Determination with Maxima of the continued fraction associated with a real number

Introduction

We suggest obtaining by means of MAXIMA, N+1 first integers of the sequence (a_n) which defines continued fraction associated with a real number x.

The sequence (x_n) of successors of x is defined by $x_0 = x$ and the relation $x_{n+1} = 1/(x_n - a_n)$ where a_n is the integer part of x_n .

The program which we are going to elaborate, comes as a supplement to the program "cf" from Maxima which applies to the rational numbers and to the square roots of integers.

It will use floating point to obtain the integer part of irrational numbers.

It is not possible to predict in advance the necessary precision to perform all calculations.
The precision required to determine the term a_n , will be estimated from calculations to determine the term a_{n-1} .

Starting precision

What is needed is a valid x approximate value for calculation. In fact x is the expression of a real number using the elementary functions.
For a precision P , bfloat (x) gives an decimal approximate value s_P of x.

In the simple cases, we have $|x - s_P| < 10^{E-(P-1)}$ where E is the integer which verifies
$10^E \leq |x| < 10^{E+1}$ for x # 0. We set E = -1, if x is an expression of 0 .
If x # 0 , s_P consists of P first digits of the decimal representation of $|x| \times 10^{-E}$, the latest which may be increased by one unit.

In general, we have to estimate the small integer L such as $|x - s_P| < 10^{E-(P-L)}$ for any $P \geq L$.
L will be called starting precision of bfloat(x) .
The valid digits of the approximate value of x will be the P+1–L first digits of bfloat (x).

Regularity index

For a quick estimate of L we use the notion of regularity of a floating point on a_n interval of precision.

This notion is built on intuitive idea that if the precision increases by a unit, we obtain a additional digit of the decimal representation of x .

For any expression of x, there is a smaller integer T such as, if the floating point is regular on an interval [A, B] containing at least T elements, then $A \geq L$.
T is called regularity index of bfloat (x).

Control the results obtained with a "convergence test"

The used test is based on the speed of convergence of convergent to x.

Application

A second part will be devoted to the determination of the "best fraction equal to x in ε near" .

PLANE OF RESEARCH

I- Elementary programs

Program in the case of a quotient of two integers

The convergent associated with the last calculated term is given by B / D.

Program CFR(x,n)

```
CFR(x,n):=(u:P:num(x), v:Q:denom(x),  A:0, B:1, C:1, D:0, U:[],
                for i:0 while i <= n and v # 0 do
                        (a:entier(u/v), U:endcons(a,U), B:A+(A:B)*a, D:C+(C:D)*a, u:-v, v:B*Q-D*P),
                U) $
```

a_n is the term of rank n of the continued fraction of x = P/Q.

The basic relations resulting from propositions A-2 and A-3 of the Appendix are:

$A_n = B_{n-1}$, $B_n = A_{n-1} + B_{n-1}\, a_n$, a_n is the integer part of $x_n = \dfrac{(C_{n-1}\, x - A_{n-1})}{(B_{n-1} - D_{n-1}\, x)}$

$C_n = D_{n-1}$, $D_n = C_{n-1} + D_{n-1}\, a_n$, $C_n\, x - A_n = -\,(B_{n-1} - D_{n-1}\, x)$, B_n/D_n is the convergent of rank n.

Relations B:A+(A:B)*a and D:C+(C:D)*a determine parameters A, B, C and D of rank i using a_i = a and parameters A, B, C and D of rank i-1 . x_i is represented by u/v .

The divisions that slow down the program are avoided.
If v = 0, x = B/D. We got the last term of the continued fraction of x. The calculations stop.

```
(%i2) (CFR(sum(1/i!,i,0,20),50),disp("q"=length(U),"B/D"=B/D),U);
```

q = 38

$$B/D = \frac{6613313319248080001}{2432902008176640000}$$

```
(%o2) [2,1,2,1,1,4,1,1,6,1,1,8,1,1,10,1,1,12,1,1,14,1,1,17,8,1,6,5,35,2,2,2,1,6,1,3,25,3]
```

q is the number of calculated terms. It is less than the requested number of terms.
37 is the rank of the last term of the continued fraction of x. x = B/D .

Program using bfloat(x)

The display precision must be sufficient to perform all calculations.

Program CFb(x,n)

```
CFb(x,n):=(Z:fpprec, t:bfloat(x), u:t, v:1, A:0, B:1, C:1, D:0, U:[],
                for i:0 while i <= n and v # 0.0b0 do (a:entier(u/v), B:A+(A:B)*a, D:C+(C:D)*a,
                                                        U:endcons(a,U), u:-v, v:B-D*t), fpprec:Z,
                U) $
```

```
(%i3) fpprec:16$ CFb(sqrt(2),30);
(%o3) [1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1]   (precision too low)

(%i5)  fpprec:100$ CFb(sqrt(2),30);
(%o5) [1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2]
```

II- Useful results in the development of autonomous programs

For a given integer n, it is impossible to predict in advance the precision to use to get the fraction continues to rank n. The autonomus program determines the precision required for the calculation of the term rank i from the calculations to rank i-1.

Moreover, it must evaluate the precision from which we obtain approximate valid values of x.

To determine the integer part of x i, it must be ensured that the valid digits of the approximate value of x_i are not all 0 or 9 in which case the entire portion is defined within one unit.

We give the necessary prerequisites . These prerequisites are developed in the appendix.

A- Conditions for obtaining an integer part with an approximate value

E(x) is the integer part of x . The decimal representation of x is noted DR of x.

Proposition C-3

Let a real number x , an integer $m > 0$ and $c_m(x) = E(10^m x) – 10^m E(x)$.

(1) The equation : $x = E(x) + c\ 10^{-m} + b\ 10^{-m}$ has a unique solution (c,b) such as c be an integer and b a real number verifying $0 \leq b < 1$: $c = c_m(x)$ and $b = 10^m x – E(10^m x)$.

(2) In basis 10, $c_m(x)$ is written: $a_1a_2…a_m$ where $a_1 , a_2 , … , a_m$ are the first m decimals of the decimal representation of x.

(3) $0 \leq c_m(x) \leq 10^m – 1$

(4) if x is integer, $c_m(x) = 0$ for any integer $m \geq 0$

(5) if x is not integer, there is an integer $p > 0$ such as for any $m \geq p$, we have $1 \leq c_m(x) \leq 10^m–2$

This condition expresses that the m first m decimals of DR of x are not all zero or all nine.

Example : $x = 5{,}947695234$, $c_6(x) = E(10^6 x) – 10^6 E(x) = 5947695 – 5000000 = 947695$

We will have to use the condition :

$3 \leq c_m(y) \leq 10^m – 3$ C(m)

Proposition C-5

x is an integer if and only if, for any integer $m>0$ and any y verifying $|x – y| < 10^{-m}$, we have:

either $c_m(y) = 0$ and $x = E(y)$

or $c_m(y) = 10^m -1$ and $x = E(y) + 1$.

Proposition C-6

Let x be a different number of an integer.

(1) Let an integer m > 0 and y be a real number such as : $|x - y| < 10^{-m}$ and $1 \le c_m(y) \le 10^m - 1$.
Then: E(x) = E(y).
If more $3 \le c_m(y) \le 10^m - 3$, for any z verifying $|x - z| < 10^{-m}$, we have E(z) = E(x) and $c_m(z) \ge 1$.

(2) There is an integer p > 0 such as :
for any integer m > p and any y verifying $|x - y| < 10^{-m}$, we have : $3 \le c_m(y) \le 10^m - 3$.

B- Starting precision and regularity index of floating point

Starting precision

Let x be a real number and E the integer which verifies $10^E \le |x| < 10^{E+1}$ if x # 0.
E = -1 if x = 0 .
For a precision P, we set t_P = bfloat(x) $s_P = \sigma \times \mathbf{m} \times 10^{\mathbf{e}+1-\mathbf{n}}$ where σ is the sign, **m** the mantissa and **e** the exponent of bfloat(x). s_n is a decimal number. s_P = round(t_P*10^(P–e–1))*10^(e+1–P) .

Under certain conditions there exists an integer $L \ge 1$ such as :
$|x - s_P| < 10^{E-(P-L)}$ for any $P \ge L$. If $P \ge L$ we have $E - 1 \le e \le E + 1$
For a sufficient value of P , e = E.

Definition

The starting precision of bfloat(x) is the smallest integer L is such as $|x - s_P| < 10^{E-(P-L)}$
for any precision $P \ge L$.

Regularity of the floating point

If $A \ge L$, $|x - s_n| < 10^{E-(n-A)}$ for any $n \in [A,B]$ is verified .

Regularity of bfloat(x) on an interval of precision [A,B]

We do not suppose any more $A \ge L$.
We say that the floating point is regular on [A,B] if $|x - s_n| < 10^{E-(n-A)}$ for any $n \in [A,B]$

Régularity index of bfloat(x)

Considering all regular intervals [A, B] such that A <L.
The number of these intervals is finished (proposition F-2). Let N be the number of elements in the one with the largest number of elements. The regularity index of bfloat(x) is T = N + 1.

Proposition F-3

If the floating-point is regular on an interval [A, B] containing at least T elements, then $A \ge L$.

P-regularity

When x is irrational, x is only accessible by values of bfloat(x) calculated using different precisions we will replace x by bfloat(x) calculated for a sufficient precision P and larger than B .

We say that bfloat(x) is P-regular on an interval [A,B] if $|s_P - s_n| < 10^{E-(n-A)}$ for any $n \in [A,B]$.

With this criterion and an estimate e of E to one unit, an estimation L_0 of L is obtained that checks $L_0 \le L \le L_0+2$ and $L_0+e+1 \ge L+E$ (proposition F-7) .

The programs use $|t_P - t_n|$ instead of $|s_P - s_n|$ .and the starting precision could be underestimated by one unit (proposition F-8) . Then $L_0+e+2 \ge L+E$.

C- Search precision V_n required to calculate a_n

Let x be a real number, x_n is the successor of rank n of x , b_n is the convergent of rank rang n of x , $H_n(t) = (C_{n-1}\, t - A_{n-1})/(B_{n-1} - D_{n-1}\, t)$, $x_n = H_n(x)$.

The principle is to determine V_n from the elements which are from the calculation a_{n-1} .
The proposition D-4 gives the elements necessary for developing the program.

Proposition D-4

Let x be an irrational real number . H_n is the function associated with the continued fraction of x. We can find two sequences (q_n) and (r_n) of real numbers and sequence (m_n) , (K_n) , (V_n) of integers and a sequence (O_n) of open intervals containing x wich verify :

(1) $r_n = H_n(q_n)$.

(2) $(D_{n-1}\, 10^{m_{n-1}})^2 \le 10^{K_n}$ for $n > 0$.

(3) $| x_n - r_n | < 10^{-m_n}$ and $3 \le c_{m_n}(r_n) \le 10^{m_n} - 3$.

(4) $K_0 \ge 0$, $V_0 \ge K_0 + m_0$ and $V_n \ge \max(K_n + m_n , V_{n-1})$ for $n > 0$.

(5) O_n is the set of numbers q such as $| x - q | < 10^{-V_n}$. $O_n \subseteq O_{n-1}$ for $n > 0$.

(6) $q_n \in O_n$ and for any $q \in O_n$, the continued fraction of q coincides with that of x at least up to the order n.

C- Choice of the precision of floating point required to calculate r_n

We choose $K_0 = \max(0,-E)$.

Let $n > 0$ and g the integer witch verifies $10^{g_{n-1}-1} < D_{n-1} \le 10^{g_{n-1}}$. We choose $K_n = 2\,(g_{n-1} + m_{n-1})$.

For a precision P, we have $|x - s_P| < 10^{E-(P-L)}$.

$|x_n - r_n| < |x - s_P|\, 10^{K_n}$ is the relation to estimate the precision required to calculate a_n (proposition D-3).

Condition $|x_n - r_n| < 10^{-m_n-1}$ is satisfied if $|x - s_P| < 10^{-V_n}$.

To obtain $10^{E-(P-L)} \le 10^{-V_n}$, just choose : $P_n \ge L+E+K_n+ m_n$ and $P_n \ge P_{n-1}$.

This choice of P_n is also valid for n = 0.

m_n is the first integer ≥ 4 determined by the program such as $3 \le c_{m_n}(r_n) \le 10^{m_n} - 3$.

The program will use parameters Q and W_n defined with $Q = L + E$ and $W_n = Q + K_n + m_n$.

Then $P_n \ge W_n$ and $P_n \ge P_{n-1}$.

We verify $P_n > L$, for any n : $P_n \ge P_0 \ge L+E+K_0 + m_0 > L$

III- Autonomus programs

The range of validity of the following programs can be expanded by increasing the value of the initial precision. There is already extensive fault with the Maxima precision.

A- Program in irrational case

The program CFI(x,n), gives the continued fraction of an irrational number . b(x) = bfloat(x) . The calculations are performed in floating point. For a precision value p which was used directly t_P = bfloat(x) instead of s_P assessing the potential loss of precision by rounding .

Estimation of the starting precsion with regulary test

Program EI(z)

```
EI(z):=(d:1, if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k))
                    else (for k:0 while d<=z do (d:10*d, e:k)))$
```

The program EI(z) determine the integer e which verifies $10^e \leq z < 10^{e+1}$ where z = |t| .

If z < 1 , the program seeks the first integer k that checks $10^k \times z \geq 1$. Then e = -k .
If z ≥ 1 ,thle program seeks the first integer k that checks $10^{k+1} > z$. Then e = k .

Program L(x)

```
L(x):=(for i:0 while i<=1 and y=1 do  (F:6*10^(e-1-i*Z), fpprec:L+i*Z, f:t-b(x),
                                                    if  -F<f  and  f<F then 1  else y:0)) $
```

L(x) is the regularity test. $s_P - s_n$ is represented by t – b(x) calculated with precision n .

The condition -F<f and f<F is used instead of abs(f) < F , which invalidates the complex values of b(x) .

To check the regularity of bfloat(x) on the interval [L, L + Z] are simply checks inequality $|t_P - t_n| < 10^{e-(n-L)}$ for the values n = L and n = L+Z where Z = fpprec .

By increasing the initial precision, is increased the length of the interval over which the regularity is checked. Thus, the sensitivity of the test is increased and the range of validity of the program is expanded.

Program ELI(x)

```
ELI(x):=(y:0,for i while y=0 do (y:1,L:L+Z,fpprec:P:L+Z+100,t:b(x),
                                    if t<0 or t>0 then (z:abs(t), EI(z) else y:0,if y=1 then L(x) else 1),
          fpprec:P) $
```

The null and complex values of t are invalided by the contition t<0 or t>0 .
The program ELI(x) coordinates the two previous programs. If t is invalided or in case of failure of regularity test, we replace L by L + 50, e is recalculated and redid the test.
The operation also renews long as the test fails. The initial value of L is 50.

Determination of terms of the continued fraction

L+e+2 is an estimate of the exact value of L+E by the program.
We set Q = L+e+2 . Then, in rank i , W = Q +K+ m where K is calculated in rank i-1 for i > 0.
K= max(0,-e) for i = 0.

Instead of calculating s_P for any vlue of n, a lump precision P is used which keeps the same value s_P as $W \leq P$. If $W > P$, P is replaced by W+100 .

Program ai(x)

```
ai(x):=(o:-o, c:0, for j while c<3 or c>s-3 do
            (m:m+m, s:s*s, W:Q+K+m,
              if W>P then (fpprec:P:W+100, t:b(x), u:C*t-A) else 1,
              if y=1 then (v:B-D*t, if o*v >0 then (r:u/v, a:entier(r), c:entier(s*r)-s*a) else (c:3,y:0))
                    else 1))$
```

The term a_i is calculated performing the test of condition C(m) to satisfy the conditions of proposition C-6 (1). s is 10^m . The number c which is $E(r\ 10^m) – 10^m E(r)$, must be greater than or equal to 3 and less than or equal to 10^m-3 (represented by s-3) .

Initially equal to 4 the value of m is doubled each failed test condition C(m). A suitable value of m will be obtained after a finite number of operations (proposition C-6 (2)).

t is used instead of s_P .There are at most three basic steps in the calculation of u/v by the floating point, which can cause a loss of precision.

For $L \geq 16$, the value P estimated in the theoretical part is sufficient to absorb the losses of precision from floating point.

o is $(-1)^i$. The inequality o*(v:B-D*q) > 0 reflects the condition $(-1)^{i-1}(t- b_{i-1}) > 0$ that expresses that t belongs to the domain of definition of the function H_i (proposition A-3) .

Program g(D)

```
g(D):=(for k:g+1 while D>d do (d:10*d, g:k)) $
```

g is the first value of k such that $D \leq 10^k$ (Proposition D-4).

Initially, we set g = 0, K = 0 and d = 1.

The approach of the program AI(x) is similar to that of elementary programs. After each application of ai (x), g(D) determines the new value of g necessary to calculate the next term.
K = 2 (g+m) where g = g(D) .
The condition a> 0 for i > 0 translates the property $a_n \geq 1$ for $n \geq 1$ of continued fractions.

Program AI(x)

```
AI(x):=(o:-1, Q:L+e+2, g:0, K:max(0,-e), d:1, A:0, B:1, C:1, D:0, u:t, U:[],
        for i:0 while y=1 and i<=n do (m:2, s:100, ai(x), if i=0 or (y=1 and a>0)
                                                    then (B:A+(A:B)*a, D:C+(C:D)*a,  g(D),
                                                          K:2*(g+m), u:-v, U:endcons(a,U))
                                                    else y:0)) $
```

Convergence test

The test is based on the speed of convergence of the convergents to x when n tends to infinity.
It is necessary to first calculate the term of order n +1 of the continued fraction by applying ai(x).

Let t = float(x) calculated for a precision V = P + 4 Z, where P is the last displayed precision.
This allows to increase the sensitivity of the test by increasing the initial precision.
According to the proposition D-4, the continued fraction of s_V coincides with those of x up to rank n+1 . Using the proposition A-5, the covergence test is obtained .

$|\sigma - b_n| = \frac{1}{(q_{n-1}+q_n\sigma_{n+1})q_n}$ where $\sigma = s_V$ and $b_n = p_n/q_n$. We have $|x_{n+1} - r_{n+1}| < 10^{Q+K-P} \le 10^{-m_{n+1}}$.

More, $|\sigma_{n+1} - x_{n+1}| < 10^K |\sigma - x| < 10^K\, 10^{e+1-(V-L-1)} = 10^{Q+K-P-4z}$.

As $|\sigma_{n+1} - r_{n+1}| \le |\sigma_{n+1} - x_{n+1}| + |x_{n+1} - r_{n+1}|$, we have $|\sigma_{n+1} - r_{n+1}| < 10^{Q+K-P}(10^{-4z} + 1) < 2\, 10^{Q+K-P}$.

Then: $1/(q_{n-1} + q_n (r_{n+1} + h)) < |q_n s_V - p_n| < 1/(q_{n+-1} + q_n (r_{n+1} - h))$ where $h = 2\, 10^{Q+K-P}$.

When n tends to infinity, the terminals of the frame tend to 0 and the precision used for calculating b(x), tends to infinity.

If any of the terms of the continued fraction of x is miscalculated, the sequence (B_n/D_n) will tends to x '# x (proposition A-10). The seqhence $|b(x) - B_n/D_n|$ tend to x '- x # 0 and there will be a rank n from which at least one of inequalities will not be checked.

Program TI(x)

```
TI(x):=(m:10, s:10^10, ai(x), if y=1 and a>0 then (F:C+D*r, f:2*D*10^(Q+K-P), fpprec:V:P+4*Z,
                                                    d:o*(B-D*b(x)),
                                                    if  1 < (F+f)*d  and  (F-f)*d <1 then 1  else  y:0)
                                                 else y:0) $
```

B/D is b_n , b(x) is s_V , a is a_{n+1} , F is $q_{n-1} + q_n r_{n+1}$ and f is $2\, D\, 10^{Q+K-P}$.
The minimum value of m is 20.

The double inequality is equivalent to: $1 < (F+f)|D\, b(x) - B|$ and $(F-f)|D\, b(x) - B| < 1$.
o*(B-D*b(x)) is $|D\, b(x) - B|$.

The results are all the more reliable as the number of calculated terms is big..

