Functions approximable by E-fractions

Nicolas Brisebarre\textsuperscript{1} Jean-Michel Muller\textsuperscript{2}
\textsuperscript{1}Laboratoire LARAL, Université Jean Monnet (Saint-Étienne)
and Laboratoire LIP (CNRS/ENS Lyon/INRIA/Univ. Lyon 1),
\textsuperscript{2}CNRS, Laboratoire LIP,
Projet Arénaire, 46 allée d’Italie, 69364 Lyon Cedex 07,
FRANCE

Abstract — After a brief reminder of Ercegovac’s E-method,
we introduce the notion of E-fraction (which is a fraction com-
putable, in a given interval, by the E-method). We characterize
the fractions that are E-fractions and give an algorithm for
checking whether a given function is approximable by an E-
fraction.

I. THE E-METHOD

We all know that, in general, rational approximations of a
given degree (say, the same for numerator and denominator)
to a function are much more accurate than polynomial approxi-
mations of the same degree. And yet, rational approximations
are rather seldom used for approximating elementary functions
in the libraries of current use, because floating-point division is
much slower than floating-point multiplication. The situation
becomes quite different if we use the E-method.

The E-method, introduced in [2], [3], allows efficient so-
lution of diagonally dominant systems of linear equations on
simple and highly regular hardware. Since the evaluation of
polynomials and certain rational functions can be achieved by
solving the corresponding linear systems, the E-method is an
attractive general approach for function evaluation. Consider
evaluation of

\[ R(x) = \frac{p_n x^n + p_{n-1} x^{n-1} + \cdots + p_0}{q_k x^k + q_{k-1} x^{k-1} + \cdots + q_1 x + 1}, \]

where the \( p_i \) and \( q_j \) are real numbers, and define \( n = \max\{m, k\} \), \( p_j = 0 \) for \( m + 1 \leq j \leq n \), and \( q_j = 0 \) for \( k + 1 \leq j \leq n \).

One can show that \( R(x) \) is equal to \( y_0 \), where

\[
\begin{bmatrix}
1 & -x & 0 & \cdots & 0 \\
q_1 & 1 & -x & 0 & \cdots & 0 \\
q_2 & 0 & 1 & -x & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
q_k & 0 & \cdots & 0 & 1 & -x & 0 \\
q_{n-1} & \cdots & 0 & 1 & 0 & \cdots & 0 \\
q_{n} & \cdots & 0 & 0 & 1 & \cdots & 0 \\
\end{bmatrix}
\begin{bmatrix}
y_0 \\
y_1 \\
y_2 \\
\vdots \\
y_{n-1} \\
y_n \\
\end{bmatrix}
= \begin{bmatrix}
p_0 \\
p_1 \\
p_2 \\
\vdots \\
p_{k-1} \\
p_k \\
\end{bmatrix}
\]

The radix-2 E-method consists in solving this linear system
by using the following basic recursion (where \( A \) is the matrix
of the above linear system):

\[ w^{(j)} = 2 \times \left[ w^{(j-1)} - Ad^{(j-1)} \right] \]

i.e., for \( i = 1, \ldots, n - 1 \),

\[ w_i^{(j)} = 2 \times \left[ w_i^{(j-1)} - q_i d_0^{(j-1)} - d_i^{(j-1)} + d_{i+1}^{(j-1)} x \right], \]

and

\[ w_0^{(j)} = 2 \times \left[ w_0^{(j-1)} - d_0^{(j-1)} + d_1^{(j-1)} x \right], \]

and

\[ w_n^{(j)} = 2 \times \left[ w_n^{(j-1)} - d_n^{(j-1)} \right] \]

with \( w^{(0)} = [p_0, p_1, \ldots, p_n]^t \), where the values \( d_i^{(j)} \in \{-1, 0, 1\} \) define the数码 representative of \( D^{(j)} \). One can show that if the sequence \( |w_i^{(j)}| \) is bounded, then \( D^{(j)} \) goes to \( y_0 \) as \( j \) goes to infinity.

