ON THE USE OF CONTINUED FRACTIONS FOR DIGITAL COMPUTER ARITHMETIC*

Kishor S. Trivedi
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

Summary

Recently, there has been some interest in the use of continued fractions for digital hardware calculations. We require that the coefficients of the continued fractions be integral powers of two. As a result well known continued fraction expansions of functions cannot be used. Methods of expansion of a large number of functions are presented.

We show that the problem of selection of coefficients of the continued fractions does not have practical solution in most of the cases we have considered. We conjecture that the solution of a polynomial equation is the only problem that can be solved in our formulation.

1. Introduction

In this study, we have investigated the possibility of using continued fractions to evaluate elementary functions in hardware. A continued fraction is represented by:

$$\frac{p_1}{q_1} + \frac{p_2}{q_2} + \dots$$

where p_i is known as a partial numerator and q_i is known as a partial denominator. The recursions to evaluate such a continued fraction are given by [1]:

$$\left.\begin{aligned} P_0 &= Q_{-1} = 0,\ Q_0 = P_{-1} = 1 \\ P_{i+1} &= p_{i+1} P_{i-1} + q_{i+1} P_i \\ Q_{i+1} &= p_{i+1} Q_{i-1} + q_{i+1} Q_i \ . \end{aligned}\right\} \quad (1.1)$$

In order to reduce the four multiplications in the above recursions to shifts, we require that the partial numerators and denominators be integral powers of two. As a result of this restriction we are not able to use well known continued fraction expansions of the functions to be evaluated. For example, to evaluate tan h x, we may not use the expansion:

$$\tanh x = \frac{x}{1} + \frac{x^2}{3} + \dots + \frac{x^2}{2n+1} + \dots \ .$$

The first step in this direction was taken by deriving a method of expansion for the solution to a quadratic equation [2]. We present a method of expansion for $\log_{b_{-1}} b_0$ in this paper. The class of Riccati differential equations is closed under a bilinear transformation [3]. In this paper we show that using this approach, a large number of elementary functions can be expanded into a continued fraction.

We would like to keep the set of allowable values of the partial numerators and denominators small. Once these two sets of allowable values are chosen, the range of numbers representable as continued fractions is fixed and finite. This introduces a restriction on the possible values of p_i and q_i at an iterative step. Furthermore, since the value of the function to be evaluated is known only implicitly through some coefficients, the selection of p_i and q_i is a non-trivial problem. It is also desirable that the selection procedure be computationally simple in the sense that it may use add, subtract and shift operations only. In general, this requires the use of an approximation in the selection procedure [2].

A selection procedure was obtained for the solution of a quadratic equation [2]. This was later extended to higher degree polynomials [4]. In this paper, we show that for functions expandable using the Riccati equation approach and for the algorithm for evaluating $\log_{b_{-1}} b_0$, a simple selection procedure does not exist. More explicitly, we show that the maximum error allowable in the selection procedure is of the same order of magnitude as the error in the solution.

In section 2, we derive the expansions of functions into continued fractions. In section 3, we investigate the selection problem.

2. Methods of Expansion

Let the function to be expanded into a continued fraction be denoted by $f(\underline{a}_0)$ where $\underline{a}_0$ is a vector of arguments. We expand $f(\underline{a}_i)$ (for i=0,1,2,...) using the following bilinear transformation:

$$f(\underline{a}_i) = \frac{p_{i+1}}{q_{i+1} + f(\underline{a}_{i+1})} \quad (2.1)$$

It is required that the vector of coefficients $\underline{a}_{i+1}$ be obtainable from $\underline{a}_i$, p_{i+1}, q_{i+1}, $\underline{a}_{i-1}$, p_i and q_i by means of simple recursions. A recursion is said to be simple if it uses shift, addition and subtraction operations only. Let us denote this system of recursions by:

$$\underline{a}_{i+1} = \underline{G}(\underline{a}_i, \underline{a}_{i-1}, p_{i+1}, p_i, q_{i+1}, q_i) \ .$$

Next we show that many functions fall in this category.

2.1 Solution of a Quadratic Equation [2]

Let $\underline{a}_i = (b_i, c_i)$, $f(\underline{a}_i) = \frac{c_i}{b_i + x}$ and $x = c_0/(b_0+c)$ then $f(\underline{a}_0)$ is a solution to the quadratic $x^2 + b_0 x - c_0 = 0$. In reference [2], a system of simple recursions $\underline{G}$ is derived, which may be written as:

$$b_{i+1} = q_{i+1} c_i - q_i c_{i-1} + b_{i-1}$$

$$c_{i+1} = q_{i+1}(b_i - b_{i+1}) + c_{i-1}$$

In reference [4], this method has been extended to higher degree polynomials.

Another method of expansion for the solution of a quadratic equation $b_0 x_0^2 + c_0 x_0 - d_0 = 0$ is obtained

*This work was supported in part by the National Science Foundation under Grant No. NSF DCR 73-07998.

by letting $\underline{a}_i = (b_i, c_i, d_i)$, $f(\underline{a}_i) = x_i$ where $b_i x_i^2 + c_i x_i - d_i = 0$. Applying the transformation (2.1), the system $\underline{G}$ can be written as:

$$b_{i+1} = d_i / p_{i+1}^2$$

$$c_{i+1} = 2d_i \frac{q_{i+1}}{p_{i+1}^2} - \frac{c_i}{p_{i+1}}$$

$$d_{i+1} = b_i + \frac{c_i \, q_{i+1}}{p_{i+1}} - d_i \left(\frac{q_{i+1}}{p_{i+1}}\right)^2 .$$

2.2 Expansion of Logarithm

Let $\underline{a}_i = (b_i, b_{i-1})$ and $f(\underline{a}_i) = \log_{b_{i-1}} b_i$. Applying the transformation (2.1), we have [5],

$$b_{i+1} = \frac{(b_{i-1})^{p_{i+1}}}{(b_i)^{q_{i+1}}} \qquad (2.2)$$

However, we note that this recursion is not simple. To solve this problem we can easily establish by induction that [6],

$$b_i = \left(\frac{(b_{-1})^{c_i}}{(b_0)^{d_i}} \right)^j \qquad (2.3)$$

where $j = 1$ if i is odd, $j = -1$ if i is even and the recursions for c_{i+1} and d_{i+1} are:

$$\left.\begin{aligned} c_{-1} &= d_0 = 1, \; c_0 = d_{-1} = 0 \\ c_{i+1} &= p_{i+1} \, c_{i-1} + q_{i+1} \, c_i \\ d_{i+1} &= p_{i+1} \, d_{i-1} + q_{i+1} \, d_i \end{aligned}\right\} \qquad (2.4)$$

Comparing the recursion (1.1) and (2.4), we see that $c_i = P_i$ and $d_i = Q_i$ for all i. Therefore, if we let $\underline{a}_i = (P_i, Q_i)$, we have, $f(\underline{a}_0) = \log_{b_{-1}} b_0$.