Program CFI(x,n)

```
CFI(x,n):= (Z:fpprec, b(x):=bfloat(x),
            L:0, y:0, for i while y=0 do (ELI(x), AI(x) , if y=1 then TI(x) else 1), fpprec:Z, U) $
```

```
(%i2) CFI(10^10*log(1+10^-10),17);
(%o2) [0,1,20000000000,3,10000000000,5,6666666666,1,4,4,555555555,2,1,8,2,1,444444443,1]

(%i3) CFI(%pi,30);
(%o3) [3,7,15,1,292,1,1,1,2,1,3,1,14,2,1,1,2,2,2,2,1,84,2,1,1,15,3,13,1,4,2]

(%i4)  CFI(sin(exp(-10))-exp(-10)+ exp(-30)/6,20);
(%o4) [0,62216466346098148020976O,19,5,5,2,2,4,4,3,2,6,1,35,1,6,28,3,2,2,6]
```

Summary program

```
CFI(x,n):=(Z:fpprec, b(x):=bfloat(x),

EI(z):=(d:1, if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k))
                    else (for k:0 while d<=z do (d:10*d, e:k))),

L(x):=(for i:0 while i<=1 and y=1 do  (F:10^(e-i*Z), fpprec:L+i*Z, f:t-b(x),
                                        if  -F<f  and  f<F then 1  else y:0)),

ELI(x):=(y:0,for i while y=0 do (y:1,L:L+Z,fpprec:P:L+Z+100,t:b(x),
                                  if t<0 or t>0 then (z:abs(t), EI(z)) else y:0,if y=1 then L(x) else 1),
          fpprec:P) ,

g(D):=(for k:g+1 while D>d do (d:10*d, g:k)),

ai(x):=(o:-o, c:0, for j while c<3 or c>s-3 do
            (m:m+m, s:s*s, W:Q+K+m,
             if W>P then (fpprec:P:W+100, t:b(x), u:C*t-A) else 1,
             if y=1 then (v:B-D*t, if o*v >0 then (r:u/v, a:entier(r), c:entier(s*r)-s*a) else (c:3,y:0))
                    else 1)),

AI(x):=(o:-1, Q:L+e+2, g:0, K:max(0,-e), d:1, A:0, B:1, C:1, D:0, u:t, U:[],
        for i:0 while y=1 and i<=n do (m:2, s:100, ai(x), if i=0 or (y=1 and a>0)
                                                  then (B:A+(A:B)*a, D:C+(C:D)*a,  g(D),
                                                        K:2*(g+m), u:-v, U:endcons(a,U))
                                                  else y:0))  ,

TI(x):=(m:10, s:10^10, ai(x), if y=1 and a>0 then (F:C+D*r, f:2*D*10^(Q+K-P), fpprec:V:P+4*Z,
                                                    d:o*(B-D*b(x)),
                                                     if  1 < (F+f)*d  and  (F-f)*d <1 then 1  else  y:0)
                                                else y:0) ,

L:0, y:0, for i while y=0 do (ELI(x), AI(x) , if y=1 then TI(x) else 1), fpprec:Z, U) $
```

B- General program

The program CF(x,n) gives the continued fraction of the real part of a complex number.
To work in the real domain or the complex domain, choose b(x) as follows:

```
if domain=complex then b(x):=bfloat(realpart(x)) else  b(x):=realpart(bfloat(x)),
```

Program E(z)

```
E(z):=( if z>0 then (d:1,if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k),
                                        if e<Y then (Y:e-1, j:j-1,y:1-j) else 1)
                                    else (for k:0 while d<=z do (d:10*d,e:k)))
                    else j:0) $
```

The changes to the program EI(t) allow to consider the case where x is an expression of 0 (for example x = log(27) – 3 log(3)) .

If t is zero, we set j = 0 .

If t is non-zero approximate value of 0. The estimated value of E will be much less than 0 and the regularity test that must work with e = -1 will fail.
Apply a Y minimum value for the value of e.
Initially Y = -50. At the start calculations we set j = 2.

If E <Y is obtained, Y is substituted by e-1, j is replaced by j-1 and is performed the regularity test. If the test fails, recalculating the value of e . If again e <Y we set e = -1 and performing the regularity test.

It resets the value of j any time the regularity test fails with j = 0.

The regularity test remains unchanged.

Program EL(x)

```
EL(x):=(y:0, h:2, for i while y<=0 do (y:1, L:L+Z, fpprec:P:L+Z+100, t:b(x), z:abs(t), E(z),
                                        if h=0 then e:-1 else 1, L(x),
                                        if y=1 then 1 elseif h=0 then h:2 else 1),
          fpprec:P) $
```

Calculation of a_i

When x is rational the calculated number of terms is finite.
If n+1 is greater than or equal to the rank p of the last term of the continued fraction of x, the program ai(x) will be to determine the integer part of x_p which is an integer.
The condition C(m) will never be verified. The test of the condition C(m) will run indefinitely.

To avoid this situation, it will cap the number of loops in this test.
The number of loops of the test of the condition C(m) will be limited to 2^J . Each failure of test of convergence, J will be doubled . The initial value of J is 2 .

In some cases, insufficient precision of the convergence test can not invalidate the wrong result. It will broaden the range of validity of the program by increasing the initial precision, which will have the effect of increasing the precision of the convergence test.

Program a(x)

```
a(x):=(o:-o, c:0,
        for j while (c<3 or c>s-3) and j<=J do
            (m:m+m, s:s*s, W:Q+K+m ,
              if W>P  then (fpprec:P:W+100, t:b(x), u:C*t-A) else 1,
              if y=1 then (v:B-D*t, if o*v >0 then (r:u/v, a:entier(r), c:entier(s*r)-s*a) else (c:3,y:0))
                      else 1),
        if y>0 then (if c=0 then y:2 elseif c=s-1 then (y:2, a:a+1) else 1) else 1) $
```

Case where x_i is integer

According to proposition C-5 if condition c= 0 is vérified with j = J, then a_i = a.
If condition c= s-1 is vérified with j = J, then a_i = a+1 .

To stop calculations we set y = 2.

Program A(x)

```
A(x):=(o:-1, Q:L+e+2, g:0, K:max(0,-e), d:1, A:0, B:1, C:1, D:0, u:t, U:[],
        for i:0 while y=1 and i<=n do (m:2, s:100, a(x), if i=0 or (y=1 and a>0) or (y=2 and a>1)
                                                    then (B:A+(A:B)*a, D:C+(C:D)*a, g(D),
                                                          K:2*(g+m), u:-v, U:endcons(a,U))
                                                    else y:0)) $
```

Convergence test

Program T(x)

```
T(x):=(if y=1 then (m:10, s:10^10, a(x), if (y=1 and a>0) or (y=2 and a>1)
                              then (fpprec:V:P+4*Z, F:C+D*r, f:2*D*10^(Q+K-P), fpprec:V:P+4*Z,
                                    d:o*(B-D*b(x)),  if 1 < (F+f)*d  and  (F-f)*d <1 then 1  else  y:0)
                              else y:0)
                  else (fpprec:V:P+4*Z, if 10^(V-Q)*abs(D*b(x)-B) < D then 1 else y:0)) $
```

If y=1, the term is calculated to rank n + 1 and the elements of TI(x) are used ..

If y = 2, let p be the order of the last calculated term. x_p is integer. According to proposition A-5,
$x = B_p / D_p$. b(x) is calculated for a precision V = P+4 Z . We set $H=10^{V-Q}$.

Equality is led by inequality $|s_V - b_p| < 1/H$

When V tends to infinity with fp, 1/H tends to 0 and b(x) tends to x. If $b_p \neq x$ there will be a value of V from which inequality will not be checked.

x = B/D .The condition to check is H*abs(D*b(x)-B) < D.

Program CF(x,n)

```
CF(x,n):=
(Z:fpprec, if domain=complex then b(x):=bfloat(realpart(x)) else  b(x):=realpart(bfloat(x)),
J:1, L:0, Y:-50, y:0, for i while y=0 do (J:J+J, EL(x),  A(x), if y>0 then T(x) else 1), fpprec:Z, U) $
```

```
(%i2)  CF(log(3/2),20);
(%o2) [0,2,2,6,1,11,2,1,2,2,1,4,3,1,1,7,2,1,1,4,1]

(%i3) CF(sum((-1)^(i+1)*2^-i/i,i,1,50),20);
(%o3) [0,2,2,6,1,11,2,1,2,2,1,4,3,1,1,7,2,1,1,4,1]

(%i4) CF(48915654/985389+log(8)-3*log(2),20);
(%o4) [49,1,1,1,3,1,1,1,9,11,1,6,3,3]
```

Requires increasing the initial precision

```
(%i6) fpprec:16$CF(1+sin(exp(-1000)),10);
(%o6) [1]
```

```
(%i8) fpprec:70$ CF(1+sin(exp(-1000)),10);
(%o8)  [1,19700711140170469938887935224[375 digits]9597058441895090500470742175 68,4,
         2,2,3,1,1,1,1,11]
```

Direct result with CFI(x,n).

Summary program

```
CF(x,n):=(Z:fpprec,   if domain=complex then  b(x):=bfloat(realpart(x)) else  b(x):=realpart(bfloat(x)),
E(z):=( if z>0 then (d:1,if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k),
                                        if e<Y then (Y:e-1, h:h-1,y:1-h) else 1)
                                    else (for k:0 while d<=z do (d:10*d,e:k)))
                 else h:0),

L(x):=(for i:0 while i<=1 and y=1 do  (F:10^(e-i*Z), fpprec:L+i*Z, f:t-b(x),
                                                          if  -F<f  and  f<F then 1  else y:0)),

EL(x):=(y:0, h:2, for i while y<=0 do (y:1, L:L+Z, fpprec:P:L+Z+100, t:b(x), z:abs(t), E(z),
                                        if h=0 then e:-1 else 1, L(x),
                                        if y=1 then 1 elseif h=0 then h:2 else 1),
          fpprec:P) ,

g(D):=(for k:g+1 while D>d do (d:10*d, g:k)),

a(x):=(o:-o, c:0,
        for j while (c<3 or c>s-3) and j<=J do
            (m:m+m, s:s*s, W:Q+K+m ,
              if W>P  then (fpprec:P:W+100, t:b(x), u:C*t-A) else 1,
              if y=1 then (v:B-D*t, if o*v >0 then (r:u/v, a:entier(r), c:entier(s*r)-s*a) else (c:3,y:0))
                     else 1),
        if y>0 then (if c=0 then y:2 elseif c=s-1 then (y:2, a:a+1) else 1) else 1),

A(x):=(o:-1, Q:L+e+2, g:0, K:max(0,-e), d:1, A:0, B:1, C:1, D:0, u:t, U:[],
         for i:0 while y=1 and i<=n do (m:2, s:100, a(x), if i=0 or (y=1 and a>0) or (y=2 and a>1)
                                                          then (B:A+(A:B)*a, D:C+(C:D)*a, g(D),
                                                                K:2*(g+m), u:-v, U:endcons(a,U))
                                                          else y:0)),

T(x):=(if y=1 then (m:10, s:10^10, a(x), if (y=1 and a>0) or (y=2 and a>1)
                                then (fpprec:V:P+4*Z, F:C+D*r, f:2*D*10^(Q+K-P), fpprec:V:P+4*Z,
                                      d:o*(B-D*b(x)),   if 1 < (F+f)*d  and  (F-f)*d <1 then 1  else  y:0)
                                else y:0)
                 else (fpprec:V:P+4*Z, if 10^(V-Q)*abs(D*b(x)-B) < D then 1 else y:0)),

J:1, L:0, Y:-50,  y:0, for i while y=0 do (J:J+J, EL(x), A(x), if y>0 then T(x) else 1), fpprec:Z, U) $
```

Case of complex number

Obtained continued fractions of real and imaginary parts of x with:

```
CFc(x,n):=(disp("CFr" = CF(x,n) , "CFi" = CF(-%i*x,n))) $
```

Whatever the domain, we get:

```
(%i10) CFc(exp(2)*exp(%i*%pi/12),20);
CFr=[7,7,3,1,1,14,1,2,9,7,2,4,1,2,1,1,1,20,2,6,1]
CFi=[1,1,10,2,2,1,1,2,8,6,1,2,3,1,59,2,1,70,1,61,1]
(%o10) done
```

Case where the expression x has several determinations

If x has not a real determination, CFc(x) choose the main determination whatever the domain.
The real domain is the default domain of Maxima.

If x has a real determination, CFc(x) choose the real determination in real domain , CFc(x) choose the main determination in complex domain as the command realpart(x) calls the main determination of x.

```
(%i11) x:tan(2)$

(%i12) CFc(log(x),20);
CFr=[0,1,3,1,1,2,1,1,1,4,1,2,8,1,14,1,12,1,5,1,2]
CFi=[3,7,15,1,292,1,1,1,2,1,3,1,14,2,1,1,2,2,2,2,1]          (continued fraction of π)
(%o12) done

(%i13) CFc(sqrt(x),20);
CFr=[0]
CFi=[1,2,10,1,25,24,1,1,2,2,12,1,3,18,1,33,1,3,13,1,5]       (continued fraction of √|tan(2)| )
(%o13) done
```

$\tan(2)^{1/3}$ has three determinations which are the solutions of the equation $z^3 = \tan(2)$.
One is real, it is equal to $-\sqrt[3]{|\tan(2)|}$ (car $\tan(2) < 0$) .

```
(%i14) CFc(x^(1/3),20);
CFr=[-2,1,2,2,1,3,1,1,7,2,7,41,3,2,1,3,1,13,5,1,6]
CFi=[0]
(%o14) done
```

In this domain the other determinations are accessible with CFc($e^{2i\pi/3}\,x^{1/3}$,n) and CFc($e^{4i\pi/3}\,x^{1/3}$,n) (main determination of $x^{1/3}$).

```
(%i15) domain:complex$

(%i16) CFc(x^(1/3),20);
CFr=[0,1,1,1,5,1,1,3,1,3,4,3,1,1,20,6,1,2,1,1,27]
CFi=[1,8,12,1,2,1,103,1,3,1,2,6,42,1,1,2,1,42,1,7,2]
(%o16) done
```

In this domain the other determinations are accessible with CFc($e^{2i\pi/3}\,x^{1/3}$,n) (real determination of $x^{1/3}$) and CFc($e^{4i\pi/3}\,x^{1/3}$,n) .

C- Extending the validity domain of floating point

Program BFLOAT(x)

```
BFLOAT(x):=(Z:fpprec,  if domain=complex then  b(x):=bfloat(realpart(x)) else  b(x):=realpart(bfloat(x)),

Fv(x):=(

  E(z):=( if z>0 then (d:1,if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k),
                                        if e<Y then (Y:e-1, j:j-1,y:1-j) else 1)
                                  else (for k:0 while d<=z do (d:10*d,e:k)))
                 else j:0),

  L(x):=(for i:0 while i<=1 and y=1 do  (F:10^(e-i*Z), fpprec:L+i*Z, f:t-b(x),
                                                    if  -F<f  and  f<F then 1  else y:0)),

  EL(x):=(y:0, j:2, for i while y<=0 do (y:1, L:L+Z, fpprec:P:L+Z+100, t:b(x),if t<0 or t>=0
                                        then (z:abs(t), E(z), if j=0 then e:-1 else 1, L(x),
                                              if y=1 then 1 elseif j=0 then j:2 else 1)
                                        else y:0),
            fpprec:P) ,

   L:0, Y:-50, EL(x), fpprec:Z, t:bfloat(t)),

Fv(x), t0:t , Fv(-%i*x), t0+t*%i) $
```

Let u the real or imaginary part of x.

For a precision Z , BFLOAT(u) gives a decimal aproached value of u, denoted σ_Z , such as:
$|u - \sigma_Z| < k\ 10^{E-(Z-1)}$ with $k < 1 + 10^{-101}$.

$(|u - \sigma_Z| \le |u - s_P| + |s_P - \sigma_Z| < 10^{E-(P-L)} + 10^{E-(Z-1)} \le 10^{E-(Z-1)}\ [1 + 10^{-101}]\ )$

```
(%i2) (fpprec:30, BFLOAT(sin(exp(-10))-exp(-10)+exp(-30)/6));
(%o2) 1.607291539891053514899475341 61b-24

(%i3) ( domain:complex, x:tan(2), fpprec:30, BFLOAT(x^(1/3)));
(%o3) 1.123786351408711291861843038 53b0*%i+6.488183524974468393305963944 72b-1

(%i4)  (domain:real,x:tan(2), fpprec:30, BFLOAT(x^(1/3)));
(%o4) -1.297636704994893678661192788 94b0

(%i5)  (fpprec:30,BFLOAT(log(x)));
(%o5)  3.141592653589793238462643383 28b0*%i+7.816340724367478139927096307 37b-1

(%i6)  (fpprec:30,BFLOAT(sqrt(x)));
(%o6)  1.478188033797297045229029913 27b0*%i
```

D- Program determing the Starting precision and the regularity index

The program CFL(x,S) determines:
- The integer E whitch verifies $10^E \leq |x| < 10^{E+1}$.
- The starting precision L of bfloat(x).
- The index of regularity T .

S+1 is an estimate by excess of regularity index.
The initial precision can be increased to approximate the conditions of the proposition F-4.
The programs EI(z), L(x) and the following are used :
ELL(x) replaces ELI(x) (see summary program).

Program EE(x)

```
EE(x):=(Q:L+S+max(e+1,0), m:2, s:100, c:0, for j while  c=0 or c=s-1 do
                (m:m+m, s:s*s, if (W:Q+m)>P then (fpprec:P:W+100, t:b(x), z:abs(t), EI(z)) else 1,
                  r:z*10^-e, a:entier(r), c:entier(s*r)-s*a), sp:round(t*10^(P-e-1))*10^(e+1-P), E:e)  $
```

The program EE (x) determines the first non-zero digit "a" of the DR of | x | , which allows to obtain the exact value of E.

Program LL(A,B)

```
LL(A,B):=(for i:A while i<=B and  y=1 do
                                    (p:i-A+1, F:10^(E+A-i), fpprec:i, u:b(x), if u<0 or u>0
                                      then (z:abs(u),EI(z), si:round(u*10^(i-e-1))*10^(e+1-i),
                                              f:sp-si, if -F<f and f<F then 1 else y:0)
                                      else y:0)) $
```

The program L(A,B) is the regularity test on an interval [A,B]

Program CL(x)

```
CL(x):=(L:0, y:0,  for i while y=0 do (y:1, L:L+1, LL(L,L+S))) $
```

The program CL(x) determines the exact value of the starting precision.
It détermines the first value of L for whitch bfloat(x) is regular on the interval [L,L+Z].

Program CT(x)

```
CT(x):=(T:1, j:1, for j while j <= L-1 do (y:1,  LL(j,j+S), T:max(p,T))) $
```

For any value of j in the l'interval [1,L-1] , The program CT(x) seeks first value p for which the interval [j, j + p-1] is not regular. The greatest value of p is the index of regularity.

Program CFL(x,S)

```
CFL(x,S):=(Z:fpprec, b(x):=bfloat(x), L:0, ELL(x), EE(x), sp:round(t*10^(P-E-1))*10^(E+1-P),
                CL(x), CT(x), fpprec:Z, disp("E"=E, "L"=L, "T"=T)) $
```

Summary program

```
CFL(x,S):=(Z:fpprec, b(x):=bfloat(x),
EI(z):=(d:1, if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k))
                    else (for k:0 while d<=z do (d:10*d, e:k))),

L(x):=(for i:0 while i<=1 and y=1 do  (F:10^(e-i*S), fpprec:L+i*S, f:t-b(x),
                                        if  -F<f  and  f<F then 1  else y:0)),

ELL(x):=(y:0,for i while y=0 do (y:1,L:L+Z,fpprec:P:L+S+100,t:b(x),
                                  if t<0 or t>0 then (z:abs(t), EI(z)) else y:0,if y=1 then L(x) else 1),
          fpprec:P),

EE(x):=(Q:L+S+max(e+1,0), m:2, s:100, c:0, for j while  c=0 or c=s-1 do
              (m:m+m, s:s*s, if (W:Q+m)>P then (fpprec:P:W+100, t:b(x), z:abs(t), EI(z)) else 1,
               r:z*10^-e, a:entier(r), c:entier(s*r)-s*a), sp:round(t*10^(P-e-1))*10^(e+1-P), E:e) ,

LL(A,B):=(for i:A while i<=B and  y=1 do
                                    (p:i-A+1, F:10^(E+A-i), fpprec:i, u:b(x), if u<0 or u>0
                                      then (z:abs(u),EI(z), si:round(u*10^(i-e-1))*10^(e+1-i),
                                              f:sp-si, if -F<f and f<F then 1 else y:0)
                                      else y:0)),

CL(x):=(L:0, y:0, for i while y=0 do (y:1, L:L+1, LL(L,L+S))),

CT(x):=(T:1, j:1, for j while j <= L-1 do (y:1, LL(j,j+S), T:max(p,T))),

L:0, ELL(x), EE(x),  CL(x), CT(x), fpprec:Z, disp("E"=E, "L"=L, "T"=T)) $
```

```
(%i3)  x:10^5*log(1+exp(-50))$ CFL(x,100);
 E = -17
 L = 22
 T = 1
(%o3) done

(%i5) fpprec:51$ bfloat(x); BFLOAT(x);
(%o4) 1.92874984796391778301715681273 025534361374155865392b-17
(%o5) 1.92874984796391778301715681272 821153295468469035706b-17
                                        21 chiffres corrects
(%i6)  CFL(exp(1000),100);
 E = 434
 L = 2
 T = 1
(%o6) done
(%i7)  CFL(sin(exp(1000)),100);
 E = -1
 L = 435
 T = 3
(%o7) done

(%i8)  CFL(sin(exp(-10))-exp(-10)+exp(-30)/6,100);
E = -24
 L = 20
 T = 52
(%o8) done
```

E-Program of integer part of a different number of integer number

Program ENTIER(x)

```
ENTIER(x):=(Z:fpprec, b(x):=bfloat(x),

EI(z):=(d:1, if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k))
                    else (for k:0 while d<=z do (d:10*d, e:k))),

L(x):=(for i:0 while i<=1 and y=1 do  (F:10^(e-i*Z), fpprec:L+i*Z, f:t-b(x),
                                        if  -F<f  and  f<F then 1  else y:0)),

ELI(x):=(y:0,for i while y=0 do (y:1,L:L+Z,fpprec:P:L+Z+100,t:b(x),
                                  if t<0 or t>0 then (z:abs(t), EI(z)) else y:0,if y=1 then L(x) else 1),
          fpprec:P) ,

a0(x):=(if E<-1 then a:entier(t)
                else (Q:L+e+2, m:2, s:100, c:0, for j while c=0 or c=s-1 do
                      (m:m+m, s:s*s, if (W:Q+m) > P then (fpprec:W+100, t:b(x)) else 1,
                                                          a:entier(t), c:entier(s*t)-s*a))) ,

L:0,y:0, for i while y=0 do (ELI(x), if y=1 then a0(x) else 1),  fpprec:Z, a) $
```

```
(%i5)  x:10^5*sin(exp(450))$entier(bfloat(x));
(%o5) 47113

(%i7)  ENTIER(x);"L"=L;
(%o6)  –63274
(%o7) L = 208

(%i8)  CFL(x,100,10);
E = 4
L = 196
T = 2
(%o8) done

(%i10)  fpprec:195+4$entier(bfloat(x));
(%o10) -63271

(%i12)  fpprec:196+4$entier(bfloat(x));
(%o12) -63274
```

VI- Best fraction equal to x in ε near

Problem

Let x be a nonzero real number and ε a positive real number which will be chosen rational in the programs which follow.