The problem at step \( j \) is to find a selection function that gives a
value of the terms \( d_i^{(j)} \) from the terms \( w_i^{(j)} \) such that the values
\( w_i^{(j+1)} \) will remain bounded. In [3], the following selection
function (a form of rounding) is proposed

\[ s(x) = \begin{cases} \text{sign } x \times \lfloor x + 1/2 \rfloor, & \text{if } |x| \leq 1 \\ \text{sign } x \times \lfloor x \rfloor, & \text{otherwise} \end{cases} \]

and applied to the following cases:

1) \( d_i^{(j)} = s(w_i^{(j)}) \), i.e., the selection requires non-redundant \( w_i^{(j)} \);

2) \( d_i^{(j)} = s(w_i^{(j)}) \), where \( w_i^{(j)} \) is an approximation of \( w_i^{(j)} \)
   (in practice, \( w_i^{(j)} \) is deduced from a few digits of \( w_i^{(j)} \)
   by the means of a rounding or a truncation)

Assume

\begin{align*}
\forall i, |p_i| &\leq \xi, \\
\forall i, |x| + |q_i| &\leq \alpha, \\
|w_i^{(j)} - w_i^{(j)}| &\leq \Delta/2.
\end{align*}

The E-method gives a correct result provided that the above
defined bounds \( \xi, \alpha, \) and \( \Delta \) satisfy

\begin{align*}
\xi \leq &\frac{1}{2}(1 + \Delta), \\
0 &< \Delta < 1, \\
\alpha &\leq \frac{1}{2}(1 - \Delta).
\end{align*}

(4)

For instance, if \( \Delta = 1/4 \), one can evaluate \( R(x) \) for \( |x| \leq 1/16 \),
\( \max |p_i| \leq \frac{3}{4} \) and \( \max |q_i| \leq \frac{1}{16} \). Those bounds may seem
quite restrictive, but in practice:

\begin{itemize}
\item if we only wish to evaluate polynomials (i.e., \( q_1 = q_2 = \cdots = q_n = 0 \)),
there exist scaling techniques that make it possible to evaluate any polynomial, in any domain;
\end{itemize}
• if we wish to evaluate rational functions, of course some “scaling” is possible: we can multiply $R(x)$ by a power of 2, so that the $p_i$ are multiplied by the same power of 2. Also, multiplying $x$ by $2^k$, one computes the same function, with $p_i$ and $q_i$ multiplied by $2^{-i}$, but we cannot evaluate all rational functions. In the following, we call \(E\)-fractions the functions that are computable using the E-method (a more formal definition is given in the next section).

II. E-FRACTIONS

Definition 1 \((n, p)\)-fractions: In the following, we call \((n, p)\)-fraction a rational function whose numerator is of degree less than or equal to \(n\), and whose denominator is of degree less than or equal to \(p\).

A. Motivation

As we have seen previously, there is a change of variables that makes it possible to evaluate any polynomial in any domain using the E-method. This is not true for rational functions. And yet, using rational approximations of functions could sometimes be more interesting than using polynomial approximations. The reasons for that are the following:

- firstly, evaluating with the E-method (i.e., using iteration (2)) an \((n, n)\)-fraction is only slightly more expensive than evaluating a degree-\(n\) polynomial;
- secondly, in practice, the best approximation to a given function with an \((n, p)\)-fraction is as accurate as the best approximation with a polynomial of degree very close to \(n + p\). This is illustrated by Table I.

### Table I

<table>
<thead>
<tr>
<th>Function</th>
<th>Domain</th>
<th>(\epsilon)</th>
<th>(n_{\text{pol}})</th>
<th>(n_{\text{frac}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\exp(x))</td>
<td>([0, 1])</td>
<td>10^{-10}</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>(\exp(x))</td>
<td>([-1/128, 1/128])</td>
<td>10^{-20}</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>(\arctan(x))</td>
<td>([-1, 1])</td>
<td>10^{-2}</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>(\log(1 + x))</td>
<td>([-1/4, 1/4])</td>
<td>2^{-24}</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>(\sin(x))</td>
<td>([0, \pi/4])</td>
<td>2^{-16}</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>(\cos(x))</td>
<td>([0, \pi/8])</td>
<td>2^{-53}</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>(\log(1 + 2^x))</td>
<td>([-1/2, 1/2])</td>
<td>2^{-53}</td>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

