| \leq \frac{u}{2}$$

- if $x_h = 2$ and $x < x_h$ (which implies $-u \leq x_\ell < 0$) then

  $$\left|\sqrt{x} - \sqrt{x_h}\right| \leq \frac{u}{2\sqrt{2 - u}},$$

  which is less than $u/\sqrt{2}$ for all $u \leq 1$;

- if $x_h > 2$ or ($x_h = 2$ and $x_\ell \geq 0$) then the term $\sqrt{\xi}$ in (14) is larger than $\sqrt{2}$, so that

  $$\left|\sqrt{x} - \sqrt{x_h}\right| \leq \frac{u}{\sqrt{2}}.$$

Combined with (12) and (13), this gives

$$\left|\sqrt{x} - s_h\right| \leq \begin{cases} \frac{3u}{2} & \text{if } x_h \leq 2 - 2u, \\ u\left(1 + \frac{\sqrt{2}}{2}\right) & \text{otherwise.} \end{cases} \tag{15}$$

Let us now estimate the error committed at Line 3 of Algorithms 8 and 9. We have

$$|x_\ell + \rho_1| \leq |x_\ell| + \left|x_h - s_h^2\right| \leq |x_\ell| + \left|\sqrt{x_h} - s_h\right| \cdot \left(\sqrt{x_h} + s_h\right),$$

therefore,

$$|x_\ell + \rho_1| < |x_\ell| + u \cdot (2s_h + u). \tag{16}$$

- if $x_h \leq 2 - 2u$ then $|x_\ell| \leq u$ and

  $$s_h < \sqrt{2 - 2u} + u < \sqrt{2} \cdot (1 - u/2) + u < \sqrt{2} + u/2,$$

  so that $2s_h + u < 2\sqrt{2} + 2u$. Hence $|x_\ell + \rho_1| < u \cdot (1 + 2\sqrt{2}) + 2u^2$. That bound is less than $4u$ as soon as $u \leq 1/16$ (i.e., as soon as $p \geq 4$);

- if $x_h \geq 2$ then $|x_\ell| \leq 2u$ and $s_h \leq 2$, so that $2s_h + u \leq 4 + u$. Hence $|x_\ell + \rho_1| < 6u + u^2$.

All this gives

$$\rho_2 = \mathrm{RN}(x_\ell + \rho_1) = \mathrm{RN}(x - s_h^2) = x - s_h^2 + \epsilon_1, \tag{17}$$

with

$$|\epsilon_1| \leq \begin{cases} 2u^2 & \text{if } x_h \leq 2 - 2u, \\ 4u^2 & \text{otherwise.} \end{cases} \tag{18}$$

We now deal with Line 4 of Algorithms 8 and 9. From (16), we obtain

$$|\rho_2| \leq |x_\ell| + u \cdot (2s_h + u) + |\epsilon_1|,$$

so that

$$\left|\frac{\rho_2}{2s_h}\right| \leq \left|\frac{x_\ell}{2s_h}\right| + u \cdot \left(1 + \frac{u}{2s_h}\right) + \left|\frac{\epsilon_1}{2s_h}\right|.$$

Therefore,

- if $x_h \leq 2 - 2u$ then $s_h \geq 1$ and

$$\left|\frac{\rho_2}{2s_h}\right| \leq \frac{u}{2} + u \cdot \left(1 + \frac{u}{2}\right) + u^2 \leq \frac{3u}{2} + \frac{3u^2}{2}, \tag{19}$$

- if $x_h \geq 2$ then $s_h > \sqrt{2} - u$, and

$$\left|\frac{\rho_2}{2s_h}\right| \leq \frac{2u}{2 \cdot (\sqrt{2} - u)} + u \cdot \left(1 + \frac{u}{2 \cdot (\sqrt{2} - u)}\right) + 2u^2. \tag{20}$$

The bounds (19) and (20) are less than $2u - 2u^2$ as soon as $u \leq 1/32$ (i.e., as soon as $p \geq 5$). This is straightforward for bound (19). For bound (20) let us consider functions

$$\varphi(u) = \frac{1}{\sqrt{2} - u} + \left(1 + \frac{u}{2 \cdot (\sqrt{2} - u)}\right) + 2u$$

and

$$h(u) = 2 - 2u.$$

Function φ is an increasing function of u, whereas h is a decreasing function. One easily checks that $\varphi(1/32) < h(1/32)$. Therefore, for any $u \leq 1/32$, $\varphi(u) < h(u)$.

All this gives (as soon as $p \geq 5$)

$$|s_\ell| \leq 2u - 2u^2 \text{ and } s_\ell = \frac{\rho_2}{2s_h} + \epsilon_2, \text{ with } |\epsilon_2| \leq u^2. \tag{21}$$

This and $s_h \geq 1$ show that Fast2Sum can be used at Line 5 of Algorithm 8. By combining (17) and (21), we finally obtain

$$\begin{aligned} s_h + s_\ell &= s_h + \frac{x - s_h^2 + \epsilon_1}{2s_h} + \epsilon_2 \\ &= \frac{1}{2} \cdot \left(s_h + \frac{x}{s_h}\right) + \frac{\epsilon_1}{2s_h} + \epsilon_2. \end{aligned}$$

(not surprisingly, we find again the Heron/Newton-Raphson formula for the square root in the first term), and therefore

$$\begin{aligned} s_h + s_\ell - \sqrt{x} &= \frac{1}{2} \cdot \left(s_h - 2\sqrt{x} + \frac{x}{s_h} \right) + \frac{\epsilon_1}{2s_h} + \epsilon_2 \\ &= \frac{1}{2s_h} \cdot \left(s_h^2 - 2s_h\sqrt{x} + x \right) + \frac{\epsilon_1}{2s_h} + \epsilon_2 \\ &= \frac{1}{2s_h} \cdot \left(s_h - \sqrt{x} \right)^2 + \frac{\epsilon_1}{2s_h} + \epsilon_2. \end{aligned}$$

This gives

$$\left| s_h + s_\ell - \sqrt{x} \right| \leq \frac{1}{2s_h} \cdot \left(s_h - \sqrt{x} \right)^2 + \left| \frac{\epsilon_1}{2s_h} \right| + |\epsilon_2| \,. \tag{22}$$

Let us first quickly eliminate the case $x_h = 1$, which is the only case for which x (and therefore $\sqrt{x}$) may be less than 1. In that case, we easily find $s_h = 1$, $\rho_1 = 0$, $\rho_2 = x_\ell$, $\epsilon_1 = 0$, and $|\rho_2/(2s_h)| = |x_\ell/2| \leq u/2$, resulting in $|\epsilon_2| \leq u^2/4$. Using (22) we obtain $|s_h + s_\ell - \sqrt{x}| \leq 11u^2/8$. The relative error of Algorithm 8 is therefore bounded by $11u^2/(8\sqrt{1-u})$ which is always less than $\frac{25u^2}{8}$. Now consider Algorithm 9. Since $\mathrm{ulp}(\sqrt{x}) \geq u$, we have $|s_h + s_\ell - \sqrt{x}| \leq 11u \cdot \mathrm{ulp}(\sqrt{x})/8$. The error of the addition of Line 5 of Algorithm 9 is less than or equal to $\frac{1}{2}\mathrm{ulp}(s_h + s_\ell)$. Since $s_h = 1$ and $s_\ell = x_\ell/2$, $s_h + s_\ell < 1$ if and only if $x < 1$, i.e., if and only if $\sqrt{x} < 1$. Therefore, $\mathrm{ulp}(s_h + s_\ell) = \mathrm{ulp}(\sqrt{x})$. As a consequence, the total absolute error of Algorithm 9 is bounded by $(\frac{1}{2} + \frac{11}{8}u)\mathrm{ulp}(\sqrt{x})$, which is less than the bound of Theorem 3.6. Also, the fact that $s_h = 1$ and $|s_\ell| \leq u/2$ implies that Algorithm 9 returns $z = 1$, which proves Remark 3.7 .

In the following, we can therefore assume that $x_h \geq 1 + 2u$, which implies $x \geq 1 + u$. Using (15), (22) and the bounds (18) and (21) on ϵ_1 and ϵ_2, we deduce

- if $x_h \leq 2 - 2u$ then

$$\left| s_h + s_\ell - \sqrt{x} \right| \leq \frac{1}{2} \left(\frac{3u}{2} \right)^2 + 2u^2 = \frac{25u^2}{8} = 3.125u^2, \tag{23}$$

- if $x_h \geq 2$ then

$$\begin{aligned} \left| s_h + s_\ell - \sqrt{x} \right| &\leq \frac{1}{2 \cdot (\sqrt{2} - u)} \cdot \left[u \cdot \left(1 + \frac{\sqrt{2}}{2} \right) \right]^2 + \frac{4u^2}{2 \cdot (\sqrt{2} - u)} + u^2 \\ &= u^2 \cdot \frac{11 + 6\sqrt{2} - 4u}{4\sqrt{2} - 4u}, \end{aligned} \tag{24}$$

  which is less than $3.5u^2$ as soon as $u \leq 1/32$.

This immediately gives the bound in ulps of Theorem 3.6, by noting that $\mathrm{ulp}(\sqrt{x}) = 2u$ and that the error due to the addition of Line 5 of Algorithm 9 is bounded by $\frac{1}{2}\mathrm{ulp}(s_h + s_\ell)$, which is less than or equal to $\frac{1}{2}\mathrm{ulp}(\sqrt{x})$:

- first, $1 + 2u \leq x_h \leq 4 - 4u$ implies $1 + u \leq x \leq 4 - 2u$, so that $1 < \sqrt{x} < 2$, which implies $\mathrm{ulp}(\sqrt{x}) = 2u$;
- $1 + 2u \leq x_h \leq 4 - 4u$ also implies $1 \leq s_h \leq 2 - 2u$, therefore, using (21), $1 - 2u + u^2 \leq s_h + s_\ell \leq 2 - u^2$, hence $\mathrm{ulp}(s_h + s_\ell)$ equals u or $2u$.

Let us now prove the relative error bounds given by Theorems 3.5 and 3.6. We have

- if $1 + 2u \leq x_h \leq 2 - 2u$ then $x > 1$, and, from (23), the relative error of Algorithm 8 is bounded by $\frac{25u^2}{8} = 3.125u^2$;
- if $x_h \geq 2$, then $x \geq 2 - u$, so that $\sqrt{x} \geq \sqrt{2-u}$, and the relative error of Algorithm 8 is bounded by
$$\frac{3.5u^2}{\sqrt{2-u}},$$
which is less than $\frac{25u^2}{8}$ as soon as $u \leq 1/2$ (i.e., $p \geq 1$).

Hence the relative error of Algorithm 8 is bounded by $\frac{25u^2}{8}$. Concerning the relative error of Algorithm 9 (i.e., the relative error bound of Theorem 3.6), we have just shown that
$$\left|s_h + s_\ell - \sqrt{x}\right| \leq \frac{25}{8}u^2\sqrt{x},$$
and we have the classical bound [24]:
$$|\mathrm{RN}(s_h + s_\ell) - (s_h + s_\ell)| \leq \frac{u}{1+u} \cdot (s_h + s_\ell).$$
Therefore,
$$\begin{aligned}
\left|z - \sqrt{x}\right| &= \left|\mathrm{RN}(s_h + s_\ell) - \sqrt{x}\right| \\
&\leq |\mathrm{RN}(s_h + s_\ell) - (s_h + s_\ell)| + \left|s_h + s_\ell - \sqrt{x}\right| \\
&\leq \frac{u}{1+u} \cdot (s_h + s_\ell) + \frac{25}{8}u^2\sqrt{x} \\
&\leq \frac{u}{1+u} \cdot \left(1 + \frac{25}{8}u^2\right) \cdot \sqrt{x} + \frac{25}{8}u^2\sqrt{x} \\
&= u \cdot \left(\frac{8 + 25u + 50u^2}{8 + 8u}\right) \cdot \sqrt{x}.
\end{aligned}$$
The bound $u + \frac{17}{8}u^2 + \frac{33}{8}u^3$ of Theorem 3.6 is proved by noting that
$$1 + \frac{17}{8}u + \frac{33}{8}u^2 - \left(\frac{8 + 25u + 50u^2}{8 + 8u}\right) = \frac{33u^3}{8 + 8u}$$
is always positive.

There remains to show that the bound $25u^2/8$ of Theorem 3.5 is asymptotically optimal. This is done by considering the case

$$\left\{ \begin{array}{rcl} x_h &=& 1+6u \\ x_\ell &=& u-6u^2, \end{array} \right.$$

for which one easily checks that Algorithm 8 returns $z_h + z_\ell = 1 + \frac{7}{2}u - 6u^2$ as soon as $p \geq 3$, resulting in a relative error

$$\left| \frac{\sqrt{1+7u-6u^2} - \left(1+\frac{7}{2}u-6u^2\right)}{\sqrt{1+7u-6u^2}} \right| = \frac{25}{8}u^2 - \frac{343}{8}u^3 + \mathcal{O}(u^4).$$

□

Intuitively, since the square root of a huge number is less than that number, and the square root of a tiny number is larger than that number, overflows and underflows are not much of a concern when evaluating square roots. This does not mean that *intermediate calculations* in Algorithms 8 and 9 cannot underflow or overflow. Let us now quickly address this issue.

Remark 3.8. *Under the conditions of theorems 3.5 and 3.6 (and assuming $e_{\max} \geq 2$, which always holds in practice[8]), no overflow can occur in Algorithms 8 and 9.*

Proof. It suffices to show that all intermediate values remain below Ω.

Since $e_{\max} \geq 2$, we have $\Omega > 4$ and therefore $\sqrt{\Omega} < \Omega/2$. Hence, since $x_h \leq \Omega$,

$$s_h = \mathrm{RN}\left(\sqrt{x_h}\right) \leq \mathrm{RN}\left(\sqrt{\Omega}\right) \leq \mathrm{RN}(\Omega/2) = \Omega/2.$$

This also implies that $2s_h$ (needed at Line 4 of both algorithms) does not overflow.