If x > 0,we suggest determiningthe smallest integers $p \geq 0$ and $q > 0$ such as $|x - p/q| < \varepsilon$.
In this case p/q is the best fraction equal to x in ε near.

Il x < 0, the best fraction equal to x in ε near. the opposite of the best fraction equal to |x| in ε near.

A- First approach

We consider the case where x is a positive number or zero and ε rational.
We can elaborate a direct program in case x is a fraction.

We set x = P/Q . The best fraction I/J equal to x in ε near have to verify : |P/Q-I/J| < ε .

By supposing P = 0, Q> 0, I = 0 and J> 0, this inequality is equivalent in:

$$(P - \varepsilon\, Q)\, J/Q \; < I < (P + \varepsilon\, Q)\, J/Q \; .$$

The idea is to give in J increasing values from 1.
For any value of J we determine the smallest value of $I \geq 0$ which verifies the first inequality.
Because I is an integer it results from it that I = max(0,entier((P - ε Q) J/Q)+1) .

The double inequality is thus equivalent in: max(0,entier((P-ε Q) J/Q)+1) < (P + ε Q)J/Q .

If D is the first value of J for which this condition is realized, then the best fraction is
max(0,entier((P - ε Q) D/Q)+1)/D .
ε is represented by ec.

Program BFA(x,ec)

```
BFA(x,ec):=(P:num(x),Q:denom(x), E:(P-ec*Q)/Q, F:(P+ec*Q)/Q, b:0, B:1,for J:1 while B>= b do
                (D:J, B: max(0,entier(E*J)+1) , b:F*J), B/D) $
```

We placed in front of any result the continued fraction of the obtained fraction.

(%i2) BFA(50149/23778,10^-9);
(%o2) $\frac{50149}{23778}$ [2,9,5,1,7,3,8,2]

(%i3) BFA(50149/23778,10^-8);
(%o3) $\frac{23653}{11215}$ [2,9,5,1,7,3,8]

(%i4) BFA(50149/23778,3*10^-8);
(%o4) $\frac{17967}{8519}$ [2,9,5,1,7,3,6]

(%i5) BFA(50149/23778,10^-6);
(%o5) $\frac{1934}{917}$ [2,9,5,1,7,2]

(%i6) BFA(50149/23778,10^-5);
(%o6) $\frac{793}{376}$ [2,9,5,1,6]

In the previous calculations we notice that the terms of the continued fraction of the best fractions, except maybe the last term, simultaneous with the first terms of the continued fraction of $\frac{50149}{23778}$.

For $\varepsilon = 10^{-8}$ the best fraction is a convergent.

In a general way that x is rational or irrational, the best fractions will be obtained from convergents associated in the continued fraction of x.

We give the elements who will be used in the elaboration of the programs and which were developed in appendix.

Must be used the function G_n defined by $G_n(z) = \frac{A_{n-1} + z\,B_{n-1}}{C_{n-1} + z\,D_{n-1}}$ with $B_{n-1}.C_{n-1} - A_{n-1}.D_{n-1} = (-1)^n$

as well as the function H_n already used before.

Results (proposition B-2, B-3)

Let $x \geq 0$.

(1) There is a smaller integer n such as $|x - b_n| < \varepsilon$. If $\varepsilon \geq 1$, $n = 0$.

(2) b_n is the best convergent equal to x in ε near.

(3) There is a smaller integer d positive or null such as $|x - G_n(d)| < \varepsilon$.

(4) $G_n(d)$ is the best fraction equal to x in ε near
$G_n(d)$ restore the fraction defined by the sequence $[a_0, a_1, a_2, \ldots, a_{n-1}, d]$.

(5) Let H_n be the inverse function of G_n . Then $d = \max(\text{Entier}(H_n(x-(-1)^n\,\varepsilon))+1,0)$.

B- Program in rational case

ε is the quotient of two positive integers. ε will be represented in the program by ec = M/N.
This program, noted BFR(x,ec), applies only to expressions of x is the quotient of two integers.
The calculation is that the program CFR(x,n).
We will proceed in three stages.

(1) Determination of the rank n of the best convergent equal in |x| in ε near.

(2) Search for the first value of the integer d such as $||x|- G_n((d)| < \varepsilon$ with resukt (5) .

(3) Calculation of $G_n(d)$ which is the best fraction equal in |x| in ε near.

Determination of the rank n of the best convergent

```
AR(x):= (u:P:num(X), v:Q:denom(X), A:0, B:1, C:1, D:0, c:M*Q , o:1,
          for i:0 while o*N*v >= c*D and v # 0 do
                              (o:-o, n:i, a:entier(u/v), B:A+(A:B)*a, D:C+(C:D)*a, u:-v, v:B*Q-D*P),
          A:B-(B:A)*a, C:D-(D:C)*a) $
```

X is |x| . $X = P/Q$. $o = (-1)^{i+1}$. The first integer n such that is sought such as $|P/Q - B/D| < M/N$.
This inequality is equivalent to $N*o*(B*Q - D*P) < M*Q*D$.
The terms of the continued fraction are computed as long as $N*o*(B*Q-D*P) >= M*Q*D$.

When the integer n is reached, the parameters A, B , C , D in rank n, should be demoted to the rank n-1 to define the function H_n and G_n : A:B-(B:A)*a, C:D-(D:C)*a .
$H_n(t) = (C*t-A)/(B-D*t)$, $G_n(z) = (A+z*B)/(C+z*D)$.

Search of d

d = max(E($H_n(X-(-1)^n \varepsilon)$)+1,0) . w = $X-(-1)^n \varepsilon$ = X+o*ec.

```
DR(x):=( w:X+o*ec,  d:max(entier((C*w-A)/(B-D*w))+1,0)) $
```

Program BFR(x,ec)

```
BFR(x,ec):=(M:num(ec), N:denom(ec),sg:if x<0 then -1 else 1, X:sg*x , AR(x), DR(x),
                                                                    sg*(A+d*B)/(C+d*D)) $
```

sg is the sign of x. |x|= sg*x . The best fraction is: sg*(A+d*B)/(C+d*D) which is $G_n(d)$.

Summary program

```
BFR(x,ec):=(
AR(x):= (u:P:num(X), v:Q:denom(X), A:0, B:1, C:1, D:0, c:M*Q , o:1,
          for i:0 while o*N*v >= c*D and v # 0 do
                              (o:-o, n:i, a:entier(u/v), B:A+(A:B)*a, D:C+(C:D)*a, u:-v, v:B*Q-D*P),
          A:B-(B:A)*a, C:D-(D:C)*a),

DR(x):=( w:X+o*ec,  d:max(entier((C*w-A)/(B-D*w))+1,0)) ,

M:num(ec), N:denom(ec),sg:if x<0 then -1 else 1, X:sg*x , AR(x), DR(x), sg*(A+d*B)/(C+d*D))
$
```

(%i2) BFR(50149/23778,3*10^-8);

(%o2) $\frac{17967}{8519}$

(%i3) BFR(50149/23778,10^-10);

(%o3) $\frac{50149}{23778}$

C- Irrational case

Proposition B-7

Let x be a positive irrational number. Let p be the smallest integer such as $1/(q_{p-1} + q_p)\, q_p \leq \varepsilon$. Then the smallest integer n such as $|x - b_n| < \varepsilon$ is p-1 or p .

To determine the rank n of the best fraction, we proceed in two stages.

(1)We first determine the integer p of the proposition B-7.

(2) Then we proceed by elimination by testing inequality $|x - b_{p-1}| < \varepsilon$.

Always used the CFI(x,n) program's approach
The programs EI(t), L(x), ELI(x), ai(x) , TI(x) are kept.

Determination of p

We establish a program Ae(x) analogous to AI(x).

Program Ae(x)

```
Ae(x):=(o:-1,Q:L+e+2, g:0, K:max(0,-e), d:1, A:0,B:1,C:1,D:0, u:t,
          for i:0 while y=1 and  N>(C+D)*D*M  do
              (p:i, a0:a, m:2, s:100, ai(x), if i=0 or (y=1 and a>0) then (B:A+(A:B)*a, D:C+(C:D)*a,
                                                                          g(D), K:2*(g+m), u:-v)
                                                                  else y:0)$
```

The condition $1/((q_{i-1} + q_i)\, q_i) \le \varepsilon$ is written $1/((C+D)\, D) \le ec$ or $N \le (C+D)*D*M$.

The first value of p is determined for which $N \le (C+D)*D*M$.

If we get a non-zero value of p, $a0 = a_{p-1}$ and $a = a_p$. We set $o1 = o = (-1)^p$.

At rank p-1, the parameters, noted A1 , B1 , C1 , D1 are obtained by demoting the parameters A, B, C, D :

A1:B-(B1:A)*a, C1:D-(D1:C)*a

Determination of the rank n of the best convergent

Case p = 0

n = 0 . We set o1=1 , A1=0 , B1=1 , C1=1 , D1=0 .
These are the parameters used by the program to define the H_0 and G_0 function and determine the number d .

Case p > 0

We set X = |x| . The double inequality $0 \le |X - b_{p-1}| < ec$ is equivalent to $0 \le |X - b_{p-1}|/ec < 1$..

This means that the integer part of $(-1)^{p-1}(X - b_{p-1})/ec$ is zero.

Precision required to calculate the integer part

We set $F(X) = (-1)^{p-1}(X - b_{p-1})/ec$. Let K an integer such as : $1/ec \le 10^K$.

We have: $10^{Kp+1} \ge (C_p + D_p\, s)^2 > (C_p + D_p)\, D_p \ge N/M = 1/ec$. We may choose $K = K_{p+1}$.

Let q be an approximate value of X such as $|X - q| < 10^{Q-P}$.

We check $|F(X) - F(q)| = (-1)^p\, (X - q)/ec\ |X-q|\ 10^K$. Then $|F(X) - F(q)| < 10^{Q+K-P}$.

To obtain $|F(X) - F(q)| < 10^{-m}$, it is enough to choose $P \ge Q+K_{p+1}+m = P_{p+1}+m$.

Program SEL(x)

```
SEL(x):=(if p=0
            then (o1:1, A1:0, B1:1, C1:1, D1:0)
            else (m:1, s:10, c:0, for j while c=0 or c=s-1 do
                    (m:m+m, s:s*s, W:P+m , if V<W then (fpprec:V:W+100, t:b(x)) else 1,
                     r:(B1/D1-t)*N/M, if o1*r >0 then (a:entier(r), c:entier(s*r)-s*a)  else (c:1,y:0)),
                  if y=1 then (if a=0 then (o1:-o1, A1:B1-(B1:A1)*a0, C1:D1-(D1:C1)*a0) else  1)
                       else 1)) $
```

Is used the value t calculated for the precision R. Initially $o1 = (-1)^p$

F(q) is r = o1*(B1/D1-t)/ec and searching for the integer part of r.

Next we determine the parameters which define Hn and Gn .

If a = 0 , n = p-1 . We have $(-1)^n = (-1)^{p-1}$ and o1 is substituted by -o1 .

The parameters A1, B1, C1, D1 are demoted to the rank p-2 with :

A1:B1-(B1:A1)*a0, C1:D1-(D1:C1)*a0 .

If a # 0 , n = p . We have $(-1)^n = (-1)^p$. The parameters o1 , A1 , B1 , C1 , D1 are unchanged .

.

Determination of d

Estimate of the precision required to calculate d

Recall that $d = \max(E(H_n(X-(-1)^n \varepsilon))+1,0)$.

Let q be an aproached value of X such as $|X – q| < 110^{Q-P}$.

For example let us suppose n odd .

The calculation give: $|H_n(X+\varepsilon) – H_n(q+\varepsilon)| = |X – q|/[D_{n-1}{}^2 |(X+\varepsilon) – b_{n-1}||(q+\varepsilon) – b_{n-1}|]$

We have $|b_{n-1} – (X+\varepsilon)| = X+\varepsilon – b_{n-1} > X – b_{n-1} > 0$.

Similarly $|b_{n-1} – (q+\varepsilon)| = q+\varepsilon – b_{n-1} > q – b_{n-1} \geq X – b_{n-1} > 0$.

Then $|H_n(X+\varepsilon)–H_n(q+\varepsilon)| < |X-q|/[D_{n-1}{}^2 (X – b_{n-1})^2] = |H_n(X) – H_n(q)| < |X–q| \, 10^{Kn}$.

Then $|H_n(X+\varepsilon) – H_n(q+\varepsilon)| < 10^{Q+Kn-P}$.

To obtain $|H_n(X+\varepsilon) – H_n(q+\varepsilon)| < 10^{-m}$ it is enough to choose $P \geq Q+K_n+m = P_n+m$.

This choice is also valid if n is even .

Program D(x)

```
D(x):=(m:1 , s:10 , c:0 ,
        for j while c=0 or c=s-1 do
                        (m:m+m, s:s*s, W:P+m, if V<W then (fpprec:V:W+100, t:b(x))  else 1,
                         w:t-o1*ec , v:B1-D1*w , if o1*v > 0
                                                then (a:entier(r:(C1*w-A1)/v) , c:entier(s*r)-s*a)
                                                else (c:1,y:0)),
        if y=1  then d:max(a+1,0) else 1) $
```

Is used the value t calculated for the precision P .

$q-(-1)^n \varepsilon$ is w = t-o1/h-o1*ec .

$H_n(q-(-1)^n \varepsilon)$ = is r = (C1*w-A1)/(B1-D1*w) .

Program BF(x,ec)

```
BF(x,ec):=(Z:fpprec, b(x):=bfloat(x),
           M:num(ec), N:denom(ec), L:0, y:0 ,
                for i while y=0 do (ELI(x), if t<0 then (t:-t,sg:-1) else sg:1, X:sg*x, Ae(X),
                                    if y=1 then TI(X) else 1, if y=1 then SEL(X) else 1,
                                    if y=1 then D(X) else 1),
           fpprec:Z, sg*(A1+d*B1)/(C1+d*D1)) $
```

The best fraction equal to x in ε near is: sg*(A1+d*B1)/(C1+d*D1) = sg $G_n(d)$

```
(%i3) (x:sin(exp(100)),"y"=y:BF(x,10^-20));
```

$$(\%o3) \qquad y = \frac{1214130659}{8538302952}$$

```
(%i4) BFLOAT(x-y);
(%o4)  1.519172621677087b-21

(%i5) (x:exp(20), disp(entier(x) , BF(x,10^5)));
485165195
485065196
(%o5) done

(%i6) BF(10^10*log(1+10^-10),10^-20);
```

$$(\%o6) \qquad \frac{19999999997}{19999999998}$$

Summary program

```
BF(x,ec):=(Z:fpprec, b(x):=bfloat(x),

EI(z):=(d:1, if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k))
                    else (for k:0 while d<=z do (d:10*d, e:k))),

L(x):=(for i:0 while i<=1 and y=1 do  (F:10^(e-i*Z), fpprec:L+i*Z, f:t-b(x),
                                        if  -F<f  and  f<F then 1  else y:0)),

ELI(x):=(y:0,for i while y=0 do (y:1,L:L+Z,fpprec:P:L+Z+100,t:b(x),
                                  if t<0 or t>0 then (z:abs(t),EI(z)) else y:0,if y=1 then L(x) else 1),
         fpprec:P) ,

g(D):=(for k:g+1 while D>d do (d:10*d, g:k)),

ai(x):=(o:-o, c:0, for j while c<3 or c>s-3 do
           (m:m+m, s:s*s, W:Q+K+m,
            if W>P then (fpprec:P:W+100, t:b(x), u:C*t-A) else 1,
            if y=1 then (v:B-D*t, if o*v >0 then (r:u/v, a:entier(r), c:entier(s*r)-s*a) else (c:3,y:0))
                   else 1)),

TI(x):=(m:10, s:10^10, ai(x), if y=1 and a>0 then (F:C+D*r, f:2*D*10^(Q+K-P), fpprec:V:P+4*Z,
                                                    d:o*(B-D*b(x)),
                                                     if  1 < (F+f)*d  and  (F-f)*d <1 then 1  else  y:0)
                                              else y:0) ,

Ae(x):=(o:-1,Q:L+e+2, g:0, K:max(0,-e), d:1, A:0,B:1,C:1,D:0, u:t,
        for i:0 while y=1 and  N>(C+D)*D*M  do
           (p:i, a0:a, m:2, s:100, ai(x), if i=0 or (y=1 and a>0) then (B:A+(A:B)*a, D:C+(C:D)*a,
                                                                        g(D), K:2*(g+m), u:-v)
                                                                 else y:0),
        if y=1 then (o1:o, A1:B-(B1:A)*a, C1:D-(D1:C)*a) else 1) ,

SEL(x):=(if p=0
           then (o1:1, A1:0, B1:1, C1:1, D1:0)
           else (m:1, s:10, c:0, for j while c=0 or c=s-1 do
                    (m:m+m, s:s*s, W:P+m , if V<W then (fpprec:V:W+100, t:b(x)) else 1,
                     r:(B1/D1-t)*N/M, if o1*r >0 then (a:entier(r), c:entier(s*r)-s*a)  else (c:1,y:0)),
                 if y=1 then (if a=0 then (o1:-o1, A1:B1-(B1:A1)*a0, C1:D1-(D1:C1)*a0) else  1)
                        else 1)),

D(x):=(m:1 , s:10 , c:0 ,
       for j while c=0 or c=s-1 do
                          (m:m+m, s:s*s, W:P+m, if V<W then (fpprec:V:W+100, t:b(x))  else 1,
                           w:t-o1*ec , v:B1-D1*w , if o1*v > 0
                                                   then (a:entier(r:(C1*w-A1)/v) , c:entier(s*r)-s*a)
                                                   else (c:1,y:0)),
       if y=1  then d:max(a+1,0) else 1),

M:num(ec), N:denom(ec), L:0, y:0 , M:num(ec), N:denom(ec), L:0, y:0 ,
for i while y=0 do (ELI(x), if t<0 then (t:-t, sg:-1) else sg:1, X:sg*x,  Ae(X),
                    if y=1 then TI(X) else 1, if y=1 then SEL(X) else 1, if y=1 then D(X) else 1) ,
fpprec:Z, sg*(A1+d*B1)/(C1+d*D1)) $
```

D- Program for determining the best fractions of rank n

Proposition B-5

Let x > 0. If n = 0 the best fractions are the integers of the interval $[0,a_0]$.
If n > 0 $G_n(d)$ is a best fraction of rank n if and only if d belong to the interval $[a,a_n]$ where a is the smallest integer which satisfies $(-1)^n (G_n(a) + b_{n-1} -2 x) > 0$.

Rational case

Let X = |x| .

Program ARn(x)

```
ARn(x):= (u:P:num(X), v:Q:denom(X), A:0, B:1, C:1, D:0,  for i:0 while i <= n and v # 0 do
                                        (p:i, a:entier(u/v),  B:A+(A:B)*a,D:C+(C:D)*a, u:-v, v:B*Q-D*P),
                                         an:a, A1:B-(B1:A)*a,C1:D-(D1:C)*a) $
```

Determining the terms of the continued fraction up to order n and the parameters of the functions H_n and G_n demoting parameters of rank n: A1:B-(B1:A)*a , C1:D-(D1:C)*a .The program ARn(x) determine the terms of the continued fraction up to order n and the parameters of the functions H_n and G_n demoting parameters of rank n:

A1:B-(B1:A)*a , C1:D-(D1:C)*a .