2.3 The Riccati Equation [3,7]

Consider the first order differential equation:

$$y_i' + \sum_{j=-m}^{n} (a_i)_j \, y_i^j = 0 \qquad (2.5)$$

We apply the bilinear transformation

$$y_i = p_{i+1} / (q_{i-1} + y_{i+1})$$

to equation (2.5) and require that y_{i+1} satisfy a similar differential equation, i.e.,

$$y_{i+1}' = \sum_{j=-m}^{n} (a_{i+1})_j \, y_{i+1}^j .$$

After some tedious algebra, it is easily shown that m=0 and n=2. Now equation (2.5) is seen to be the well known Riccati Equation. Let

$$y_i' = j(a_i \, y_i^2 + b_i \, y_i + c_i) \qquad (2.6)$$

where $j=1$ if i is even, $j=-1$ if i is odd and the initial conditions are, $y_i(0) = t_i$. Applying the bilinear transformation, we obtain the system of recursions [7]:

$$\left.\begin{aligned} a_{i+1} &= c_i / p_{i+1} \\ b_{i+1} &= b_i + 2c_i \, q_{i+1} / p_{i+1} \\ c_{i+1} &= a_i \, p_{i+1} + b_i \, q_{i+1} - c_i \, q_{i+1}^2 / p_{i+1} \\ t_{i+1} &= p_{i+1} / t_i - q_{i+1} \end{aligned}\right\} \qquad (2.7)$$

Now if we let $\underline{a}_i = (a_i, b_i, c_i, t_i, x)$ and $f(\underline{a}_i) = y_i(x)$ then we have a method of expansion of $y_0(x) = f(\underline{a}_0)$. The system $\underline{G}$ is given by the set of recursions (2.7). We note that the recursions for a_{i+1}, b_{i+1} and c_{i+1} are simple since we have assumed that p_{i+1} and q_{i+1} are integral powers of two. However, the recursion for t_{i+1} is not simple. This problem can be solved by letting $t_i = d_i / e_i$, $d_0 = t_0$, $e_0 = 1$ and

$$\begin{aligned} d_{i+1} &= k_{i+1}(p_{i+1} \, e_i - q_{i+1} \, d_i) \\ e_{i+1} &= k_{i+1}(d_i) \end{aligned} \qquad (2.8)$$

We adjoin the recursions (2.8) to (2.7) after removing the recursion for t_{i+1}. Also, the vector $\underline{a}_i$ is redefined so that $\underline{a}_i = (a_i, b_i, c_i, d_i, e_i, x)$. By choosing the initial coefficients a_0, b_0, c_0, d_0 and e_0 appropriately, many different functions can be expanded using this approach.

Let L denote the set of all Riccati equations and L_0 be a subset of L formed by the set of all Riccati equations with constant coefficients. Consider $\ell_0 \in L_0$ given by, $y_0' = a_0 \, y_0^2 + b_0 \, y_0 + c_0$. Depending on the sign of $\Delta = b_0^2 - 4a_0 \, c_0$, the solution $y_0(x)$ of ℓ_0 can be written as,

$$y_0(x) = \frac{\sqrt{-\Delta}}{2a_0} \left(\tan\left(\frac{\sqrt{-\Delta}}{2} x + A_0\right) - \frac{b_0}{\sqrt{-\Delta}}\right)$$

if $\Delta < 0$ and $a_0 \neq 0$;

$$y_0(x) = -\frac{1}{a_0 x} - \frac{b_0}{2a_0} + A_0 \qquad \text{if } \Delta = 0, \; a_0 \neq 0;$$

$$y_0(x) = \frac{\sqrt{\Delta}}{-2a_0} \quad \tanh\left(\frac{\sqrt{\Delta}}{+2} x + A_0\right) - \frac{b_0}{\sqrt{\Delta}})$$

if $\Delta > 0$, $a_0 \neq 0$;

$$y_0(x) = A_0 \, e^{b_0 x} + c_0 \, x \qquad \text{if } a_0 = 0 .$$

Depending on the values of the coefficients a_0, b_0, c_0 and the initial condition $t_0 = y_0(0)$, many different functions may be expanded as shown in the following table.

a_0	b_0	c_0	Δ	t_0	$y_0(x)$
1	0	1	-4	0	tan x
-1	0	-1	-4	∞	cot x
-1	0	0	0	∞	1/x
-1	0	1	4	∞	cot h x
-1	0	1	4	0	tan h x
0	± 1	0	>0	1	$e^{\pm x}$

Table 2.1

Next consider a subset L_1 of L such that,

$$L_1 = \{y' = a(x)\, y^2 + b(x)\, y + c(x) \mid a(x) = k(x)\, \bar{a},\ b(x) = k(x)\, \bar{b},\ c(x) = k(x)\, \bar{c}, \text{ and } \bar{a}, \bar{b}, \bar{c} \text{ are constants}\}.$$

Recursions for $\bar{a}_{i+1}$, $\bar{b}_{i-1}$ and $\bar{c}_{i+1}$ can be derived from the recursions (2.7) and are as follows:

$$\left.\begin{aligned} \bar{a}_{i+1} &= \bar{c}_i / p_{i+1} \\ \bar{b}_{i+1} &= \bar{b}_i + 2\bar{c}_i\, q_{i+1}/p_{i+1} \\ \bar{c}_{i+1} &= \bar{a}_i\, p_{i+1} + \bar{b}_i\, q_{i+1} + \bar{c}_i\, q_{i+1}^2/p_{i+1}. \end{aligned}\right\} \quad (2.9)$$

Depending on the sign of $\bar{\Delta}_0 = \bar{b}_0^2 - 4a_0\, \bar{c}_0$, the solution to $\ell_0 \in L_1$ is given by:

$$y_0(x) = \frac{-\bar{\Delta}_0}{2\bar{a}_0} \left(\tan\left(\frac{\sqrt{-\bar{\Delta}_0}}{2} \int k(x)\, dx + A_0\right) - \frac{\bar{b}_0}{\sqrt{-\bar{\Delta}_0}}\right) \quad \text{if } \bar{\Delta}_0 < 0,\ \bar{a}_0 \neq 0;$$

$$y_0(x) = -\frac{1}{\bar{a} \int k(x)\, dx} - \frac{\bar{b}_0}{2\bar{a}_0} - A_0 \quad \text{if } \bar{\Delta}_0 = 0,\ \bar{a}_0 \neq 0;$$

$$y_0(x) = -\frac{\sqrt{\bar{\Delta}_0}}{2\bar{a}_0} \left(\tan h\left(\frac{\sqrt{\bar{\Delta}_0}}{2} \int k(x)\, dx + A_0\right) - \frac{\bar{b}_0}{\sqrt{\bar{\Delta}_0}}\right) \quad \text{if } \bar{\Delta}_0 > 0,\ \bar{a}_0 \neq 0;$$

$$y_0(x) = A_0\, e^{\bar{b}_0 \int k(x)\, dx} - \frac{\bar{c}_0}{\bar{b}_0} \quad \text{if } a_0 = 0,\ \bar{b}_0 \neq 0;$$

and

$$y_0(x) = \bar{c}_0 \int k(x)\, dx + A_0 \quad \text{if } \bar{a}_0 = \bar{b}_0 = 0 .$$

Clearly, a large class of functions can be expanded with this method.

We will now consider a subset L_{10} of L_1 such that, $L_{10} = \{\ell \in L_1 \mid \bar{\Delta}_0 = 0\}$. Any $\ell \in L_{10}$ can be rewritten as: $y' = k(x)(a^*y+b^*)^2$ where, $a^* = \sqrt{\bar{a}}$, $b^* = a^*(\frac{\bar{b}}{2\bar{a}})$.