We have seen that $x_h - s_h^2$ is computed exactly. Hence,

$$\begin{aligned} |\rho_1| = |(\sqrt{x_h} - s_h)(\sqrt{x_h} + s_h)| &\leq u\sqrt{x_h}(\sqrt{x_h} + \sqrt{x_h}(1+u)) \\ &\leq (2u+u^2)x_h \\ &\leq (2u+u^2)\Omega \\ &\leq \tfrac{65}{1024}\Omega \end{aligned}$$

(since $p \geq 5$ implies $u \leq 1/32$). Therefore, $|x_\ell + \rho_1| \leq (3u+u^2)x_h$, so that

$$|\rho_2| \leq (3u+u^2)(1+u)x_h = (3u+4u^2+u^3)x_h \leq \tfrac{3201}{32768}\Omega.$$

Hence,

$$\left|\frac{\rho_2}{2s_h}\right| \leq \frac{(3u+4u^2+u^3)x_h}{2(1-u)\sqrt{x_h}} = \frac{3u+4u^2+u^3}{2-2u}\sqrt{x_h} \leq \tfrac{3201}{63488}\sqrt{x_h} < \Omega.$$

[8] In fact, Condition $e_{\max} \geq 2$ can be deduced from the conditions of theorems 3.5 and 3.6. We know that $x \geq 2^{e_{\min}+p}$, so that $x_h \geq 2^{e_{\min}+p}$, which can be representable only if $e_{\max} \geq e_{\min} + p$. And since $p \geq 5$ and $e_{\min} = 1 - e_{\max}$, we have $e_{\max} \geq 6 - e_{\max}$. Therefore, $e_{\max} \geq 3$.

Multiplying by $(1+u)$ we get a bound on s_ℓ still less than Ω. We can now add s_h and s_ℓ:

$$\begin{aligned} s_h + s_\ell &\leq \sqrt{x_h} \cdot \left(1 + u + \frac{3u + 4u^2 + u^3}{2 - 2u}(1+u)\right) \\ &= \frac{2 + 3\,u + 5\,u^2 + 5\,u^3 + u^4}{2 - 2u}\sqrt{x_h} \leq \tfrac{2200737}{2031616}\sqrt{\Omega} \leq \tfrac{2200737}{2031616} \cdot \frac{\Omega}{2} < \Omega. \end{aligned}$$

□

Remark 3.9. *Under the conditions of theorems 3.5 and 3.6, with the additional assumption $p + 3 \leq e_{\max}$, underflows in Algorithms 8 and 9 are impossible or harmless.*

Proof. Since $x \geq 2^{2k}$, with $k \geq (e_{\min} + p)/2 > e_{\min}$, $x_h \geq 2^{2k}$ and therefore $\sqrt{x_h}$ and s_h are larger than or equal to $2^{e_{\min}}$. Therefore, we always have

$$|s_h - \sqrt{x_h}| \leq \frac{u}{1+u}\sqrt{x_h}. \tag{25}$$

We have

$$\begin{aligned} \left|s_h - \sqrt{x}\right| &\leq |s_h - \sqrt{x_h}| + \left|\sqrt{x_h} - \sqrt{x}\right| \\ &\leq \frac{u}{1+u}\sqrt{x_h} + \frac{1}{2\min\{\sqrt{x_h}, \sqrt{x}\}}\,|x_h - x| \end{aligned} \tag{26}$$

(using the mean value theorem). If $x_h \geq x$, then (26) gives

$$\begin{aligned} \left|s_h - \sqrt{x}\right| &\leq \frac{u}{1+u}\sqrt{x(1+u)} + \frac{1}{2\sqrt{x}}\left(\frac{u}{1+u}\right)x \\ &= \sqrt{x} \cdot u \cdot G_1(u), \end{aligned}$$

with

$$G_1(u) = \frac{2\sqrt{1+u} + 1}{2(1+u)} < \frac{2\left(1 + \frac{u}{2}\right) + 1}{2(1+u)} < \frac{3}{2}.$$

Now, if $x_h < x$, then (26) gives

$$\begin{aligned} \left|s_h - \sqrt{x}\right| &\leq \frac{u}{1+u}\sqrt{x} + \frac{1}{2\sqrt{x(1-u)}}\left(\frac{u}{1+u}\right)x \\ &= \sqrt{x} \cdot u \cdot G_2(u), \end{aligned}$$

with

$$G_2(u) = \frac{1}{1+u}\left(1 + \frac{1}{2\sqrt{1-u}}\right) = \frac{3}{2} + \frac{1 - (1+3u)\sqrt{1-u}}{2(1+u)\sqrt{1-u}}.$$

We have

$$(1+3u)^2(1-u) - 1 = u(5 + 3u - 9u^2) = -9u \cdot \left(u - \frac{1}{6} - \frac{\sqrt{21}}{6}\right) \cdot \left(u - \frac{1}{6} + \frac{\sqrt{21}}{6}\right),$$

therefore, for $u \in \left[0, \frac{1}{6} + \frac{\sqrt{21}}{6}\right] \approx [0, 0.93]$, $(1+3u)\sqrt{1-u} > 1$, so that $G_2(u) < \frac{3}{2}$.

Hence, in all cases,

$$\left|\sqrt{x} - s_h\right| \leq \frac{3}{2} u \sqrt{x},$$

and therefore

$$\left(\sqrt{x} - s_h\right)^2 \leq \frac{9}{4} u^2 x.$$

This would also allow one to directly deduce that $|x - s_h^2| \leq \left(3u + \frac{9}{4}u^2\right) \cdot x$, but we can compute a slightly better bound: Define $v = u/(1+u)$. From (25), we have

$$\begin{aligned}
\left|x - s_h^2\right| &= \left|x_h + x_\ell - s_h^2\right| \\
&\leq \left|x_h - s_h^2\right| + |x_\ell| \\
&\leq |(\sqrt{x_h} - s_h) \cdot (\sqrt{x_h} + s_h)| + v \cdot x \\
&\leq (v\sqrt{x_h}) \cdot (\sqrt{x_h} + (\sqrt{x_h} + v\sqrt{x_h})) + v \cdot x \\
&= x_h \cdot (2v + v^2) + v \cdot x \\
&\leq \left[(1+v)(2v+v^2) + v\right] \cdot x \\
&= \frac{3u + 9u^2 + 7u^3}{(1+u)^3} \cdot x \\
&= 3u \cdot \frac{1 + 3u + \frac{7}{3}u^2}{1 + 3u + 3u^2 + u^3} \cdot x \\
&\leq 3ux.
\end{aligned}$$

Since s_h is a normal FPN larger than 2^k, it can be written $S_h \cdot 2^{e-p+1}$, where $e \geq k$ is an integer and $|S_h| \leq 2^p - 1$. Lemma 1.5 therefore implies that $x_h - s_h^2$ is a FPN, so that $\rho_1 = x_h - s_h^2$ (ρ_1 may be subnormal, but this does not have any consequence on its accuracy).

Let us now deal with Lines 3 and 4 of Algorithms 8 and 9. Let us first note that if

$$\left| s_\ell - \frac{x - s_h^2}{2s_h} \right| \leq \epsilon,$$

then

$$\begin{aligned}
\left|(s_h + s_\ell) - \sqrt{x}\right| &\leq \left| s_h + \frac{x - s_h^2}{2s_h} - \sqrt{x} \right| + \left| s_\ell - \frac{x - s_h^2}{2s_h} \right| \\
&\leq \frac{1}{2s_h} \left(s_h - \sqrt{x}\right)^2 + \epsilon \\
&\leq \frac{9u^2}{8} \frac{x}{s_h} + \epsilon \\
&\leq \frac{9u^2}{8} \cdot \frac{x}{\sqrt{x}\left(1 - \frac{3}{2}u\right)} + \epsilon = \frac{9u^2\sqrt{x}}{8 - 12u} + \epsilon.
\end{aligned} \tag{27}$$

This last bound is less than $\frac{5}{4}u^2\sqrt{x} + \epsilon$ for $u \leq 1/32$.

- if $|\rho_2|$ and $|\rho_2/(2s_h)|$ are above the underflow threshold $2^{e_{\min}}$ then no underflow occurs in the calculation, so that the proof of Theorems 3.5 and 3.6 still holds;
- if $|\rho_2| < 2^{e_{\min}}$, then Lemma 1.3 implies $\rho_2 = x_\ell + \rho_1 = x - s_h^2$, and therefore
  1. if $|\rho_2/(2s_h)| \geq 2^{e_{\min}}$ (so that s_ℓ is in the normal domain) then

$$\begin{aligned}\left|s_\ell - \frac{\rho_2}{2s_h}\right| &\leq u \cdot \left|\frac{\rho_2}{2s_h}\right| \\ &= u \cdot \left|\frac{x - s_h^2}{2s_h}\right| \\ &\leq \frac{3}{2}u^2\frac{x}{s_h} \\ &\leq \frac{3}{2}u^2\frac{x}{\sqrt{x}\left(1-\frac{3}{2}u\right)} \\ &= \frac{3u^2\sqrt{x}}{2-3u},\end{aligned}$$

     and from

$$\frac{3u^2}{2-3u} - \frac{3}{2}u^2 - 3u^3 = -3u^3 \cdot \left(\frac{1-6u}{4-6u}\right),$$

     we easily deduce that $\left|s_\ell - \frac{\rho_2}{2s_h}\right|$ is less than $\left(\frac{3}{2}u^2 + 3u^3\right)\sqrt{x}$ for $u \leq 1/32$. Therefore, we can take ϵ equal to that value, and obtain

$$|s_h + s_\ell - \sqrt{x}| < \left(\frac{11}{4}u^2 + 3u^3\right)\sqrt{x},$$

     which is a better bound than that of Theorem 3.5;
  2. if $|\rho_2/(2s_h)| < 2^{e_{\min}}$ then $|s_\ell - \rho_2/(2s_h)| \leq 2^{e_{\min}-p}$, so that we can use (27) with $\epsilon = 2^{e_{\min}-p}$, and obtain

$$|s_h + s_\ell - \sqrt{x}| < \frac{5}{4}u^2\sqrt{x} + 2^{e_{\min}-p}.$$

     From $\sqrt{x} \geq 2^{(e_{\min}+p)/2}$ we deduce $2^{e_{\min}-p} \leq 2^{(e_{\min}-3p)/2}\sqrt{x}$, and $e_{\max} = 1 - e_{\min} \geq p + 3$ implies $2^{(e_{\min}-3p)/2} \leq 2^{-2p-1} = \frac{u^2}{2}$, therefore

$$|s_h + s_\ell - \sqrt{x}| < \frac{7}{4}u^2\sqrt{x},$$

     which is a better bound than that of Theorem 3.5.
- If $|\rho_2| \geq 2^{e_{\min}}$ and $|\rho_2/(2s_h)| < 2^{e_{\min}}$ this means that $s_h > 1/2$ and therefore $\sqrt{x_h} > 1/2$, which implies $x_h > 1/4$, so that $x > 1/4$ and $\sqrt{x} > 1/2$. We have

$|\rho_2 - (x - s_h^2)| = |\rho_2 - (x_\ell + \rho_1)| \leq u|\rho_2|$ and $|s_\ell - \rho_2/(2s_h)| \leq 2^{e_{\min}-p}$. Therefore,

$$\left| s_\ell - \frac{x - s_h^2}{2s_h} \right| \leq u \cdot \frac{|\rho_2|}{2s_h} + 2^{e_{\min}-p} < 2^{e_{\min}-p+1} < 2^{e_{\min}-p+2}\sqrt{x}.$$

Since $e_{\max} = 1 - e_{\min} \geq p + 3$, we obtain $2^{e_{\min}-p+2} \leq 2^{-2p} = u^2$. Therefore, (27) gives

$$|s_h + s_\ell - \sqrt{x}| < \frac{9}{4}u^2\sqrt{x},$$

which is a better bound than that of Theorem 3.5.

□

Now, let us assume that the input values (x_h, x_ℓ) of Algorithm SQRTDWtoFP (Algorithm 9) approximate some number x with a known relative error bound. Let us see how SQRTDWtoFP (x_h, x_ℓ) approximates $\sqrt{x}$. We have

Theorem 3.10. *If (x_h, x_ℓ) approximates a positive number x with relative error bounded by νu^2 with*

$$\nu u^2 < 1 \tag{28}$$

and if no underflow/overflow occurs, then

$$R = \mathrm{SQRTDWtoFP}(x_h, x_\ell)$$

approximates $\sqrt{x}$ with a relative error bounded by

$$\left(u + \frac{17}{8}u^2 + \frac{33}{8}u^3\right) \cdot \left(1 + \frac{\nu \cdot u^2}{1 - \nu \cdot u^2}\right) + \frac{\nu \cdot u^2}{1 - \nu \cdot u^2} \tag{29}$$

Furthermore, under the more stringent condition

$$2u\nu < 1, \tag{30}$$

we have

$$\left|R - \sqrt{x}\right| \leq \left(\frac{1}{2} + u \cdot \left(\frac{7}{4} + \frac{\nu}{1 - \nu \cdot u^2}\right)\right) \mathrm{ulp}\left(\sqrt{x}\right). \tag{31}$$

Proof. We wish to bound the distance between R and $\sqrt{x}$, both in terms of relative error and in terms of error in ulps. We have

$$\left|R - \sqrt{x}\right| \leq \left|R - \sqrt{x_h + x_\ell}\right| + \left|\sqrt{x_h + x_\ell} - \sqrt{x}\right|. \tag{32}$$

Theorem 3.6 gives

$$\left|R - \sqrt{x_h + x_\ell}\right| \leq \left(u + \frac{17}{8}u^2 + \frac{33}{8}u^3\right) \cdot \sqrt{x_h + x_\ell}. \tag{33}$$

Since (x_h, x_ℓ) approximates x with relative error $< \nu u^2$, we have

$$|(x_h + x_\ell) - x| \leq \nu \cdot u^2 \cdot x. \tag{34}$$

Hence, using the mean value theorem, we deduce

$$\left|\sqrt{x_h + x_\ell} - \sqrt{x}\right| = \frac{|(x_h + x_\ell) - x|}{2\sqrt{\xi}} \leq \frac{\nu \cdot u^2 \cdot x}{2\sqrt{\xi}}$$

for some ξ between $x_h + x_\ell$ and x. From (34), we have

$$\xi \geq x \cdot \left(1 - \nu \cdot u^2\right),$$

therefore, assuming (28), we have (just by using the relation $\sqrt{1-t} \geq 1 - t$ for $t \in [0, 1]$)

$$\sqrt{\xi} \geq \sqrt{x} \cdot \left(1 - \nu \cdot u^2\right). \tag{35}$$

This gives:

$$\left|\sqrt{x_h + x_\ell} - \sqrt{x}\right| \leq \frac{\nu \cdot u^2}{1 - \nu \cdot u^2} \cdot \sqrt{x}. \tag{36}$$

Combining (32), (33), and (36), we obtain the relative error bound (29) of the theorem.

