Laboratoire de l'Informatique du Parallélisme

Ecole Normale Supérieure de Lyon
Unité de recherche associée au CNRS n°1398

An Algorithm that Computes a Lower Bound on the Distance Between a Segment and $\mathbb{Z}^2$

Vincent Lefèvre

June 1997

Research Report N$^{\text{O}}$ 97-18

Ecole Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, France
Téléphone : (+33) (0)4.72.72.80.00 Télécopieur : (+33) (0)4.72.72.80.80
Adresse électronique : lip@lip.ens−lyon.fr

An Algorithm that Computes a Lower Bound on the Distance Between a Segment and $\mathbb{Z}^2$

Vincent Lefèvre

June 1997

Abstract

We give a fast algorithm for computing a lower bound on the distance between a straight line and the points of a regular grid. This algorithm is used to find worst cases when trying to round the elementary functions correctly in floating-point arithmetic, which consists in returning the machine number that is the closest (there are other rounding modes) to the exact result.

Keywords: elementary functions, floating-point arithmetic, rounding

Résumé

Nous donnons un algorithme rapide permettant de calculer une minoration de la distance entre un segment de droite et les points d'une grille régulière. Cet algorithme est utilisé pour trouver les pires cas lorsque l'on arrondit exactement les fonctions élémentaires en arithmétique à virgule flottante, ce qui consiste à renvoyer le nombre machine le plus proche (il existe d'autres modes d'arrondi) du résultat exact.

Mots-clés: fonctions élémentaires, arithmétique virgule-flottante, arrondi

An Algorithm that Computes a Lower Bound on the Distance Between a Segment and $\mathbb{Z}^2$

Vincent Lefèvre

June 1997

1 Introduction

Our goal is to provide exactly rounded elementary functions in floating-point arithmetic. That is, when computing $f(x)$, where f is exp, log, sin, cos, etc... and x is a "machine number", we want to always get the machine number that is the closest[1] to $f(x)$ [1]. To do this, we first compute (with a precision that is somewhat higher than the "target" precision) an approximation to $f(x)$. Then we round the approximation. The problem is to know if we get the same result as if we had rounded the exact value $f(x)$. Indeed, if the approximation to $f(x)$ is not accurate enough, we cannot ensure that $f(x)$ is correctly rounded; this problem is known as the *Table Maker's Dilemma*. To solve this problem, we must know with which precision we must carry out the intermediate calculations; that is, we must know the smallest possible non-zero value of $|f(x) - y|$ where x is a machine number and y is either a machine number or the average of two consecutive machine numbers, depending on the rounding mode.

For a given x, the above distance $|f(x) - y|$ is denoted d. We split the considered domain into very small intervals, and in each interval I, we look for the set S_I of machine numbers x for which d is less than a given real ε_I; we can choose ε_I small enough such that S_I is generally empty, but large enough so that we can approximate the function by degree-1 polynomials, i.e., segments. Approximating the function can be performed quickly enough; the part that takes most of the time is the tests themselves. As the total number of points x is very large, e.g., of the order of 10^{20} for the double-precision numbers, we need a very fast algorithm.

In the chosen domains, the machine numbers are regularly spaced, so that we can multiply the numbers by powers of two to consider that they are, in fact, integers. Thus the problem is now: what are the points on the given segment such that the x-coordinate is an integer and the distance between the

[1] We also consider other "rounding modes", e.g., we may want to get the largest machine number that is less than or equal to $f(x)$.

y-coordinate and the integers is less than a given ε? We recall that in most cases, there are no such points (from the choice of ε).

The naive approach consists in testing each point whose x-coordinate is an integer: each iteration requires an addition and a comparison by calculating modulo 1 (this is possible if one uses the integer arithmetic of the processor), and shifting the segment by ε upwards: with only one unsigned comparison, we can test if a given point is in the interval $[0, 2\varepsilon]$. The time required by the other operations can be neglected.

If the number of points to test, denoted N, is large enough (e.g., 1000 or larger), there exists a faster method, using the fact that the set S_I is generally empty: we can look for a lower bound on the distance d, and if d is larger than ε, then S_I is empty; otherwise, we can split the interval into subintervals and use this method with different parameters or use the naive approach.

The segment has an equation of the form $y = ax - b$, where x is restricted to a given interval, e.g., $0 \leq x < N$. In Section 2, we give some mathematical preliminaries and notations. In Section 3, we study the distribution of the points $k.a$ modulo 1, where k is an integer satisfying an inequality $0 \leq k < n$; in particular, we mention a theorem known as the three-distance theorem [2, 3, 4]. In Section 4, we give the algorithm, based on the properties described in Section 3.