Détermination de a

```
DRa(x):=(w:2*X-B1/D1, a:entier((A1-C1*w)/(D1*w-B1))+1) $
```

$2 x - b_{n-1}$ est représenté par w = 2*X-B1/D1 .

$H_n(2 x - b_{n-1})$ est représenté par (A1-C1*w)/(D1*w-B1) .

Program BFRn(x,n)

```
BFRn(x,n):=(sg: if x < 0 then -1 else 1, X:sg*x , ARn(x), M:[],
                if n=0 then M:makelist(i,i,0,a)
                        elseif n <= p then (DRa(x), M:[B/D],
                                                for i:a while i < an do (B:B-B1, D:D-D1, M:cons(B/D,M)))
                        else 1 ,
sg*M) $
```

For n = 0 , the best fractions are the integers of the interval $[0, a_0]$.
To determine the best fractions for n>0 , we use the sequence:

M:[B/D], for b:a while b<an do (B:B-B1, D:D-D1,M:cons(B/D,M),

If n is greater than the rank of the last term of the continued fraction of x, M = [] .

Summary Program

```
BFRn(x,n):=(
ARn(x):= (u:P:num(X), v:Q:denom(X), A:0, B:1, C:1, D:0,  for i:0 while i <= n and v # 0 do
                                    (p:i, a:entier(u/v),  B:A+(A:B)*a,D:C+(C:D)*a, u:-v, v:B*Q-D*P),
                                     A1:B-(B1:A)*a,C1:D-(D1:C)*a, an:a),

DRa(x):=(w:2*X-B1/D1, a:entier((A1-C1*w)/(D1*w-B1))+1),

sg: if x < 0 then -1 else 1, X:sg*x , ARn(x), M:[],
if n = 0  then M:makelist(i,i,0,a)
          elseif n <= p then (DRa(x), M:[B/D], for i:a while i < an do
                                                    (B:B-B1, D:D-D1, M:cons(B/D,M)))
                        else 1 ,
 sg*M) $
```

(%i2) BFRn(sum(1/i!,i,0,5),0);

(%o2) [0,1,2]

(%i3) BFRn(sum(1/i!,i,0,5),1);

(%o3) [3]

(%i4) BFRn(sum(1/i!,i,0,5),2);

(%o4) $[\frac{5}{2},\frac{8}{3}]$

(%i5) BFRn(sum(1/i!,i,0,5),3);

(%o5) $[\frac{11}{4}]$

(%i6) BFRn(sum(1/i!,i,0,5),4);

(%o6) $[\frac{19}{7}]$

(%i7) BFRn(sum(1/i!,i,0,5),5);

(%o7) $[\frac{87}{32},\frac{106}{39},\frac{125}{46},\frac{144}{53},\frac{163}{60}]$

(%i8) BFRn(sum(1/i!,i,0,5),6);
(%o8) [ ]

We obtain the list the best fractions associated in x. In fat, the convergents.

$[\ 0\ ,\ 1\ ,\ \mathbf{2}\ ,\ \mathbf{3}\ ,\ \frac{5}{2}\ ,\ \mathbf{\frac{8}{3}}\ ,\ \mathbf{\frac{11}{4}}\ ,\ \mathbf{\frac{19}{7}}\ ,\ \frac{87}{32}\ ,\ \frac{106}{39}\ ,\ \frac{125}{46}\ ,\ \frac{144}{53}\ ,\ \mathbf{\frac{163}{60}}\ ]$

$x = \frac{163}{60}$

Irrational case

The program denoted BFn(x,n) give the best rank n fractions of rank n of |x|.

They will be assigned by the sign "–" if x <0. It retains the programs EI(x) , L(x) and ELI(x).

Program An(x)

```
An(x):=(o:-1,Q: L+e+2 , g:0, K:max(0,-e), d:1,  A:0, B:1, C:1, D:0, u:t ,
          for i:0 while y=1 and i <= n do
                    (m:2, s:100, ai(x), if i=0 or (y=1 and a>0) then (B:A+(A:B)*a, D:C+(C:D)*a, g(D),
                                                                      K:2*(g+m), u:-v)
                                                                 else y:0),
          if y=1 then (an:a, Bn:B, Dn:D, o1:o , A1:B-(B1:A)*a, C1:D-(D1:C)*a) else 1) $
```

We determine the terms of the continued fraction up to order n.

In rank n is stored: an , Bn , Dn and $o1 = (-1)^n$.

We determine the parameters of the functions H_n and G_n demoting parameters of rank n :

A1:B-(B1:A)*a , C1:D-(D1:C)*a .

Determination of a

As in the rational case for n> 0, we have to determine the integer a defined by:

$$a = \text{Entier}(H_n(2\,x - b_{n-1})) + 1 .$$

Precision needed to calculate a for n > 0

Let q be an approximate value of X such as $|X - q| < 10^{Q-P}$.

Calculation give:

$|H_n(2\,X - b_{n-1}) - H_n(2\,q - b_{n-1})| = |X - q| /(4\,D_{n-1}^2\,|X - b_{n-1}|\,|q - b_{n-1}|) =$
$|H_n(X) - H_n(q)| /4 < |X-q|\,10^{Kn} < 10^{Q+Kn-P}$.

To obtain $|H_n(2\,X - b_{n-1}) - H_n(2\,q - b_{n-1})| < 10^{-m}$ it is enough to choose $P \geq Q+K_n+m = P_n+m$.

Program Da(x)

```
Da(x):=(if n=0 then 1
                else (m:1, s:10, c:0,  for j while c=0 or c=s-1 do
                                (m:m+m, s:s*s, W:P+m , if V<W then (fpprec:V:W+100, t:b(x)) else 1,
                                 w:2*t-B1/D1,  v:B1-D1*w,
                                 if  o1*v > 0 then (a:entier(r:(C1*w-A1)/v),c:entier(s*r)-s*a)
                                              else (c:1,y:0)),
                        if  y=1 and a>= 0  then a:a+1 else y:0)) $
```

Is used the value t calculated for the precision P . We have $(-1)^n = o1$.

We set w = 2*t – B1/D1 and we determine the integer part of r = (w*C1-A1)/(B1-D1*w).

We verify : o1*(B1-D1*w) = 2*o1*(B1-D1*t) > 0 . a = entier(r)+1.

Program BFn(x,n)

```
BFn(x,n):=(Z:fpprec,  b(x):=bfloat(x),
            L:0  , y:0 , for i while y=0 do (ELI(x), if t<0 then (t:-t, sg:-1) else sg:1, X:sg*x, An(X),
                                              if y=1 then TI(X) else 1 , if y=1 then Da(X) else 1),
            if n = 0 then M:makelist(i,i,0,an)
                  else (M:[Bn/Dn] ,
                        for b:a while b<=an do (Bn:Bn-B1 , Dn:Dn-D1, M:cons(Bn/Dn,M))),
            fpprec:Z, sg*M ) $
```

For n = 0 the best fractions are given by: makelist(i,i,0,an) .

For n > 0 the best fractions are given by the sequence:

M:[Bn/Dn] , for b:a while b < an do (Bn:Bn-B1 , Dn:Dn-D1, M:cons(Bn/Dn,M),

The final result is given by sg*M .

(%i2) BFn(%pi,0);

(%o2) [0,1,2,3]

(%i3) BFn(exp(1),5);

(%o3) $[\ \frac{30}{11}\ ,\ \frac{49}{18}\ ,\ \frac{68}{25}\ ,\ \frac{87}{32}\ ]$

Example illustrating the result 5

Let x = π . For n = 2 , we have a2 = 15 , b1 = 22/7 , b2 = 333/106 .
We must determine the smallest integer a such that: $G_2(a) + b_1 - 2\pi > 0$
We give the 15 values of $G_2(c)$ for $c \in [1,15]$ in the increasing order.

$$[\frac{25}{8},\frac{47}{15},\frac{69}{22},\frac{91}{29},\frac{113}{36},\frac{135}{43},\frac{157}{50},\frac{179}{57},\frac{201}{64},\frac{223}{71},\frac{245}{78},\frac{267}{85},\frac{289}{92},\frac{311}{99},\frac{333}{106}]$$

We have 157/50 + 22/7 < 2 π and 179/57 + 22/7 > 2 π . Then a = 8 .
Alone better possible fractions in rank 2 are:

$$[\frac{179}{57},\frac{201}{64},\frac{223}{71},\frac{245}{78},\frac{267}{85},\frac{289}{92},\frac{311}{99},\frac{333}{106}]$$

(%i4) BFn(%pi,2);

(%o4) $[\frac{179}{57},\frac{201}{64},\frac{223}{71},\frac{245}{78},\frac{267}{85},\frac{289}{92},\frac{311}{99},\frac{333}{106}]$

Summary pogram

```
BFn(x,n):=(Z:fpprec, b(x):=bfloat(x),

EI(z):=(d:1, if z<1 then (for k:1 while d*z<1 do (d:10*d, e:-k))
                    else (for k:0 while d<=z do (d:10*d, e:k))),

L(x):=(for i:0 while i<=1 and y=1 do  (F:10^(e-i*Z), fpprec:L+i*Z, f:t-b(x),
                                        if  -F<f  and  f<F then 1  else y:0)),

ELI(x):=(y:0,for i while y=0 do (y:1,L:L+Z,fpprec:P:L+Z+100,t:b(x),
                                    if t<0 or t>0 then (z:abs(t), EI(z)) else y:0,if y=1 then L(x) else 1),
          fpprec:P) ,

g(D):=(for k:g+1 while D>d do (d:10*d, g:k)),

ai(x):=(o:-o,  m:2, s:100, c:0, for j while c<3 or c>s-3 do
            (m:m+m, s:s*s, W:Q+K+m,
             if W>P then (fpprec:P:W+100, t:b(x), u:C*t-A) else 1,
             if y=1 then (v:B-D*t, if o*v >0 then (r:u/v, a:entier(r), c:entier(s*r)-s*a) else (c:3,y:0))
                    else 1)),

An(x):=(o:-1,Q: L+e+2 , g:0, K:max(0,-e), d:1,  A:0, B:1, C:1, D:0, u:t ,
          for i:0 while y=1 and i <= n do
                (m:2, s:100, ai(x), if i=0 or (y=1 and a>0) then (B:A+(A:B)*a, D:C+(C:D)*a, g(D),
                                                                  K:2*(g+m), u:-v)
                                                              else y:0),
         if y=1 then (an:a, Bn:B, Dn:D, o1:o , A1:B-(B1:A)*a, C1:D-(D1:C)*a) else 1) ,

TI(x):=(m:10, s:10^10, ai(x), if y=1 and a>0 then (F:C+D*r, f:2*D*10^(Q+K-P), fpprec:V:P+4*Z,
                                                    d:o*(B-D*b(x)),
                                                    if  1 < (F+f)*d  and  (F-f)*d <1 then 1  else  y:0)
                                               else y:0) ,

Da(x):=(if n=0 then 1
                else (m:1, s:10, c:0,  for j while c=0 or c=s-1 do
                                (m:m+m, s:s*s, W:P+m , if V<W then (fpprec:V:W+100, t:b(x)) else 1,
                                 w:2*t-B1/D1,  v:B1-D1*w,
                                 if  o1*v > 0 then (a:entier(r:(C1*w-A1)/v),c:entier(s*r)-s*a)
                                              else (c:1,y:0)),
                          if  y=1 and a>= 0  then a:a+1 else y:0)),

 L:0, y:0,  for i while y=0 do (ELI(x), if t<0 then (t:-t, sg:-1) else sg:1, X:sg*x, An(X),
                                  if y=1 then TI(X) else 1 , if y=1 then Da(X) else 1),
 if n = 0 then M:makelist(i,i,0,an)
          else (M:[Bn/Dn],  for i:a while i<an do (Bn:Bn-B1 , Dn:Dn-D1, M:cons(Bn/Dn,M))),
 fpprec:Z, sg*M )  $
```

The following calculations show a value of ε to which $\frac{179}{57}$ is the best fraction. It also shows that $\frac{157}{50}$ can not be a best fraction for π .

```
(%i5) float(%pi-179/57);
(%o5) 0.00124177639681066 8
```

```
(%i6) BF(%pi,1242*10^-6);
```

(%o6) $\frac{179}{57}$

For ε = 1242 10^{-6} , $\frac{179}{57}$ is the best fraction equal to π in ε near.

```
(%i7)  float(%pi-157/50);
(%o7)  0.00159265358979299 2
```

```
(%i8) BF(%pi,1592*10^-6);
```

(%o8) $\frac{22}{7}$

We have π -157/50 > 1592 0^{-6} . We can conclude that there is no number ε such as $\frac{157}{50}$ is the best fraction equal in π in ε near.

A case where the best fraction and the best convergent are identical

Proposition B-6

Let x be a positive real number such as b_{n+1} # x . For ε = 1/ (q_{n+1} q_n), b_n is the best convergent equal to x in ε near; it is also the best fraction equal to x in ε near.

```
(%i10) x:sqrt(1+sin(50*exp(20))^2);
```

(%o10) $\sqrt{\sin(50\ \%e^{20})^2 + 1}$

We propose to prove that the best fraction is the convergent in rank 3 for ε = 1/(q_4 q_3).
By making CF(x,4), the convergent in rank 3 is A/C , q_3 = C and q_4 = D . We make:

```
(%i11) (CF(x,4) , disp(A/C) , BF(x,1/(D*C)));
```

$\frac{234}{167}$

(%o11) $\frac{234}{167}$

Case where x is rational and where $b_{n+1} = x$

Result

If $b_{n+1} = x$, x is rational and $| x - b_n| = 1/(q_{n+1}\, q_n)$. For $\varepsilon = 1/(q_{n+1}\, q_n)$, the best convergent is b_{n+1} and the best fraction is the first best fraction in the rank n+1.

We choose, for example, as value of x the convergent of rank 3 of the previous example:

$$x = \frac{234}{167} \quad .$$

```
(%i3) (x:234/167,CFR(x,10));
(%o3) [1,2,2,33]
```

3 is the rank of the last term of the continued fraction of x.
We suggest calculating the best convergent and the best fraction for $\varepsilon = 1/(q_3\, q_2)$.
We determine all the best fraction of rank 3.

```
(%i4) BFRn(x,3);
(%o4)
```

$$\left[\frac{122}{87}, \frac{129}{92}, \frac{136}{97}, \frac{143}{102}, \frac{150}{107}, \frac{157}{112}, \frac{164}{117}, \frac{171}{122}, \frac{178}{127}, \frac{185}{132}, \frac{192}{137}, \frac{199}{142}, \frac{206}{147}, \frac{213}{152}, \frac{220}{157}, \frac{227}{162}, \frac{234}{167}\right]$$

By applying CFR(x,3), we obtain $C = q_2$ and $D = q_3$ because the program stops in the term of rank 3. Then $\varepsilon = 1/(D\, C)$. We apply the program BFR(x,1/(D*C)) .

```
(%i5) (CFR(x,3) , BFR(x,1/(D*C)));
```

$$(\%o5) \qquad \frac{122}{87}$$

The best fraction is the first best fraction in the rank 3.

VII- Appendix:

A- Useful results concerning the continued fractions

Continued fraction associated with a real number

Definitions

N indicate the set of positive or null integers, E(x) is the integer part of x .

We associate in a real number x the sequences (x_n) and (a_n) defined asfollows:
For n = 0 , we set $x_0 = x$, $a_0 = E(x)$.
For n > 0, and $x_{n-1} \# a_{n-1}$, we set $x_n = 1/(x_{n-1} - a_{n-1})$ and $a_n = E(x_n)$.
If it exists p such as $x_p = a_p$, x_n and a_n are not defined for n > p .

M indicate the set of integers n for which x_n and a_n are defined.
If it exists p such as $x_p = a_p$, **M** is the interval [0,p] of integers. Otherwise **M** = **N** .

The sequence (a_n) is the continued fraction associated with x.
The sequence (x_n) is the sequence of successors of x .

For any $q \in \mathbf{M}$, the sequence $(a_q, ...)$ obtained by deleting q first terms of the sequence (a_n) , is the continued fraction associated with x_q .

Proposition A-1

(1) For any $n \in \mathbf{M}$ with n > 0 , $x_n > 1$ and $a_n \geq 1$.

(2) If **M** = [0,p] and if p> 0, then $a_p \geq 2$.

(3) If $n+1 \in \mathbf{M}$, then $x_n = a_n + 1/x_{n+1}$.

(4) X is rational, if and only if the sequence (a_n) is finished.

Convergent of order n associated with x

For $n \in \mathbf{M}$, the convergent of order n associated with x is the rational number defined by:

$$b_n = a_0 + 1/(a_1 + 1/(a_2 + \ldots + 1/(a_{n-1} + 1/a_n) \ldots))$$

We also say that b_n is the converent defined by the finite sequence $(a_0, \ldots, a_n)$.

For example, for n = 4 , $b_4 = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4}}}}$

Proposition A-2

Let x be a non-zero real number and b_n the convergent of order n associated with x.
We set $b_n = p_n / q_n$ where p_n / q_n is an irreducible fraction with $q_n > 0$.
We set $p_{-2} = 0$, $q_{-2} = 1$, $p_{-1} = 1$ and $q_{-1} = 0$.

(1) For any $n \in \mathbf{M}$, we have : $p_n = p_{n-2} + p_{n-1}\, a_n$
$q_n = q_{n-2} + q_{n-1}\, a_n$

(2) For any $n \in \mathbf{M} \cup \{-1\}$, we have: $p_n\, q_{n-1} - p_{n-1}\, q_n = (-1)^{n+1}$.

(3) Sequence (q_n) is increasing, it is it strictly from the rank n = 1 .

(4) If x is irrational, the sequence (q_n) tends to the infinite.

Sequences (G_n) and (H_n) of functions associated with sequence (a_n)

Proposition A-3

We set $G_n(z) = \dfrac{p_{n-2} + z\, p_{n-1}}{q_{n-2} + z\, q_{n-1}}$ and $H_n(t) = \dfrac{t\, q_{n-2} - p_{n-2}}{p_{n-1} - t\, q_{n-1}}$

(1) For n = 0 , G_0 is defined on **R** by $G_0(z) = z$.
For n = 1 , G_1 is defined on $]0,+\infty[$ by $G_1(z) = a_0 + 1/z$.
For n > 1 , G_n is defined on $]-q_{n-2}/q_{n-1}, +\infty[$.
G_n is continuous strictly increasing if n is even and decreasing if n is odd .

(2) H_n is the inverse map of G_n .
For n = 0 , H_0 is defined on $G_0(\mathbf{R}) = \mathbf{R}$ by $H_0(t) = t$.
For n = 1 , H_1 is defined on $G_1(]0,+\infty[) =]a_0, +\infty[$ by $H_1(t) = 1/(t - a_0)$.
For n > 1 , H_n is defined on $G_n(]-q_{n-2}/q_{n-1}, +\infty[)$.
$G_n(]-q_{n-2}/q_{n-1}, +\infty[) =]-\infty, b_{n-1}[$ if n is even and $]b_{n-1}, +\infty[$ if n is odd.
For n > 0, t belongs to domain of definition of H_n , if and only if $(-1)^{n-1}(t - b_{n-1}) > 0$.
H_n is continuous strictly increasing if n is even and decreasing if n is odd .

(3) We have $b_n = G_n(a_n)$, $x = G_n(x_n)$, $x_n = H_n(x)$ and $a_n = H_n(b_n)$

Proposition A-4

Let x > 0 .

For n = 0, we set $F_0 = G_0([0, a_0])$. Then : $F_0 = [0, a_0]$
For n > 0, we set $F_n = G_n(]0, a_n])$. Then: $F_n =]b_{n-2}, b_n]$ if n is even and $[b_n, b_{n-2}[$ if n is n odd.

If n is even , $F_n \subset [0,x]$. If n is odd , $F_n \subset [x,+\infty[$.

The intervals F_n are pairwise disjoint.

If x is irrational, the intervals F_n constitute a partition of $[0,x[\cup]x,+\infty[$.

If x is rational with $x = b_p$, the intervals F_n constitute a partition of
$[0,x] \cup [b_{p-1},+\infty[$ if p is even
$[0,b_{p-1}] \cup [x,+\infty[$ if p is odd

Framing of $|x - b_n|$

Proposition A-5

Let $n+1 \in \mathbf{M}$.

(1) $$x - b_n = \frac{(-1)^n}{(q_{n-1}+q_n x_{n+1})q_n} \quad \text{and} \quad |x - b_n| = (-1)^n (x - b_n)$$

(2) If $n+1 \in \mathbf{M}$, $$\frac{1}{(q_n+q_{n+1})q_n} < |x - b_n| \leq \frac{1}{q_{n+1}q_n} \leq \frac{1}{(q_{n-1}+q_n)q_n} .$$

We have $|x - b_n| = 1/ (q_{n+1} q_n)$ only if x is rational and if $b_{n+1} = x$.