Let us now focus on the error bound in ulps. Using Theorem 3.6, we obtain

$$\left|R - \sqrt{x_h + x_\ell}\right| \leq \left(\frac{1}{2} + \frac{7}{4}u\right) \text{ulp}\left(\sqrt{x_h + x_\ell}\right)$$

Using (36) and Property 1.1, we deduce that $\sqrt{x_h + x_\ell}$ is within

$$\frac{\nu \cdot u}{1 - \nu \cdot u^2} \text{ulp}\left(\sqrt{x}\right)$$

from $\sqrt{x}$.

- if $\sqrt{x_h + x_\ell}$ and $\sqrt{x}$ belong to the same binade (which will be the case in general since (36) implies that they are very close), or if $x_h + x_\ell \leq x$ then

  $$\text{ulp}\left(\sqrt{x_h + x_\ell}\right) \leq \text{ulp}\left(\sqrt{x}\right),$$

  and therefore

  $$\left|R - \sqrt{x}\right| \leq \left(\frac{1}{2} + u \cdot \left(\frac{7}{4} + \frac{\nu}{1 - \nu \cdot u^2}\right)\right) \text{ulp}\left(\sqrt{x}\right).$$

- if $\sqrt{x_h + x_\ell}$ and $\sqrt{x}$ are not in the same binade and $x_h + x_\ell > x$, then, under condition (30), which is a much stronger condition than (28), we have (using (34))

  $$x < 2^k \leq x_h + x_\ell \leq (1 + \nu u^2) \cdot x < \left(1 + \frac{u}{2}\right) \cdot x,$$

  where k is an *even* integer (otherwise, $\sqrt{x_h + x_\ell}$ and $\sqrt{x}$ would be in the same binade). This implies

  $$2^k \leq x_h + x_\ell < \left(1 + \frac{u}{2}\right) \cdot 2^k,$$

so that $x_h = 2^k$ and $x_\ell \geq 0$. Using Remark 3.7, we conclude that we obtain $R = 2^{k/2}$. We also have

$$\frac{2^k}{1+\frac{u}{2}} < x \leq 2^k$$

and therefore

$$\frac{2^{k/2}}{\sqrt{1+\frac{u}{2}}} < \sqrt{x} \leq 2^{k/2} \tag{37}$$

The number

$$\frac{2^{k/2}}{\sqrt{1+\frac{u}{2}}}$$

is always larger than $2^{k/2} \cdot \left(1 - \frac{u}{2}\right)$. This comes from

$$\frac{1}{1+\frac{u}{2}} - \left(1 - \frac{u}{2}\right)^2 = \frac{u \cdot (4 + 2u - u^2)}{8 + 4u} > 0.$$

Therefore, from (37) we deduce

$$\mathrm{RN}\left(\sqrt{x}\right) = 2^{k/2}.$$

Hence, R is the correctly rounded result, and therefore $|R - \sqrt{x}| \leq \frac{1}{2}\mathrm{ulp}\left(\sqrt{x}\right)$, which is stronger than (31).

□

Very similarly to what we have done with Algorithm SQRTDWtoFP (Algorithm 9), let us now assume that the input values (x_h, x_ℓ) of Algorithm SQRTDWtoDW (Algorithm 8) approximate some number x with a known relative error bound. Let us see how SQRTDWtoDW (x_h, x_ℓ) approximates $\sqrt{x}$. We have

Theorem 3.11. *If (x_h, x_ℓ) approximates a positive number x with relative error bounded by νu^2 with*

$$\nu u^2 < 1 \tag{38}$$

and if no underflow/overflow occurs, then

$$R = \mathrm{SQRTDWtoDW}(x_h, x_\ell)$$

approximates $\sqrt{x}$ with a relative error bounded by

$$u^2 \cdot \left(\frac{25}{8} + \frac{\nu}{1-\nu u^2} + \frac{25}{8} \cdot \frac{\nu u^2}{1-\nu u^2}\right). \tag{39}$$

Proof. The bound is derived almost immediately from Theorem 3.5, (32), and (36). □

3.3 Formalization

To formalize the results of this section, in particular the proofs of correctness of the proposed square root algorithm and the computation of the error bounds, we have used the Coq proof assistant with the help of the Flocq library [8, 9] on the arithmetic of floating-point numbers.

The Flocq library offers various models for representing floating-point numbers. For the proof under the hypothesis that the algorithm runs without overflow and without underflow (theorems 3.5, 3.6, 3.10 and 3.11), we used the model of representation of floating-point numbers without restrictions on the exponents (the FLX format). This make the proofs easier, and close to the paper proof, avoiding tedious verification about the exponents.

On the other hand, in the case of overflow studies (Remark 3.8), to verify that the algorithm does not generate overflows, we have had to prove that all numbers involved in the algorithm remained, in absolute value, below the largest floating-point number) Ω. Indeed, in the Flocq library, the choice was made not to put an upper bound on the exponents.

Finally, concerning underflow (Remark 3.9), the FLT format of Flocq allows one to give a lower bound to the exponents and the library develops a whole "theory" (i.e., definitions and properties) for this format of floating-point numbers, in particular for the normal and subnormal numbers. It is thus this format that we used to prove Remark 3.9.

Note that, for the rounding function, we have chosen *not to specify* the tie-breaking rule for round-to-nearest (the default, as said above, is ties-to-even). This makes the proofs more general: they will for instance be still valid with the less frequently used (but specified by IEEE 754) "ties-to-away" tie-breaking rule.

4 Our algorithms for computing Euclidean norms

4.1 Computing a Euclidean norm assuming no underflow or overflow occurs

In this section, we assume that all the terms a_i of (1) are in MED, so that no underflow/overflow occurs and a_i^2 is exactly representable by a DW number for all i. We first approximate the sum of squares $\sum_{i=0}^{n-1} a_i^2$ by a DW number (S_h, S_ℓ), with some relative error νu^2, and then use Algorithm SQRTDWtoFP (Algorithm 9) to approximate the square-root of $S_h + S_\ell$ by a floating-point number R. The final error will be deduced from ν using Theorem 3.10.

Let us now first present two different ways of computing (S_h, S_ℓ).

4.1.1 Sequential computation of the sum of squares

Let us first consider the following, sequential algorithm.

Algorithm 10 Sequential computation of $\sum_{i=0}^{n-1} a_i^2$ assuming no underflow/overflow occurs. It takes $13n-5$ FP operations.

1. For $i = 0 \dots n-1$, express the terms a_i^2 as double-word numbers (y_i^h, y_i^ℓ), defined as
$$(y_i^h, y_i^\ell) = \mathrm{Fast2Mult}(a_i, a_i) \tag{40}$$
We have $a_i^2 = y_i^h + y_i^\ell$.

2. Accumulate the terms y_i^h using the DWPlusFP algorithm (Algorithm 4). More precisely, define
$$(x_1^h, x_1^\ell) = 2\mathrm{Sum}(y_0^h, y_1^h)$$
first, then, iteratively compute, for $i = 2 \dots n-1$, the terms
$$(x_i^h, x_i^\ell) = \mathrm{DWPlusFP}(x_{i-1}^h, x_{i-1}^\ell, y_i^h).$$

3. Accumulate the terms y_i^ℓ using the conventional "recursive" summation, i.e., for $i = 0 \dots n-2$, compute
$$\sigma_{i+1} = \mathrm{RN}(\sigma_i + y_{i+1}^\ell),$$
with $\sigma_0 = y_0^\ell$.

4. Obtain the approximation to $\sum_{i=0}^{n-1} a_i^2$ with one call to DWPlusFP:
$$(S_h, S_\ell) = \mathrm{DWPlusFP}(x_{n-1}^h, x_{n-1}^\ell, \sigma_{n-1}).$$

Algorithm 10 assumes $n \geq 2$. It can be applied to the special case $n = 1$ if it stops after step 1 and returns $(S_h, S_\ell) = (y_0^h, y_0^\ell)$, taking only 2 FP operations.

We have,

Lemma 4.1. *Assuming no underflow/overflow occurs, the double-word number (S_h, S_ℓ) returned by Algorithm 10 satisfies*

$$\begin{aligned} &\left| (S_h + S_\ell) - \sum_{i=0}^{n-1} a_i^2 \right| \\ &\leq \left((2n-1)u^2 + (n-1)u^3 + (2n-2)u^4 + (n-1)u^5 \right) \cdot \sum_{i=0}^{n-1} a_i^2. \end{aligned} \tag{41}$$

If $n \leq 1/u$, this implies

$$\left| (S_h + S_\ell) - \sum_{i=0}^{n-1} a_i^2 \right| < \left((2n-1)u^2 + (n+1)\,u^3 \right) \cdot \sum_{i=0}^{n-1} a_i^2. \tag{42}$$

Proof. Thanks to Remark 3.2. Lemma 1.4 can applied with $\mathbb{F}$ being the set of double-word numbers, and with the bound u^2 given by Theorem 3.1 on the terms $|\epsilon_i|$. This gives:

$$\left|\left(x^h_{n-1} + x^\ell_{n-1}\right) - \sum_{i=0}^{n-1} y^h_i\right| \leq (n-1)\cdot u^2 \cdot \sum_{i=0}^{n-1} y^h_i \leq (n-1)\cdot u^2 \cdot (1+u)\cdot \sum_{i=0}^{n-1} a_i^2. \quad (43)$$

Very similarly, using $|y^\ell_i| \leq u \cdot a_i^2$, the same lemma applied to the summation of the terms y^ℓ_i (with now $\mathbb{F}$ being the set of floating-point numbers) gives

$$\left|\sigma_{n-1} - \sum_{i=0}^{n-1} y^\ell_i\right| \leq (n-1)\cdot u \cdot \sum_{i=0}^{n-1} \left|y^\ell_i\right| \leq (n-1)\cdot u^2 \cdot \sum_{i=0}^{n-1} a_i^2. \quad (44)$$

Adding (43) and (44), we deduce

$$\left|\left(x^h_{n-1} + x^\ell_{n-1} + \sigma_{n-1}\right) - \sum_{i=0}^{n-1} a_i^2\right| \leq \left(2\cdot(n-1)\cdot u^2 + (n-1)\cdot u^3\right)\cdot \sum_{i=0}^{n-1} a_i^2. \quad (45)$$

There remains to take into account the error bound on the last DWPlusFP operation. We have

$$\begin{aligned}\left|(S_h + S_\ell) - \left(x^h_{n-1} + x^\ell_{n-1} + \sigma_{n-1}\right)\right| &\leq u^2 \cdot \left(x^h_{n-1} + x^\ell_{n-1} + \sigma_{n-1}\right) \\ &\leq u^2 \cdot \left(1 + 2\cdot(n-1)\cdot u^2 + (n-1)\cdot u^3\right) \cdot \sum_{i=0}^{n-1} a_i^2. \quad (46)\end{aligned}$$

Adding (45) and (46) we finally obtain (41). Furthermore, if $n \leq 1/u$ then $(2n-2)u^4 + (n-1)u^5 \leq 2nu^4 - 2u^4 + \frac{1}{u}\cdot u^5 \leq 2u^3 - u^4 < 2u^3$, which gives (42).

□

4.1.2 Blockwise computation of the sum of squares

Now, assume that $n = k \times m$, and that we separate the input numbers a_i into k blocks of m numbers, either for parallelizing the calculation or for obtaining (as it will be clear later on) a more accurate result. Block number j ($j = 0, \ldots, k-1$) contains the elements

$$a_{mj}, a_{mj+1}, a_{mj+2}, \ldots, a_{m(j+1)-1}.$$

We will separately sum the elements of each block using Algorithm 10. The results of these "partial" summations are DW numbers (Z^h_j, Z^ℓ_j). A solution could be to sum these numbers using Algorithm SloppyDWPlusDW (Algorithm 5). We will obtain, however, a better error bound by summing these terms in the same way as we have summed the terms a_i^2 in Algorithm 10, i.e., we will first compute a DW approximation to the sum of the "higher" terms Z^h_j using DWPlusFP (Algorithm 4) iteratively, we will then accumulate the "lower" terms Z^ℓ_j using naive summation, and we will finally add

the obtained results with one call to DWPlusFP. This gives Algorithm 11, presented below.

It could be possible to repeat that block decomposition recursively, resulting in better error bounds. We doubt this would be efficient (except possibly for huge values of n).

For analyzing Algorithm 11, we need the following lemma.

Lemma 4.2. *Let n be a positive integer.*

- *the maximum possible value of $k+m$, where k and m are integers larger than or equal to 2 satisfying*
$$k \cdot m = n, \tag{47}$$
is $\frac{n}{2} + 2$;
- *the minimum possible value of $k+m$, where k and m are positive integers satisfying (47) is $r_n + n/r_n$, where r_n is the largest divisor of n less than or equal to $\sqrt{n}$. That bound is always larger than or equal to $2\sqrt{n}$.*

Proof. Straightforward by considering the variation of function $t \to t + n/t$. □

Algorithm 11 Blockwise computation of $\sum_{i=0}^{n-1} a_i^2$ assuming no underflow/overflow occurs. It takes $13n + 6k - 5$ FP operations.