2 Mathematical Preliminaries – Notations

$\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$, $\mathbb{N}$ respectively denote the sets of real numbers, rational numbers, integers and non-negative integers.

$\mathbb{R}/\mathbb{Z}$ is the additive group of the real numbers modulo 1. This set can be viewed as a circle, or the segment $[0, 1]$ where both points 0 and 1 are identified (i.e., reals 0 and 1 represent the same point). With this second representation, the point represented by 0 (or 1) can be regarded as an origin. If $a \in \mathbb{R}/\mathbb{Z}$ and $k \in \mathbb{N}$, k is said to be the (group) index of $k.a$ (in the group generated by a).

If $a \in \mathbb{R}$, its image in $\mathbb{R}/\mathbb{Z}$ will also be denoted a, as there is no possible confusion.

$[x, y]$ represents an interval of real numbers (open, if one has round brackets). $[[x, y]]$ represents an interval of integers. The symbol # denotes the cardinality of a finite set.

3 Properties of $k.a$ mod 1

In this section, we study the properties of the points $y = k.a$ modulo 1, where a is a given real number and k is an integer restricted to a given interval, e.g., satisfying $0 \leq k < N$ (where N is a given positive integer). Since $\mathbb{R}\backslash\mathbb{Q}$ is dense into $\mathbb{R}$ and thanks to topological properties, we can suppose that $a \notin \mathbb{Q}$ for the mathematical study; thus we avoid casual equalities (see below). The numbers a and y may be regarded as elements of $\mathbb{R}/\mathbb{Z}$. Let us take for $n \in \mathbb{N}$:

$$E_n = \{k.a \in \mathbb{R}/\mathbb{Z} : k \in \mathbb{N},\ k < n\}.$$

Since $a \notin \mathbb{Q}$, the set E_n has exactly n elements, i.e., there is no multiple-order point. On examples, we can see that the distribution of the points of E_n has very interesting properties. In particular, we will look for a construction of E_n based on distances between consecutive points on $\mathbb{R}/\mathbb{Z}$.

An example is given on the following figure. We chose a rational number $(17/45)$ for a to make the notations simpler and multiplied the rational numbers by 45 to get integers, and instead of dealing with $\mathbb{R}/\mathbb{Z}$, we deal with $\mathbb{Z}/45\mathbb{Z}$. Of course, in our example, we have chosen n small enough to avoid the problems mentioned above (casual equalities…).

For $0 \leq i < n$, the $e_{n,i}$'s denote the images of the points of E_n in $[0, 1)$, in increasing order. We define $e_{n,n} = 1$, which represents the same point as $e_{n,0} = 0$. The distances between two consecutive points on the segment $[0, 1]$ (or the circle $\mathbb{R}/\mathbb{Z}$) are the values $e_{n,i+1} - e_{n,i}$ for $0 \leq i < n$.

We now give a new construction of E_n (the equivalence will be proved later), based on distances. For all $n \geq 2$, we define a sign $s_n \in \{-1, +1\}$ and a sequence S_n of n 4-tuples