The recursions for a_i^*, b_i^* can now be written as follows:

$$\begin{aligned} a_{i+1}^* &= b_i^* / \sqrt{p_{i+1}}\ , \\ b_{i+1}^* &= (a_i^*\, p_{i+1} + b_i^*\, q_{i+1}) / \sqrt{p_{i+1}} \end{aligned} \quad (2.10)$$

The solution to $\ell_0 \in L_{10}$ is given by,

$$y_0(x) = \frac{1}{(a_0^*)^2 (A_0 - \int k(x)\, dx)} - \frac{b_0^*}{a_0^*} \quad \text{if } a_0^* \neq 0,$$

$$y_0(x) = (b_0^*)^2 \int k(x)\, dx + A_0 \quad \text{if } a_0^* = 0 .$$

Note that we can integrate the given function k(x) by this method by setting $a_0^* = 0$ and $b_0^* = 1$.

We conclude this section by presenting a schema for the evaluation of a function using continued fractions. The problem of selection, which is hidden in the procedure "select" of the schema, will be discussed in the next section.

Schema A:

Step 1 [Initialize]:

$P_0 \leftarrow Q_{-1} \leftarrow 0$; $P_{-1} \leftarrow Q_0 \leftarrow 1$; $i \leftarrow 0$;

Initialize the coefficient vector $\underline{a}_0$ depending on the function to be evaluated;

Step 2 [Selection]:

$(p_{i+1}, q_{i+1}) \leftarrow$ select ($\underline{a}_i$, function to be evaluated);

Step 3 [Recursions]:

$\underline{a}_{i+1} \leftarrow \underline{G}(\underline{a}_i, \underline{a}_{i-1}, p_i, p_{i+1}, q_i, q_{i+1})$;

$P_{i+1} \leftarrow p_{i+1}\, P_{i-1} + q_{i+1}\, P_i$;

$Q_{i+1} \leftarrow p_{i+1}\, Q_{i-1} + q_{i+1}\, Q_i$;

Step 4 [Test]:

After a sufficient number of iterations GOTO Step 5; otherwise set $i \leftarrow i+1$ and return to Step 2;

Step 5 [Evaluate]:

$f(\underline{a}_0) \simeq \dfrac{P_{i+1}}{Q_{i+1}}$;

End A;

3. The Selection Problem

Let the set of allowable values of partial numerators be denoted by S_p and the set of allowable values of partial denominators be denoted by S_q. We will assume that both S_p and S_q are finite subsets of positive reals. Let $p_{min} = \min S_p$, $p_{max} = \max S_p$, $q_{min} = \min S_q$ and $q_{max} = \max S_q$. Let the set of numbers representable as infinite continued fractions (i.c.f.) using the sets S_p and S_q be denoted by $R(S_p, S_q)$. Let

$$m = \frac{p_{min}}{q_{max} + \frac{p_{max}}{q_{min} + m}} \quad \text{and}$$

let

$$M = \frac{p_{max}}{q_{min} + \frac{p_{min}}{q_{max} + M}} \; .$$

It is clear that,

$$m = \inf\, (R(S_p,S_q)),$$

$$M = \sup\, (R(S_p,S_q)), \text{ and}$$

$$R(S_p,S_q) \subseteq [m,M] \; .$$

We would like to impose some conditions on the sets S_p and S_q so that $R(S_p,S_q) = [m,M]$. As a result, any number in the interval $[m,M]$ can be represented as an i.c.f. Let $m(p,q) = \frac{p}{q+M}$, $M(p,q) = \frac{p}{q+m}$, $I(p,q) = [m(p,q),M(p,q)]$ and $I(S_p,S_q) = \bigcup_{\substack{p\in S_p \\ q\in S_q}} I(p,q)$. Note that, $I(p,q)$ is a closed interval of the positive real numbers. It can be shown that the following theorem holds [6]:

Theorem 1:

$$R(S_p,S_q) = [m,M] \text{ iff } I(S_p,S_q) = [m,M] \; .$$

Given the sets S_p and S_q, if the conditions of Theorem 1 are satisfied then we say that $R(S_p,S_q)$ is a number system (NS). Given an $f_0 \in [m,M]$ we can expand it into an i.c.f. by letting

$$f_{i-1} = \frac{p_i}{q_i + f_i} \qquad i=1,2,3,\ldots$$

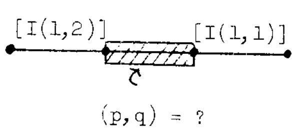
The method of selection of the pair (p_i,q_i) is as follows:

Search for a pair (p_i,q_i) such that

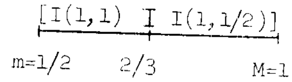
$$p_i \in S_p,\; q_i \in S_q \text{ and } f_{i-1} \in I(p_i,q_i) \; .$$

Note that this search will always succeed provided $R(S_p,S_q)$ is an N.S. Furthermore the definition of $I(p,q)$ guarantees that $f_i \in [m,M]$; therefore, the above procedure can be applied repetitively.

As an example, let $S_p = \{1\}$ and $S_q = \{1,2\}$. In this case a simple computation reveals that the conditions of Theorem 1 are not satisfied and $R(S_p,S_q)$ is not an N.S. The gap between selection intervals $I(1,1)$ and $I(1,2)$ is the reason for trouble as shown by the following figure:



As another example, let $S_p = \{1\}$ and $S_q = \{1,1/2\}$. In this case there are no gaps as shown by the following figure:



Therefore, $R(S_p,S_q)$ forms an N.S. In this case, the selection procedure can be specified as follows:

(a) If $f_{i-1} \in [1/2,2/3)$ then $p_i = 1$, $q_i = 1$.

(b) If $f_{i-1} \in (2/3,1]$ then $p_i = 1$, $q_i = 1/2$.

(c) If $f_{i-1} = 2/3$ then $p_i = 1$ and $q_i = 1/2$ or 1.

Note that two choices are possible for q_i if $f_{i-1} = 2/3$. Let an interval $I(p,q)$ be known as a selection interval. The reason for multiple choice is seen to be the non-null intersection of adjacent selection intervals. As a result of this, some numbers in $[m,M]$ will have multiple i.c.f. representations. Let us define an N.S. $R(S_p,S_q)$ to be nonredundant provided for any two distinct pairs (p,q) and (p',q'), $I(p,q) \cap I(p',q')$ is either null or is a singleton. In such a case it is easy to see that multiple choice of (p_i,q_i) results for only a finite number of points f_{i-1} $[m,M]$. We see that for $S_p = \{1\}$ and $S_q = \{1,1/2\}$, $R(S_p,S_q)$ is a nonredundant N.S. An example of a redundant N.S. is obtained by letting $S_p = \{1\}$ and $S_q = \{1,1/2,1/4\}$. In this case we note that [8],

$$I(1,1/2) \cap I(1,1) = [0.485,\; 0.72] \text{ and}$$

$$I(1,1/2) \cap I(1,1/4) = [0.553,\; 1.124] \; .$$

Thus far, we have outlined a selection procedure when the number to be expanded is known explicitly. However, when using the schema of section 2, the number to be expanded at the i^{th} step (i.e., $f(\underline{a}_i)$) is known only implicitly via the coefficient vector $\underline{a}_i$. Therefore, we should specify the selection procedure in terms of $\underline{a}_i$. Recall that in terms of $f(\underline{a}_i)$, the condition for selecting $(p_{i+1},q_{i+1}) = (p,q)$ is that $f(\underline{a}_i) \in I(p,q)$. This condition must, somehow, be translated in terms of $\underline{a}_i$. Even after such a transformation, it turns out that a prohibitive amount of computation is needed in the selection procedure. We may, however, reduce the computation by use of an approximation. By using a redundant N.S., we hope that the error introduced in the selection due to the use of an approximation will be corrected by the redundancy of the N.S.