Definition 2: Let \(I\) be the interval \([-a, a]\), and let \(\Delta\) be a parameter, \(0 < \Delta < 1\).

\[
R(x) = \frac{p_0 + p_1 x + \cdots + p_m x^m}{q_0 + q_1 x + \cdots + q_k x^k}
\]

is an \(E\)-fraction for interval \(I\) and parameter \(\Delta\) if there exists another fraction

\[
R'(x) = \frac{p'_0 + p'_1 x + \cdots + p'_m x^m}{1 + q'_1 x + \cdots + q'_k x^k}
\]
such that

1) there exist two integers \(j_1\) and \(j_0\) such that

\[
\frac{p_0 + p_1 x + \cdots + p_m x^m}{q_0 + q_1 x + \cdots + q_k x^k} = 2^{j_1} R'(2^{j_0} x);
\]

2) the coefficients of \(R'\) satisfy

\[
\begin{align*}
|p'_i| & \leq \frac{1}{2}(1 + \Delta), \\
|q'_i| + 2^{j_0} |a| & \leq \frac{1}{2}(1 - \Delta),
\end{align*}
\]

for any \(i\).

It is worth being noticed that the fraction \(R'\) of Definition 2 is immediately computable by the \(E\)-method, with parameter \(\Delta\), in the interval \([-2^{j_0} a, 2^{j_0} a]\). Hence, Definition 2 defines the rational functions that will be computable in interval \(I\) by the \(E\)-method with a simple change of variable.

B. Characterization of \(E\)-fractions

The following result shows that almost all rational functions will be computable, if interval \(I\) is small enough.

**Theorem 1:** Let

\[
R(x) = \frac{p_0 + p_1 x + \cdots + p_m x^m}{q_0 + q_1 x + \cdots + q_k x^k}
\]

be a rational function, and let \(\Delta\) be a parameter, \(0 < \Delta < 1\). If \(q_0 \neq 0\) then there exists \(a > 0\) such that \(R\) is an \(E\)-fraction for interval \([-a, a]\) and parameter \(\Delta\).

**Proof.** We will proceed by successive transformations of the initial fraction. Assume a (momentarily) arbitrary value \(a > 0\). First, define

\[
\xi = \frac{1}{2}(1 + \Delta),
\]

and

\[
\alpha = \frac{1}{4}(1 - \Delta).
\]

1) we first divide all coefficients by the degree-0 coefficient of the denominator of \(R\). This gives

\[
R^{(1)}(x) = \frac{p_0^{(1)} + p_1^{(1)} x + \cdots + p_m^{(1)} x^m}{1 + q_1^{(1)} x + \cdots + q_k^{(1)} x^k};
\]

with, for any \(i\), \(p_i^{(1)} = p_i / q_0\) and \(q_i^{(1)} = q_i / q_0\). This first step is not really a “transformation”, since, obviously, \(R^{(1)}(x) = R(x)\). Being able to perform that step requires that \(q_0\) be nonzero.

2) Let \(j_0\) be the largest integer such that

\[
|2^{j_0}a| \leq \frac{\alpha}{2},
\]

and define, for any \(i\),

\[
\begin{align*}
p_i^{(2)} & = 2^{-j_0} p_i^{(1)} \\
q_i^{(2)} & = 2^{-j_0} q_i^{(1)}
\end{align*}
\]

The rational function

\[
R^{(2)}(x) = \frac{p_0^{(2)} + p_1^{(2)} x + \cdots + p_m^{(2)} x^m}{1 + q_1^{(2)} x + \cdots + q_k^{(2)} x^k}
\]

satisfies

\[
R(x) = R^{(2)}(2^{j_0} x).
\]
Notice that
\[ \max_{i=1,\ldots,k} \left| q_i^{(2)} \right| = \max_{i=1,\ldots,k} 2^{-j_0 i} \left| \frac{q_i}{q_0} \right|. \]