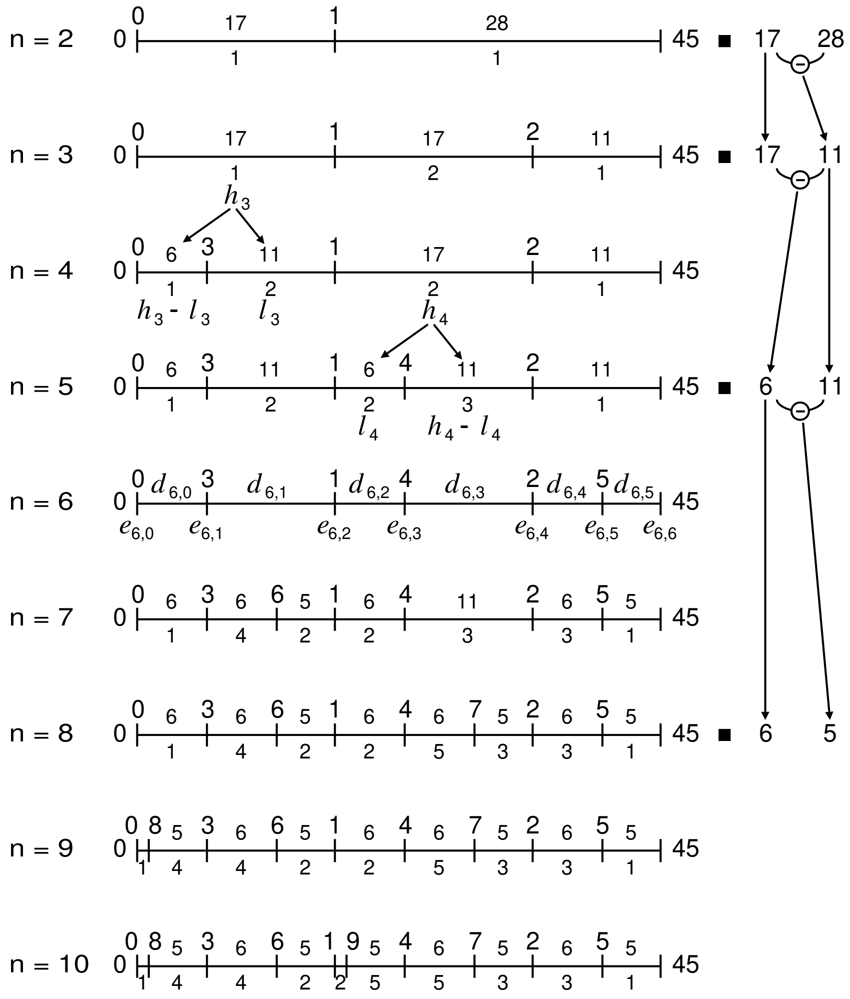
$$S_n = (d_{n,i}\ ,\ r_{n,i}\ ,\ j_{n,i}\ ,\ k_{n,i})_{0 \leq i < n}$$

where $d_{n,i}$ is a positive real number representing a *distance*, $r_{n,i}$ is a positive integer representing a *rank* associated with the distance, and both $j_{n,i}$ and $k_{n,i}$ are elements of $\mathbb{N}$ representing *group indices*. Let us take $D_n = \{d_{n,i} : 0 \leq i < n\}$, which is the set of the distances in S_n, $h_n = \max D_n$ and $\ell_n = \min D_n$; we will show that D_n has two or three elements only (this is the three-distance theorem). The sequence S_n and the sign s_n are defined by:

$$d_{2,0} = a, \quad d_{2,1} = 1 - a, \quad r_{2,0} = r_{2,1} = 1, \quad s_2 = \mathrm{sign}(1 - 2a),$$

$$j_{2,0} = 0, \quad k_{2,0} = 1, \quad j_{2,1} = 1, \quad k_{2,1} = 0,$$

and the following transformation. Let i be the unique index such that $d_{n,i} = h_n$ and the rank $r_{n,i}$ is minimal. The 4-tuple $(d_{n,i}\ ,\ r_{n,i}\ ,\ j_{n,i}\ ,\ k_{n,i})$ is replaced by two consecutive 4-tuples defined below; the other terms of the sequence remain unmodified and in the same order. The distances of the two 4-tuples are ℓ_n and $h_n - \ell_n$ but the order is determined by s_n: ℓ_n then $h_n - \ell_n$, if $s_n = +1$; $h_n - \ell_n$ then ℓ_n, if $s_n = -1$. The new ranks are the smallest positive integers such that all the ranks associated with the distance are different, i.e., all the couples (d, r) in the sequence are different; note that $h_n - \ell_n \neq \ell_n$ since a is irrational. The group indices $(j_{n,i}\ ,\ k_{n,i})$ are replaced by $(j_{n,i}\ ,\ n)$ and $(n\ ,\ k_{n,i})$. Finally, we take $s_{n+1} = s_n \cdot \mathrm{sign}(h_n - 2\ell_n)$, i.e., the sign of s_n changes if and



only if $\ell_{n+1} < \ell_n$; this choice ensures that intervals having the same length are split in the same way (see figure).

We can associate a function $f_n : [[0, n-1]] \to \mathbb{R}/\mathbb{Z}$ with each sequence S_n, such that each function f_n is a restriction of a function $f : \mathbb{N} \to \mathbb{R}/\mathbb{Z}$ with $f(0) = 0$ and $f(k) - f(j) = d \pmod 1$ for each (d, r, j, k) of a sequence S_n.