---

1. for $j = 0, 1, \ldots, k-1$, compute an approximation (Z_j^h, Z_j^ℓ) to $\sum_{i=mj}^{m(j+1)-1} a_i^2$ using Algorithm 10 (the sequential summation algorithm) applied to $a_{mj}, a_{mj+1}, a_{mj+2}, \ldots, a_{m(j+1)-1}$;
2. accumulate the terms Z_j^h using Algorithm DWPlusFP (Algorithm 4). More precisely, defining
$$\left(\Sigma_1^h, \Sigma_1^\ell\right) = 2\mathrm{Sum}(Z_0^h, Z_1^h),$$
iteratively compute, for $j = 2 \ldots k-1$ the terms
$$\left(\Sigma_j^h, \Sigma_j^\ell\right) = \mathrm{DWPlusFP}\left(\Sigma_{j-1}^h, \Sigma_{j-1}^\ell, Z_j^h\right);$$
3. accumulate the terms Z_j^ℓ using the conventional "recursive" summation, i.e., for $j = 0 \ldots k-2$, compute
$$\tau_{j+1} = \mathrm{RN}(\tau_j + Z_{j+1}^\ell),$$
with $\tau_0 = Z_0^\ell$;
4. obtain the approximation (S_h, S_ℓ) to $\sum_{i=0}^{n-1} a_i^2$ as
$$(S_h, S_\ell) = \mathrm{DWPlusFP}\left(\Sigma_{k-1}^h, \Sigma_{k-1}^\ell, \tau_{k-1}\right).$$

Algorithm 11 assumes $k \geq 2$. It can be applied to the special case $k = 1$ if it stops after step 1 and returns $(S_h, S_\ell) = (Z_0^h, Z_0^\ell)$, so that it reduces to Algorithm 10, taking only $13n - 5$ FP operations.

Lemma 4.3. *Assuming no underflow/overflow occurs, the double-word number (S_h, S_ℓ) returned by Algorithm 11 satisfies*

$$\frac{\left|(S_h + S_\ell) - \sum_{i=0}^{n-1} a_i^2\right|}{\sum_{i=0}^{n-1} a_i^2} \leq \beta(k) + \beta(m) + \beta(k)\beta(m), \tag{48}$$

where $\beta(t) = (2t-1)u^2 + (t-1)u^3 + (2t-2)u^4 + (t-1)u^5$. Furthermore, if $n < 1/u$ and $u \leq 1/32$, we obtain

$$\frac{\left|(S_h + S_\ell) - \sum_{i=0}^{n-1} a_i^2\right|}{\sum_{i=0}^{n-1} a_i^2} \leq (2k + 2m - 2)\cdot u^2 + (0.129n + 0.939(k+m) + 2.03)\cdot u^3. \tag{49}$$

Proof. Lemma 4.1 applied to the computation of $\sum_{i=mj}^{m(j+1)-1} a_i^2$ with Algorithm 10 gives

$$\left|\left(Z_j^h + Z_j^\ell\right) - \sum_{i=mj}^{m(j+1)-1} a_i^2\right| \leq \beta(m) \cdot \sum_{i=mj}^{(m+1)j-1} a_i^2,$$

which implies

$$\left|\sum_{j=0}^{k-1}\left(Z_j^h + Z_j^\ell\right) - \sum_{i=0}^{n-1} a_i^2\right| \leq \beta(m) \cdot \sum_{i=0}^{n-1} a_i^2. \tag{50}$$

Now we need to take into account the error due to summing the terms $\left(Z_j^h + Z_j^\ell\right)$ to obtain (S_h, S_ℓ). One easily notes that these terms are added exactly as the terms a_i^2 were added in Algorithm 10, hence we can readily adapt the proof of Lemma 4.1 and obtain

$$\left|(S_h + S_\ell) - \sum_{j=0}^{k-1}\left(Z_j^h + Z_j^\ell\right)\right| \leq \beta(k) \cdot \sum_{j=0}^{k-1}\left(Z_j^h + Z_j^\ell\right). \tag{51}$$

Combining (50) and (51), we obtain

$$\left|(S_h + S_\ell) - \sum_{i=0}^{n-1} a_i^2\right| \leq (\beta(k) + \beta(m) + \beta(k)\beta(m)) \cdot \sum_{i=0}^{n-1} a_i^2, \tag{52}$$

which gives (48). When we develop this expression using the definition of β, we obtain a fairly complex expression (whose first term of the series in u is $(2m + 2k - 2)u^2 + \mathcal{O}(u^3)$), but with the further assumptions $n < 1/u$ (that will be needed anyway in Section 4.2) and $u \leq 1/32$ (which always holds in practice and is needed anyway for the correctness of the DW square root algorithms), we can obtain a much simpler expression. We have seen at the end of the Proof of Lemma 4.1 (to deduce (42)) that if

$t \leq 1/u$ then $\beta(t,u) \leq (2t-1)u^2 + (t+1)\,u^3$. Of course, $n < 1/u$ implies $k < 1/u$ and $m < 1/u$. With this simpler bound for β, the term $\beta(k) + \beta(m) + \beta(k)\beta(m)$ in (52) becomes

$$\begin{aligned}(2\,k - 2 + 2\,m)\,u^2 + (k + 2 + m)\,u^3 + (2\,k-1)\,(2\,m-1)\,u^4 \\ + ((2\,k-1)\,(m+1) + (k+1)\,(2\,m-1))\,u^5 + (k+1)\,(m+1)\,u^6 \quad (53)\end{aligned}$$

Let us now bound the sum of the terms of order u^3, u^4, u^5 and u^6 in (53). Since $u \leq 1/32$, $u^4 \leq u^3/32$, $u^5 \leq u^3/1024$, and $u^6 \leq u^3/32768$. Therefore, the sum of these terms is less than

$$\begin{aligned}&(k + 2 + m)\,u^3 + (2\,k-1)\,(2\,m-1)\,\frac{u^3}{32} \\ &\quad + ((2\,k-1)\,(m+1) + (k+1)\,(2\,m-1))\,\frac{u^3}{1024} + (k+1)\,(m+1)\,\frac{u^3}{32768} \\ &\qquad = \frac{(4225\,km + 30753\,k + 30753\,m + 66497)}{32768} \cdot u^3,\end{aligned}$$

which is less than $(0.129n + 0.939(k+m) + 2.03) \cdot u^3$. □

Let us now compare the bounds of Algorithm 10 and Algorithm 11, i.e., the bounds (41) and (48), and the bound of Graillat et al's algorithm [13] (which comes from the relative error bound $3u^2$ of Algorithm 5), namely

$$\frac{3(n-1)u^2}{1 - 3(n-1)u^2}. \quad (54)$$

We have the following property:

Property 4.4.

- *if $k = 1$ and $m = n$ or $k = n$ and $m = 1$ then Algorithm 11 boils down to Algorithm 10;*
- *as soon as $n \geq 3$, $u \leq 1/32$, and $3(n-1)u^2 < 1$ (which is necessary for (54) to make sense), the bound (54) is larger than the bound (42);*
- *if $k \geq 2$ and $m \geq 2$, assuming $u \leq 1/32$ and $n \leq 1/u$, the bound (42) is larger than the bound (49).*

Property 4.4 shows that in all practical cases, our blockwise algorithm has a better error bound than our sequential algorithm, which has a better error bound that Graillat et al.'s algorithm.

Proof. Let us first show that, under the conditions $n \geq 3$,$u \leq 1/32$, and $3(n-1)u^2 < 1$, the bound (54) is larger than the bound (42). Let us subtract the second bound from

the first one, we have

$$\begin{aligned}\frac{3(n-1)u^2}{1-3(n-1)u^2} - \left((2n-1)u^2 + (n+1)\,u^3\right) & \\ \geq 3(n-1)u^2 - \left((2n-1)u^2 + (n+1)\,u^3\right) & \\ = (n-2)u^2 - (n+1)\,u^3 & \\ \geq (n-2)u^2 - \left(\frac{n}{32} + \frac{1}{32}\right)u^2 &\end{aligned}$$

(since $u \leq 1/32$) Therefore, the difference of both bounds is larger than

$$\left(\frac{31n}{32} - \frac{65}{32}\right)u^2,$$

which is positive as soon as $n \geq 3$.

Now, let us show that if $k \geq 2$, $m \geq 2$, $u \leq 1/32$ and $n \leq 1/u$, the bound (42) is larger than the bound (49). The difference of both bounds is

$$\begin{aligned}&\left((2n-1)u^2 + (n+1)\,u^3\right) \\ &\qquad - \left((2k+2m-2)\cdot u^2 + (0.129n + 0.939(k+m) + 2.03)\cdot u^3\right)\end{aligned}$$

which is equal to

$$(2n - 2(k+m) + 1)\,u^2 + (0.871n - 0.939(k+m) - 1.03)\,u^3. \tag{55}$$

Lemma 4.2, implies that $k+m \leq \frac{n}{2}+2$. Using that in (55), we find that the difference "bound (42) minus bound (49)" is larger than

$$(n-3)u^2 + (0.4015n - 2.908)u^3 \geq (n-3)u^2 - 2.908u^3.$$

Since $u \leq 1/32$, this difference is larger than $(n - 3 - 2.908/32)u^2$, which is positive as soon as $n \geq 4$ (incidentally, $n \geq 4$ is implied by our hypothesis $k \geq 2$ and $m \geq 2$). □

Assuming $nu \ll 1$, so that the bound (48) is essentially $(2k+2m-2)\cdot u^2$, a direct consequence of Lemma 4.2 is that an approximate minimum of the bound (48) is reached when $k = r_n$ or $k = n/r_n$, where r_n is the largest divisor of n less than or equal to $\sqrt{n}$, resulting, if $r_n \approx \sqrt{n}$ in a relative error less than around $(4\sqrt{n}-2)\cdot u^2$. For instance, in IEEE 754 binary64 arithmetic ($u = 2^{-53}$), with $n = 6000$, the obtained relative error bounds are:

- $1.3322 \times 10^{-12}\,u$ with our sequential algorithm;
- $1.9981 \times 10^{-12}\,u$ with Graillat et al's algorithm;
- $3.3506 \times 10^{-14}\,u$ with the blockwise summation algorithm, with the near-optimal choices $k = 60$ and $m = 100$ or $k = 100$ and $m = 60$.

Note that even with a very unbalanced block splitting the blockwise summation is significantly more accurate than the sequential summation. Still with the same values of n and u, the error bound becomes 3.3374×10^{-13} in the case $k = 4$ and $m = 1500$. Beyond being more accurate, the blockwise version exhibits more parallelism than the sequential version, which may lead to better overall performance. Of course, if n is prime or has divisors that do not allow for a balanced splitting, it may be worth using the blockwise algorithm with a number of elements slightly larger than n (by appending additional zero elements to the vector $(a_0, a_1, \ldots, a_{n-1})$).

4.1.3 Obtaining the Euclidean norm barring underflow/overflow

We can now combine Lemma 4.3 and Theorem 3.10, and obtain.

Theorem 4.5. *Assume that for all i, $a_i \in \mathrm{MED}$. Assume that Algorithm 10 (sequential summation) or Algorithm 11 (blockwise summation) is used to compute the approximation (S_h, S_ℓ) to $\sum_{i=0}^{n-1} a_i^2$ (with k blocks of m elements, where $km = n$) and Algorithm SQRTDWtoFP (Algorithm 9) is used to approximate the square-root of $S_h + S_\ell$ by a floating-point number R. Let $\beta(t) = (2t-1)u^2 + (t-1)u^3 + (2t-2)u^4 + (t-1)u^5$, and define a parameter ν as follows:*

$$\nu = \frac{\beta(n)}{u^2} \quad \textit{if the sequential summation algorithm is used} \tag{56a}$$

$$\nu = \frac{\beta(k) + \beta(m) + \beta(k)\beta(m)}{u^2} \quad \textit{if the blockwise summation algorithm is used (general case)} \tag{56b}$$

If $\nu < \frac{1}{2u}$, we have:

$$\left| R - \sqrt{\sum_{i=0}^{n-1} a_i^2} \right| \leq \left(\frac{1}{2} + u \cdot \left(\frac{7}{4} + \frac{\nu}{1 - \nu \cdot u^2} \right) \right) \mathrm{ulp}\left(\sqrt{\sum_{i=0}^{n-1} a_i^2} \right). \tag{57}$$

Note that if $n < 1/u$ and $u \leq 1/32$, we can use in Theorem 4.5 the following simpler expressions for ν (see Lemmas 4.1 and 48):

$$\nu = (2n - 1) + (n + 1)\, u \tag{58}$$

with the sequential summation algorithm, and

$$\nu = (2k + 2m - 2) + (0.129n + 0.939(k + m) + 2.03) \cdot u \tag{59}$$

with the blockwise summation algorithm. Condition $\nu < 1/(2u)$ is not that restrictive: in the binary32/single precision format of the IEEE 754 Standard, assuming we use the sequential algorithm for summing the terms a_i^2, it is satisfied for $n \leq 4194304$, and in the binary64/double precision format, it is satisfied for $n \leq 2,251,799,813,685,248$. Even larger values of n can be reached if we use the blockwise algorithm: in the binary32 format, choosing k equal to the largest divisor of n less than or equal to $\sqrt{n}$, the condition is satisfied for $n = 500,000,000,000$.

If we use our algorithm for computing the Euclidean norm of a vector of 10000 binary64 elements, the returned result will be within 0.500000000002221 ulp from the exact result with the sequential summation, and within 0.5000000000000444 ulp from the exact result with the blockwise summation and the choice $k = m = 100$. This means that we will almost always obtain a correctly rounded result.

4.2 Computing Euclidean norms in the general case

4.2.1 Choice of the parameters minmed, maxmed, t_{tiny}, and t_{big}

Let us compute Euclidean norms, still using DW arithmetic, now in the general case (i.e., we no longer assume that underflow and overflow cannot occur). We will use the three-class approach presented in Section 2.1, very much like what is done by Graillat et al. [13], but with different choices for the parameters minmed, maxmed, t_{tiny}, and t_{big}, for two reasons:

- first, from (4), if we wish to express the squares of the elements of MED as DW numbers without error, we need to replace Constraint (6a) of Section 2.1 by

$$\mathrm{minmed} \geq \eta;$$

- then, to make the necessity of scaling as infrequent as possible, and to make as small as possible the error committed when we neglect the elements of TINY because BIG is nonempty, we try to have maxmed as large as possible and minmed as small as possible.

Exactly as is done by Graillat et al., we choose maxmed and minmed equal to powers of 2, and to avoid introducing additional rounding errors in the scalings, we choose t_{big} and t_{tiny} equal to *even* powers of 2. To simplify the analysis we also choose $t_{\mathrm{big}} = 1/t_{\mathrm{tiny}}$. We also assume $n_{\max} = 1/u = 2^p$, i.e., we wish to guarantee a correct behavior of the algorithms for vectors of dimension up to 2^p. To simplify the analysis, we also assume $u \leq 1/32$, which always holds in practice. In particular, this makes it possible to use the simpler expressions (58) and (59) for variable ν in Theorem 4.5.