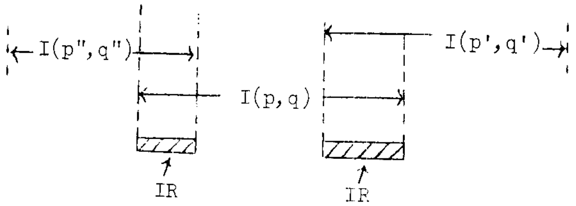
3.1 Selection for the Quadratic [2,8]

The (p,q) selection condition can be written as:

$$m(p,q) \le \frac{c_i}{b_i + x} \le M(p,q)$$

$$\text{or} \quad (b_i+x)\, m(p,q) \le c_i \le (b_i+x)\, M(p,q) \tag{3.1}$$

Note that $f(\underline{a}_0) = x$ is the unknown to be expanded, therefore, we must use an approximation to x. Let us assume that three adjacent selection intervals $I(p,q)$, $I(p',q')$ and $I(p'',q'')$ are as shown in the following figure:



Thus, IL and IR are selection intersection intervals. Let us assume that we have an approximation $\tilde{x}$ of x (and $\tilde{x}$ is simple to compute from b_0 and c_0), and $z\ell$ and zr are properly chosen constants such that $z\ell \in IL$ and $zr \in IR$. We may now use the following (p,q) selection rule:

$$(b_i+\tilde{x}) * z\ell \le c_i \le (b_i+\tilde{x}) * zr \qquad (3.2)$$

It is clear that the selection rule (3.2) may only be used provided the interval of c_i specified by (3.2) is contained in the interval specified by (3.1). In other words,

$$(b_i+x) * m(p,q) \le (b_i+\tilde{x}) * z\ell$$

and

$$(b_i+\tilde{x}) * zr \le (b_i+x) * M(p,q) .$$

Thus we have a restriction on the maximum error allowable in approximating x by $\tilde{x}$. In references [2,8], an approximation $\tilde{x}$ satisfying these conditions was derived. Thus we have an algorithm for the solution of a quadratic equation. This was later extended to higher degree polynomials.

Selection for the second method of the solution to a quadratic is even simpler. Since the (p,q) selection rule in terms of x_i is that $x_i \in I(p,q)$. An approximation $\tilde{x}_i$ to x_i can easily be obtained from the coefficients b_i, c_i and d_i.

3.2 Selection for Logarithm [6]

The (p,q) selection condition in terms of $f(\underline{a}_i)$ is given by:

$$m(p,q) \le \log_{b_{i-1}} b_i \le M(p,q)$$

or

$$b_{i-1}^{m(p,q)} \le b_i \le b_{i-1}^{M(p,q)} .$$

However, this selection rule requires exponentiation and it is in terms of b_i and b_{i-1}. A desirable selection rule will be simple and will be in terms of P_i, Q_i, P_{i-1} and Q_{i-1}. This can be done by rewriting the selection condition as:

$$m(p,q) \le \frac{\log_{b_{-1}} b_i}{\log_{b_{-1}} b_{i-1}} \le M(p,q) .$$

Next we claim that $\log_{b_{-1}} b_{i-1} > 0$ since $0 < m \le \log_{b_{i-1}} b_i \le M$ which implies

$$b_{-1}^{m^{i+1}} \le b_{i-1}^{m} \le b_i \le b_{i-1}^{M} \le b_{-1}^{M^{i+1}}$$

which implies

$$0 < m^{i+1} \le \log_{b_{-1}} b_i \le M^{i+1} .$$

We can, therefore, rewrite the selection condition as,

$$m(p,q) \log_{b_{-1}} b_{i-1} \le \log_{b_{-1}} b_i \le M(p,q) \log_{b_{-1}} b_{i-1} .$$

Now from equation (2.3) we have,

$$\log_{b_{-1}} b_i = j(i) \, [P_i \log_{b_{-1}} b_{-1} - Q_i \log_{b_{-1}} b_0]$$

$$= j(i) \, [P_i - u \, Q_i]$$

where $u = \log_{b_{-1}} b_0$ and $j(i) = 1$ if i is odd and -1 otherwise. Now the selection condition is,

$$j(i-1)[P_{i-1}-uQ_{i-1}] \, m(p,q) \le j(i)[P_i-uQ_i]$$
$$\le j(i-1)[P_{i-1}-uQ_{i-1}] \, M(p,q).$$

Noting that $j(i-1) = - j(i)$ and transposing, we have,

$$j(i) \frac{M(p,q) \, P_{i-1} + P_i}{M(p,q) \, Q_{i-1} + Q_i} \le u \, j(i) \le j(i) \frac{m(p,q) \, P_{i-1} + P_i}{m(p,q) \, Q_{i-1} + Q_i}.$$

Since u is the unknown to be evaluated, we cannot use this as our selection rule. Since b_{-1}^{x} is a monotone increasing function of x (note $b_{-1} > 1$), we can rewrite the selection condition as:

$$j(i) \, b_{-1}^{ARG_i(M(p,q))} \le j(i) \, b_0 \le b_{-1}^{ARG_i(m(p,q))}$$

where

$$ARG_i(s) = \frac{s \, P_{i-1} + P_i}{s \, Q_{i-1} + Q_i} .$$

To reduce the computation, we use an approximation $T_i(s)$ of $b_{-1}^{ARG_i(s)}$. Assume that $I(p,q)$, $I(p',q')$ and $I(p'',q'')$ are three selection intervals as in section 3.1. Then using $T_i(s)$, $z\ell$ and zr, (p,q) selection rule can be specified as:

$$j(i) \, T_i(zr) \le j(i) \, b_0 \le j(i) \, T_i(z\ell) .$$

This selection rule is valid provided

$$j(i) \, x_i(M(p,q)) \le j(i) \, T_i(zr)$$

and

$$j(i) \, T_i(z\ell) \le j(i) \, x_i(m(p,q))$$

where

$$x_i(x) = b_{-1}^{ARG_i(s)} .$$

This implies that the maximum error allowable in the approximation $T_i(s)$ to $x_i(s)$ is of the form $|x_i(s_1) - x_i(s_2)|$ for some $s_1 \ne s_2$ and $s_1, s_2 \in [m,M]$. Note that,

$$x_i(s_1) - x_i(s_2) \quad b_{-1}^{ARG_i(s_2)} [b_{-1}^{ARG_i(s_1) - ARG_i(s_2)} - 1].$$

Since $ARG_i(s)$ and b_{-1} are both finite, it is sufficient to study the difference $ARG_i(s_1) - ARG_i(s_2)$. In particular, if this difference approaches zero then the difference $x_i(s_1) - x_i(s_2)$ also approaches zero. After some algebra, it can be shown that [6],

$$|ARG(s_1) - ARG(s_2)| = \frac{|(s_1 - s_2)| \left(\frac{P_{i-1}}{Q_{i-1}} - \frac{P_i}{Q_i}\right)}{\left(s_1 + \frac{Q_i}{Q_{i-1}}\right)\left(s_2 \frac{Q_{i-1}}{Q_i} + 1\right)}$$

$$< \frac{|s_1 - s_2|}{s_1} \left|\frac{P_{i-1}}{Q_{i-1}} - \frac{P_i}{Q_i}\right|$$

The quantity $\left|\frac{P_{i-1}}{Q_{i-1}} - \frac{P_i}{Q_i}\right|$ is a measure of the rate of convergence of the continued fraction $u = \ell og_{b_{-1}} b_0$. Let us assume that it is $\leq \alpha^{-i}$ for some $\alpha > 1$. Then $|ARG_i(s_1) - ARG_i(s_2)| < \alpha^{-i} \frac{|s_1 - s_2|}{s_1}$. But this means that the maximum error allowable in approximating $x_i(s)$ by $T_i(s)$ rapidly approaches zero. In other words, no approximation can be allowed in the selection rule.