(3) If $n+1 \in \mathbf{M}$, $|q_n x - p_n| > |q_{n+1} x - p_{n+1}|$ and $|x - b_n| > |x - b_{n+1}|$.

(4) The sequence (b_{2k}) is strictly increasing. The sequence (b_{2k+1}) is strictly decreasing.

(5) If x is irrational, the sequence (bn) tends to x , the sequences (b_{2k}) and (b_{2k+1}) are adjacent.

(6) If x is rational, the last convergent b_p is equal to x .

Some algebraic properties of the continued fractions and the convergents

Proposition A-6

Let $(x_0 , x_1 , x_2 , x_3 , \dots , x_n , \dots)$ be the sequence of successors of x ,
$(a_0 , a_1 , a_2 , a_3 , \dots , a_n , \dots)$ be the continued fraction associated with x ,
and $(b_0 , b_1 , b_2 , b_3 , \dots , b_n , \dots)$ be the sequence of convergents associated with x .

Let $a \in \mathbf{Z}$.
Sequence of successors of $x + a$: $(x_0 +a, x_1 , x_2 , x_3 , \dots , x_n , \dots)$.
Continued fraction associated with $x + a$: $(a_0 +a , a_1 , a_2 , a_3 , \dots , a_n , \dots)$.
Sequence of the convergents associated with $x + a$: $(b_0 +a , b_1 +a , b_2 +a , b_3 +a ,\dots , b_n +a , \dots)$.

Let $0 < x < 1$.
Sequence of successors of $1/x$: $(x_1 , x_2 , x_3 , \dots , x_n , \dots)$.
Continued fraction associated with $1/x$: $(a_1 , a_2 , a_3 , \dots , a_{n+1} , \dots)$.
Sequence of the convergents associated with $1/x$: $(1/b_1 , 1/b_2 , 1/b_3 , \dots , 1/b_{n+1} , \dots)$.

Let $x > 1$.
Sequence of successors of $1/x$: $(1/x_0 , x_0 , x_1 , x_2 , x_3 , \dots , x_n , \dots)$.
Continued fraction associated with $1/x$: $(0 , a_0 , a_1 , a_2 , a_3 , \dots , a_{n-1} , \dots)$.
Sequence of the convergents associated with $1/x$: $(0 ,1/b_0 ,1/b_1 ,1/b_2 ,1/b_3 , \dots ,1/b_{n-1} , \dots)$.

Let $x > 0$ and $x - a_0 > 1/2$ $(a1 = 1)$.
Sequence of successors of $-x$: $(-x_0 , x_2 +1 , x_3 , \dots , x_n , \dots)$.
Continued fraction associated with $-x$: $(-a_0-1, a_2 +1, a_3 , \dots , a_n , \dots)$.
Sequence of the convergents associated with $-x$: $(-b_1 , -b_2 , -b_3 , \dots , -b_{n+1} , ...)$.

Let $x > 0$ and $x - a_0 < 1/2$ $(a1 > 1)$.
Sequence of successors of $-x$: $(-x_0 , 1+1/(x_1 -1) , x_1 -1 , x_2 , x_3 , \dots , x_n , \dots)$.
Continued fraction associated with $-x$: $(-a_0-1, 1 , a_1-1, a_2 , a_3 , \dots , a_n , \dots)$.
Sequence of the convergents associated with $-x$: $(-b_0-1, -b_0 , -b_1 , -b_2 , -b_3 , \dots , -b_{n-1} , \dots)$.

Theorem A-7 (Best approximation.)

Let x be an irrational number.

For any integer p and any integer q such as $0 < q < q_{n+1}$ we have : $| q\,x - p | \geq | q_n\,x - p_n |$.

The inequality is an equality if and only if $p = p_n$ and $q = q_n$.

Proposition A-8

Let x be an irrational number. Let p and $q \geq 0$ be two integers.

For any integer n one of convergent $b_n = p_n/q_n$ or $b_{n+1} = p_{n+1}/q_{n+1}$ satisfies: $| x - p/q | < 1/2q^2$.

Proposition A-9

Let x be an irrational number. Let p and q be two integers such as $q > 0$.

The relation $| x - p/q | < 1/2q^2$ implies that p/q is a convergent of x .

Determining a convergent defined by its sequence of integers with Maxima

Let b_n be a convergent defined by the list $F = [a_0, a_1, a_2, \ldots, a_{n-1}, a_n]$.

$$b_n = a_0+1/(a_1 + 1/(a_2 + \ldots + 1/(a_{n-1} + 1/a_n)...)) .$$

A first program uses the recurrence relation: $\beta_n = a_n$, $\beta_i = a_i + 1/\beta_{i+1}$ for $n > i \geq 0$. Then $b_n = \beta_0$.

```
R0(F):=(b:last(F) , F:rest(F,-1) , for j while F # [] do (a:last(F) , b:a+1/b , F:rest(F,-1) ) , b) $
```

Another program, reconstitutes the coefficients A_i , B_i , C_i , D_i by beginning by i = 0.

For the last calculated values of B and D, we have x = B/D.

```
R(F):=(A:0 , B:1 , C:1 , D:0 , for j while F # [] do
                                (a:first(F) , B:A + (A:B)* a , D :C + (C:D) * a , F:rest(F,1) ) , B/D) $
```

```
(%i3)  R([0,1,2,3,4,5,6,7,8,9,10]);
(%o3)  5225670/7489051
```

Proposition A-10

(1) Let $(a_0, \ldots, a_m)$ be a finite sequence of integers such as $a_n \geq 1$ for any $n > 0$ and $a_m \geq 2$.
Then the sequence is the continued fraction of a rational number.

(2) Let $(a_n)_{n \in \mathbf{N}}$ be an infinite sequence of integers such as $a_n \geq 1$ for any $n > 0$.
Then the sequence is the continued fraction of an irrational number.

Proof

(1) We proceed by induction on m.
For m = 0, $x = a_0$. $E(x) = a_0$. (a_0) is the continued fraction of the integer x.
Let us assume the true property at rank m.
Let$(a_0, \ldots, a_{m+1})$ et $x = a_0+1/(a_1 + 1/(a_2 + \ldots + 1/(a_m + 1/a_{m+1})...))$.
We set $x_1 = a_1 + 1/(a_2 + \ldots + 1/(a_m + 1/a_{m+1})...)$.
According to the induction hypothesis, $(a_1, \ldots, a_{m+1})$ is the continued of x_1 .
More $x = a_0+1/x_1$. We verify : $x_1 > 1$. Then $E(x) = a_0$ and $x_1 = 1/(x-a_0)$.
Then $(a_0, \ldots, a_{m+1})$ is the continued of x .

(2) We set $p_{-2} = 0$, $q_{-2} = 1$, $p_{-1} = 1$ and $q_{-1} = 0$.
For any $n \in \mathbf{N}$, we set :
$p_n = p_{n-2} + p_{n-1} a_n$, $q_n = q_{n-2} + q_{n-1} a_n$ and $b_n = a_0+1/(a_1 + 1/(a_2 + \ldots + 1/(a_{n-1} + 1/a_n)...))$.
We check the following properties which depend only on the definitions of p_n , q_n , b_n and Properties of the sequence (a_n) :

(a) $b_n = p_n /q_n$.

(b) For any $n \in \mathbf{N} \cup \{-1\}$, $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n+1}$.

(c) The sequence (q_n) is increasing, with $q_{n+1} > q_n$ for $n \geq 1$.

(d) La suite (q_n) tend vers l'infini .

(e) $b_{n+2} - b_n = (-1)^n a_{n+2} /q_{n+2} q_n$ et $0 < |b_{n+2} - b_n| < 1/q_{n+1} q_n$ pour $n \geq 0$.

(f) $b_{n+1} - b_n = (-1)^n /q_{n+1} q_n$ for $n \geq 0$.

It follows that the sequence (b_{2k}) is strictly increasing, the sequence (b_{2k+1}) is strictly decreasing and their difference tends to 0. These two adjacent sequences converge to a real number x which satisfies $b_{2k} < x < b_{2k+1}$. Then the sequence (b_n) converges to x.

Let $s > n+1$ be an integer . If $a_s > 1$, then $(a_0, \ldots, a_s)$ is the continued of b_s .
If $a_s = 1$, then $(a_0, \ldots, a_{s-2}, a_{s-1}+1)$ is the continued of b_s .
It follows from the proposition A-5 : $(-1)^n (q_n b_s - p_n) > (-1)^{n+1} (q_{n+1} b_s - p_{n+1}) > 0$ for $n \geq -1$.
When s tends to infinity, we obtain: $(-1)^n (q_n b_s - p_n) \geq (-1)^{n+1} (q_{n+1} b_s - p_{n+1}) > 0$.
(The equality $q_{n+1} x - p_{n+1} = 0$ contradicts the relation $b_{2k} < x < b_{2k+1}$) .
We set $x_n = H_n(x)$. $H_n(x) = (q_{n-2} x - p_{n-2})/(p_{n-1} - q_{n-1} x)$.

It remains to prove that $E(x_n) = a_n$ et $x_{n+1} = 1/(x_n - a_n)$ for any integer n.
Calculation gives : $x_n - a_n = H_n(x) - a_n = (q_n x - p_n)/(p_{n-1} - q_{n-1} x)$.
Then $x_{n+1} = 1/(x_n - a_n)$.
As $q_n x - p_n$ and $q_{n-1} x - p_{n-1}$ have opposite signs, we have $x_n - a_n > 0$.
More $(q_n x - p_n)/(p_{n-1} - q_{n-1} x) \leq 1$. If $x_n - a_n < 1$, then $E(x_n) = a_n$.

What happens if $x_n - a_n = 1$? Let us suppose, for example, n is even.
We have $x_n = a_n + 1$, $x = G_n(a_n+1) = G_{n+1}(1) \geq G_{n+1}(a_{n+1}) = b_{n+1}$. What contradicts $x < b_{n+1}$.

Then the sequence (a_n) is the continued of x and x is an irrational number.

B- Best fraction equal to x in ε near

In what follows ε denotes a positive real number.

Let x be a nonzero real number.

If $x > 0$,we suggest determiningthe smallest integers $p \geq 0$ and $q > 0$ such as $|x - p/q| < \varepsilon$.
In this case p/q is the best fraction equal to x in ε near.

Il $x < 0$, the best fraction equal to x in ε near. the opposite of the best fraction equal to |x| in ε near.

In what follows, we suppose $x \geq 0$

Proposition B-1

Let A, B, C, D be positive or null integer such as $|AD - BC| = 1$.
We have the following properties:

(1) B and D be relatively prime. A and C are relatively prime.

(2) Let d be an integer. Then A + B d and C + D d are relatively prime.

(3) Let r/s be an irreducible fraction with r> 0 and s> 0. Then:

(a) A s + B r and C s + D r are relatively prime.

(b) If r/s > d-1 where d is positive or null integer, we have:
$A\,s + B\,r \geq A + B\,d$ and $C\,s + D\,r \geq C + D\,d$.
If none of the integers A, B, C and D is zero and if r/s # d, these inequalities are strict.

Proof

(1) , (2) and (3) (a) are obtained with Bezout' theorem.
Let us demonstrate (3) (b), Let us suppose r/s # d. It is enough to demonstrate $r \geq d$.
It is evident in the following cases d = 0 , d = 1 , s = 1 .
If $d \geq 2$ and $s \geq 2$ we have $r > s(d-1) \geq d + d - 2 \geq d$.

Existence of the best fraction

We will use the function G_n defined by $G_n(z) = \dfrac{p_{n-2} + z\,p_{n-1}}{q_{n-2} + z\,q_{n-1}}$ with $p_{n-1}.q_{n-2} - p_{n-2}.q_{n-1} = (-1)^n$.

Proposition B-2

Let $x \geq 0$.

(1) There is a smallest integer n such as $|x - b_n| < \varepsilon$.

(2) There is a smallest positive or null integer d such as $|x - G_n(d)| < \varepsilon$.

(a) If n = 0, d is an element of the interval $[0, a_0]$.

(b) If n>0, d is an element of the interval $[1, a_n]$.

(c) $G_n(d)$ is the best fraction equal to x in ε near.

Proof

(1) Let **A** be the set of the integers n such as $|x - b_n| < e$. If x is irrational, **A** is not empty because the sequence $(|x - b_n|)$ tends to 0. If x is rational, **A** is not empty because it contains the rank m of the last term of the continued fraction, since $b_m = x$ .Thus **A** admits a smaller element.

(2) Let **B** be the set of the integers p such as $|x-G_n(p)| < e$. **B** is lower bound by 0 and is not empty because it contains a_n ($G_n(a_n) = b_n$). Thus **B** admits a smaller element d belonging to $[0,a_n]$.

(b) Lat $n > 0$. If $n = 1$, $d \# 0$ as $G_1(0)$ is not défined .
In other cases, $G_n(0) = b_{n-2}$ and $|x - b_{n-2}| \geq \varepsilon$. Then $d \# 0$.
Then d belong to the interval $[1, a_n]$.

(c) If 0 belongs to $]x-\varepsilon,x+\varepsilon[$, we have $n = 0$, $G_0(0) = 0 = 0/1$ is the best fraction.
Suppose that 0 does not belong to $]x-\varepsilon,x+\varepsilon[$. We verify $]x-\varepsilon,x+\varepsilon[\subset G_n(]0, +\infty[)$:
for example, for n even > 0 , we have $x - b_{n-2} \geq \varepsilon$ and $b_{n-1} - x \geq \varepsilon$.
Then $]x-\varepsilon,x+\varepsilon[\subset]b_{n-2}, b_{n-1}[= G_n(]0, +\infty[)$ (proposition A-6) .

Let p/q be a fraction belonging to $]x-\varepsilon,x+\varepsilon[$.
There is an irreducible fraction $r/s > 0$ such that $s>0$ and $p/q = G_n(r/s)$.
As G_n is monotonic and d is the smallest integer such that $G_n(d)$ belongs to $]x-\varepsilon,x+\varepsilon[$, we have

$r/s > d-1$. $G_n(r/s) = (s\, p_{n-2} + r\, p_{n-1})/(s\, q_{n-2} + r\, q_{n-1})$ and $G_n(d) = (q_{n-2} + d\, q_{n-1})/(q_{n-2} + d\, q_{n-1})$.

According to the proposition B-1 , the fraction $(s\, p_{n-2} + r\, p_{n-1})/(s\, q_{n-2} + r\, q_{n-1})$ is irreducible and

We have : $s\, p_{n-2} + r\, p_{n-1} \geq p_{n-2} + d\, p_{n-1}$ and $s\, q_{n-2} + r\, q_{n-1} \geq q_{n-2} + d\, q_{n-1}$.

Then $p \geq p_{n-2} + d\, p_{n-1}$ and $q \geq q_{n-2} + d\, q_{n-1}$.

Then $G_n(d)$ is the best fraction.

Proposition B-3

Let H_n be the reverse function of G_n . Then $d = \max(E(H_n(x-(-1)^n \varepsilon))+1,0)$.

Proof:

d is the smaller integer positive or null such as $|x - G_n(d)| < \varepsilon$.
Let H_n be the reverse function of G_n.

Let $n = 0$. $G_0(d) = d$. For $d \in [0, a_0]$ we have $0 < x - d < \varepsilon$.
d is the smaller integer such as $d > x - \varepsilon$ and $d \geq 0$.
Then $d = \max(\text{Entier}(X-\varepsilon) + 1,0)$ . .

Let $n > 0$ and n impair. G_n is a strictly increasing function on $[0, a_n]$.
d is the smaller integer which verifie: $0 < x - b_n = x - G_n(a_n) \leq x - G_n(d) < \varepsilon$
d is the smaller integer which verifie: $G_n(d) > x - \varepsilon$. Then $d = \text{Entier}(H_n(x-\varepsilon)) + 1$.

For $n > 0$ and n odd , is obtained in the same manner : $d = \text{Entier}(H_n(x+\varepsilon)) + 1$.

In any case, we have: $d = \max(\text{Entier}(H_n(x-(-1)^n \varepsilon)+1,0)$.

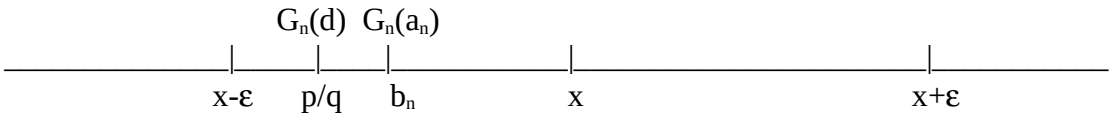
Remarks

(1) Let n be the integer defined in the proposition B-2.
For any k > n, we have $|x - b_k| < \varepsilon$, $p_n \leq p_k$ and $q_n \leq q_k$ (these inequalities are strict for n>0).
We will say that b_n is the best convergent equal to x in ε near .

(2) The best convergent equal to x in ε near is not always the best fraction equal to x in ε near.

(3) The best convergent is defined by the sequence $(a_0 , a_1 , a_2 , a_3 , \dots , a_{n-1} , a_n)$.
The best fraction is defined by the sequence $(a_0 , a_1 , a_2 , a_3 , \dots , a_{n-1} , d)$.

Interpretation in the case where n is even



G_n being strictly increasing, it is sometimes possible to find an integer $d \in [0,a_n[$ such as $G_n(d) \in]x-\varepsilon, b_n[$.
In that case the best fraction equal to x in ε near is different of b_n .
The objective of what follows is to characterize, for a given value of n, the integers d such as $G_n(d)$ is the best fraction equal to x in ε near, for a value suitably chosen of ε.

For n = 0, $G_0(d) = d$ and $a_0 = b_0$. The suitable values of d are integers belonging to the interval $[0,a_0]$. It is enough to choose $\varepsilon = x-d+1/2$.

We deduce from the proposition A-4 the following result.

Proposition B-4

Let x > 0 . Let p/q be the best fraction equal to x in ε near .
The pair (n,d) which verifies $d \in [1, a_n]$ and $p/q = G_n(d)$ is unique.

Remark

Let x > 0 .
If n is even the inequality $b_{n-1} - x > x - G_n(d)$ is equivalent in $G_n(d) + b_{n-1} > 2x$.
If n is odd the inequality $x - b_{n-1} > G_n(d) - x$ is equivalent in $G_n(d) + b_{n-1} < 2x$.
Is verified $(-1)^n (G_n(a_n) + b_{n-1} - 2x) > 0$.

Proposition B-5

Let x > 0 and a_n the term of rank n > 0 of the continued fraction associated with x .

(1) the smallest positive integer a such as: (C) $(-1)^n (G_n(a) + b_{n-1} - 2x) > 0$
is defined by: $a = E(H_n(2x - b_{n-1})) + 1$

(2) Let d an integer. We can find ε such as $G_n(d)$ be the best fraction equal to x in ε near if and only if d is element of $[a,a_n]$.

Proof

Let us place we in case n is even.
(1) Condition (C) is written $G_n(a) > 2x - b_{n-1}$.
As G_n is strictly increasing, we have $a = E(H_n(2x - b_{n-1})) + 1$

(2) Let us note at first that $x - b_n < b_{n-1} - x$ and $G_n(a_n) = b_n$. Let d be an element of $[a,a_n]$.
If d = a, according to the condition (C), the definition of a, and by the fact that G_n is increasing we have: $0 \le x - b_n \le x - G_n(a) < b_{n-1} - x \le x - G_n(a-1)$.
It is enough to choose ε such as $x - G_n(a) < \varepsilon < b_{n-1} - x$.
If d > a , we have: $0 \le x - b_n \le x - G_n(d) < x - G_n(d-1) \le x - G_n(a) < b_{n-1} - x$.
It is enough to choose ε such as $x - G_n(d) < \varepsilon < x - G_n(d-1)$.

Prove that if d belongs to the interval [1,a[, $G_n(d)$ can not be a best fraction.
For d element of [1,a[, we have: $0 < b_{n-1} - x \le x - G_n(a-1) \le x - G_n(d)$.
If there was ε such as $G_n(d)$ be the best fraction equal to x in ε near, we would have $0 < b_{n-1} - x < \varepsilon$.
There would be an integer $p \le n-1$ and $d' \in [1, a_p]$ such as $G_n(d) = G_p(d')$, this contradicts proposition B-4.

Remark

Let x > 0 . Any integer $d \in]0 , a_0]$ is the best fraction equal to x for $\varepsilon \in]x-d , x-d+1]$,
0 is the best fraction equal to x for ε > x.

A case where the best fraction and the best convergent are identical

Proposition B-6

Let x be a positive real number such as $b_{n+1} \neq x$. Let $\varepsilon = 1/q_{n+1} q_n$. Then:
(1) bn is the best convergent equal to x in ε near.
(2) bn is the best fraction equal to x in ε near.