Let us take the last example. For $n = 2$, we have two points on the circle $\mathbb{Z}/45\mathbb{Z}$, with respective coordinates $f(0) = 0$ and $f(1) = 17$ (modulo 45). These two points form two intervals. The first interval has length 17, the left end is point 0, the right end is point 1, and the rank is 1 (initial interval); thus $(d, r, j, k) = (17, 1, 0, 1)$. The second interval has length $45 - 17 = 28$, the left end is point 1, the right end is point 0, and the rank is also 1; thus $(d, r, j, k) = (28, 1, 1, 0)$. Now, let us consider the $n = 3$ to 4 iteration. For

$n = 3$, we have $f(0) = 0$, $f(1) = 17$, and $f(2) = 34$. The interval of length $h_3 = 17$ and the minimal rank is $I = (17, 1, 0, 1)$. This 4-tuple is replaced by $I' = (6, 1, 0, 3)$ and $I'' = (11, 2, 3, 1)$ respectively. Since $d' + d'' = d$, $j' = j$, $k'' = k$, and $k' = j'' = n$, this transformation defines a new point $f(3) = 6$.

We now give the theorem showing that both constructions (E_n and S_n) are equivalent. It will be proved later.

Theorem 1 *For all $n \geq 2$ and $0 \leq i < n$, we have: $d_{n,i} = e_{n,i+1} - e_{n,i}$, $e_{n,i} = j_{n,i}.a$ and $e_{n,i+1} = k_{n,i}.a$, i.e., $\forall\, k \geq 0,\ f(k) = k.a$.*

Let us take $C_n = \#\{i : d_{n,i} = h_n\}$ and define a sequence (γ_i, δ_i):

$$(\gamma_0, \delta_0) = (a, 1 - a), \quad (\gamma_{i+1}, \delta_{i+1}) = (\min\{\gamma_i, \delta_i\}, |\gamma_i - \delta_i|),$$

i.e., at each iteration, one keeps the smaller element and replaces the larger one by the difference.

The following theorem says that some sequences S_n contain only two different distances, and the next pair of distances is obtained by replacing the larger distance with the difference. Between such two sequences, there is a transient period, where the three distances (both distances of the initial sequence, and the difference) are present.

Theorem 2 *There exists a strictly increasing function $\varphi : \mathbb{N} \to \mathbb{N}$ such that $\varphi(0) = 2$, and for all $i \geq 0$:*

$$D_{\varphi(i)} = \{\gamma_i\,,\,\delta_i\}, \quad \textit{and for } \varphi(i) < n < \varphi(i+1),\ D_n = \{\gamma_i\,,\,\delta_i\,,\,\delta_{i+1}\}.$$

For all i and n such that $\varphi(i) \leqslant n < \varphi(i+1)$, one has $\varphi(i+1) = n + C_n$. In particular, $\varphi(i+1) - \varphi(i) = C_{\varphi(i)}$.

◇ *Proof.* Theorem 2 is a direct consequence of the construction of the sequences S_n: we only use the fact that, at each iteration, an interval of length h_n is replaced by two intervals of lengths ℓ_n and $h_n - \ell_n$. □

Theorem 1 will be deduced from the following lemma.

Lemma 1 *For all n such that $\#D_n = 2$, i.e., $n \in \varphi(\mathbb{N})$:*

1. *$r_{n,0} = r_{n,n-1} = 1$;*
2. *if $s_n = +1$, then $d_{n,0} = \ell_n$ and $d_{n,n-1} = h_n$;*
   if $s_n = -1$, then $d_{n,0} = h_n$ and $d_{n,n-1} = \ell_n$;
3. *$j_{n,0} = k_{n,n-1} = 0$,*
   $k_{n,0} = \#\{i : d_{n,i} = d_{n,n-1}\}$,
   $j_{n,n-1} = \#\{i : d_{n,i} = d_{n,0}\}$;

4. for all (d, r, j, k), the values $j - r$ and $k - r$ only depend on the value of d.

◇ *Proof.* This lemma can be proved by induction on n. The main points are given in this proof. Details are left to the reader.

For the proof, (i, m) denotes the point i of the lemma for $n = m$ ($1 \leq i \leq 4$, $m \geq 2$).

We can easily verify that, from the definition of S_2, the lemma is true for $n = 2$.

Assume that the lemma is true for a given $n \in \varphi(\mathbb{N})$. Let us prove that it is still true for the next value $n' \in \varphi(\mathbb{N})$. We recall that, from the initial value n to the next value n', each interval of length h_n is split into two intervals.