3.3 Selection for the Riccati-Approach [9]

We have seen that the form of the solution to a Riccati equation depends on the sign of the discriminant Δ. It is also clear that the selection procedure will be different for different forms of the solution, i.e., depending on the sign of Δ. Therefore, if Δ remains invariant under the bilinear transformation then hopefully the same selection procedure can be used consistently during the iterative evaluation of a function. It can be easily shown that [7] this is indeed the case, i.e., $\Delta_i = \Delta_{i-1} = \ldots = \Delta_0$.

In section 3.3.1 we consider selection procedures for Riccati equations with constant coefficients, and in section 3.3.2, we consider the more general case of variable coefficients.

3.3.1 Constant Coefficients

We will consider two subcases separately depending upon the value of the discriminant Δ.

3.3.1.1 The Case with $\Delta < 0$

Consider ℓ such that $y_i' = j(a_i y_i^2 + b_i y_i + c_i)$ where $a_i \neq 0$ and $j = 1$ if i is even and -1 otherwise. The solution to this equation is given by,

$$y_i(x) = \frac{j\sqrt{-\Delta}}{2a_i} \left[\tan\left(\frac{\sqrt{-\Delta}}{2} x + A_i\right) - \frac{jb_i}{\sqrt{-\Delta}}\right] \ldots \qquad (3.3)$$

If we let the initial condition be, $y_i(0) = d_i/e_i$ then we can evaluate the arbitrary constant A_i by substituting the initial condition in equation (3.3). Thus,

$\frac{d_i}{e_i} = j \frac{\sqrt{-\Delta}}{2a_i} (\tan(A_i) - jb_i/\sqrt{-\Delta})$ from which,

$$A_i = j \arctan\left(\frac{2a_i d_i - b_i e_i}{e_i \sqrt{-\Delta}}\right).$$

Substituting in (3.3), we get,

$$y_i(x) = j \frac{\sqrt{-\Delta}}{2a_i} \left[\frac{\tan\left(\frac{\sqrt{-\Delta}}{2} x\right) + j \frac{2a_i d_i + b_i e_i}{e_i \sqrt{-\Delta}}}{1 - j \tan\left(\frac{\sqrt{-\Delta}}{2} x\right) \frac{2a_i d_i + b_i e_i}{e_i \sqrt{-\Delta}}}\right] - \frac{b_i}{2a_i}$$

$$= \frac{j \tan\left(\frac{\sqrt{-\Delta}}{2} x\right)[-e_i \Delta + b_i(2a_i d_i + b_i e_i)] + \sqrt{-\Delta}(2a_i d_i)}{2a_i[e_i \sqrt{-\Delta} - j \tan\left(\frac{\sqrt{-\Delta}}{2} x\right)(2a_i d_i b_i e_i)]}$$

$$= \frac{j r_i u + (\sqrt{-\Delta}) d_i}{(\sqrt{-\Delta}) e_i - j h_i u} \qquad (3.4)$$

where $r_i = 2c_i e_i + b_i d_i$, $h_i = 2a_i d_i + b_i e_i$ and $u = \tan\left(\frac{\sqrt{-\Delta}}{2} x\right)$. It is clear that the process of selection will involve r_i, h_i, d_i and e_i but not a_i, b_i, and c_i. Therefore, if we could obtain recursions for r_i and h_i which are free of a_i, b_i and c_i then we will avoid the computation of a_i, b_i and c_i. The recursions for r_i and h_i can be easily derived [9] and are given by:

$$\left.\begin{aligned} h_{i+1} &= k_{i+1} r_i \\ r_{i+1} &= k_{i+1}(p_{i+1} h_i + q_{i+1} r_i) \\ d_{i+1} &= k_{i+1}(p_{i+1} e_i - q_{i+1} d_i) \\ e_{i+1} &= k_{i+1} d_i \end{aligned}\right\} \qquad (3.5)$$

The condition for the selection of a (p,q) pair is given by: $y_i(x) \in I(p,q)$. In other words, the selection condition is: If $m(p,q) \leq \frac{j r_i u + \sqrt{-\Delta}}{\sqrt{-\Delta} e_i - j h_i u} \leq M(p,q)$ then choose (p,q). Note that we cannot use this condition directly since u is an unknown, therefore, we would like to rewrite the selection condition as follows:

$$\text{arc tan}(ARG_i m(p,q)) \leq \frac{\sqrt{-\Delta} jx}{2} \leq \text{arc tan}(ARG_i M(p,q)) \quad (3.6)$$

where

$$ARG_i(s) = \frac{\sqrt{-\Delta} e_i s - \sqrt{-\Delta} d_i}{r_i + s h_i}$$

Note that such a rewriting is valid if both of the following conditions are satisfied: (1) $ARG_i(s)$ is a monotone-increasing function of s, and (2) arc tan(z) is a monotone-increasing function of z. Since condition (2) is already known to be satisfied, we only have to verify condition (1). To do this, note that,

$$\frac{\partial ARG_i(s)}{\partial s} = \frac{(r_i + h_i s)(\sqrt{-\Delta} e_i) - h_i \sqrt{-\Delta}(e_i s - d_i)}{(r_i + h_i s)^2}$$

$$= \sqrt{-\Delta}(r_i e_i + h_i d_i)/(r_i + h_i s)^2.$$

Now

$$r_{i+1} e_{i+1} + h_{i+1} d_{i+1} = k_{i+1}(p_{i+1}h_i+q_{i+1}r_i) k_i d_i + k_{i+1}^2 r_i(p_{i+1}e_i-q_{i+1}d_i)$$

$$= k_{i+1}^2(p_{i+1}h_i d_i+p_{i+1}r_i e_i)$$

$$= p_{i+1} k_{i+1}^2(r_i e_i+h_i d_i) .$$

Therefore,

$$r_i e_i + h_i d_i = (\prod_{j=1}^{i} (p_j k_j^2))(r_0 e_0+h_0 d_0) .$$

Therefore, $ARG_i(s)$ is a montone-increasing function of s provided $r_0 e_0 + h_0 d_0 > 0$. Observe that there is no loss of generality in assuming that $r_0 e_0 + h_0 d_0 > 0$. Since if $r_0 e_0 + h_0 d_0 < 0$ then $ARG_i(s)$ will be a monotone-decreasing function of s and we can turn the inequality (3.6) around and follow very similar arguments. Also note that the condition $r_0 e_0 + h_0 d_0 = 0$ will not occur, since this implies that either t_0 (the initial condition) is complex or $d_0 = e_0 = 0$ or $a_0 = 0$.