Proof
(1) According to the proposition A-4, we have $|x - b_n| < \varepsilon$. If n = 0, b_0 is the best convergent.
Let n > 0. $|x - b_{n-1}| > 1/ (q_{n-1} (q_{n-1} + q_n)) \ge 1/ (q_{n-1} q_{n+1}) \ge 1/(q_{n+1} q_n) = \varepsilon$.
Then b_n is the best convergent.
(2) If $n \neq 0$ and $a_n = 1$, the property is evident. In other cases,
$|x - G_n(a_n - 1)| = (x_n - a_n + 1)/((q_n - q_{n-1})(x_n q_{n-1} + q_{n-2})) > 1/((q_n - q_{n-1})(q_n + q_{n-1})) \ge$
$1/((q_n - q_{n-1}) q_{n+1}) \ge 1/(q_{n+1} q_n) = \varepsilon$. Then b_n is the best fraction.

Remark

We have $b_{n+1} = x$ when x is rational. In that case, $| x - b_n| = 1/(q_{n+1} q_n) = \varepsilon$.
Then the best convergent is b_{n+1} and the best fraction is the first best fraction in the rank n+1 .

Proposition B-7

Let x be a positive irrational number. Let p be the smallest integer such as $\dfrac{1}{(q_{p-1} + q_p) q_p} \le \varepsilon$.

Then the smallest integer n such as $|x - b_n| < \varepsilon$ is p-1 or p .

Proof
We use the following result: $1/((q_p+q_{p+1}) q_p) < |x - b_p| < 1/((q_{p-1}+q_p) q_p)$.
From this relation we deduct $n \le p$.
If p = 0 or p = 1, the property is obvious.
Let i >1 . we have $|x - b_{p-2}| > 1/ ((q_{p-2} + q_{p-1}) q_{p-2}) > 1/((q_{p-2} + q_{p-1}) q_{p-1}) > \varepsilon$. Then n > p-2 .

C- Decimal representation of a real number

Proposition C-0

(1) $9\,(10^{m-1} + 10^{m-2} + ... + 10 + 1) = 10^m - 1$.

(2) $9\,(1/10 + 1/10^2 + ... + 1/10^m) = 1 - 1/10^m$.

(3) $9\,(1/10 + 1/10^2 + ... + 1/10^m)$ tends to 1 when m tends to infinity .

Let $x = a_0 + a_1/10 + a_2/10^2 + \ldots + a_p/10^p$.
a_0 is an integer, a_1 , a_2 , … , a_p are integers between 0 and 9 .
x is a decimal number as 10^p x is an integer.

Let m be an integer such as $0 \le m \le p$. We set : $S_m = a_0 + a_1/10 + a_2/10^2 + \ldots + a_m/10^m$.

Then : $x = S_m + b/10^m$ where $b = a_{m+1}/10^1 + \ldots + a_p/10^{p-m}$.

We verify: $S_m = E(10^m x)/10^m$, $0 \le b \le 9\,(1/10 + 1/10^2 + ... + 1/10^{p-m}) = 1 - 1/10^{p-m} < 1$, $a_0 = E(x)$ and

$$0 \le x - (a_0 + a_1/10 + a_2/10^2 + \ldots + a_m/10^m) < 1/10^m$$

This simply express that for $x = 5{,}947695234$, we have $0 \le x - 5{,}947695 < 1/10^6$

We can complete the sequence $(a_0 , a_1 , a_2 , .. , a_p)$ with $a_n = 0$ for $n > p$.
The previous relation is still valid for any integer $m \ge 0$.

The objective is to determine, for any real number x, an infinite sequence (a_m) having the same property.

This sequence will be decimal representation of x (denoted DR of x).

Proposition C-1

Soient un nombre réel x , un entier $m \ge 0$. On pose $d_m = E(10^m x)/10^m$.

(1) There is only one decimal $d \in]\, x - 1/10^m, x\,]$ such as $d\,10^m$ be integer. It is equal to d_m .

(2) There is only one decimal $d \in]\, x, x + 1/10^m\,]$ such as $d\,10^m$ be integer. It is equal to $d_m + 1/10^m$.

(3) The sequences (d_m) and $(d_m + 1/10^m)$ converge to x.

(4) There is an unique infinite sequence (a_m) of integer which verifies :

$$0 \le a_m \le 9 \quad \text{and} \quad 0 \le x - \sum_{i=0}^{m} a_i/10^i < 1/10^m \text{ , for any integer } m \ge 0$$

It is defined with $a_0 = E(x)$ and $a_m = E(10^m x) - 10\,E(10^{m-1} x)$ for $m > 0$.

(5) If x is a decimal number, the terms a_m are zero from a certain rank.

(6) There is no rank from which all terms a_m are equal to 9 .

Proof :

(1) Let d such as 10^m d be an integer. The double inequality $x - 10^{-m} < d \leq x$ is equivalent to $d \leq x < d + 10^{-m}$, next to $10^m d \leq 10^m x < 10^m d + 1$, next to $E(10^m x) = 10^m d$, next to $d = d_m$.
$10^m d_m$ is an integer.

(2) The double inequality $x < d \leq x+1/10^m$ is equivalent to $x - 10^{-m} < d - 10^{-m} \leq x$.
Assuming $10^m d$ integer, $10^m (d - 10^{-m})$ is integer. According to (1) , $d-10^{-m} = d_m$.
Then $d = d_m + 1/10^m$. $10^m (d_m + 10^{-m})$ is an integer.

(3) results from (1) , (2) and that $1/10^m$ tends to 0 when m tends to infinity.

(4) For such a sequence we set $S_m = \sum_{i=0}^{m} a_i /10^i$. $10^m S_m$ is an integer and $S_m \in]\, x-1/10^m, x\,]$.
It follows that $S_m = d_m$.
$a_0 = S_0 = d_0 = E(x)$. For m > 0, calculation gives $a_m = 10^m (d_m - d_{m-1}) = E(10^m x) - 10\, E(10^{m-1}x)$.
This ensures the uniqueness of the sequence (a_m) .

Let $y = 10^{m-1} x$. $a_m = E(10\, y) - 10\, E(y)$. $E(y) \leq y < E(y) + 1$. $10\, E(y) \leq 10\, y < 10\, E(y) + 10$.
Then : $10\, E(y) \leq E(10\, y) < 10\, E(y) + 10$, next : $0 \leq E(10\, y) - 10\, E(y) < 10$.
As a_m is integer, we have $0 \leq a_m \leq 9$.
In the following, $a_0 = E(x)$ and $a_m = E(10^m x) - 10\, E(10^{m-1}x)$ for m > 0 .
Prove by induction: $S_m = d_m$.
For m = 0, $S_0 = a_0 = E(x) = d_0$.
Let m > 0 . We assume $S_{m-1} = d_{m-1}$. Then $S_m = S_{m-1} + a_m /10^m = d_{m-1} + 10^m (d_m - d_{m-1})/10^m = d_m$.
This ensures the existence of the sequence (a_m) .

The sequence (a_m) is the decimal representation of x .

To express that the sequence (S_m) converges to x, we write : $x = \sum_{i=0}^{\infty} a_i /10^i$.

(5) If x is a decimal number, there is an integer p such as 10^p x is an integer.
Then, for any integer m > p , we have $a_m = E(10^m x) - 10\, E(10^{m-1} x) = 10^m x - 10 \times 10^{m-1} x = 0$.

(6) Let bus assume that there is an integer p such as $a_m = 9$ for any m > p .
In that case x is not decimal.

$x = \sum_{i=0}^{p} a_i /10^i + 1/10^p \sum_{i=1}^{\infty} 9/10^i = \sum_{i=0}^{p} a_i /10^i + 1/10^p$ (proposition C-0 (3)). x would be decimal.

Which is contradictory.

Writing

Let x > 0 . We set $x = a_0,a_1\, a_2 \ldots a_m \ldots$ and $-x = -a_0 ,a_1\, a_2 \ldots a_m \ldots$.
Example : $\pi = 3{,}141592653...$. $-\pi = -3{,}141592653...$.

However $(-a_0 , a_1 , a_2 , \ldots , a_m , \ldots)$ is not the DR of – x .
Let (a'_m) be the DR of – x. If x is integer, E(-x) = – x , if not, E(– x) = –E(x) – 1 . Let m > 0.

If 10^{m-1} x and 10^m x are not integer : $a'_m = 9 - a_m$.
If 10^{m-1} x is not integer and 10^m x integer : $a'_m = 10 - a_m$.
If 10^{m-1} x and 10^m x are integers : $a'_m = 0$.

DR of $-\pi$: (- 4,8,5,8,4,0,7,3,4,6, …) . $-\pi$ = -4 +0,858407346… .
For x = 51,289652395, DR of – x : (-52,7,1,0,3,4,7,6,0,5,0, … ,0, …). -x= -52+0,710347605 .

Rounding of a_m to the nearest unit

Using the propositions C-0 and C-1, we proof :

Proposition C-2

(1) If $a_{m+1} \le 4$, we set $D_m = S_m$. Then $0 \le x - D_m < 5/10^{m+1}$.
(2) If $a_{m+1} \ge 5$, we set $D_m = S_m + 1/10^m$. Then $0 < D_m - x \le 5/10^{m+1}$.
In any case : $|x - D_m| \le 5/10^{m+1}$.

Proposition C-3

We assume that $d\,10^m$ is an integer.
If $0 \le x - d < 5/10^{m+1}$, then $d = S_m$ and $a_{m+1} \le 4$.
If $0 < d - x < 5/10^{m+1}$, then $d = S_m + 1/10^m$ and $a_{m+1} \ge 5$.
If $|d - x| = 5/10^{m+1}$, then $d = S_m$ or $d = S_m + 1/10^m$ and $x = S_m + 5/10^{m+1}$.

The first m decimals of the decimal representation of x

Proposition C-4

Let a real number x , an integer $m > 0$ and $c_m(x) = E(10^m x) - 10^m E(x) = E(10^m (x - E(x))$.

(1) The equation : $x = E(x) + c\,10^{-m} + b\,10^{-m}$ has a unique solution (c,b) such as c be an integer and b a real number verifying $0 \le b < 1$: $c = c_m(x)$ and $b = 10^m x - E(10^m x)$.

(2) In basis 10, $c_m(x)$ is written: $a_1 a_2 \ldots a_m$ where a_1 , a_2 , … , a_m are the first m digits of the decimal representation of x.

(3) $0 \le c_m(x) \le 10^m - 1$.

(4) If x is integer, $c_m(x) = 0$ for any integer $m \ge 0$.

(5) If x is not integer, there is an integer $p > 0$ such as for any $m \ge p$, we have $1 \le c_m(x) \le 10^m - 2$.

Proof:

(1) We set $d = E(x) + c\,10^{-m}$. The equation is equivalent to $0 \le x - d < 1/10^m$ where $10^m d$ is an integer. Then $E(x) + c\,10^{-m} = d_m = E(10^m x)/10^m$ (Proposition C-1 (1)).
Calculation gives $c = c_m(x)$, next $b = 10^m x - E(10^m x)$. c is an integer and $0 \le b < 1$.

(2) $c_m(x) = 10^m (S_m - a_0) = \sum_{i=1}^{m} a_i\, 10^{m-i}$.

(3) $0 = \sum_{i=1}^{m} 0\; 10^{m-i} \le \sum_{i=1}^{m} a_i\, 10^{m-i} \le \sum_{i=1}^{m} 9\; 10^{m-i} = 10^m - 1$.

(4) If x is integer, $10^m x$ is integer and $c_m(x) = E(10^m x) - 10^m E(x) = 10^m x - 10^m x = 0$.

(5) According to the proposition C-1, there is a rank $u > 0$ such as $a_u \# 9$.
Then for any $m \ge u$, $c_m(x) \# 10^m - 1$.

If x is not integer, there is a rank $v > 0$ such as $a_v \# 0$. Then for any $m \ge v$, $c_m(x) \# 0$.

Then for any $m \ge \max(u,v) = p$, we have $1 \le c_m(x) \le 10^m - 2$.

Propriété C: $c_{m+p}(y) = 10^p\, c_m(y) + c_p(y)$ and $10^p\, c_m(y) \le c_{m+p}(y) \le 10^p\, c_m(y) + 10^p - 1$

Obtaining condition of an integer part

Proposition C-5

Let x be an integer.
Then for any integer $m>0$ and any y verifying $|x - y| < 10^{-m}$ we have:

either $c_m(y) = 0$ and $x = E(y)$

or $c_m(y) = 10^m - 1$ and $x = E(y) + 1$.

Proof :

We have $x = y = x + a\ 10^{-m}$ with $-1 < a < 1$.

If $0 \le a < 1$, we have $c_m(y) = 0$ and $E(y) = x$.

If $-1 < a < 0$, $y = x - 1 + (1+a\ 10^{-m})$ where $0 < 1+a\ 10^{-m} < 1$. Then $E(y) = x - 1$.

Furthermore, $y = x - 1 + (10^m - 1)\ 10^{-m} + (1+a)\ 10^{-m}$ with $0 < 1+a < 1$. Then $c_m(y) = 10^m - 1$.

Proposition C-6

Let x be a different number of an integer.

(1) Let an integer $m > 0$ and y be a real number such as : $|x - y| < 10^{-m}$ and $1 \le c_m(y) \le 10^m - 1$.
Then: $E(x) = E(y)$.
If more $3 \le c_m(y) \le 10^m - 3$, for any z verifying $|x - z| < 10^{-m}$, we have $E(z) = E(x)$ and $c_m(z) \ge 1$.

(2) There is an integer $p > 0$ such as :
for any integer $m > p$ and any y verifying $|x - y| < 10^{-m}$, we have : $3 \le c_m(y) \le 10^m - 3$.

Proof

(1) $x = y + a\ 10^{-m}$ with $-1 < a < 1$.

$y = E(y) + c_m(y)\ 10^{-m} + b\ 10^{-m}$ with $0 \le b < 1$. Then $x = E(y) + c_m(y)\ 10^{-m} + (a+b)\ 10^{-m}$.

We set : $d = c_m(y) + a + b$. We have : $0 < d < 10^m$. Then $0 < d\ 10^{-m} < 1$ and $E(x) = E(y)$.

Assume $3 \le c_m(y) \le 10^m - 3$. Let z such as $|x - z| < 10^{-m}$. $z = x + e\ 10^{-m}$ with $-1 < e < 1$.

$z = E(x) + (c_m(y) + a + b + e)\ 10^{-m}$. We set $f = c_m(y) + a + b + e$. We have $1 < f < 10^m$.

Then $0 < f\ 10^{-m} < 1$ and $E(z) = E(x)$. More $1 \le E(f) = c_m(z) \le 10^m - 1$.

(3) As x is not integer, there is $q > 0$ such as $1 \le c_q(x) \le 10^q - 2$.

Let an integer $r \ge 1$ and $m = q + r$.

We have: $10^r \le c_{q+r}(x) \le 10^{q+r} - 10^r - 1$ (property C).

$x = E(x) + c_{q+r}(x)\ 10^{-q-r} + b\ 10^{-q-r}$ with $0 \le b < 1$.

Let y such as $|x - y| < 10^{-q-r}$.

$y = x + a\ 10^{-q-r}$ with $-1 < a < 1$. $y = E(x) + c_{q+r}(x)\ 10^{-q-r} + (a+b)\ 10^{-q-r}$.

$d = c_{q+r}(x) + a+b$. $10^r - 1 < d < 10^{q+r} - 10^r + 1$.

Then $3 < 10^r - 1 \le E(d) = c_m(y) \le 10^m - 10^r < 10^m - 3$. Just choose $p = q+1$.

D- Necessary precision in the calculation of a_n

In what follows, x is an irrational number. For any n , x_n is not integer.

The functions G_n and H_n inverses of each other are defined by :

$$G_n(z)=\frac{(A_{n-1}+B_{n-1}z)}{(C_{n-1}+D_{n-1}z)} \quad \text{and} \quad H_n(t)=\frac{(C_{n-1}t- A_{n-1})}{(B_{n-1}- D_{n-1}t)}$$

From proposition A-3, is deduced the following results:

For $n > 0$, q belongs to domain of definition of H_n , if and only if $(-1)^{n-1}(q - b_{n-1}) > 0$.

For any integer $p > 0$ and any integer n, G_n is defined on interval $I =]x_n - 10^{-p} , x_n + 10^{-p} [$.
It's true for $n = 0$. It's also true for $n > 0$ as $x_n > 1$.

Let J the set of numbers q such as $|x_n - H_n(q)| < 10^{-p}$. As G_n is continuous strictly monotone, J is an open interval which contains x defined by $J = G_n(I)$.

We deduct from the above and the proposition C-6 :

Proposition D-1

Let (m_n) be a sequence (m_n) of positive integers and (r_n) be a sequence of real numbers such as, $|x_n - r_n| < 10^{-m_n}$ and $3 \leq c_{m_n}(r_n) \leq 10^{m_n} - 3$.

(1) Let I_n the set of the numbers z such as $|x_n - z| < 10^{-m_n}$.
Let J_n the set of the numbers q such as $|x_n - H_n(q)| < 10^{-m_n}$ We have : $J_n = G_n(I_n)$.
J_n is an open interval containing x . For any$q \in J_n$, $E(H_n(q)) = a_n$.

(2) Let N_n the intersection of the intervals J_i for $0 \leq i \leq n$. N_n is an open interval containing x.
For any $q \in N_n$, the continued fraction of q coincides with that of x at least up to the order n.

Calculation of x_n and a_n is iterative .The value of m_n is not known a priori. We must determine the necessary precision with which we must choose the approximate value q_n of x to obtain a value of m_n satisfying the property (1) with $r_n = H_n(q_n)$.We will use elements from the calculation of a_{n-1} .

Proposition D-2

Let $n > 0$, x be an irrational number and q be a real number such as $(-1)^{n-1}(q - b_{n-1}) > 0$.
We assume that the continued fraction of q coincides with that of x up to rank n-1. Then :

$$| x_n - H_n(q) | = |x - q | \ (C_{n-1} + D_{n-1} x_n)(C_{n-1} + D_{n-1} H_n(q)) \qquad (1)$$

Proof:
As $q \# b_{n-1}$, the continued fraction of q is defined at least up to rank n.
The Function H_n defined with parameters of rank n-1, is the same for x and q .

We have $x_n = H_n(x)$ and we set $r = H_n(q)$. By calculation :

$$|x_n - r| = |H_n(x) - H_n(q)| = \frac{|x-q|}{|D_{n-1}x - B_{n-1}||D_{n-1}q - B_{n-1}|} = \frac{|x-q|}{D_{n-1}^2|x - b_{n-1}||q - b_{n-1}|}$$

As $|A_{n-1} D_{n-1} - B_{n-1} C_{n-1}| = 1$ and using proposition A-5, we obtain :

$|x - b_{n-1}| = 1/[(C_{n-1} + D_{n-1} x_n) D_{n-1}]$ and $|q - b_{n-1}| = 1/[(C_{n-1} + D_{n-1} r) D_{n-1}]$. Relation (1) results.

Proposition D-3

Let $n > 0$ and q be a real number such as $(-1)^{n-1}(q - b_{n-1}) > 0$.
We assume that the continued fraction of q coincides with that of x up to rank n-1.

Let an integer $m > 0$ and r a number such as : $| x_{n-1} - r | < 10^{-m}$ and $3 \le c_m(r) \le 10^m - 3$.

Letr K_n be an integer verifying : $(D_{n-1}\, 10^m)^2 \le 10^{Kn}$.

We assume : $| x_{n-1} - H_{n-1}(q)| < 10^{-m}$.
Then:

$$(C_{n-1} + D_{n-1}\, x_n\ )\ (C_{n-1} + D_{n-1}\, H_n(q)) < (D_{n-1}\, 10^m)^2 \quad (2)$$

$$| x_n - H_n(q) | < |x - q|\ 10^{Kn} \quad (3)$$

Proof
We set $c_m = c_m(r)$.

According to proposition C-6 (1) , we have $E(r) = E(x_{n-1}) = a_{n-1}$.
$r = a_{n-1} + c_m\, 10^{-m} + b\, 10^{-m}$ with $0 \le b < 1$.

By definition, $x_n = 1/(x_{n-1} - a_{n-1})$.
We have $x_{n-1} = r + a\, 10^{-m}$ with $-1 < a < 1$.
Then $x_{n-1} - a_{n-1} = (c_m + a + b)10^{-m} > 2\, 10^{-m}$, next $x_n < 10^m /2$.

Then $C_{n-1} + D_{n-1}\, x_n < C_{n-1} + D_{n-1}\, 10^m /2 \le D_{n-1}\, (1+ 10^m/2)$.

As the continued fraction of q coincides with that of x up to rank n-1, the function H_n is the same for x and q .

We set $H_{n-1}(q) = z$ and $H_n(q) = s$. Then $s = 1/(z - a_{n-1})$.

As $| x_{n-1} - z| < 10^{-m}$, we have $c_m(z) \ge 1$ and $E(z) = E(x_{n-1}) = a_{n-1}$ (proposition C-6 (1)) .
$z = a_{n-1} + c_m(z)10^{-m} + d\, 10^{-m}$ with $0 \le d < 1$. D'où $z - a_{n-1} \ge 10^{-m}$, next $s \le 10^m$.