First, let us see what happens for $n + 1$. The interval of length h_n and rank 1 is split into two intervals of lengths ℓ_n and $h_n - \ell_n$. From point $(1, n)$, this interval is adjacent to an end point of $[0, 1]$. From the construction of S_{n+1}, the two new intervals are placed in such a way that the interval of length $h_n - \ell_n$ is adjacent to the end point. This proves point $(1, n+1)$, therefore point $(1, n')$. Using point $(2, n)$ and the fact that s has changed if and only if ℓ has changed, this also proves point $(2, n')$.

Now, let us consider point 3. From the construction of the sequences, both indices $j_{m,0}$ and $k_{m,m-1}$ will still be zero for $m = n'$, and one of the indices $k_{m,0}$ and $j_{m,m-1}$, depending on the value of s_n, will not change from $m = n$ to $m = n'$. Let us denote this index by $\iota(m)$ and the other index by $\iota'(m)$. From points $(2, n)$ and $(3, n)$, this index $\iota(m)$ is equal to $\#\{i : d_{n,i} = h_n\}$, which is equal to $\#\{i : d_{n',i} = h_n - \ell_n\}$; this proves the part of point $(3, n')$ concerning $\iota(n')$. The other index $\iota'(n')$ will be added when $m = n + 1$, therefore it will be equal to n. And we have

$$\#\{i : d_{n',i} = \ell_n\} = \#\{i : d_{n,i} = \ell_n\} + \#\{i : d_{n,i} = h_n\} = \#S_n = n,$$

which proves the last part of point $(3, n')$.

Concerning point 4, let us start with $d = \ell_n$. By symmetry, we can take $s_n = +1$ (the opposite case is similar). For all (ℓ_n, r, j, k), where $1 \leq r \leq j_{n,n-1}$, we have $j = r - 1$ and $k = k_{n,0} + r - 1$. The new 4-tuple $(\ell_n, j_{n,n-1} + 1, j, k)$ in S_{n+1} will satisfy $j = j_{n,n-1}$ and $k = n$. Thus we still have $j = r - 1$ and $k = \#S_n = k_{n,0} + j_{n,n-1} = k_{n,0} + r - 1$. Distance $d = h_n - \ell_n$ (the other distance in $S_{n'}$) is new in S_{n+1}, so that there is nothing else to verify for the moment. We have just proved point $(4, n + 1)$. Points $(4, n + 2)$ to $(4, n')$ can be deduced from point $(4, n + 1)$ and the construction of the sequences. □

Now, we can prove Theorem 1.

◇ *Proof of Theorem 1.* We prove Theorem 1 by induction, in a way similar to the proof of the lemma. For $n = 2$, Theorem 1 is true.

Assume that the theorem is true for a given $n \in \varphi(\mathbb{N})$. Let us prove that it is true for $n+1$, then for the other values up to the next value in $\varphi(\mathbb{N})$.

By symmetry, we assume that $s_n = +1$. According to the lemma, we have $j_{n,n-1} + k_{n,0} = n$. Therefore

$$f(n) = j_{n,n-1}.a + \ell_n = j_{n,n-1}.a + d_{n,0} = j_{n,n-1}.a + k_{n,0}.a = n.a,$$

and the theorem is true for $n+1$. Considering the interval $(h_n, r+1, j, k)$, we have $j = j_{n,n-1} + r$ since $j - r$ is a constant (according to the lemma). Thus

$$f(n+r) = j.a + \ell_n = (j_{n,n-1} + r).a + k_{n,0}.a = (n+r).a$$

and the theorem is true for $n+r+1$. □

4 Algorithm

We will consider the successive $D_{\varphi(i)}$, and memorize the position of point b in the interval that contains this point (the distance from b to the lower bound of the interval) and the way in which the intervals are split, i.e., the values h_n, ℓ_n and s_n, where $n = \varphi(i)$. We recall that, at each iteration, the intervals of length h_n are split into two intervals of lengths ℓ_n and $h_n - \ell_n$ (in the order given by s_n), and the intervals of length ℓ_n remain unchanged. We stop when $n \geq N$, where N is the initial number of values to be tested. Then we can calculate the distance from b to the two ends of the interval.

In fact, we want to know whether the distance between the segment and $\mathbb{Z}^2$ is larger than ε or not. To avoid calculating the distance from b to the upper end of the interval, we apply the algorithm to $b+\varepsilon$ instead of b, i.e., the segment is shifted by ε downwards, and we only need to know the distance from b to the lower end of the interval, which is directly given by the algorithm.