In theory, the selection condition (3.6) can be used to select the (p,q) pair during each iterative step, but the amount of computation involved is clearly excessive. It is, therefore, clear that we would like to use an approximation to arc tan $(ARG_i(s))$ which is "easy" enough to compute from the available coefficients h_i, r_i, d_i, e_i and the known value of s. We note that the use of an approximation in the selection procedure implies the use of redundancy in the digit sets since otherwise we cannot guarantee correct selection. Let us denote the approximate value of arc tan $(ARG_i(s))$ by $AT_i(s)$ and let $z\ell$ and zr have the same meaning as in section 3.1, then the selection rule to be used can be specified by:

$$\text{If } AT_i(z\ell) \le \frac{\sqrt{-\Delta}}{2} j x \le AT_i(zr) \text{ then choose } (p,q) \tag{3.7}$$

In order to guarantee correct selection using condition (3.7), we have to show that the region specified by condition (3.7) is a subset of the region specified by the condition (3.6). From this, we can say that the maximum error allowable in the computation of arc tan $(ARG_i(s))$, denoted by E_i, is given by:

$$E_i \le \text{arc tan}(ARG_i(s_2)) - \text{arc tan}(ARG_i(s_1)) \quad \text{for some } s_1 < s_2$$

such that $s_1, s_2 \in [m,M]$. Now we note that, arc tan(z) satisfies the Lipschitz condition, i.e.,

$$|\text{arc tan}(z_2) - \text{arc tan}(z_1)| \le L|z_2 - z_1|$$

for $L > 0$ and $L \le \Pi$. Therefore,

$$E_i \le L(ARG_i(s_2) - ARG_i(s_1)) . \tag{3.8}$$

Now let,

$$H_i = ARG_i(s_2) - ARG_i(s_1)$$

$$= \frac{\sqrt{-\Delta}(e_i s_2-d_i)}{(r_i+h_i s_2)} - \frac{\sqrt{-\Delta}(e_i s_1-d_i)}{(r_i+h_i s_1)}$$

$$= \frac{(r_i e_i+h_i s_i)(\sqrt{-\Delta})(s_2-s_1)}{(s_1 h_i+r_i)(s_2 h_i+r_i)}$$

Using an expression derived for $r_i e_i + h_i d_i$ earlier, we have

$$H_i = \frac{\sqrt{-\Delta}(s_2-s_1)(\prod_{j=1}^{i} p_j k_j^2)(r_0 e_0+h_0 d_0)}{(s_1 h_i+r_i)(s_2 h_i+r_i)} \tag{3.9}$$

We are now interested in eliminating h_i and r_i from the expression of H_i. Towards this end, we will show that,

$$r_i = r_0 K_i Q_i + h_0 K_i P_i$$

where

$$K_i = \prod_{j=1}^{i} (k_j) .$$

We proceed to prove this result by induction on i. Since $P_0 = 0$, $Q_0 = 1$ and $K_0 = 1$, we have $r_0 = r_0 \cdot 1 \cdot 1 + h_0 \cdot 1 \cdot 0 = r_0$. Now from recursions (3.5), we have,

$$r_1 = k_1(r_0 q_1+h_0 p_1) = r_0 K_1 Q_1 + h_0 K_1 P_1 .$$

Now assume that the required result is true for r_j for $j \le i$. Again from recursions (3.5),

$$r_{i+1} = k_{i+1}(p_{i+1}h_i+q_{i+1}r_i)$$

$$= k_{i+1}(p_{i+1}k_i r_{i-1}+q_{i+1}r_i)$$

$$= k_{i+1}(p_{i+1}k_i(r_0 K_{i-1}Q_{i-1}+h_0 K_{i-1}P + q_{i+1}(r_0 K_i Q_i+h_0 K_i P_i))$$

$$= r_0 K_{i+1}(p_{i+1}Q_{i-1}+q_{i+1}Q_i) + h_0 K_{i+1}(p_{i+1}P_{i-1}+q_{i+1}P_i)$$

$$= r_0 K_{i+1}Q_{i+1} + h_0 K_{i+1}P_{i+1} .$$

Thus, we have the required result. It follows from this that

$$h_i = k_i r_{i-1} = K_i(r_0 Q_{i-1}+h_0 P_{i-1})$$

Now substituting these expressions for h_i and r_i in the equation (3.9), we have,

$$H_i = \frac{(\prod_{j=1}^{i} p_j(K_i^2(r_0 e_0+h_0 d_0)\sqrt{-\Delta}(s_2-s_1)}{K_i^2[s_1(r_0 Q_{i-1}+h_0 P_{i-1})+r_0 Q_i+h_0 P_i]*[s_2(r_0 Q_{i-1}+h_0 P_{i-1})+r_0 Q_i+h_0 P_i]}.$$

Substituting this in the expression (3.8), we have,

$$E_i \le \frac{(\prod_{j=1}^{i} p_j) L(r_0 e_0+h_0 d_0)\sqrt{-\Delta}(s_2-s_1)}{[s_1(r_0 Q_{i-1}+h_0 P_{i-1})+r_0 Q_i+h_0 P_i][s_2(r_0 Q_{i-1}+h_0 P_{i-1})+r_0 Q_i+h_0 P_i]}.$$

Now we consider two cases, depending upon the value of r_0. If $r_0 \neq 0$ then we have,

$$E_i \leq B_1 \frac{(\prod_{j=1}^{i} p_j)}{Q_i\ Q_{i-1}} \tag{3.10}$$

since P_i, Q_i, P_{i-1}, Q_{i-1}, s_1, s_2 are all greater than zero and where

$$B_1 = L\left(\frac{r_0 e_0 + h_0 d_0}{r_0^2}\right)\left(\frac{s_2 - s_1}{s_2}\right)\sqrt{-\Delta} > 0 .$$

On the other hand if $r_0 = 0$

$$E_i \leq \frac{(\prod_{j=1}^{i} p_j)\ L\ h_0\ d_0 \sqrt{-\Delta}(s_2 - s_1)}{h_0^2 (s_1 P_{i-1} + P_i)(s_2 P_{i-1} + P_i)}$$

$$\leq \frac{(\prod_{j=1}^{i} p_j)}{P_i\ P_{i-1}}\left(\frac{s_2 - s_1}{s_2}\right)\sqrt{-\Delta}\ \frac{d_0}{h_0} \tag{3.11}$$

We will now obtain a bound on $P_i\ P_{i-1}$ in terms of $Q_i\ Q_{i-1}$. A well known property of the convergents of an infinite continued fraction, f, can be written as [1]:

$$\frac{P_0}{Q_0} \leq \frac{P_2}{Q_2} \leq \cdots \leq f \leq \cdots \leq \frac{P_3}{Q_3} \leq \frac{P_1}{Q_1} .$$

Therefore, if i is odd, $\frac{P_i}{Q_i} \geq m$. If $i \geq 2$ is even,

$$\frac{P_i}{Q_i} \geq \frac{P_2}{Q_2} \geq \frac{p_{min}}{q_{max} + \frac{p_{max}}{q_{min}}} .$$ Therefore,

$$\frac{P_i}{Q_i} \cdot \frac{P_{i-1}}{Q_{i-1}} \geq \frac{m\ p_{min}}{q_{max} + \frac{p_{max}}{q_{min}}}$$