Lemma 3.3 implies that if maxmed is a power of 2, when summing, using Algorithm 10 and/or Algorithm 11, n numbers less than maxmed^2, with $n \leq 1/u = 2^p$, the computed result is less than $n \cdot \mathrm{maxmed}^2$. This gives the following constraint on maxmed:

$$2^p \cdot \mathrm{maxmed}^2 \leq \Omega.$$

This leads us to the following choices:

- minmed is the power of 2 just above or equal to η, i.e.,

$$\mathrm{minmed} = 2^{\lceil (e_{\min}+p)/2 \rceil}.$$

  (with this choice we still can observe rare cases where the low-order element of a DW number generated by the computation of a square is subnormal, but this has no influence on accuracy, even if it can have one on performance)

- maxmed is the power of 2 just below $\sqrt{\Omega/2^p}$, i.e.,

$$\mathrm{maxmed} = 2^{\lfloor (e_{\max}-p)/2 \rfloor}.$$

With these definitions, we have

$$\frac{\mathrm{minmed}}{\mathrm{maxmed}} = 2^{-e_{\max}+p+1}, \tag{60}$$

and

$$\mathrm{minmed} \cdot \mathrm{maxmed} \in \{1, 2\}, \tag{61}$$

which will be useful later on.

- Concerning t_{tiny} and t_{big}, the possible values of these parameters are induced by the choices of minmed and maxmed, the constraints (6c), (6d), (6e), and (6f) presented in Section 2.1, and the additional constraint that $t_{\mathrm{tiny}} = 1/t_{\mathrm{big}}$ is an even power of 2.

We will assume

$$3p + 1 \leq e_{\max}, \text{ i.e., } e_{\min} \leq -3p. \tag{62}$$

Table 4 gives the various parameters and constraints associated with our algorithm for the binary16, binary32, binary64 and binary128 formats of the IEEE 754-2019 Standard for FP Arithmetic [21], and the bfloat16 format [16]. One can see that among all these formats, binary16 is the only one for which the various constraints required by our algorithm are not satisfied.

Our choice $t_{\mathrm{tiny}} = 1/t_{\mathrm{big}}$ constraints even more the possible values of t_{big} and t_{tiny}. Table 5 compares the obtained values for the binary32, binary64, and binary128 formats with the values used by Graillat et al. [13].

Table 4: The various parameters of our algorithm for the binary16 format of IEEE 754-2019 [21], the bfloat16 format [16], and the binary32, binary64, and binary128 formats of IEEE 754-2019.

parameters	binary16	bfloat16	binary32	binary64	binary128
p	11	8	24	53	113
$e_{\max}$	15	127	127	1023	16383
α	2^{-24}	2^{-133}	2^{-149}	2^{-1074}	2^{-16494}
η	$2^{-3/2}$	2^{-59}	2^{-51}	$2^{-969/2}$	$2^{-16269/2}$
Ω	65504	$2^{128}-2^{120}$ $\approx 3.390\times10^{38}$	$2^{128}-2^{104}$ $\approx 3.403\times10^{38}$	$2^{1024}-2^{971}$ $\approx 1.798\times10^{308}$	$2^{16384}-2^{16271}$ $\approx 1.190\times10^{4932}$
minmed	$1/2$	2^{-59}	2^{-51}	2^{-484}	2^{-8134}
maxmed	4	2^{59}	2^{51}	2^{485}	2^{8135}
Constraints on t_{big} ((6c), (6d) and even power of 2)	$1/4 \le t_{\text{big}} \le 2^{-14}$ (IMPOSSIBLE)	2^{-118} $\le t_{\text{big}}$ $\le 2^{-70}$	2^{-102} $\le t_{\text{big}}$ $\le 2^{-78}$	2^{-968} $\le t_{\text{big}}$ $\le 2^{-540}$	2^{-16268} $\le t_{\text{big}}$ $\le 2^{-8250}$
Constraints on t_{tiny} ((6e), (6f) and even power of 2)	$2^{24} \le t_{\text{tiny}} \le 4$ (IMPOSSIBLE)	2^{74} $\le t_{\text{tiny}}$ $\le 2^{118}$	2^{98} $\le t_{\text{tiny}}$ $\le 2^{102}$	2^{590} $\le t_{\text{tiny}}$ $\le 2^{968}$	2^{8360} $\le t_{\text{tiny}}$ $\le 2^{16268}$
Constraint $3p+1 \le e_{\max}$	NOT SATISFIED	satisfied	satisfied	satisfied	satisfied

Table 5: The parameters minmed, maxmed, t_{big}, and t_{tiny} in binary32 arithmetic, for our algorithm and for Graillat et al.'s algorithm [13].

	binary32		binary64		binary128	
	our solution	Graillat et al.	our solution	Graillat et al.	our solution	Graillat et al.
minmed	2^{-51}	2^{-60}	2^{-484}	2^{-376}	2^{-8134}	2^{-5536}
maxmed	2^{51}	2^{34}	2^{485}	2^{324}	2^{8135}	2^{5424}
t_{tiny}	2^{γ}, where $\gamma \in \{98, 100, 102\}$	2^{94}	2^{γ}, where γ is even and $590 \leq \gamma \leq 968$	2^{700}	2^{γ}, where γ is even and $8360 \leq \gamma \leq 16268$	2^{10960}
t_{big}	$\frac{1}{t_{\mathrm{tiny}}}$	$\frac{1}{t_{\mathrm{tiny}}} = 2^{-94}$	$\frac{1}{t_{\mathrm{tiny}}}$	$\frac{1}{t_{\mathrm{tiny}}} = 2^{-700}$	$\frac{1}{t_{\mathrm{tiny}}}$	$\frac{1}{t_{\mathrm{tiny}}} = 2^{-10960}$

4.2.2 Obtaining the result from the intermediate sums of squares

The sum $S_{\mathrm{med}} = \sum_{a_i \in \mathrm{MED}} a_i^2$ of the elements of MED is approximated by a double-word $(S_{\mathrm{med}}^h, S_{\mathrm{med}}^\ell)$, obtained using Algorithm 10 (sequential summation) or Algorithm 11 (blockwise summation with k blocks of m elements, where $km = n$). This approximation satisfies

$$\left| \left(S_{\mathrm{med}}^h + S_{\mathrm{med}}^\ell\right) - \sum_{a_i \in \mathrm{MED}} a_i^2 \right| \leq \nu u^2 \sum_{a_i \in \mathrm{MED}} a_i^2,$$

where ν is defined in (58) if we use Algorithm 10, and (59) if we use Algorithm 11. These algorithms are also applied to the elements of BIG and TINY pre-multiplied by t_{big} and t_{tiny} respectively. Similarly, this gives double words $(S_{\mathrm{big}}^h, S_{\mathrm{big}}^\ell)$ and $(S_{\mathrm{tiny}}^h, S_{\mathrm{tiny}}^\ell)$ that satisfy

$$\left| \left(S_{\mathrm{big}}^h + S_{\mathrm{big}}^\ell\right) - \sum_{a_i \in \mathrm{BIG}} (t_{\mathrm{big}} a_i)^2 \right| \leq \nu u^2 \sum_{a_i \in \mathrm{BIG}} (t_{\mathrm{big}} a_i)^2$$

and

$$\left| \left(S_{\mathrm{tiny}}^h + S_{\mathrm{tiny}}^\ell\right) - \sum_{a_i \in \mathrm{TINY}} (t_{\mathrm{tiny}} a_i)^2 \right| \leq \nu u^2 \sum_{a_i \in \mathrm{TINY}} (t_{\mathrm{tiny}} a_i)^2.$$

This gives

$$\begin{aligned} \left| \frac{1}{t_{\mathrm{tiny}}^2} \left(S_{\mathrm{tiny}}^h + S_{\mathrm{tiny}}^\ell\right) + \left(S_{\mathrm{med}}^h + S_{\mathrm{med}}^\ell\right) + \frac{1}{t_{\mathrm{big}}^2} \left(S_{\mathrm{big}}^h + S_{\mathrm{big}}^\ell\right) - \sum_{i=0}^{n-1} a_i^2 \right| \\ \leq \nu u^2 \sum_{i=0}^{n-1} a_i^2. \end{aligned} \tag{63}$$

Now, we can follow a reasoning similar to that of Graillat et al. [13]. Denote $S_{\mathrm{med}} = S_{\mathrm{med}}^h + S_{\mathrm{med}}^\ell$, $S_{\mathrm{big}} = S_{\mathrm{big}}^h + S_{\mathrm{big}}^\ell$, and $S_{\mathrm{tiny}} = S_{\mathrm{tiny}}^h + S_{\mathrm{tiny}}^\ell$. As the expression

$$\frac{1}{t_{\mathrm{tiny}}^2} \left(S_{\mathrm{tiny}}^h + S_{\mathrm{tiny}}^\ell\right) + \left(S_{\mathrm{med}}^h + S_{\mathrm{med}}^\ell\right) + \frac{1}{t_{\mathrm{big}}^2} \left(S_{\mathrm{big}}^h + S_{\mathrm{big}}^\ell\right)$$

cannot be used directly, we operate on a case-by-case basis according to whether BIG, MED, and TINY contain elements or not. These cases can be represented by triplets $(a, b, c) \in \{0, 1\}^3$, where 0 means "empty" and 1 means "nonempty" for BIG, MED, and TINY, respectively. For instance, "$(1, 0, 1)$" means "MED is empty, and BIG and TINY are nonempty". In all generality, there are eight cases to consider, but this number can be reduced thanks the following remarks.

- As soon as BIG is not empty,[9] we can neglect the elements of S_{tiny}. So Case $(1, 1, 1)$ reduces to Case $(1, 1, 0)$, and Case $(1, 0, 1)$ reduces to Case $(1, 0, 0)$.

[9] As pointed out by Graillat et al., there is no need to preliminarily check whether BIG is empty or not. One progressively accumulates sums of squares in two registers, initially dedicated to the elements of MED and TINY, and as soon as an element of BIG is met, the accumulation of the terms of TINY is abandoned, and the very same register is now used for accumulating the elements of BIG.

Indeed, in these cases, we have

$$\sum_{i=0}^{n-1} a_i^2 > \mathrm{maxmed}^2,$$

and

$$\sum_{a_i \in \mathrm{TINY}} a_i^2 \leq (n-1) \cdot \mathrm{minmed}^2.$$

Hence,

$$\sum_{a_i \in \mathrm{TINY}} a_i^2 < \frac{(n-1) \cdot \mathrm{minmed}^2}{\mathrm{maxmed}^2} \cdot \sum_{i=0}^{n-1} a_i^2 < 2^p \cdot \frac{\mathrm{minmed}^2}{\mathrm{maxmed}^2} \cdot \sum_{i=0}^{n-1} a_i^2. \quad (64)$$

Using (60), the term

$$2^p \cdot \frac{\mathrm{minmed}^2}{\mathrm{maxmed}^2} = 2^{-2e_{\max}+3p+2}$$

bounds the relative error committed by neglecting the elements of TINY in the summation. From (62), we deduce that it is less than or equal to u^3.

We therefore easily obtain

$$\left| \left(S_{\mathrm{med}} + \frac{1}{t_{\mathrm{big}}^2} S_{\mathrm{big}} \right) - \sum_{i=0}^{n-1} a_i^2 \right| \leq u^2 \cdot (\nu + u) \cdot \sum_{i=0}^{n-1} a_i^2. \quad (65)$$

Therefore $\sqrt{S_{\mathrm{med}} + \frac{1}{t_{\mathrm{big}}^2} S_{\mathrm{big}}}$ will be a good approximation to $\sqrt{\sum_{i=0}^{n-1} a_i^2}$ (the error of that approximation will be given later on). Hence we will compute

$$\sqrt{S_{\mathrm{med}} + \frac{1}{t_{\mathrm{big}}^2} S_{\mathrm{big}}}. \quad (66)$$

- Saying that MED is empty is equivalent to saying that $S_{\mathrm{med}} = 0$. So there is no need to develop Case $(1, 0, 0)$ further provided that what we do on Case $(1, 1, 0)$ is still correct when $S_{\mathrm{med}} = 0$.
- Likewise, saying that TINY is empty is equivalent to saying that $S_{\mathrm{tiny}} = 0$. So there is no need to develop Case $(0, 1, 0)$ further provided that what we do on Case $(0, 0, 1)$ is still correct when $S_{\mathrm{tiny}} = 0$.

We are therefore left with only four cases to consider: $(1, 1, 0)$, $(0, 1, 1)$, $(0, 0, 1)$, and $(0, 0, 0)$.

1. **If BIG is nonempty (Case $(1, 1, 0)$),** the computation must be carried on without underflows or overflows. More precisely, concerning underflow, we must make sure that no term becomes less than $2^{e_{\min}+p}$, otherwise it could not be represented accurately by a double-word number.