3) Choose \( j_1 \) equal to the smallest integer such that
\[ \max_{i=1,\ldots,m} \left| p_i^{(2)} \right| / 2^{j_1} \leq \xi, \]
and define, for any \( i \),
\[ \begin{align*}
q_i^{(3)} &= p_i^{(2)}/2^{j_1}, \\
q_i^{(2)} &= q_i^{(3)}. \end{align*} \]

Define \( R' \) as
\[ R'(x) = p_0^{(3)} + p_1^{(3)} x + \cdots + p_m x^m \]
\[ 1 + q_1^{(3)} x + \cdots + q_k^{(3)} x^k. \]

This rational function satisfies
\[ R(x) = 2^{j_1} R'(2^{j_0} x). \]

Therefore, if
\[ \max_{i=1,\ldots,k} 2^{-j_0 i} \left| \frac{q_i}{q_0} \right| \leq \frac{\alpha}{2^j} \]
then \( R' \) is an E-fraction for interval \([-a, a]\) and parameter \( \Delta \).

From the definition of \( j_0 \), we have
\[ 2^{j_0} a \leq \frac{\alpha}{2} < 2^{j_0+1} a, \]
therefore,
\[ \frac{\alpha}{4a} < 2^{j_0}, \]

hence, for any \( i \),
\[ \left| q_i^{(3)} \right| \leq \left( \frac{4a}{\alpha} \right)^i \left| \frac{q_i}{q_0} \right|. \]

Equation (6) shows that if \( \alpha \) is small enough, all values \( \left| q_i^{(3)} \right| \) will be less than \( \alpha/2 \), so that \( R' \) will be an E-fraction for interval \([-a, a]\) and parameter \( \Delta \). This ends the proof of Theorem 1. \( \square \)

When the problem at stake is to approximate functions for which range reduction to a small interval is easily feasible, Theorem 1 is immediately applicable. Examples are the exponential, logarithm and trigonometric functions. Let us examine an example with more details.

C. Application: exponential function in \([-1, 1]\)

Let us consider rational approximations to the exponential function in \([-1, 1]\), with numerators and denominators of degree 3. Let us choose \( \Delta = 1/2 \). Consider the (3,3)-Padé approximant to \( \exp(x) \):
\[ R(x) = \frac{1 + 1/2 x + 1/10 x^2 + 1/120 x^3}{1 - 1/2 x + 1/10 x^2 - 1/120 x^3}. \]

This rational fraction is not an E-fraction for interval \([-1, 1]\) and \( \Delta = 1/2 \). And yet, it is an E-fraction for interval \([-1/128, 1/128]\) and \( \Delta = 1/2 \). The corresponding fraction transformation is
\[ R'(x) = 2^2 R'(2^2 x) \]
with
\[ R'(x) = \frac{1/4 + 1/64 x + 1/2560 x^2 + 1/245760 x^3}{1 - 1/16 x + 1/640 x^2 - 1/61440 x^3}. \]

The approximation error is \( 1.78 \times 10^{-20} \) which is quite good. Getting a similar error in the same interval with a minimax polynomial approximation would require a polynomial of degree 6. Range reduction to \([-1/128, 1/128]\) is done rather easily, if we assume that the values \( \exp(x/128) \) are pre-computed and stored for \( i = -128, \ldots, 128 \).

It is possible to get an even better rational approximation to the exponential function, that is also an E-fraction for interval \([-1/128, 1/128]\) and \( \Delta = 1/2 \), by starting from the minimax rational approximation of degree-3 numerator and denominator to \( \exp(x) \) in \([-1/128, 1/128]\). The approximation error becomes \( 2.75 \times 10^{-22} \).

In the appendix, we give a Maple program that computes the best rational approximation to a given function \( f \) in an interval \( I = [-\text{xmax}, +\text{xmax}] \) and checks if the obtained approximation is an E-fraction. Using that program, we have for instance obtained the following results:

- the best (3,3)-fraction for \( \sin(x) \) in \( I = [-\pi/64, +\pi/64] \) is an E-fraction in \( I \). The approximation error is \( 1.83 \times 10^{-17} \), which corresponds to around 57 bits of accuracy. Reaching the same accuracy with a polynomial would require degree 6: that would correspond to an operator twice as large;
- the best (2,2)-fraction for \( \log(1 + x) \) in \( I = [-\log(2)/256, +\log(2)/256] \) is an E-fraction in \( I \). The approximation error is \( 1.56 \times 10^{-18} \), which corresponds to around 59 bits of accuracy. Reaching the same accuracy with a polynomial would require degree 5.