Substituting this in (3.11) we have,

$$E_i \leq \frac{(\prod_{j=1}^{i} p_j)}{Q_i\ Q_{i-1}} \cdot B_2 \tag{3.12}$$

where $B_2 = \frac{s_2 - s_1}{s_2} \cdot \sqrt{-\Delta} \cdot \frac{d_0}{h_0} \cdot \frac{m\ p_{min}}{q_{max} + \frac{p_{max}}{q_{min}}}$. From (3.11) and (3.12), we have,

$$E_i \leq B \frac{\prod_{j=1}^{i} p_j}{Q_i Q_{i-1}}$$

where $B = B_1$ if $r_0 \neq 0$ and B_2 otherwise. Note that B is a fixed, finite and bounded constant independent of the value of i. The factor $(\prod_{j=1}^{i} p_j)/Q_i\ Q_{i-1}$ can be interpreted as the error in the solution, since it equals the difference in values of the successive convergents P_{i-1}/Q_{i-1} and P_i/Q_i [1]. Therefore, if we demand linear convergence then we must have,

$$\frac{\prod_{j=1}^{i} p_j}{Q_i Q_{i-1}} = \text{constant} \cdot \alpha^{-i}$$

for a small positive constant and some $\alpha > 1$. As a result, we have,

$$E_i \leq B' \cdot \alpha^{-i} .$$

But this implies that the computation of arc tan $(ARG_i(s))$ must be carried out to nearly the same precision as that of the desired precision of the function being evaluated. Thus we conclude that we cannot obtain a computationally simple selection procedure for the functions that can be evaluated using the Riccati equation with constant coefficients and $\Delta < 0$.

3.3.1.2 The Case with $\Delta > 0$

Consider the following Riccati equation:

$$y_i' = j(a_i y_i^2 + b_i y_i + c_i)$$

such that $\Delta = \Delta_i > 0$ and $j = 1$ if i is even and -1 otherwise. The solution to this equation can be written as,

$$y_i(x) = \frac{\sqrt{\Delta}}{2a_i} \coth\left(\frac{jx\sqrt{\Delta}}{2} + A_i\right) - \frac{b_i}{2a_i} \tag{3.13}$$

where A_i is an arbitrary constant of integration. Using the initial condition $y_i(0) = t_i = d_i/e_i$, we obtain $\tanh A_i = -\frac{\sqrt{\Delta} e_i}{2a_i d_i + b_i e_i}$. For the sake of brevity, we let $h_i = 2a_i\ d_i + b_i\ e_i$ and after substituting for A_i in (3.13), we get,

$$y_i(x) = \frac{1}{2a_i}\left\{\frac{j\Delta e_i \tanh(\frac{\sqrt{\Delta}x}{2}) - jb_i h_i \tanh(\frac{\sqrt{\Delta}x}{2}) - \sqrt{\Delta}2a_i d_i}{jh_i \tanh(\frac{\sqrt{\Delta}x}{2}) - \sqrt{\Delta}e_i}\right\}$$

From which, we get,

$$j\ \tanh\left(\frac{\sqrt{\Delta}x}{2}\right) = \frac{\sqrt{\Delta}(y_i e_i - d_i)}{(y_i h_i + r_i)} \tag{3.14}$$

where $r_i = b_i\ d_i + 2c_i\ e_i$. From equation (3.13), we note that if $e_0 = 1$, $d_0 = 0$, $h_0 = 0$ and $r_0 = \sqrt{\Delta}$ then $y_0(x) = \tan h(\frac{\sqrt{\Delta}x}{2})$. If $e_0 = 0$, $d_0 = 1$, $h_0 = -\sqrt{\Delta}$ and $r_0 = 0$ then $y_0(x) = \coth(\frac{\sqrt{\Delta}x}{2})$. If $c_0 = 0$ and $a_0 = 0$ then we have $y_0(x) = A_0 e^{b_0 x}$.

From the form of the equation (3.14) and the definitions of r_i and h_i, it is clear that we can follow the same arguments as in section 3.3.1.1 and prove that a computationally simple selection procedure cannot be obtained in the case that $\Delta > 0$ or $a_0 = 0$. Thus we have shown the negative results for the Riccati equation with constant coefficients, i.e., for the subset L_0 of L.

3.3.2 Variable Coefficients

We will only consider the case with $\bar{\Delta}_0 = 0$, i.e., we consider the subset L_{10} of L. Consider the equation

$$y_i' = j\, k(x)(a_i y_i + b_i)^2 \qquad (3.15)$$

where j = 1 if i is even and -1 otherwise. Let g(x) = k(x) dx. We will assume that g(0) = 0, the function $g^{-1}(z)$ exists, is a monotone in z and is Lipschitz continuous with a "small" value of the Lipschitz constant L. We will use the following set of recursions for a_i, b_i, d_i and e_i:

$$\begin{aligned} a_{i+1} &= b_i/\sqrt{p_{i+1}} \\ b_{i+1} &= a_i\sqrt{p_{i+1}} + b_i\, q_{i+1}/\sqrt{p_{i+1}} \\ d_{i+1} &= e_i\sqrt{p_{i+1}} - d_i\, q_{i+1}/\sqrt{p_{i+1}} \\ e_{i+1} &= d_i/\sqrt{p_{i+1}} \end{aligned} \qquad (3.16)$$

The solution to this equation is given by:

$$y_i(x) = \frac{d_i + j(g(x)-g(0))\, b_i(a_i b_i + d_i e_i)}{e_i - j(g(x)-g(0))\, a_i(a_i b_i + d_i e_i)} \,. \qquad (3.17)$$

To simplify the equation (3.17), we can easily prove by induction on i, that

$$a_i\, b_i + d_i\, e_i = a_0\, b_0 + d_0\, e_0 \triangleq r_0 \,.$$

Note that (using the recursions 3.16),

$$\begin{aligned} & a_{i+1}\, b_{i+1} + d_{i+1}\, e_{i+1} \\ &= b_i/\sqrt{p_{i+1}}(e_i\sqrt{p_{i+1}} - d_i q_{i+1}/\sqrt{p_{i+1}}) + \\ &\quad (a_i\sqrt{p_{i+1}} + b_i q_{i+1}/\sqrt{p_{i+1}})\, d_i/\sqrt{p_{i+1}} \\ &= a_i\, d_i + b_i\, e_i \,. \end{aligned}$$

Using this, we get

$$y_i(x) = \frac{d_i + j(g(x)-g(0))\, b_i\, r_0}{e_i + j(g(x)-g(0))\, a_i\, r_0} \,.$$

The selection condition can now be written as: If

$$m(p,q) \le \frac{d_i + j(g(x)-g(0))b_i r_0}{e_i + j(g(x)-g(0))a_i r_0} \le M(p,q) \text{ then choose } (p,q).$$

Since g(x) is the unknown we want to transform the selection condition to:

$$g^{-1}\left(\frac{j\, M(p,q)e_i - d_i + j\, M(p,q)g(0)a_i r_0}{b_i r_0 + M(p,q)a_i r_0}\right) \le x$$

$$\le g^{-1}\left(\frac{j\, m(p,q)e_i - d_i + j\, m(p,q)g(0)a_i r_0}{r_0(b_i + m(p,q)a_i)}\right) \qquad (3.18)$$

But this transformation is valid provided, $ARG_i(s)$ is a montone-increasing function of s and $g^{-1}(z)$ is a monotone-increasing function of z. Note that,

$$ARG_i(s) = \frac{s\, e_i - d_i + j\, s\, g(0)\, a_i\, r_0}{r_0(b_i + s a_i)} \,.$$

Therefore,

$$\frac{\partial ARG_i(s)}{\partial s} = \frac{r_0(b_i + s a_i)(e_i + jg(0)a_i r_0) - (s e_i - d_i + jsg(0)a_i r_0) r_0 a_i}{r_0^2(b_i + s a_i)^2}$$

$$= \frac{1 + j\, g(0)\, a_i\, b_i}{(b_i + s a_i)^2}$$

Since by assumption, g(0) = 0 then $ARG_i(s)$ is a monotone-increasing function of s. Observe that there is no loss of generality in assuming that $g^{-1}(z)$ is a monotone-increasing function of z. Since if it is monotone-decreasing then we can turn the inequality (3.18) around and follow very similar arguments.