Then $C_{n-1} + D_{n-1}\, s < C_{n-1} + D_{n-1}\, 10^m \le D_{n-1}\, (1+ 10^m)$.
Next $(C_{n-1} + D_{n-1}\, x_n\ )\ (C_{n-1} + D_{n-1}\, s\ ) \le (D_{n-1})^2\, (10^{2m} +3\, 10^m+2)/2 < (D_{n-1})^2\, 10^{2m}$ (as $m \ge 1$).
(3) results of the relations (1) and (2).

Proposition D-4

Let x be an irrational real number . H_n is the function associated with the continued fraction of x.
We can find two sequences (q_n) and (r_n) of real numbers and sequence (m_n) , (K_n) , (V_n) of integers and a sequence (O_n) of open intervals containing x wich verify :

(1) $r_n = H_n(q_n)$.

(2) $(D_{n-1}\, 10^{m_{n-1}})^2 \le 10^{K_n}$ for $n > 0$.

(3) $| x_n - r_n | < 10^{-m_n}$ and $3 \le c_{m_n}(r_n) \le 10^{m_n} - 3$.

(4) $K_0 \ge 0$, $V_0 \ge K_0 + m_0$ and $V_n \ge \max(K_n + m_n , V_{n-1})$ for $n > 0$.

(5) O_n is the set of numbers q such as $| x - q | < 10^{-V_n}$. $O_n \subseteq O_{n-1}$ for $n > 0$.

(6) $q_n \in O_n$ and for any $q \in O_n$, the continued fraction of q coincides with that of x at least up to the order n.

Proof:

For an integer n, J_n is the set of the numbers q such as $|x_n - H_n(q)| < 10^{-m_n}$ and $N_n = \bigcap_{i=0}^{i=n} J_i$.

We proceed by induction on n jointly proving property : $O_n \subseteq N_n$.

Let n = 0 .

As x is not integer, there is an integer m_0 such as for any y verifying $|x - y| < 10^{-m_0}$, we have $3 \leq c_{m_0}(y) \leq 10^{m_0} - 3$ (proposition C-6 (2)) .

For an integer $K_0 \geq 0$, we choose $V_0 \geq K_0 + m_0$.
We choose q_0 (possibly decimal) such as : $|x - q_0| < 10^{-V_0}$. We have $|x - q_0| < 10^{-m_0}$.
Then $3 \leq c_{m_0}(q_0) \leq 10^{m_0} - 3$. According to C-6 (1), $E(q_0) = a_0$ ($q_0 \in O_0$ and $r_0 = q_0$) .

For any $q \in O_0$, $E(q) = a_0$ (proposition C-6 (1)) . We have $O_0 \subseteq J_0 = N_0$

Let n > 0 .

Assume the sequence (q_n) defined up to order n-1 .

Let $q \in O_{n-1}$. We have $O_{n-1} \subseteq N_{n-1}$.

The continued fraction of q coincides with that of x at least up to the order n-1 (proposition D-1).

We have $|x_{n-1} - r_{n-1}| < 10^{-m_{n-1}}$, $3 \leq c_{m_{n-1}}(r_{n-1}) \leq 10^{m_{n-1}} - 3$ and $|x_{n-1} - H_{n-1}(q)| < 10^{-m_{n-1}}$.

According to the proposition D-3 , we have : $|x_n - H_n(q)| < |x - q| \, 10^{K_n}$ (7) .

As x_n is not integer, there is an integer m_n such as
for any y verifying $|x - y| < 10^{-m_n}$, we have $3 \leq c_{m_n}(y) \leq 10^{m_n} - 3$ (proposition C-6 (2)).

We choose an integer $V_n \geq \max(K_n + m_n , V_{n-1})$. We have $O_n \subseteq O_{n-1}$.

Prove that any $q \in O_n$ belongs to domain of definition of H_n

$(-1)^{n-1}(q - b_{n-1}) = (-1)^{n-1}(x - b_{n-1}) + (-1)^{n-1}(q - x) = |x - b_{n-1}| + (-1)^{n-1}(q - x) \geq |x - b_{n-1}| - |q - x|$.

Note that $C_{n-1} + D_{n-1} x_n < D_{n-1}(1 + 10^{m_{n-1}}/2) < D_{n-1} 10^{m_{n-1}}$ (see proof of proposition D-3).
Alors $|x - b_{n-1}| > 1/[10^{m_{n-1}} (D_{n-1})^2]$ (proposition A-5). De plus $|q - x| < 1/10^{K_n} \leq 1/(D_{n-1} 10^{m_{n-1}})^2$.

Then $(-1)^{n-1}(q - b_{n-1}) > 1/[10^{m_{n-1}} (D_{n-1})^2] - 1/(D_{n-1} 10^{m_{n-1}})^2 > 0$.

We choose q_n (possibly decimal) in O_n and we set $r_n = H_n(q_n)$.
Using (7), we obtain $|x_n - r_n| < 10^{-m_n}$. Then $3 \leq c_{m_n}(r_n) \leq 10^{m_n} - 3$.

The same for any $q \in O_n$, we have $|x_n - H_n(q)| < 10^{-m_n}$. Then $O_n \subseteq O_{n-1} \cap J_n \subseteq N_{n-1} \cap J_n = N_n$.

According to the proposition D-1 the continued fraction of any $q \in O_n$ coincides with that of x at least up to the order n.

Remarks

The calculation program proposes increasing values of m_n until the condition
$3 \leq c_{m_n}(r_n) \leq 10^{m_n} - 3$ holds.
The proposition C-6 (2) assures that a suitable value of m_n will be obtained after a finite number of operations.

If x is rational, the above result is valid for values of n below the rank p of the last term of the continued fraction of x. For n=p , if $c_{m_p}(r_p) = 0$, $E(r_p) = a_p$, if not $E(r_p) = a_p - 1$.

E- Starting Precision

Let x be a real number and E the integer which verifies $10^E \le |x| < 10^{E+1}$ if x # 0.
E = -1 if x = 0 .

For a precision P, we set t_P = bfloat(x) and $s_P = \boldsymbol{\sigma} \times \mathbf{m} \times 10^{\mathbf{e}+1-\mathbf{P}}$ where $\boldsymbol{\sigma}$ is the sign, **m** the mantissa which has P digits and **e** the exponent of bfloat(x). s_P is a decimal number.
For a sufficient value of P , e = E.
With Maxima, s_P = round(t_P*10^(P–e–1))*10^(e+1–P) .

We set X =|x| $\times 10^{-E}$. Let S_{P-1} be the decimal representation of X limited to the rank P-1 .
If x is a decimal number it is assumed that:
(1) Either $|s_P| \times 10^{-E} = S_{P-1}$, or $|s_P| \times 10^{-E} = S_{P-1} + 10^{1-P}$.
(2) If P is greater than the rank of the last non-zero digit of the DR de X then $|s_P| \times 10^{-E} = S_{P-1} = X$.

By the proposition C-1 we prove :
- If $|s_P| \times 10^{-E} = S_{P-1}$, then $0 \le |x| - |s_P| < 10^{E-(P-1)}$.
- If $|s_P| \times 10^{-E} = S_{P-1} + 10^{1-P}$, then $0 < |s_P| - |x| < 10^{E-(P-1)}$. The condition (2) excludes $|s_P| - |x| = 10^{E-(P-1)}$.
In all cases, we have $|x - s_P| < 10^{E-(P-1)}$.

If x is any one we get, at best $|x - s_P| < 10^{E-(P-1)}$ for any $P \ge 1$.
If more $|x - s_P| \le (5/10)10^{E-(P-1)}$ the rounding of the last term is done at the nearest unit .

Generally , under certain conditions there exists an integer $L \ge 1$ such as :
$|x - s_P| < 10^{E-(P-L)}$ for any $P \ge L$. If $P \ge L$ we have $E - 1 \le e \le E + 1$.

Definition

The starting precision of bfloat(x) is the smallest integer L such as $|x - s_P| < 10^{E-(P-L)}$
for any precision $P \ge L$.

If x is the quotient of two integers, the starting precision is estimated to be 1.

Estimation of starting precision using the mean value theorem

Function of elementary variables

The expression of x can contain operations between diverse elements as integers, elementary functions applied to rational numbers.
These numbers are denoted $t_1, t_2, ..., t_n$.

Then x appears as the value of a function
$(y_1, y_2, ..., y_n) \rightarrow F(y_1, y_2, ..., y_n)$ in the point $(t_1, t_2, ..., t_n)$ and $y_1, y_2, ..., y_n$ are elementary variables.

For a precision P we set y_i = bfloat(t_i) the starting precision of which is known.

Calculation of $F(y_1, y_2, ..., y_n)$ gives an evaluation of bfloat(x) .

Mean value theorem

Are (a,b) a point of $\mathbf{R}^2$, let U an open set of R^2 containing (a, b).
(y,z)→ F(y,z) a numeric function of class C1 on U and whose partial derivatives of F are bounded on U.
Let A and B posqitive numbers such as, for any $M \in U$, $|\partial F/\partial y\ (M)| \le A$ and $|\partial F/\partial z(M)| \le B$.
Let h and k real numbers such as (a+h,b+k) is in U .

Then $|F(a+h,b+k) - F(a,b)| \le A|h| + B|k|$

Remark

The mean value theorem gives an estimate of L that do not include roundings by the floating point any step of calculating bfloat (x). However, if the number of intermediate steps required to calculate bfloat(x) is not too large, the estimate obtained by this method which is often broad, allows to obtain a suitable estimation of L.

Example

$x = \sqrt{\sqrt{2} - 1414/10^3}$
We consider the elementary variable $t = \sqrt{2}$. We set $a = 1414/10^3$.
For a precision $P \geq 4$, a is a constant for the flotting point .
It is assumed that the starting of bfloat(t) is 1 .
We verify $14142\ 10^{-4} < t < 14143\ 10^{-4}$, $a+2\ 10^{-4} < t < a+3\ 10^{-4}$, $2\ 10^{-4} < t-a < 3\ 10^{-4}$. Then E=-2.

We set $F(y) = \sqrt{(y-a)}$. Then $F'(y) = -1/(2\sqrt{(y-a)})$. F is of class C^1 on $]a,+\infty[$.

The function $y \rightarrow |F'(y)|$ is decreasing and bounded on the interval $U =]a+\alpha,+\infty[$ where $\alpha = 10^{-4}$.
We verify that t is in U.

For a precision P, we have $|t - bfloat(t)| < 10^{0-(P-1)}$.
For $P \geq 5$ we have $bfloat(t) > t - 10^{-4} > a + 10^{-4} = a+\alpha$.
Then, for any $P \geq 5$, bfloat(t) is in U .

As $y \rightarrow |F'(y)|$ is decreasing on U, we have $|F'(y)| \leq |F'(a+\alpha)| = 50$.

By the mean value theorem we have :

$$|F(t) - F(bfloat(t))| \leq A\,|t - bfloat(t)| < 50\,|t - bfloat(t)| < 10^{-(P-3)} = 10^{E-(P-3+E)}$$

As E = -2, L = 5 is an accurate estimate of the starting precision of bfloat(x).

```
(%i2) CFL(sqrt(sqrt(2)-1414/10^3),100,10);
E=-2
L=4
T=1
(%o2) done
```

Loss of precision by rounding in a intermediate calculation of floating point

Floating point proceeds by basic steps in which it makes a single rounding. It thus determines the values $bfloat(z_1)$, $bfloat(z_2)$, ... , $bfloat(z_q)$ associated with a sequence of numbers $z_1, z_2, \ldots, z_q$ to finally get $bfloat(x) = bfloat(z_q)$.
Each step is accompanied by a loss of precision by rounding.

Proposition E-1

Let d be the decimal approximate value of x_i calculated for a precision $P > M > 0$ by floating point before rounding such as $|x_i - d| < 10^{Ei-(P-M)}$ where E_i is the integer which verifies $10^{Ei} \leq |x_i| < 10^{Ei+1}$.

Then $|x_i - s_P| < 10^{Ei-(P-M-\alpha)}$ where $\alpha = \log_{10}(1 + 10^{2-M})$.
α is the loss of precision by rounding. It does not depend on the applied precision.
For $M = 1$, $\alpha \approx 1$ For $M = 16$, $\alpha \approx 4{,}3\ 10^{-15}$.

Proof

Let e the integer which verifies $10^e \le |d| < 10^{e+1}$. We have $E_i - 1 \le e \le E_i + 1$.

We have : $|d - \sigma_i| < 10^{e-(P-1)}$. Then: $|z_i - \sigma_i| \le |z_i - d| + |d - \sigma_i| < 10^{Ei-(P-M)} + 10^{e-(P-1)} \le 10^{Ei-(P-M)} (1+10^{2-M})$.

We set $\alpha = \log_{10}(1+10^{2-M})$. Then : $|z_i - \sigma_i| < 10^{Ei-(P-M-\alpha)}$.

As in the previous examples, it can be prove that the starting precision is defined for expressions of x whose function of the basic variables, is of class C^1 near $(t_1, t_2,..., t_n)$, knowing that the loss of precision of the floating-point by rounding does not depend on the applied precision but only the number of intermediate steps in calculating bfloat (x).

Let x = log(27)-3*log(3). We can observe a failure with CFL(exp(1)+x^(1/3),100) , while CFL(exp(1)+x^(4/3),100) gives L = 1 .
This is due to the fact that the derivative of the function $t \to t^{1/3}$ tends to infinity when t tends to 0.

Loss of precision in the calculation of u/v by the program CFI(x, n) for x> 0 and n> 1

The basic steps of the calculation of u/v by the floating point can be described as follows:
Calculation of u , calculation of v, calculation of u/v.

In the calculation of v we assume that the floating-point makes two rounding:
one in the calculation of the approximate value of $z_1 = D x$ and another in that the approximate value of $z_2 = B - z1$. (It may not carried out a single rounding).

The calculation of the approximate value of $z_0 = C x - A$, is analogous to the case of v.

In the calculation of the approximate value of $z_3 = z_0/z_2$, the floating-point makes a single rounding.

Values of paramèters

$L \ge 16$. $Q = L+E+2$, $C_{n-1} = C$, $D_{n-1} = D$, $K_n = 2 g_{n-1}+2 m_{n-1}$.

We recall that the sequence (D_n) is increasing.

g_{n-2} and g_{n-1} are integers whitch verify $10^{g_{n-2}-1} < C \le 10^{g_{n-2}}$ and $10^{g_{n-1}-1} < D \le 10^{g_{n-1}}$. We have $g_{n-2} \le g_{n-1}$.

E , E_0 , E_1 , E_2 , E_3 are integers whitch verify :

$10^E \le |x| < 10^{E+1}$, $10^{E_0} \le |z_0| < 10^{E_0+1}$, $10^{E_1} \le |z_1| < 10^{E_1+1}$, $10^{E_2} \le |z_2| < 10^{E_2+1}$, $10^{E_3} \le |z_3| < 10^{E_3+1}$.

Wer have $P \ge Q + K_n + m_n = L+E+2+2 g_{n-1}+2 m_{n-1} + m_n$.

If $E < 0$, then $D \ge D_1 = a_1 = E(1/x)$. Then $g_{n-1} > -E - 1$.

Then $E+1+g_{n-1} > 0$ regardless of the sign of E.

Loss of precision in the calculation of z_1

We have $E+g_{n-1}-1 \le E_1 \le E+g_{n-1}$ .Then $0 \le E+g_{n-1}-E_1 \le 1$.

$|z_1 - D s_P| = D |x - s_P| < D\ 10^{E-(P-L)} \le 10^{E+g_{n-1}-(P-L)} = 10^{E_1-(P-L-E-g_{n-1}+E_1)}$

We have $P \ge L+E+2+2 g_{n-1}+2 m_{n-1} + m_n > L+1 \ge L+E+g_{n-1}-E_1 = M_1 \ge L \ge 16$.

The rounding results in a loss of precision less than $\alpha = \log_{10}(1+10^{-14})$ (proposition E-1).

$|z_1 - \sigma_1| \le 10^{E+g_{n-1}-(P-L-\alpha_1)}$ where σ_1 is the decimal number defined by bfloat(D s_P).

Loss of precision in the calculation of z_2

$|z_2 - (B - \sigma_1)| = |\, z_1 - \sigma_1 \,| < 10^{E+g_{n-1}-(P-L-\alpha_1)}$. $|z_2| = |B - D\,x| = 1/(C+D\,x_n)$.

Using the elements of proof of Proposition D-3, is obtained $x_n < 10^{m_{n-1}}/2$,

next $1/(D\,10^{m_{n-1}}) < |z_2| < 1/D$.

Then $-g_{n-1}-m_{n-1} \le E_2 \le -g_{n-1}$ and $g_{n-1} \le -E_2 \le g_{n-1}+m_{n-1}$

$|z_2 - (B - \sigma_1)| < 10^{E_2-(P-L-E-g_{n-1}+E_2-\alpha_1)}$.

We have $P \ge L+E+2 + 2g_{n-1}+2m_{n-1} + m_n > L+E+g_{n-1}-E_2 + \alpha_1 = M_2 > L-1 \ge 15$.

The rounding results in a loss of precision less than $\alpha_2 = \log_{10}(1+10^{-13})$

$|z_2 - \sigma_2| < 10^{E+g_{n-1}-(P-L-\alpha_1-\alpha_2)}$ where σ_2 is the decimal number defined by bfloat$(B - \sigma_1)$.

$10^{E_2-1} \le |\sigma_2| < 10^{E_2+2}$.

Loss of precision in the calculation of z_0

The loss of precision in the calculation of u,approximate value of $z_0 = C\,x - A$, is analogous to the case of v.

$|z_0 - \sigma_0| < 10^{E+g_{n-2}-(P-L-\alpha_1-\alpha_2)}$. (For n = 2, $|z_0 - \sigma_0| = |x - s_p| < 10^{E-(P-L)}$).

Loss of precision in the calculation of z_3

As $10^{-E_2-2} < 1/|\sigma_2| \le 10^{-E_2+1}$ and $|z_0|/|z_2| < 10^{E_3+1}$, we have:

$|z_0/z_2 - \sigma_0/\sigma_2| \le (|z_0|/|z_2|)\,|z_2 - \sigma_2|/|\sigma_2| + |z_0 - \sigma_0|/|\sigma_2| \le 10^{E_3-E_2+2}\,|z_2 - \sigma_2| + 10^{-E_2+1}|z_0 - \sigma_0|$.

$|z_2/z_0 - \sigma_2/\sigma_0| < 10^{E_3-(P-L-E+E_2-2-g_{n-1}-\alpha_1-\alpha_2)} + 10^{E_3-(P-L-E+E_2-1+E_3-g_{n-2}-\alpha_1-\alpha_2)}$.

As $E_3 \ge 0$ and $g_{n-2} \le g_{n-1}$ we have $|z_2/z_0 - \sigma_2/\sigma_0| < (1+10^{-1})10^{E_3-(P-L-E+E_2-2-g_{n-1}-\alpha_1-\alpha_2)}$.

Then $|z_3 - \sigma_2/\sigma_0| \le 10^{E_3-(P-L-E+E_2-2-g_{n-1}-\beta-\alpha_1-\alpha_2)}$ where $\beta = \log_{10}(1+10^{-1})$.

We have: $P \ge L+E+2 + 2g_{n-1}+2m_{n-1} + m_n > L+E-E_2+2+g_{n-1}+\beta+\alpha_1+\alpha_2 = M_3 > L+1 \ge 17$.

The rounding results in a loss of precision less than $\alpha_3 = \log_{10}(1+10^{-15})$.

Then $|z_3 - \sigma_3| < 10^{E_3-(P-L-E+E_2-2-g_{n-1}-\beta-\alpha_1-\alpha_2-\alpha_3)} = 10^{-(P-L-E-E_3+E_2-2-g_{n-1}-\beta-\alpha_1-\alpha_2-\alpha_3)}$ where σ_3 is the decimal number defined by bfloat(σ_2/σ_0) .

It must also verify that the condition $|z_3 - \sigma_3| < 10^{-m_n}$ is satisfied.

As $x_n < 10^{m_{n-1}}/2$, we have $E_3 \le m_{n-1} - 1$. Then $-E_3 \ge -m_{n-1} + 1$. More $E_2 \ge -g_{n-1}-m_{n-1}$.

$P-L-E-E_3+E_2-2-g_{n-1}-\beta-\alpha_1-\alpha_2-\alpha_3 \ge L+E+2 + 2g_{n-1}+2m_{n-1}+m_n-L-E-E_3+E_2-2-g_{n-1}-\beta-\alpha_1-\alpha_2-\alpha_3$

$\ge 2g_{n-1}+2m_{n-1}+m_n-m_{n-1}+1-g_{n-1}-m_{n-1}-g_{n-1}-\beta-\alpha_1-\alpha_2-\alpha_3 \ge m_n+1-\beta-\alpha_1-\alpha_2-\alpha_3 > m_n$.