CONCLUSION

We are able to determine if a rational function is an E-fraction in a given domain. Using that, we are able to find good rational approximations to most usual functions, that can be evaluated using the E-method.

REFERENCES

Our Maple program

The following program computes the best rational approximation to a given function $f$ in an interval $I = [-xmax, +xmax]$ and checks if the obtained approximation is an E-fraction.

```
Efraction := proc(f,xmax,degnum,degden,Delta);
Digits := 45:
with(numapprox):
# computation of best rational approximation of f
# in [0,xmax]
# with degree degnum numerator
# and degree degden denominator
R := minimax(f(x),x=0..xmax,[degnum,degden],1,'err');
# xmax is divided by the smallest value 2^kx
# such that xmax/2^kx is less than 1/8(1-Delta)
# the 1/8(1-Delta) is arbitrary
# it comes from xmax + max|qi| < 1/4(1-Delta)
# I have cut the 1/4(1-Delta) in two parts
kx := floor(log(xmax)/log(2.0));
boundx := (1/8)*(1-Delta);
xmaxnew := evalf(xmax/2^kx);
while xmaxnew > boundx do
  xmaxnew := xmaxnew/2; kx := kx+1 od;
Rup := numer(R); Rdown := denom(R);
# we divide the coefficients by the degree-0
# coefficient of the denominator
for i from 0 to degnum do numerator[i] :=
  coeff(Rup,x,i)/coeff(Rdown,x,0); od;
for i from 0 to degden do denominator[i] :=
  coeff(Rdown,x,i)/coeff(Rdown,x,0); od;
# we take into account the scaling on x
for i from 0 to degnum do
  numerator[i] := 2^(kx*i)*numerator[i] od;
for i from 0 to degden do
  denominator[i] := 2^(kx*i)*denominator[i] od;
scalmaxnum := floor(log(abs(numerator[0]))/log(2.0));
for i from 1 to degnum do
  tempmax := floor(log(abs(numerator[i]))/log(2.0));
  if tempmax > scalmaxnum then scalmaxnum := tempmax; fi
od;
twopscalmaxnum := 2^(scalmaxnum+2);
for i from 0 to degnum do
  numerator[i] := numerator[i]/twopscalmaxnum od;
OK := true;
boundqi := (1/8)*(1-Delta);
for i from 1 to degden do
  if abs(denominator[i]) > boundqi then
    OK := false fi
od;
if OK then
  printf("** The obtained approximation is an 
E-fraction **\n");
  printf("f(x) = 2^%a R(2^%ax), where R is\n",scalmaxnum+2,-kx);
  printf("Error: %a, which means %a bits of 
accuracy\n", err,evalf(-log(abs(err))/log(2.,2)));
  printf("Numerator: 
");
  for i from 0 to degnum do printf("Degree %a : %a\n",i,numerator[i]) od;
  printf("Denominator: 
");
  for i from 0 to degden do printf("Degree %a : %a\n",i,denominator[i]) od;
else printf("** The obtained approximation is NOT an E-fraction **\n"); fi
end;
```

An example using our program

```
> Efraction(x -> sin(x),Pi/64,3,3,1/2);
** The obtained approximation is an E-fraction **
f(x) = 2^1 R(2^0x), where R is
Error: .18307873943985292141956088010e-16,
which means 57. bits of accuracy
Numerator:
Degree 0 : .915393697069904615929920426083131e-17
Degree 1 : .499999999999999999999999999999999e-1
Degree 2 : .964805085688877689143893172480277e-3
Degree 3 : -.58340931613327113256598660327032e-1
Denominator:
Degree 0 : 1.00000000000000000000000000000000
Degree 1 : .192961015344390635659656886237157e-2
Degree 2 : .49984087800149918509058938899444e-1
Degree 3 : .3215169326520224840568937313155e-3
```