The inequality (3.18) can be split up into two parts depending upon the value of i. We will only consider the case when i is even, the other case being very similar. Then the selection condition is:

$$g^{-1}(ARG_i\; m(p,q)) \le x \le g^{-1}(ARG_i\; M(p,q)) \,.$$

Now since $g^{-1}(ARG_i(s))$ is difficult to compute in general, therefore, we would like to use an approximation. The maximum error allowable in such an approximation can be written as,

$$E_i = g^{-1}(ARG_i(s_2)) - g^{-1}(ARG_i(s_1))$$

where $m \le s_1 < s_2 \le M$. Since by assumption, g^{-1} satisfies the Lipschitz condition with the Lipschitz constant L. Then

$$E_i \le L[ARG_i(s_2) - ARG_i(s_1)] \qquad (3.19)$$

Let

$$\begin{aligned} H_i &= ARG_i(s_2) - ARG_i(s_1) \\ &= \frac{s_2\, e_i - d_i}{r_0(b_i + s_2 a_i)} - \frac{s_1\, e_i - d_i}{r_0(b_i + s_1 a_i)} \\ &= \frac{s_2 - s_1}{(b_i + s_2 a_i)(b_i + s_1 a_i)} \end{aligned}$$

From this point onwards, we can follow a procedure similar to section 3.3.1 to obtain a similar negative result.

4. Conclusion and Further Remarks

Recently, there has been some interest in the use of continued fractions for digital hardware calculations. We require that the coefficients of the continued fractions be integral powers of two. As a result well known continued fraction expansions of functions cannot be used. We have presented methods of expansion of a large number of functions into continued fractions.

Selection of coefficients of the continued fractions is, however, a difficult problem. We have shown that the selection problem can be solved for the solution of a quadratic and higher degree polynomial equations. However, this is the only class of problems for which the selection problem has been solved. We have shown that for most of the remaining functions discussed in this paper no simple selection procedure can be found. Expressions to prove this claim were derived in section 3. We now outline an intuitive but less rigorous argument to explain this behavior.

We have seen that while evaluating a function f, the selection procedure involves the computation of the inverse function f^{-1}. Since the computation of f^{-1} is generally as complex as the computation of f, we require that an approximation of f^{-1} be used in the selection procedure. Thus the whole process of evaluating f may be looked upon as an attempt to obtain a good approximation to f given a crude approximation of f^{-1}. Let us split the coefficient vector $\underline{a}_i$ into two vectors so that $\underline{a}_i = (\underline{\alpha}_i, \underline{\beta})$. Thus the vector $\underline{\alpha}_i$ consists of all the coefficients which vary with index i and $\underline{\beta}$ consists of invariant coefficients. As an example, in the case of the quadratic, $\underline{\alpha}_i = (b_i, c_i)$ and $\underline{\beta}$ is null. As another example, for the Riccati-approach, $\underline{\alpha}_i = (a_i, b_i, c_i, d_i, e_i)$ and $\underline{\beta} = (x)$. We say that the initial coefficient vector $\underline{\alpha}_0$ together with the system of recursions $\underline{G}$ determine the function to be evaluated and $\underline{\beta}$ is the vector of true arguments for which the function is to be evaluated. Note that $\underline{\beta}$ will play a role in the selection procedure. Since we have assumed that an approximation to f^{-1} is used in the selection procedure, we can find two values of $\underline{\beta}$, namely $\underline{\beta}_1$ and $\underline{\beta}_2$, such that $\underline{\beta}_1 \neq \underline{\beta}_2$ but the corresponding approximation of f^{-1} yields the same value. Note that since $(\underline{\alpha}_0)_1 = (\underline{\alpha}_0)_2$, we have that $(p_1, q_1)_1 = (p_1, q_1)_2$. With this condition we can prove by induction that $(\underline{\alpha}_i)_1 = (\underline{\alpha}_i)_2$ and $(p_i, q_i)_1 = (p_i, q_i)_2$ for all i. Therefore, $f(\underline{\alpha}_0, \underline{\beta}_1) = f(\underline{\alpha}_0, \underline{\beta}_2)$. Thus f is not able to resolve $\underline{\beta}$ values if the approximation to f^{-1} is not able to resolve the same $\underline{\beta}$ values. It is therefore clear that for our procedure to work, we must require that the $\underline{\beta}$ vector be null. Indeed, in the case of the solution to polynomial equations $\underline{\beta}$ vector is null. In the Riccati-Approach, $\underline{\beta}$ vector is always nonull. In the unmodified expansion of $\log_{b_{-1}} b_0$, $\underline{\alpha}_i = (b_i, b_{i-1})$ and $\underline{\beta}$ is null. But since the system $\underline{G}$ was not simple, we applied a transformation. As a result, we had $\underline{\alpha}_i = (P_i, Q_i)$ and $\underline{\beta} = (b_0, b_{-1})$ thus making the problem unsolvable.

Finally, we conjecture that the solution of a polynomial equation (which includes the quadratic) is the only problem that can be solved in our formulation.

Acknowledgment

The author wishes to thank Professor James E. Robertson for introducing him to the subject and for continued support and encouragement. Thanks are also due to Mrs. June Wingler for typing this paper.

References

[1] Wall, H. S., Analytic Theory of Continued Fractions, Van Nostrand, New Jersey, 1950.

[2] Robertson, J. E. and K. S. Trivedi, "The Status of Investigations into Computer Hardware Design Based on the use of Continued Fractions," IEEE Transactions on Computers, Vol. C-22, No. 6, June, 1973, pp. 555-560.

[3] Wynn, P., "On Some Recent Developments in the Theory and Application of Continued Fractions," Journal SIAM on Numerical Analysis, Vol. 1, 1964, pp. 177-197.

[4] Bracha, A., "A Method for Solving Polynomial Equations by Continued Fractions," IEEE Transactions on Computers, Vol. C-23, No. 10, October, 1974, p. 1093.

[5] Lyusternik, L. A., et al., Handbook for Computing Elementary Functions, Permagon Press, New York, 1965.

[6] Trivedi, K. S., "On a Negative Result Regarding the Use of Continued Fractions for Digital Computer Arithmetic," University of Illinois, Department of Computer Science Report UIUCDCS-R-75-693, January, 1975.

[7] Trivedi, K. S., "The Use of Riccati Equation in Digital Computer Arithmetic," University of Illinois, Department of Computer Science Report UIUCDCS-R-74-674, August, 1974.

[8] Trivedi, K. S., "An Algorithm for the Solution of a Quadratic Equation using Continued Fractions," M.S. Thesis, University of Illinois, Urbana, June, 1972; also Department of Computer Science Report 525.

[9] Trivedi, K. S., "Further Negative Results Regarding the Use of Continued Fractions for Digital Computer Arithmetic," University of Illinois, Department of Computer Science Report UIUCDCS-R-75-721, May, 1975.