- if $S^h_{\rm med} < u^2{\rm minmed}^2/t^2_{\rm big}$ then $S_{\rm med} < u^2{\rm minmed}^2/t^2_{\rm big}$ (because the bound is a FP number), hence,

$$S_{\rm med} < \frac{S_{\rm big}}{t^2_{\rm big}}u^2.$$

Therefore, the term $S_{\rm med}$ can be neglected in front of the term $S_{\rm big}/t^2_{\rm big}$. More precisely,

$$\left|\frac{S_{\rm big}}{t^2_{\rm big}} - \sum_{i=0}^{n-1} a_i^2\right| \le u^2 \cdot \left(1+\nu+u+u^2\nu+u^3\right) \cdot \sum_{i=0}^{n-1} a_i^2, \qquad (67)$$

and one can return

$$\frac{1}{t_{\rm big}} \cdot {\rm SQRTDWtoFP}(S^h_{\rm big}, S^\ell_{\rm big}) = t_{\rm tiny} \cdot {\rm SQRTDWtoFP}(S^h_{\rm big}, S^\ell_{\rm big}).$$

- if $S^h_{\rm big} > {\rm maxmed}^2 \cdot t^2_{\rm big}/u^3$ then $S_{\rm big} > {\rm maxmed}^2 \cdot t^2_{\rm big}/u^3$. Also, Lemma 3.3 implies $S_{\rm med} \le n \cdot {\rm maxmed}^2 < {\rm maxmed}^2/u$, therefore

$$S_{\rm med} < \frac{S_{\rm big}}{t^2_{\rm big}}u^2,$$

and, as previously, (67) holds and one can return

$$t_{\rm tiny} \cdot {\rm SQRTDWtoFP}(S^h_{\rm big}, S^\ell_{\rm big}).$$

- if $S^h_{\rm med} \ge u^2{\rm minmed}^2/t^2_{\rm big}$ and $S^h_{\rm big} \le {\rm maxmed}^2 \cdot t^2_{\rm big}/u^3$, then $S_{\rm med} \ge (u^2{\rm minmed}^2/t^2_{\rm big})(1-u)$ and $S_{\rm big} \le {\rm maxmed}^2 \cdot t^2_{\rm big}(1+u)/u^3$. Consider

$$\chi = \frac{1}{t_{\rm big}}S_{\rm big} + t_{\rm big}S_{\rm med} = t_{\rm tiny}S_{\rm big} + t_{\rm big}S_{\rm med}.$$

The number χ can be computed without underflow or overflow:

(a) **Overflow:** we have $t_{\rm big}S_{\rm med} \le t_{\rm big} \cdot n \cdot {\rm maxmed}^2 \le t_{\rm big} \cdot \Omega$, and

$$\begin{aligned} t_{\rm tiny}S_{\rm big} &\le \frac{t_{\rm tiny}{\rm maxmed}^2 t^2_{\rm big}(1+u)}{u^3} \\ &\le t_{\rm big}\left(\frac{1}{u}{\rm maxmed}^2\right) \cdot \frac{1+u}{u^2}, \\ &\le t_{\rm big} \cdot \Omega \cdot \frac{1+u}{u^2}. \end{aligned}$$

Therefore,

$$\begin{aligned} \chi &\le \Omega \cdot t_{\rm big} \cdot \left(1+(1+u)/u^2\right) \\ &\le {\rm maxmed} \cdot \left(1+u+u^2\right) \cdot 2^{2p} \\ &< 2^{(3p+e_{\max})/2} \cdot \left(1+u+u^2\right), \end{aligned}$$

and since $u \leq 1/32$ implies $1 + u + u^2 < \sqrt{2}$ we deduce $\chi < 2^{(3p+e_{\max}+1)/2}$, which is less than or equal to $2^{e_{\max}} < \Omega$ since $3p + 1 \leq e_{\max}$.

(b) **Underflow:** we have $t_{\text{tiny}} S_{\text{big}} > S_{\text{big}}$, therefore the term $t_{\text{tiny}} S_{\text{big}}$ is larger than minmed^2, which is larger than $\eta^2 = 2^{e_{\min}+p}$. Using (6f), we also have

$$\begin{aligned} t_{\text{big}} S_{\text{med}} &\geq (\text{minmed}^2 \cdot t_{\text{tiny}}) \cdot u^2 \cdot (1-u) \\ &\geq (\text{minmed}^3/\alpha) \cdot u^2 \cdot (1-u) \\ &\geq 2^{3(e_{\min}+p)/2-(e_{\min}-p+1)-2p-1} \\ &\geq 2^{e_{\min}/2+p/2-2}, \end{aligned}$$

and (62) implies $e_{\min}/2 + 3p/2 \leq 0$. Hence

$$t_{\text{big}} S_{\text{med}} \geq 2^{e_{\min}+2p-2} \geq 2^{e_{\min}+p}.$$

Therefore, it suffices to compute χ in double-word arithmetic by summing $t_{\text{big}} \cdot (S^h_{\text{med}}, S^\ell_{\text{med}})$ and $t_{\text{tiny}} \cdot (S^h_{\text{big}}, S^\ell_{\text{big}})$ by the means of SloppyDWPlusDW (Algorithm 5). If we call $\hat{\chi}$ the computed result, namely

$$\hat{\chi} = \text{SloppyDWPlusDW}(t_{\text{tiny}} S^h_{\text{big}}, t_{\text{tiny}} S^\ell_{\text{big}}, t_{\text{big}} S^h_{\text{med}}, t_{\text{big}} S^\ell_{\text{med}}),$$

we obtain

$$|\hat{\chi} - \chi| \leq 3u^2 \chi. \tag{68}$$

Combined with (65) this gives

$$\left| \hat{\chi} - t_{\text{big}} \sum_{i=0}^{n-1} a_i^2 \right| \leq u^2 \cdot \left(\nu + 3 + u + 3u^2\nu + 3u^3\right) \cdot \sum_{i=0}^{n-1} a_i^2. \tag{69}$$

we then take the square root R of $\hat{\chi}$ by the means of SQRTDWtoFP (Algorithm 9), and multiply R by $\sqrt{t_{\text{tiny}}}$ (this last multiplication is errorless since t_{tiny} is an even power of two).

2. **If BIG is empty, and MED and TINY are nonempty (Case $(0, 1, 1)$),** we need to compute

$$\sqrt{\frac{1}{t_{\text{tiny}}^2} S_{\text{tiny}} + S_{\text{med}}}$$

without underflows or overflows. Note that this can be rewritten

$$\frac{1}{t_{\text{tiny}}} \sqrt{S_{\text{tiny}} + \frac{1}{t_{\text{big}}^2} S_{\text{med}}}. \tag{70}$$

The square-root part in (70) is exactly as (66) (with S_{med} replaced by S_{tiny} and S_{big} replaced by S_{med}). Furthermore, the terms S_{tiny}, S_{med} and S_{big} have the same bounds. Therefore the reasoning is exactly as previously, (the error bounds are slightly smaller because we no longer have the error term due to neglecting TINY) and we obtain:

Table 6: value of the comparison constants $\mathrm{minmed}^2 u^2/t_{\mathrm{big}}^2$ and $\mathrm{maxmed}^2 \cdot t_{\mathrm{big}}^2/u^3$ (assuming t_{big} is the largest allowed value, or equivalently t_{tiny} is the smallest allowed value) needed by the algorithm, for the binary32, binary64, and binary128 formats of IEEE 754-2019.

format	binary32 (with $t_{\mathrm{tiny}} = 2^{98}$)	binary64 (with $t_{\mathrm{tiny}} = 2^{590}$)	binary128 (with $t_{\mathrm{tiny}} = 2^{8360}$)
$\frac{\mathrm{minmed}^2 u^2}{t_{\mathrm{big}}^2}$	2^{46}	2^{106}	2^{226}
$\frac{\mathrm{maxmed}^2 t_{\mathrm{big}}^2}{u^3}$	2^{-22}	2^{-51}	2^{-111}

- if $S_{\mathrm{tiny}}^h < \mathrm{minmed}^2 u^2/t_{\mathrm{big}}^2$ or $S_{\mathrm{med}}^h > \mathrm{maxmed}^2 \cdot t_{\mathrm{big}}^2/u^3$ then we can return
$$\mathrm{SQRTDWtoFP}(S_{\mathrm{med}}^h, S_{\mathrm{med}}^\ell),$$
- otherwise, we can compute
$$\chi = \frac{1}{t_{\mathrm{big}}} S_{\mathrm{med}} + t_{\mathrm{big}} S_{\mathrm{tiny}} = t_{\mathrm{tiny}} S_{\mathrm{med}} + t_{\mathrm{big}} S_{\mathrm{tiny}}$$
in double-word arithmetic with one call to SloppyDWPlusDW, take its square-root R by the means of SQRTDWtoFP, and multiply R by $\sqrt{t_{\mathrm{big}}}$.

3. **If BIG and MED are empty (Case $(0, 0, 1)$),** then we return
$$t_{\mathrm{big}} \times \mathrm{SQRTDWtoFP}(S_{\mathrm{tiny}}^k, S_{\mathrm{tiny}}^\ell),$$
and the error bound of Theorem 4.5 applies.

4. **If BIG and MED and TINY are empty (Case $(0, 0, 0)$),** then we return 0. Note that considering this case is important as algorithm 9 requires a special treatment for a null input.

Table 6 gives the value of the comparison constants $\mathrm{minmed}^2 u^2/t_{\mathrm{big}}^2$ and $\mathrm{maxmed}^2 \cdot t_{\mathrm{big}}^2/u^3$ (assuming t_{big} is the largest allowed value in Table 5) needed by the algorithm.

4.2.3 Final error bound

Consider the term
$$\nu' := \nu + 3 + u + 3u^2\nu + 3u^3$$
that appears in (69). To be able to obtain a final error in ulps we must make sure that the condition "$2u\nu' < 1$" is satisfied, which corresponds to Condition (30) of Theorem 3.10 with ν replaced by ν'. Let us consider the cases of sequential summation and blockwise summation separately.

If the terms a_i^2 have been summed-up using sequential summation
In that case, ν is given by (58), which implies
$$2u\nu' = (4n + 4)u + (2n + 4)\,u^2 + (12n - 6)u^3 + (6n + 12)\,u^4,$$

and one easily checks that it is strictly less than 1 for $n \leq \frac{1}{4u} - 2$ (which we will assume thereafter) and $u \leq \frac{1}{32}$. Theorem 3.10 implies that the error is bounded by

$$\left(\frac{1}{2} + u \cdot \left(\frac{7}{4} + \frac{\nu'}{1 - \nu' \cdot u^2}\right)\right) \text{ulp}\left(\sqrt{\sum_{i=0}^{n-1} a_i^2}\right),$$

which is equal to

$$\left(\frac{1}{2} + u \cdot \frac{N}{D}\right) \text{ulp}\left(\sqrt{\sum_{i=0}^{n-1} a_i^2}\right),$$

where

$$\begin{aligned} N &= \frac{15}{4} + 2n + (n+2)\,u + \left(\frac{5n}{2} - \frac{13}{2}\right) u^2 + \left(\frac{5n}{4} + \frac{5}{2}\right) u^3 \\ &\quad + \left(-\frac{21n}{2} + \frac{21}{4}\right) u^4 + \left(-\frac{21n}{4} - \frac{21}{2}\right) u^5 \\ &\leq \frac{15}{4} + 2n + (n+2) + \left(\frac{5n}{2} - \frac{13}{2}\right) u^2 + \left(\frac{5n}{4} + \frac{5}{2}\right) u^3 + \frac{21}{4} u^4 \end{aligned}$$

and

$$D = 1 + (-2n-2)\,u^2 + (-n-2)\,u^3 + (-6n+3)\,u^4 + (-3n-6)\,u^5.$$

Since $n \leq \frac{1}{4u} - 2$, we have

$$\begin{aligned} N &\leq \frac{15}{4} + 2\,n + \left(\left(\frac{1}{4u} - 2\right) + 2\right) u + \left(\frac{5\,\left(\frac{1}{4u} - 2\right)}{2} - \frac{13}{2}\right) u^2 \\ &\quad + \left(\frac{5\,\left(\frac{1}{4u} - 2\right)}{4} + \frac{5}{2}\right) u^3 + \frac{21}{4} u^4 \\ &\leq 4 + 2n + \frac{5}{8} u \end{aligned}$$

since $u \leq 1/32$, and

$$\begin{aligned} D &\geq 1 + \left(-2\,\left(\frac{1}{4u} - 2\right) - 2\right) u^2 + \left(-\left(\frac{1}{4u} - 2\right) - 2\right) u^3 \\ &\quad + \left(-6\,\left(\frac{1}{4u} - 2\right) + 3\right) u^4 + \left(-3\,\left(\frac{1}{4u} - 2\right) - 6\right) u^5 \\ &= 1 - \frac{u}{2} + \frac{7}{4} u^2 - \frac{3}{2} u^2 + \frac{57}{4} u^4 \\ &\geq 1 - \frac{u}{2} \end{aligned}$$

since $u \leq 1/32$.

We finally obtain,

Theorem 4.6. *If $n \leq \frac{1}{4u} - 2$ and $u \leq \frac{1}{32}$ and if the sequential algorithm (Algorithm 10) is used for the summation of squares then our algorithm computes $\sqrt{\sum_{i=0}^{n-1} a_i^2}$ with an error bounded by*

$$\left(\frac{1}{2} + \frac{(4+2n)\,u + \frac{5}{8}u^2}{1 - \frac{u}{2}}\right) \text{ulp}\left(\sqrt{\sum_{i=0}^{n-1} a_i^2}\right),$$

without any risk of spurious underflow or overflow.

If the terms a_i^2 have been summed-up using blockwise summation

Assuming k blocks of m terms, with $km = n$, the value of ν is given by (59). We obtain

$$\begin{aligned} 2u\nu' = {} & (4(k+m)+2)u \\ & + (0.258km + 1.878(k+m) + 6.06)u^2 \\ & + (12(k+m) - 12)u^3 \\ & + (0.774km + 5.634(k+m) + 18.18)u^4. \end{aligned}$$

Now, using $km = n \leq 1/u$, and defining $\ell = k + m$, we obtain

$$\begin{aligned} 2u\nu' \leq {} & (4\ell + 2.258)u + (1.878\ell + 6.06)u^2 \\ & + (12\ell - 11.226)u^3 + (5.634\ell + 18.18)u^4, \end{aligned} \tag{71}$$

and one easily checks that it is strictly less than 1 for $\ell \leq \frac{1}{4u} - 2$ (which we will assume thereafter) and $u \leq \frac{1}{32}$. We will use Theorem 3.10 with ν replaced by $1/(2u)$ times the bound (71), i.e.,

$$2.0\,\ell + 1.1290 + (0.9390\,\ell + 3.030)\,u + (6.0\,\ell - 5.6130)\,u^2 + (2.8170\,\ell + 9.090)\,u^3$$

which will give an error bounded by

$$\left(\frac{1}{2} + u \cdot \frac{N}{D}\right) \text{ulp}\left(\sqrt{\sum_{i=0}^{n-1} a_i^2}\right),$$

where

$$\begin{aligned} N = {} & 2.879 + 2\ell + (0.939\ell + 3.03)\,u + (2.5\ell - 7.58875)\,u^2 \\ & + (1.17375\ell + 3.7875)\,u^3 \\ & + (-10.5\ell + 9.82275)\,u^4 + (-4.92975\ell - 15.9075)\,u^5 \\ \leq {} & 2.879 + 2\ell + (0.939\ell + 3.03)\,u + (2.5\ell - 7.58875)\,u^2 \\ & + (1.17375\ell + 3.7875)\,u^3 \\ & + 9.82275u^4 \end{aligned}$$

and

$$D = 1 + (-2.0\,\ell - 1.1290)\,u^2 + (-0.9390\,\ell - 3.030)\,u^3 \\ + (-6.0\,\ell + 5.6130)\,u^4 + (-2.8170\,\ell - 9.090)\,u^5.$$