Note that with this algorithm, we consider more points than wanted. But the number of considered points is bounded from above by twice the initial number of points, i.e., $2N$, which is not too large for our problem, since the value of N can be chosen such that the probability that the test fails is still small.

In order to avoid copying or swapping values and testing "status variables" (such as s), we will replace the variables ℓ and h by the variables $x = d_{n,0}$ and $y = d_{n,n-1}$ (thus we avoid swapping ℓ and h each time h becomes less than ℓ) and we will remove status variables, like s, by duplicating the code: one part for $s = +1$ and one part for $s = -1$. Thus we will know the position of h and ℓ without any test: $(x, y) = (\ell, h)$ in the part where $s = +1$, and $(x, y) = (h, \ell)$ in the part where $s = -1$. Instead of comparing ℓ and h, and updating s, we will compare x and y and perform a conditional branch.

We define two new variables: u and v, which denote the number of intervals of respective lengths x and y; they are only used for calculating n. The variable

b will be modified in such a way that it always contains the distance from the considered point to the lower end of the interval.

Of course, we will apply the algorithm to rational values, whereas the mathematical study considered irrational values for practical reasons. The algorithm remains the same, but we must be careful concerning the particular cases ($h = \ell$, then $\ell = 0$) and ensure there is no infinite loop.

We have four possible states:

- $h+$: the interval containing the point has length h and $s = +1$.
- $h-$: the interval containing the point has length h and $s = -1$.
- $\ell-$: the interval containing the point has length ℓ and $s = -1$.
- $\ell+$: the interval containing the point has length ℓ and $s = +1$.

In fact, we will group $h-$ and $\ell+$ (point b in the interval of length x), as well as $h+$ and $\ell-$ (point b in the interval of length y). The algorithm given below can be implemented in different ways; an optimization may require that some instructions are moved, removed or added.

```
Initialization:  x = a;  y = 1 − a;  u = v = 1;
Infinite loop:
    if (b < x)
        while (x < y)
            if (u + v ⩾ N) exit
            y −= x;  u += v;
        if (u + v ⩾ N) exit
        x −= y;  v += u;
    else
        b −= x;
        while (y < x)
            if (u + v ⩾ N) exit
            x −= y;  v += u;
        if (u + v ⩾ N) exit
        y −= x;  u += v;
```

If b is larger than 2ε, then the distance between the segment and $\mathbb{Z}^2$ is larger than ε. Otherwise the test fails, and we need a more accurate test (e.g., by splitting the segment or using a slower algorithm).

We notice that this algorithm "contains" Euclid's algorithm, which is used to compute the development of a into a continued fraction.

When one of the first partial quotients of the continued fraction of a is very large, the above algorithm is rather slow; for instance, x is much smaller than y (a partial quotient is large) and u is small (the partial quotient is one of the first ones), thus the number u of points added at each iteration is small, and many iterations are needed. It is possible to speed up the algorithm in these cases; however it will slow it down in the general case, which occurs much more often. Different solutions are possible, but they depend on the context in which the algorithm is applied.

5 Conclusion

The algorithms have been implemented on Sun SparcStations to find the value of x, among the double-precision floating-point numbers in $[\frac{1}{2}, 1]$, for which the distance between $\exp x$ and a machine number or a number equidistant to two consecutive machine numbers is minimal. The naive algorithm required 3 cycles per argument in average. The method described in this paper allowed to deal with 30 arguments per cycle in average (with non-optimal parameters).

With the above algorithm, we obtained a speed-up of 90 over the naive algorithm. This can still be improved by choosing better parameters, and by improving the implementation (e.g., using the fact that the slope of the segment increases very slowly for this particular problem). Thanks to this algorithm, we will be able to solve the Table Maker's Dilemma in a reasonable time.

References

[1] V. Lefèvre, J.-M. Muller, and A. Tisserand. Towards correctly rounded transcendentals. In *Proceedings of the 13th IEEE Symposium on Computer Arithmetic*, Asilomar, USA, 1997. IEEE Computer Society Press, Los Alamitos, CA.

[2] V. T. Sós. On the distribution mod 1 of the sequence $n\alpha$. *Ann. Univ. Sci. Budapest, Eötvös Sect. Math.*, 1:127–134, 1958.

[3] J. Surányi. Über die Anordnung der Vielfachen einer reellen Zahl mod 1. *Ann. Univ. Sci. Budapest, Eötvös Sect. Math.*, 1:107–111, 1958.

[4] S. Swierczkowski. On successive settings of an arc on the circumference of a circle. *Fundamenta Math.*, 46:187–189, 1958.