(In other cases the loss of precision is still less than α1 + α2 + α3. In particular, it is zero for n = 0).

Conclusion

Proposition E-2

The calculation of u/v with floating point is accompanied by a loss of precision less than:
$\alpha_1+\alpha_2+\alpha_3 < 5\ 10^{-14}$.

The value P_n estimated in the theoretical part is sufficient to absorb the losses of precision from floating point.

F- Regularity of floating point on an interval

We suggest giving a method to estimate the starting precision of bfloat (x).
s_n is the décimal number determinad by bfloat(x) calculated for a precision n .
L is the starting précision of bfloat(x).

For $n \geq L$, we have $|x - s_n| < 10^{E-(n-L)}$. We easily verify:

Proposition F-1

[A,B] is an interval of **N** such as $A \geq L$.

Then, for any $n \in [A,B]$, we have $|x - s_n| < 10^{E-(n-A)}$.
In what follows we do not suppose any more $A \geq L$.

Définition

We say that the floating point is regular on [A,B] (or [A,B[) , if, for any $n \in [A,B]$ (or [A,B[) , we have $|x - s_n| < 10^{E-(n-A)}$.

To express this property, we will say simply that the interval [A,B] is regular.

If the floating point is regular on [A,B] , it is it on any interval $[C,D] \subset [A,B]$.

Proposition F-2

(1) The floating point is regular on [A,+∞ [if and only if $A \geq L$.

(2) The number of the regular intervals [A,B] verifying $A < L$ is finite .

Proof

(1) If $A \geq L$, [A,+∞ [is regular (proposition F-1).
If [A,+∞ [is regular, for any integer $n \geq A$, we have $|x - s_n| < 10^{E-(n-A)}$.

According to the definition of the starting precision, we have $A \geq L$.

(2) Let **G** be the set of all the integers B for which the interval [A,B] is regular and verifies $A < L$.
If **G** is wide, the property is evident. Let us suppose **G** not empty and let us show that **G** is bounded.

Indeed, if this set ensemble was not bounded ,The interval [L-1,+∞ [would be regular, what contradicts the property (1) .

Let M be the biggest element of **G** . Any regular interval [A,B] verifying $A < L$ is included in [1,M]. Their number is thus finite.

Regularity index

Because the number of regular intervals [A,B] verifying A < L , is finite, there is one which has most large number of elements.

Let N_0 be the number of its elements. In the absence of regular intervals [A,B] verifying A < L , in particular if L = 1, we set $N_0 = 0$.

We set $T = N_0 + 1$. T is called regularity index of bfloat(x). Let us note that M < L + T -2 .

Proposition F-3

[A,B] is a regular interval containing at least T elements. Then

(1) A ≥ L .

(2) If furthermore A – 1 = 0 or if A – 1 > 0 and [A-1,B-1] is not regular then L = A .

Example:

For x=sin(sqrt(501))*cos(sqrt(301)) , the values of bfloat(x) when fpprec goes from 1 to 15 , are obtained by program: b(x,m,n):=(for i:m while i<=n do (fpprec:i, disp([i, bfloat(x)]))) $
We make: b(x,1,15);

```
(%i2) x=sin(sqrt(501))*cos(sqrt(301))$ b(x,1,15);
[1,-1.0b-1]
[2,-3.6b-2]
[3,-2.78b-2]
[4,-2.702b-2]
[5,-2.6946b-2]
[6,-2.69378b-2]
[7,-2.693871b-2]
[8,-2.6938616b-2]
[9,-2.69386106b-2]
[10,-2.693861195b-2]
[11,-2.6938611982b-2]
[12,-2.69386119745b-2]
[13,-2.693861197392b-2]
[14,-2.6938611973976b-2]
[15,-2.69386119739711b-2]
(%o2) done
```

In view of this list it can be conjectured that the starting precision is 2 or 3.
Considering s_{15} as a suitable representative of x, E = -2.
We see that there is no regular interval [A, B] for A = 1.
Let us look for the largest regular interval [2, B] included in [1,15].

The following program makes it possible to determine the greatest regular interval [A,B] contained in [A,P]. The program stops when the inequality abs (t-u) <10^(E-(n-A) is no longer satisfied.
For precision "sufficient" P , we set t = bfloat(x). For precision n, we set u = bfloat (x).
Using the following program where t acts as x and u acts as s_n :

```
c(x,E,A,P):=(fpprec:P , t:bfloat(x), f:0 , F:1 ,
                for n:A while n <= P  and  f < F do
                        (p:n, fpprec:n, u:bfloat(x), f:abs(t-u), F:10^(E-(n-A)), disp([n,E-(n-A),f]))) $
```

```
(%i4) c(x,-1,2,15);
[2,-2,9.1b-3]
[3,-3,8.32b-4]
[4,-4,8.488b-5]
[5,-5,7.0333b-6]
[6,-6,8.19564b-7]
[7,-7,9.778887b-8]
[8,-8,4.3655746b-9]
[9,-9,1.33150024b-9]
(%o4) done
```

This calculation shows that [2,8] is regular but not [2,9] .
Moreover, it can be verified that the interval [3,15] is regular. We conjecture L = 3 and T = 8.

Is carried out: CFL(x,100,10); We obtain : *E = -2* *L = 3* *T = 8*

Remark

The value of x is not directly accessible in the programs (case of irrational numbers).
We are going to replace x by s_P for a sufficient value of P.

P-régularity

Are integers A, B and P satisfying $A \le B < P$.
We will say that the floating point is P-regular on an interval
[A,B] or that the interval [A,B] is P-regular, if $|s_P - s_n| < 10^{E-(n-A)}$ for any $n \in [A,B]$.

If the floating point is P-regular on [A,B] , it is it on any interval $[C,D] \subset [A,B]$.

Proposition F-4

Let x be an irrational number and N an integer verifying $N \ge L + T$.
Let **F** be the set of regular intervals [A,B] included in [1,N] . Then:

(1) Any regular interval [A,B] such as A < L is in **F** .

(2) There is an integer P_0 such as for any $P \ge P_0$, **F** be the set of intervals P-regular included in [1,N] .

Proof

(2) **F** is a finite set.
Let $[A,B] \in \mathbf{F}$. Set $n \in [A,B]$ such as $|x - s_n| < 10^{E-(n-A)}$.
When P tends to infinity, $|s_P - s_n|$ tends to $|x - s_n|$.
There is an integer P_n such as for any $P \ge P_n$, we have $|s_P - s_n| < 10^{E-(n-A)}$.
Let P_{AB} the maximum of P_n when n runs n runs through [A,B] and P_F the maximum of P_{AB} when [A,B] n runs through **F**. Then , for any $P \ge P_F$, any element of **F** is P-régulier.

Let **G** be the set of intervals included in [1,N] which are not regular. This set is finite.
Let $[A,B] \in \mathbf{G}$. There is $n \in [A,B]$ such as $|x - s_n| > 10^{E-(n-A)}$, as x is not décimal.
When P tends to infinity, $|s_P - s_n|$ tends to $|x - s_n|$.
There is an integer P'_{AB} such as for any $P \ge P'_{AB}$ we have $|s_P - s_n| > 10^{E-(n-A)}$.
Let P_G the maximum of P'_{AB} when [A,B] runs through **G**.
Then for eny $P \ge P_G$, any element of **G** is not P-regular.

Let $P_0 = \max(P_F,P_G)$. For any $P \ge P_0$, **F** is the set of intervals P-regular included in [1,N].

Proposition F-5

Let x an irrational number, N , L_0 and T_0 are positive integers such as $T_0 \geq T$ and $N \geq L_0 + T_0$.
P is an integer such as $P \geq P_N$ where P_N is the integer defined in proposition F-4.

(1) If the interval $[L_0, L_0+T_0-1]$ is P-régular, then $L_0 \geq L$.

(2) If, furthermore $L_0=1$ or if $[L_0-1, L_0+T_0-2]$ is not P-regular, then $L_0 = L$.

Remark

We give complementary results allowing an largest estimation of L, more easily accessible by the programs than the exact value of L under hypothesis of the proposition F-4 .

Proposition F-6

Let q is an integer such as $e = E+q$. We suppose $P > B-A+L$.
We suppose that, for any $n \in [A,B]$ we have $|s_P - s_n| < 10^{e-(n-A)}$.

If q = -1, then [A,B] is regular.

If q = 0 and $B \geq A+1$, then [A+1,B] is regular.

If q = 1 and $B \geq A+2$, then [A+2,B] is regular.

Proof

Let $n \in [A,B]$, we have $n-A-P+L \leq B-A-P+L \leq -1$.

If q = -1 we have : $|x - s_n| \leq |x - s_P| + |s_P - s_n| < 10^{E-(P-L)} + 10^{e-(n-A)}$
$= 10^{E-(n-A)} [10^{n-A-P+L} + 10^{q}] \leq 10^{E-(n-A)} [2/10] < 10^{E-(n-A)}$.

If $q \geq 0$ we have :
$|x - s_n| \leq |x - s_P| + |s_P - s_n| < 10^{E-(P-L)} + 10^{e-(n-A)} = 10^{E-(n-(A+q))} [10^{n-A-q-P+L} + 1]$.

We have $n-A-q-P+L \leq -1-q \leq -1$.

Then $|x - s_n| < 10^{E-(n-(A+q))} [10^{-1} + 1] < 10^{E-(n-(A+q+1))}$.

Proposition F-7

e , L_0 , T_0 and P are integers such as $P > L_0+T_0 + 2$.
We suppose that for any $n \in [L_0, L_0+T_0-1]$ we have $|s_P - s_n| < 10^{e-(n-L0)}$.

If $e = E-1$ and $T_0 \geq T$, then $L_0 \geq L$.

If $e = E$ and $T_0 \geq T+1$, then $L_0 \geq L-1$.

If $e = E+1$ and $T_0 \geq T+2$, then $L_0 \geq L-2$.

In any case, we have $L_0 + e + 1 \geq L + E$.

Proof

We apply propositions F-3 and F-6 for $A = L_0$ and $B = L_0+T_0-1$.

Replacing $|s_P - s_n|$ by $|t_P - t_n|$

Proposition F-8

Replacing $|s_P - s_n|$ by $|t_P - t_n|$ calculated for a precision n , may cause underestimation of L by one unit .

Proof :

Let F be the integer which verifies : $10^F \le |s_P - s_n| < 10^{F+1}$ $(s_P \neq s_n)$.

Calculated for e precision $n < P$, $|t_P - t_n|$ is an estimation of bfloat($|s_P - s_n|$) calculated for the precision n .

Let σ_n the decimal number defined by $|t_P - t_n|$.

$||s_P - s_n| - \sigma_n| < 10^{F-(n-1)} \le |s_P - s_n| 10^{1-n}$.

Then $|s_P - s_n|(1 - 10^{1-n}) \le \sigma_n \le |s_P - s_n|(1 + 10^{1-n})$
and $\sigma_n /(1+ 10^{1-n}) \le |s_P - s_n| \le \sigma_n /(1- 10^{1-n})$. We set $\alpha = -\log_{10}(1- 10^{1-n})$. $\alpha > 0$.
We set $\alpha = -\log_{10}(1- 10^{1-n})$. $\alpha > 0$.

$\sigma_n < 10^{E-(n-A)}$ implies $|s_P - s_n| < 10^{E-(n-(A+\alpha))}$.

For n = 16, $\alpha \approx 4{,}3\ 10^{-16}$.

The starting precision can be underestimated by one.

In the programs we added an additional unit to the value of L estimated by the regularity test.

D- Some curiosities

Continued fractions associated with the exponential

```
(%i1) CFI(coth(1),20);
(%o1) [1,3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,37,39,41]

(%i2) CFI(coth(1/2),20);
(%o3) [2,6,10,14,18,22,26,30,34,38,42,46,50,54,58,62,66,70,74,78,82]

(%i4) CFI(coth(1/10),20);
(%o5) [10,30,50,70,90,110,130,150,170,190,210,230,250,270,290,310,330,350,370,390,410]

(%i5) CFI(10^50*(exp(10^-50)-1),10);
(%o5) [1, 19999999999999999999999999999999999999999999999999, 2, 1,
          19999999999999999999999999999999999999999999999999, 4, 1,
          19999999999999999999999999999999999999999999999999, 6, 1]

(%i6) CFI(-10^50*(exp(-10^-50)-1),9);
(%o6) [0, 1, 19999999999999999999999999999999999999999999999999, 1, 2,
            19999999999999999999999999999999999999999999999999, 1, 4,
            19999999999999999999999999999999999999999999999999, 1, 6]
```

(%i7) CFI(exp(1),1000);

Evaluation took 0.3120 seconds (0.3120 elapsed) using 101.307 MB.

(%o7)

[2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, 1, 1, 12, 1, 1, 14, 1, 1, 16, 1, 1, 18, 1, 1, 20, 1, 1,
22, 1, 1, 24, 1, 1, 26, 1, 1, 28, 1, 1, 30, 1, 1, 32, 1, 1, 34, 1, 1, 36, 1, 1, 38, 1, 1, 40, 1, 1,
42, 1, 1, 44, 1, 1, 46, 1, 1, 48, 1, 1, 50, 1, 1, 52, 1, 1, 54, 1, 1, 56, 1, 1, 58, 1, 1, 60, 1, 1,
62, 1, 1, 64, 1, 1, 66, 1, 1, 68, 1, 1, 70, 1, 1, 72, 1, 1, 74, 1, 1, 76, 1, 1, 78, 1, 1, 80, 1, 1,
82, 1, 1, 84, 1, 1, 86, 1, 1, 88, 1, 1, 90, 1, 1, 92, 1, 1, 94, 1, 1, 96, 1, 1, 98, 1, 1, 100, 1, 1, 102, 1, 1,
104, 1, 1, 106, 1, 1, 108, 1, 1, 110, 1, 1, 112, 1, 1, 114, 1, 1, 116, 1, 1, 118, 1, 1, 120, 1, 1, 122, 1, 1,
124, 1, 1, 126, 1, 1, 128, 1, 1, 130, 1, 1, 132, 1, 1, 134, 1, 1, 136, 1, 1, 138, 1, 1, 140, 1, 1, 142, 1, 1,
144, 1, 1, 146, 1, 1, 148, 1, 1, 150, 1, 1, 152, 1, 1, 154, 1, 1, 156, 1, 1, 158, 1, 1, 160, 1, 1, 162, 1, 1,
164, 1, 1, 166, 1, 1, 168, 1, 1, 170, 1, 1, 172, 1, 1, 174, 1, 1, 176, 1, 1, 178, 1, 1, 180, 1, 1, 182, 1, 1,
184, 1, 1, 186, 1, 1, 188, 1, 1, 190, 1, 1, 192, 1, 1, 194, 1, 1, 196, 1, 1, 198, 1, 1, 200, 1, 1, 202, 1, 1,
204, 1, 1, 206, 1, 1, 208, 1, 1, 210, 1, 1, 212, 1, 1, 214, 1, 1, 216, 1, 1, 218, 1, 1, 220, 1, 1, 222, 1, 1,
224, 1, 1, 226, 1, 1, 228, 1, 1, 230, 1, 1, 232, 1, 1, 234, 1, 1, 236, 1, 1, 238, 1, 1, 240, 1, 1, 242, 1, 1,
244, 1, 1, 246, 1, 1, 248, 1, 1, 250, 1, 1, 252, 1, 1, 254, 1, 1, 256, 1, 1, 258, 1, 1, 260, 1, 1, 262, 1, 1,
264, 1, 1, 266, 1, 1, 268, 1, 1, 270, 1, 1, 272, 1, 1, 274, 1, 1, 276, 1, 1, 278, 1, 1, 280, 1, 1, 282, 1, 1,
284, 1, 1, 286, 1, 1, 288, 1, 1, 290, 1, 1, 292, 1, 1, 294, 1, 1, 296, 1, 1, 298, 1, 1, 300, 1, 1, 302, 1, 1,
304, 1, 1, 306, 1, 1, 308, 1, 1, 310, 1, 1, 312, 1, 1, 314, 1, 1, 316, 1, 1, 318, 1, 1, 320, 1, 1, 322, 1, 1,
324, 1, 1, 326, 1, 1, 328, 1, 1, 330, 1, 1, 332, 1, 1, 334, 1, 1, 336, 1, 1, 338, 1, 1, 340, 1, 1, 342, 1, 1,
344, 1, 1, 346, 1, 1, 348, 1, 1, 350, 1, 1, 352, 1, 1, 354, 1, 1, 356, 1, 1, 358, 1, 1, 360, 1, 1, 362, 1, 1,
364, 1, 1, 366, 1, 1, 368, 1, 1, 370, 1, 1, 372, 1, 1, 374, 1, 1, 376, 1, 1, 378, 1, 1, 380, 1, 1, 382, 1, 1,
384, 1, 1, 386, 1, 1, 388, 1, 1, 390, 1, 1, 392, 1, 1, 394, 1, 1, 396, 1, 1, 398, 1, 1, 400, 1, 1, 402, 1, 1,
404, 1, 1, 406, 1, 1, 408, 1, 1, 410, 1, 1, 412, 1, 1, 414, 1, 1, 416, 1, 1, 418, 1, 1, 420, 1, 1, 422, 1, 1,
424, 1, 1, 426, 1, 1, 428, 1, 1, 430, 1, 1, 432, 1, 1, 434, 1, 1, 436, 1, 1, 438, 1, 1, 440, 1, 1, 442, 1, 1,
444, 1, 1, 446, 1, 1, 448, 1, 1, 450, 1, 1, 452, 1, 1, 454, 1, 1, 456, 1, 1, 458, 1, 1, 460, 1, 1, 462, 1, 1,
464, 1, 1, 466, 1, 1, 468, 1, 1, 470, 1, 1, 472, 1, 1, 474, 1, 1, 476, 1, 1, 478, 1, 1, 480, 1, 1, 482, 1, 1,
484, 1, 1, 486, 1, 1, 488, 1, 1, 490, 1, 1, 492, 1, 1, 494, 1, 1, 496, 1, 1, 498, 1, 1, 500, 1, 1, 502, 1, 1,
504, 1, 1, 506, 1, 1, 508, 1, 1, 510, 1, 1, 512, 1, 1, 514, 1, 1, 516, 1, 1, 518, 1, 1, 520, 1, 1, 522, 1, 1,
524, 1, 1, 526, 1, 1, 528, 1, 1, 530, 1, 1, 532, 1, 1, 534, 1, 1, 536, 1, 1, 538, 1, 1, 540, 1, 1, 542, 1, 1,
544, 1, 1, 546, 1, 1, 548, 1, 1, 550, 1, 1, 552, 1, 1, 554, 1, 1, 556, 1, 1, 558, 1, 1, 560, 1, 1, 562, 1, 1,
564, 1, 1, 566, 1, 1, 568, 1, 1, 570, 1, 1, 572, 1, 1, 574, 1, 1, 576, 1, 1, 578, 1, 1, 580, 1, 1, 582, 1, 1,
584, 1, 1, 586, 1, 1, 588, 1, 1, 590, 1, 1, 592, 1, 1, 594, 1, 1, 596, 1, 1, 598, 1, 1, 600, 1, 1, 602, 1, 1,
604, 1, 1, 606, 1, 1, 608, 1, 1, 610, 1, 1, 612, 1, 1, 614, 1, 1, 616, 1, 1, 618, 1, 1, 620, 1, 1, 622, 1, 1,
624, 1, 1, 626, 1, 1, 628, 1, 1, 630, 1, 1, 632, 1, 1, 634, 1, 1, 636, 1, 1, 638, 1, 1, 640, 1, 1, 642, 1, 1,
644, 1, 1, 646, 1, 1, 648, 1, 1, 650, 1, 1, 652, 1, 1, 654, 1, 1, 656, 1, 1, 658, 1, 1, 660, 1, 1, 662, 1, 1,
664, 1, 1, 666, 1, 1]

Continued fractions associated with the golden number

(%i8) CFI((1+sqrt(5))/2,77);

(%o8)

[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

(%i9) CFI((1000000000000+sqrt(1000000000000^2+4))/2,9);

(%o9) [1000000000000, 1000000000000, 1000000000000, 1000000000000, 1000000000000,
1000000000000, 1000000000000, 1000000000000, 1000000000000, 1000000000000]

Bibliography:

(1) Beeler, M., Gosper, R.W. Et Schroeppel, R. HAKMEM.
MIT(Massachusetts Institute of Technology) AI Note 239, February 29th, 1972

(2) Jean Vuillemin
Exact Real Computer Arithmetic with Continued Fractions. 14-27 (1988)
Electronic Edition (ACMDL) BiB Tex

(3) M. Couchouron
Développement d'un réel en fractions continues. Université de Rennes 1

(4) Université Paris 7 – Denis Diderot. Année 2007/2008. Licence 2. MA 3. Compléments sur les séries. 1 Le *développement décimal* d'un nombre réel.

(5) Jean-Michel Muller
Arithmétique virgule flottante – 3 septembre 2013– CNRS-INRIA-ENS Lyon-Univ. Claude Bernard

Dominique Drux St Zacharie janvier 2017