Since $\ell \le \frac{1}{4u} - 2$, we obtain

$$\begin{aligned} N &\le 2.879 + 2\ell + \left(0.939\left(\tfrac{1}{4u} - 2\right) + 3.03\right)u + \left(2.5\left(\tfrac{1}{4u} - 2\right) - 7.58875\right)u^2 \\ &\quad + \left(1.17375\left(\tfrac{1}{4u} - 2\right) + 3.7875\right)u^3 + 9.82275u^4 \\ &= 3.11375 + 2\,\ell + 1.777\,u - 12.2953125\,u^2 + 1.44\,u^3 + 9.82275\,u^4 \\ &\le (3.12 + 2\ell) + 1.8u \end{aligned}$$

since $u \le 1/32$. Similarly,

$$\begin{aligned} D &\ge 1 + \left(-2.0\,\left(\tfrac{1}{4u} - 2\right) - 1.1290\right)u^2 + \left(-0.9390\,\left(\tfrac{1}{4u} - 2\right) - 3.030\right)u^3 \\ &\quad + \left(-6.0\,\left(\tfrac{1}{4u} - 2\right) + 5.6130\right)u^4 + \left(-2.8170\,\left(\tfrac{1}{4u} - 2\right) - 9.090\right)u^5. \\ &= 1 - \tfrac{u}{2} + 2.63625u^2 - 2.652u^3 + 16.90875u^4 - 3.456u^5 \\ &\ge 1 - \tfrac{u}{2} \end{aligned}$$

since $u \le 1/32$. This gives,

Theorem 4.7. *If $n \le 1/u$, $k+m \le \frac{1}{4u} - 2$ and $u \le \frac{1}{32}$ and if the blockwise algorithm (Algorithm 11) is used for the summation of squares, with k blocks of m elements, with $km = n$, then our algorithm computes $\sqrt{\sum_{i=0}^{n-1} a_i^2}$ with an error bounded by*

$$\left(\frac{1}{2} + \frac{(3.12 + 2(k+m))\,u + 1.8u^2}{1 - \frac{u}{2}}\right) \mathrm{ulp}\left(\sqrt{\sum_{i=0}^{n-1} a_i^2}\right),$$

without any risk of spurious underflow or overflow.

In all practical cases, if the decomposition in blocks is balanced enough, constraint "$k + m \le \frac{1}{4u} - 2$" in Theorem 4.7 is less strong than constraint $n \le 1/u$. More precisely, assume we choose $k = m = \lceil\sqrt{n}\rceil$ (i.e., we possibly extend the vector $(a_0, a_1, a_2, \ldots, a_{n-1})$ with additional zeros if n is not a perfect square). Constraint $n \le 1/u$ implies $\lceil\sqrt{n}\rceil < 1/\sqrt{u} + 1$, so that

$$k + m < \frac{2}{\sqrt{u}} + 2,$$

and one easily checks that for all precisions $p \ge 7$ (i.e., $u \le 1/128$), $\frac{2}{\sqrt{u}} + 2 < \frac{1}{4u} - 2$.

4.2.4 Examples

Table 7 gives the error bounds we obtain in the cases $n = 12$ and $n = 10000$, for the bfloat16, binary32, and binary64 formats.

Table 7: Maximum possible values of n assuming sequential and blockwise summations, and final error, expressed in ulp $\left(\sqrt{\sum_{i=0}^{n-1} a_i^2}\right)$ for $n = 12$ and $n = 10000$, for the bfloat16, binary32, and binary64 formats.

Format	bfloat16	binary32	binary64
max. value of n assuming seq. summation	62	$4,194,302$	2.25×10^{15}
max. value of n assuming blockw. summation	256	$16,777,216$	9.01×10^{15}
error bound for $n = 12$ assuming seq. summation	0.614 ulp	0.5000018 ulp	$0.5\underbrace{0000000000000}_{13}319$ ulp
error bound for $n = 12$ assuming blockw. summation with $k = 3, m = 4$	0.5699 ulp	0.50000106 ulp	$0.5\underbrace{0000000000000}_{13}197$ ulp
error bound for $n = 10000$ assuming seq. summation	N/A	0.5012 ulp	$0.5\underbrace{0000000000}_{10}2221$ ulp
error bound for $n = 10000$ assuming blockw. summation with $k = m = 100$	N/A	0.5000241 ulp	$0.5\underbrace{000000000000}_{12}45$ ulp

5 Numerical experiments

We have checked our algorithm and compared it with other solutions from the literature on two aspects: accuracy and speed. We decided to design two different implementations of our algorithm: an implementation in Julia, used for accuracy testings, and a C implementation, used for performance evaluation and comparison. The reason for this choice is simple: the versatility of Julia makes it much easier to play with different precisions. However, it is exactly the same algorithm that was implemented in both environments.

5.1 Accuracy testings

We have implemented the algorithm of Graillat et al. [13] as well as our algorithm (with the blockwise summation, i.e., Algorithm 11, with $k = 2$) in the Julia programming language, and we have measured the errors obtained with randomly chosen input arrays. We have performed experiments in the binary32 and binary64 floating-

Table 8: Median and maximum errors obtained over a random set of input values, and percentage of faithfully and correctly rounded results. For $S = 7, 8, \ldots, 14$, we generate $4096 \cdot 2^{14-S}$ arrays of random lengths between 2^{S-1} and 2^S then, for each components, we generate random exponents between $e_{\min} + p$ and $e_{\max} - p$ (to avoid non-spurious underflow/overflow) and random significand between 1 and $2 - 2u$. Each random series are uniform. An enclosure of the exact result is computed with the help of the Arb library [25] with enough precision to determine the correctly rounded floating-point number unambiguously (verified by the program).

		Error / u		Rounding	
Algorithm	p	Median	Maximum	Faithful	Correct
Ours	24	0.3415	0.9990	100 %	100 %
	53	0.2725	0.9989	100 %	100 %
Graillat et al.	24	0.3465	1.4916	100 %	87 %
	53	0.2730	1.4608	100 %	89 %

point formats of the IEEE 754 Standard. The input arrays are built as follows. For $S = 7, \ldots, 14$ we generate $4096 \times 2^{14-S}$ arrays of input values, whose length is uniformly generated between 2^{S-1} and 2^S. By doing this, very different possible array sizes are considered. The elements of the arrays are floating-point numbers. Their exponents are uniformly generated between $e_{\min} + p$ and $e_{\max} - p$, and their significands are uniformly generated between 1 and $2 - 2u$.

The reason for choosing this exponent range is that we want to avoid *non*-spurious underflow or overflow. Enclosures of the exact results are computed using Johansson's Arb library [25], with enough accuracy to allow one to determine the correct roundings of the exact results unambiguously. We initially wanted to perform experiments in the binary128 format too. For that purpose we wanted to use the GNU libquadmath library, which provides a software implementation of the binary128 format, but we discovered that its square root function (`sqrtq`) is not correctly rounded. As this is a requirement for the accuracy of the algorithm of Graillat et al. [13] as well as ours, the tests could not be performed in that format. Table 8 presents median and maximum errors, and the percentage of faithfully and correctly rounded results. Beware: the "100%" in the table can be a bit misleading: these results do not show that our algorithm always return correctly rounded values (indeed, *it cannot*), but that incorrectly rounded values are extremely unlikely in practice. Incorrectly rounded values are much more frequent with the algorithm of Graillat et al. [13], which is not surprising: that algorithm was designed to always return *faithfully rounded* values, and our tests show that this is indeed the case for all the input arrays we have built.

5.2 Performance evaluation

Our experiments have been performed in a similar way as the ones that were reported in [13]. As said above, we implemented our algorithm in the C language for evaluating its performance and for comparing it to three other algorithms:

Table 9: The four systems on which we performed our experiments.

machine	CPU	ISA	SIMD extension	k
ARM ThunderX2	Cavium ThunderX2	ARM v8.1	Neon	2
Intel Coffee Lake	Intel Core i7-8700	x86-64	AVX2	4
AMD Zen 2	AMD EPYC 7282	x86-64	AVX2	4
Intel Skylake	Intel Xeon Gold 6136	x86-64	AVX512	8

- the "*Naive* algorithm" is the straightforward implementation of (1). It does not prevent spurious overflow/underflow from happening, and can, in rare cases, be inaccurate when underflows occur;
- the "*Netlib* algorithm" (i.e., Hammarling's method presented in Section 2.1), as implemented by Graillat et al. [13],
- the *Graillat et al.* algorithm presented in [13].

All four algorithms are compared using the IEEE-754 binary64 format.

We performed our tests on four different machines, that we designate by the microarchitecture they are based on: *ARM ThunderX2*, *Intel Coffee Lake*, *AMD Zen 2* and *Intel Skylake*.

In Section 4.1.2 we have considered possible values of the number of blocks in the blockwise summation (i.e., variable k in Algorithm 11, in order to minimize the error bound. We have seen (see Property 4.4) that even $k = 2$ is a significant improvement, in terms of accuracy, compared to the sequential summation. Now, for a binary64 implementation, if we reason in terms of performance, the best choice is the maximum number of binary64 FPNs that fit in an SIMD vector. That number varies across the different extensions considered.

The main characteristics of the four used architectures are summarized in Table 9. In this table, we indicate for each system the processor name, the name of the instruction set architecture (column "ISA"), the name of the SIMD extension that was used to compile or to program the algorithms, and the chosen number k of blocks in the blockwise summation (taken equal to the number of binary64 FPNs that fit in an SIMD vector).

We retrieved the code of Graillat et al. [13] from `http://www.christoph-lauter.org/faithfulnorm.tgz`, and we directly used the plain C code they provide for the "*Naive*" and "*Netlib*" algorithms. They also provide an implementation of their Euclidean norm algorithm using intrinsics functions for manipulating AVX2 vectors. In particular, they use these intrinsics functions to make the inner-loop of their algorithm branch-free by using componentwise masking operations.

We have used the same techniques for implementing our algorithm, but we have added a small intermediate library to facilitate porting the code to different SIMD extensions. This library contains a type `vec_t` for the SIMD vectors, whose definition depends on the targeted extension. For example, when the code is compiled for the

AVX2 SIMD extension, `vec_t` is an alias for the type `__m256d`, and the componentwise addition of two SIMD vectors is defined by

```
inline vec_t vec_add(vec_t v1, vec_t v2) {
  return _mm256_add_pd(v1, v2);
}
```

On the other hand, when the code is compiled for the Neon extension of the ARM architecture, `vec_t` is now an alias for the `float64x2_t` type, and the componentwise addition of two vectors becomes

```
inline vec_t vec_add(vec_t v1, vec_t v2) {
  return vaddq_f64(v1, v2);
}
```

Note that an `__m256d` vector gathers 4 binary64 numbers, while a `float64x2_t` contains 2 binary64 numbers. Hence, the code we wrote is parameterized by a macro constant `vec_len` that takes for value 8, 4 or 2 depending on the targeted SIMD extension.

We used this small library for implementing our algorithm, and we also used it to re-implement the algorithm proposed by Graillat et al. [13]: on an AVX2 (with FMA) platform, the code we obtain is exactly the same as the one they wrote, but this allowed us to port it easily to the AVX512 and ARM Neon extensions.

The benchmark program of Graillat et al. [13] generates series of random numbers according to as specific *profile* and measures statistics on the time taken by the different algorithms. Three different profiles are considered here:

- AROUND_ONE generates floating-point numbers in a narrow range around one, with exponents between -5 and 5, so only the MED class is used in these test cases.
- FULL_RANGE generates numbers in the whole range of binary64 numbers, including subnormal ones: the exponent of each floating-point number is uniformly picked between -1074 and 1023.
- REALLY_SMALL selects floating-point numbers whose squares are subnormal numbers: the exponents of the floating-point numbers are picked between -1074 and -512. Note that subnormal inputs are expected to be much more frequent than with the FULL_RANGE profile.

Tables 10, 11, 12, and 13 present the timings obtained on these various systems. The Naive algorithm is, as expected, the fastest algorithm (but inaccurate and prone to spurious overflow/underflow). On Intel and AMD systems, our algorithm is generally slightly faster than Graillat et al.'s algorithm; the Netlib algorithm is as fast or faster with the AROUND_ONE profile, while slower with the FULL_RANGE profile. In the very particular case of the REALLY_SMALL profile, all performances are worse on Intel systems, whereas with AMD systems our algorithm and Graillat et al.'s are only slightly slower and the Netlib algorithm is faster. Results on the ARM architecture are quite different: timings do not depend on the input profile, and the Netlib algorithm

Table 10: Timing statistic comparisons of four algorithms of computation of the Euclidean norm on ARM ThunderX2, for three different array sizes, and three different profiles of input. For each entry, the mean value and standard deviation of a population of 100 000 runs is given.

ARM ThunderX2 (Neon) (pyxis-2.lyon.grid5000.fr)				
Algorithm	n	Timing averages in hundreds of nanoseconds		
		AROUND_ONE	FULL_RANGE	REALLY_SMALL
Naive	256	7(1)	7(1)	7(0)
	1024	28(1)	28(1)	28(1)
	4096	112(2)	112(3)	112(2)
Netlib	256	17(1)	17(1)	17(1)
	1024	64(2)	64(2)	65(3)
	4096	253(5)	254(4)	256(5)
Graillat et al.	256	35(1)	35(2)	35(1)
	1024	138(2)	138(2)	138(4)
	4096	552(8)	553(6)	552(8)
Ours	256	35(1)	35(1)	35(1)
	1024	136(2)	136(3)	139(4)
	4096	557(5)	557(4)	557(4)

Table 11: Timing statistic comparisons of four algorithms of computation of the Euclidean norm on Intel Skylake (AVX2), for three different array sizes, and three different profiles of input. For each entry, the mean value and standard deviation of a population of 100 000 runs is given.

Intel Skylake (AVX2) @3.2 GHz (nlbook)				
Algorithm	n	Timing averages in hundreds of nanoseconds		
		AROUND_ONE	FULL_RANGE	REALLY_SMALL
Naive	256	4(6)	4(7)	121(42)
	1024	15(8)	15(11)	380(141)
	4096	59(20)	58(20)	645(268)
Netlib	256	5(7)	13(13)	18(14)
	1024	16(8)	47(17)	68(14)
	4096	62(20)	183(37)	267(39)
Graillat et al.	256	9(10)	9(10)	20(16)
	1024	32(11)	32(15)	73(15)
	4096	120(25)	120(28)	291(40)
Ours	256	8(9)	8(11)	19(15)
	1024	27(11)	28(15)	69(16)
	4096	103(26)	104(30)	273(43)

Table 12: Timing statistic comparisons of four algorithms of computation of the Euclidean norm on AMD Zen2, for three different array sizes, and three different profiles of input. For each entry, the mean value and standard deviation of a population of 100 000 runs is given.

AMD Zen2 (AVX2) (arunch)				
Algorithm	n	Timing averages in hundreds of nanoseconds		
		AROUND_ONE	FULL_RANGE	REALLY_SMALL
Naive	256	2(0)	4(0)	4(0)
	1024	10(0)	14(1)	14(1)
	4096	39(1)	57(2)	55(2)
Netlib	256	4(0)	11(1)	8(1)
	1024	16(1)	44(1)	30(1)
	4096	65(2)	177(5)	121(4)
Graillat et al.	256	6(0)	6(0)	8(1)
	1024	21(0)	21(1)	32(1)
	4096	83(3)	83(3)	129(4)
Ours	256	5(0)	5(0)	7(1)
	1024	17(0)	17(0)	27(1)
	4096	67(2)	67(2)	106(3)

Table 13: Timing statistic comparisons of four algorithms of computation of the Euclidean norm on Intel Skylake (AVX512), for three different array sizes, and three different profiles of input. For each entry, the mean value and standard deviation of a population of 100 000 runs is given.

Intel Skylake (AVX512) @3.0 GHz (srunch)				
Algorithm	n	Timing averages in hundreds of nanoseconds		
		AROUND_ONE	FULL_RANGE	REALLY_SMALL
Naive	256	4(0)	6(1)	11(1)
	1024	15(0)	23(2)	40(3)
	4096	61(1)	91(4)	158(6)
Netlib	256	5(0)	13(3)	19(3)
	1024	16(1)	50(6)	73(6)
	4096	62(1)	196(12)	288(11)
Graillat et al.	256	5(1)	6(1)	15(2)
	1024	18(1)	18(1)	58(3)
	4096	68(1)	68(2)	228(7)
Ours	256	5(0)	6(1)	15(2)
	1024	16(0)	16(1)	54(3)
	4096	61(1)	61(2)	211(6)

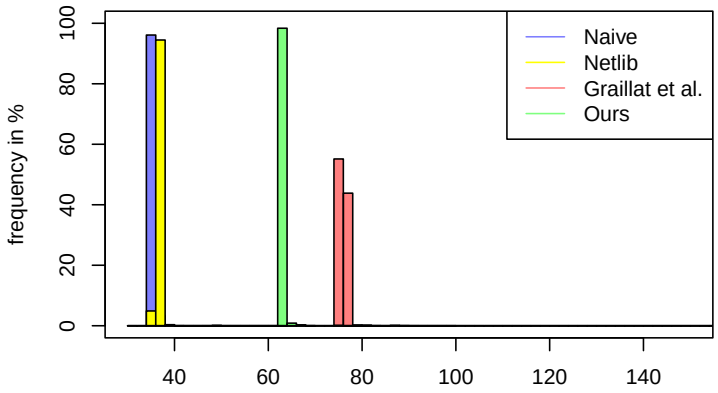

Figure 3: Repartition of obtained timings on Intel Skylake for the AROUND_ONE Profile (unit $= 10^{-7}$ seconds).

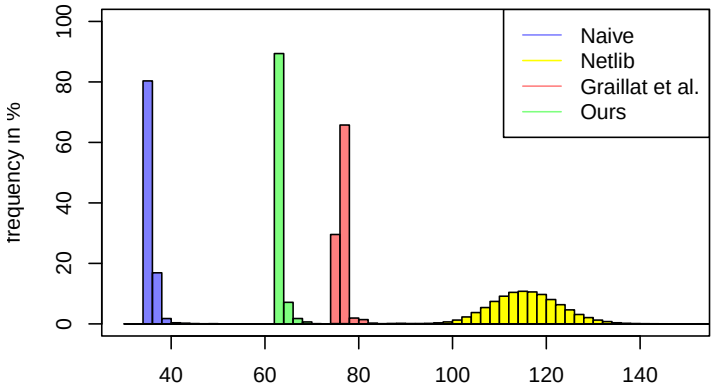

Figure 4: Repartition of obtained timings on Intel Skylake for the FULL_RANGE Profile (unit $= 10^{-7}$ seconds).

is consistently faster. The large standard deviation observed for our algorithm with $n = 256$ and the FULL_RANGE is due to a single run.

To give more insight into the timings reported in the previous tables, where only average values and standard deviations are given, we present in Figures 3 and 4 the histograms of the timings measured on the Intel Skylake system, with input vectors of size $n = 4096$. With the AROUND_ONE profile (Fig. 3), the few observed slight variations are probably due to operating system hazards. With the FULL_RANGE profile (Fig. 4), the Netlib algorithm is the only one for which the timings significantly differ from the ones of the AROUND_ONE profile. The Netlib performances are clearly degraded and more scattered: these effects may be due to frequent changes of the scaling factor.

It is clear from Tables 10, 11, 12, and 13 that the performances of the various algorithms depend much on the platform being used. However, in any case, these experiments show that our algorithm performs quite nicely compared to the other algorithms while being more accurate.

Conclusion

We have presented algorithms that make it possible to compute euclidean norms of large vectors very accurately, and without spurious underflows or overflows. Our tests show that the performance of the "blockwise" version of our algorithm is in general slightly better than the performance of the slightly less accurate algorithm of Graillat et al., and significantly better than the significantly less accurate Netlib algorithm. Our work on the computation of euclidean norms also led us to obtain results on double-word arithmetic that can be of interest in other areas:

- we have shown that when the operands are positive, the DWPlusFP algorithm has relative error bound u^2, and that bound is asymptotically optimal;
- we have shown the asymptotic optimality of the already known error bound $3u^2$ for the SloppyDWPlusDW algorithm when the operands are positive;
- we have introduced new algorithms for computing square roots of double-word numbers (SQRTDWtoDW and SQRTDWtoFP), given an asymptotically optimal relative error bound for the first one, and an error bound in ulps for the second one.

Interestingly enough, avoiding spurious underflows and overflows and computing more accurately comes at a reasonable cost: the experiments presented in Section 5.2 show that our algorithm is never more than two times slower than the naive algorithm.

References

[1] Nelson H. F. Beebe. *The Mathematical-Function Computation Handbook: Programming Using the MathCW Portable Software Library*. Springer, 2017.

[2] James L. Blue. A portable fortran program to find the euclidean norm of a vector. *ACM Trans. Math. Softw.*, 4(1):15–23, March 1978.

[3] G. Bohlender, W. Walter, P. Kornerup, and D. W. Matula. Semantics for exact floating point operations. In P. Kornerup and D. W. Matula, editors, *Proceedings of the 10th IEEE Symposium on Computer Arithmetic*, pages 22–26. IEEE Computer Society Press, Los Alamitos, CA, June 1991.

[4] S. Boldo and M. Daumas. Representable correcting terms for possibly underflowing floating point operations. In *Proceedings of the 16th Symposium on Computer Arithmetic*, pages 79–86. IEEE Computer Society Press, Los Alamitos, CA, 2003.

[5] S. Boldo, C. Q. Lauter, and J. Muller. Emulating round-to-nearest-ties-to-zero "augmented" floating-point operations using round-to-nearest-ties-to-even arithmetic. *IEEE Transactions on Computers*, 70(7):1046 – 1058, july 2021.

[6] Sylvie Boldo, Stef Graillat, and Jean-Michel Muller. On the robustness of the 2sum and fast2sum algorithms. *ACM Trans. Math. Softw.*, 44(1), July 2017.

[7] Sylvie Boldo, Catherine Lelay, and Guillaume Melquiond. Formalization of real analysis: A survey of proof assistants and libraries. *Mathematical Structures in Computer Science*, 26(7):1196–1233, 2016.

[8] Sylvie Boldo and Guillaume Melquiond. Flocq: A unified library for proving floating-point algorithms in coq. In *2011 IEEE 20th Symposium on Computer Arithmetic*, pages 243–252, 2011.

[9] Sylvie Boldo and Guillaume Melquiond. *Computer Arithmetic and Formal Proofs: Verifying Floating-point Algorithms with the Coq System.* Elsevier, 2017.

[10] Carlos F. Borges. Algorithm 1014: An improved algorithm for hypot(x,y). *ACM Trans. Math. Softw.*, 47(1), December 2020.

[11] T. J. Dekker. A floating-point technique for extending the available precision. *Numerische Mathematik*, 18(3):224–242, 1971.

[12] J. Demmel and H. D. Nguyen. Numerical reproducibility and accuracy at exascale. In *21st IEEE Symposium on Computer Arithmetic*, pages 235–237, 2013.

[13] Stef Graillat, Christoph Lauter, Ping Tak Peter Tang, Naoya Yamanaka, and Shin'ichi Oishi. Efficient Calculations of Faithfully Rounded l2-Norms of n-Vectors. *ACM Transactions on Mathematical Software*, 41(4):1–20, October 2015.

[14] Richard J. Hanson and Tim Hopkins. Remark on algorithm 539: A modern fortran reference implementation for carefully computing the euclidean norm. *ACM Trans. Math. Softw.*, 44(3), December 2017.

[15] J. R. Hauser. Handling floating-point exceptions in numeric programs. *ACM Trans. Program. Lang. Syst.*, 18(2):139–174, 1996.

[16] Greg Henry, Ping Tak Peter Tang, and Alexander Heinecke. Leveraging the bfloat16 artificial intelligence datatype for higher-precision computations. In *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pages 69–76, 2019.

[17] Y. Hida, X. S. Li, and D. H. Bailey. Algorithms for quad-double precision floating-point arithmetic. In *ARITH-16*, pages 155–162, June 2001.

[18] Y. Hida, X.S. Li, and D.H. Bailey. C++/fortran-90 double-double and quad-double package, release 2.3.17. Accessible electronically at `http://crd-legacy.lbl.gov/~dhbailey/mpdist/`, March 2012.

[19] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, 2nd edition, 2002.

[20] T. E. Hull, T. F. Fairgrieve, and P. T. P. Tang. Implementing complex elementary functions using exception handling. *ACM Transactions on Mathematical Software*, 20(2):215–244, June 1994.

[21] IEEE. *IEEE Standard for Floating-Point Arithmetic (IEEE Std 754-2019)*. July 2019.

[22] C.-P. Jeannerod, N. Louvet, J.-M. Muller, and A. Panhaleux. Midpoints and exact points of some algebraic functions in floating-point arithmetic. *IEEE Transactions on Computers*, 60(2), February 2011.

[23] C.-P. Jeannerod, J. Muller, and P. Zimmermann. On various ways to split a floating-point number. In *25th IEEE Symposium on Computer Arithmetic, Amherst, MA, USA*, pages 53–60, June 2018.

[24] Claude-Pierre Jeannerod and Siegfried M. Rump. On relative errors of floating-point operations: optimal bounds and applications. *Mathematics of Computation*, 87:803–819, 2018.

[25] Fredrik Johansson. Arb: Efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Transactions on Computers*, 66(8):1281–1292, 2017.

[26] Mioara Joldes, Jean-Michel Muller, and Valentina Popescu. Tight and rigorous error bounds for basic building blocks of double-word arithmetic. *ACM Trans. Math. Softw.*, 44(2), 2017.

[27] W. Kahan. Lecture notes on the status of IEEE-754. PDF file accessible at `http://www.cs.berkeley.edu/~wkahan/ieee754status/IEEE754.PDF`, 1996.

[28] D. Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley, Reading, MA, 3rd edition, 1998.

[29] X. Li, J. Demmel, D. H. Bailey, G. Henry, Y. Hida, J. Iskandar, W. Kahan, A. Kapur, M. Martin, T. Tung, and D. J. Yoo. Design, implementation and testing of extended and mixed precision BLAS. Technical Report 45991, Lawrence Berkeley National Laboratory, 2000. `http://crd.lbl.gov/~xiaoye/XBLAS`.

[30] O. Møller. Quasi double-precision in floating-point addition. *BIT*, 5:37–50, 1965.

[31] Jean-Michel Muller, Nicolas Brisebarre, Florent de Dinechin, Claude-Pierre Jeannerod, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, Damien Stehlé, and Serge Torres. *Handbook of Floating-Point Arithmetic*. Birkhäuser Boston, 2010.

[32] Jean-Michel Muller, Nicolas Brunie, Florent de Dinechin, Claude-Pierre Jeannerod, Mioara Joldes, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, and Serge Torres. *Handbook of Floating-Point Arithmetic, 2nd edition*. Birkhäuser Boston, 2018. ACM G.1.0; G.1.2; G.4; B.2.0; B.2.4; F.2.1., ISBN 978-3-319-76525-9.

[33] Jean-Michel Muller and Laurence Rideau. Formalization of double-word arithmetic, and comments on "Tight and rigorous error bounds for basic building blocks of double-word arithmetic". *ACM Transactions on Mathematical Software (to appear)*, 2021.

[34] Jean-Michel Muller and Laurence Rideau. Formalization of double-word arithmetic, and comments on "Tight and rigorous error bounds for basic building blocks of double-word arithmetic". *ACM Transactions on Mathematical Software*, 2021. to appear.

[35] Y. Nievergelt. Scalar fused multiply-add instructions produce floating-point matrix arithmetic provably accurate to the penultimate digit. *ACM Transactions on Mathematical Software*, 29(1):27–48, 2003.

[36] E. Jason Riedy and James Demmel. Augmented arithmetic operations proposed for IEEE-754 2018. In *25th IEEE Symposium on Computer Arithmetic, Amherst, MA, USA*, pages 45–52, June 2018.

[37] Siegfried M Rump and Marko Lange. Error estimates for the summation of real numbers with application to floating-point summation. *BIT Numerical Mathematics*, 57:927–941, 2017.

[38] Siegfried M Rump and Marko Lange. Faithfully Rounded Floating-Point Computation. *ACM Transactions on Mathematical Software*, 46(3), 2020.