Padé Approximation to the Solution of the Ricatti Equation

By Wyman Fair

Introduction. In this paper we employ the $\tau$-method, see [1], [2] and [3], to obtain the main diagonal Padé approximations to the solution of the Ricatti differential equation whose coefficients are rational. The results are applicable to first order linear differential equations. This approach is quite different than the conventional utilization of the linear fractional transformation, see [4] and [5], to develop a continued fraction representation of the solution of the Ricatti equation. There are certain advantages in each approach.

Khovanskii [4] has studied the Ricatti equation whose coefficients are at most quadratics in the independent variable. In certain instances his approximations are of the Padé type. In this event, our results are much more general. Merkes and Scott [5] develop continued fraction solutions to a much wider class of Ricatti equations than are treated here. However, no particular attention is devoted to the Padé approximations and the results are not applicable to first order linear differential equations.

In Section I we develop recurrence relations which determine the main diagonal Padé approximations to the Ricatti equation with polynomial coefficients. Section II entails a discussion of convergence of the approximations developed in Section I. In Section III we give some important examples and applications of the theory.

I. Padé Approximation to the Solution of the Ricatti Equation. Here we develop a method of obtaining recursively the main diagonal Padé approximants to the solution of

$$Py' + Qy + Ry^2 + S = 0, \quad y_0 = y(0),$$

and further that

$$y = \sum_{k=0}^{\infty} c_k x^k, \quad c_0 = y_0,$$

and further that

$$d_m = \begin{vmatrix} c_0 & c_1 & \cdots & c_m \\ c_1 & c_2 & \cdots & c_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ c_m & c_{m+1} & \cdots & c_{2m} \end{vmatrix} \neq 0, \quad m = 0, 1, 2, \cdots.$$
In this case $y$ possesses a continued fraction expansion [6] of the form

$$y = y_0 - \frac{\alpha_1 x}{1 + \beta_1 x + \frac{\alpha_2 x^2}{1 + \beta_2 x + \frac{\alpha_3 x^3}{1 + \beta_3 x + \ddots}}},$$

(1.4)

and the $n$th approximant of (1.4) is the $n$th order main diagonal Padé approximation to $y$.

In keeping with the $\tau$-method philosophy, we append a term to the right-hand side of (1.1) and consider the related equation

$$P y_n' + Q y_n + R y_n^2 + S = \frac{T_n}{B_n^2} = \frac{x^{2n}}{B_n^2} \sum_{k=0}^{m} \tau_{n,k} x^k,$$

(1.5)

$$m = \max \{p - 1, q, r, s\}, \quad y_n = \frac{A_n}{B_n}, \quad A_n = \sum_{k=0}^{n} a_{n,k} x^k,$$

$$B_n = \sum_{k=0}^{n} b_{n,k} x^k \quad \text{and} \quad y_0 = \frac{A_0}{B_0} = \frac{y_0}{1}.$$

That $y_n$ is the main diagonal Padé approximation to $y$ can be shown by using analysis similar to that of Wall [6, p. 412–413]. Since $y_n$ is the main diagonal Padé approximation to $y$, we have from the theory of continued fractions, $A_n$ and $B_n$ both satisfy

$$A_n = \rho_n A_{n-1} + \alpha_n x^2 A_{n-2}, \quad \rho_n = 1 + \beta_n x.$$  

(1.6)

For convenience let

$$\alpha_{k,j} = \alpha_k \alpha_{k-1} \cdots \alpha_j,$$

(1.7)

$$\alpha_{k,k} = \alpha_k, \quad \alpha_{k-1,k} = 1 \text{ and}$$

$$\alpha_{k,j} = 0 \text{ for } k < j - 1.$$

Then

$$\Delta_n = B_n A_{n-1} - A_n B_{n-1} = -\alpha_n x^2 \Delta_{n-1} = (-)^{n-1} \alpha_n x^{2n-1} \Delta_{n-1}$$

(1.8)

and

$$\Delta_1 = \alpha_1 x.$$

After multiplying both sides of (1.5) by $B_n^2$ and repeatedly applying (1.6) and (1.8) to the resulting equation, we get

$$\rho_n^2 T_{n-1} + \alpha_n x^4 T_{n-2} + \alpha_n \Delta_{n-1} x P$$

(1.9)

$$+ 2 \alpha_n \rho_n x^2 \left( \rho_{n-1} T_{n-2} + \sum_{k=1}^{n-3} \alpha_{n-1,n-k} \rho_{n-k-1} x^{2k} T_{n-k-2} \right) + \alpha_{n,2} \rho_n x^{2n-2} U = T_n,$$

(1.10)

$$U = (A_0 B_1 + A_1 B_0) Q + 2 A_0 A_1 R + 2 B_1 S = \sum_{k=0}^{m+1} u_k x^k.$$
Equating coefficients of powers of $x$ in (1.9) we get the system of equations which determine directly the values $\alpha_n$, $\beta_n$ and $\tau_{n,k}$. We have

\begin{equation}
\alpha_n = -\tau_{n-1,0} \left\{ (-)^{n} \alpha_{n-1,1} p_0 + \alpha_{n-1,2} u_0 + 2 \sum_{j=3}^{n} \alpha_{n-1,j} \tau_{j-2,0} \right\},
\end{equation}

\begin{equation}
\beta_n = \frac{1}{\tau_{n-1,0} + (-)^{n-1} \alpha_{n,1} p_0} \left\{ \tau_{n-1,1} + (-)^{n} \alpha_{n,1} p_1 + \alpha_{n,2} u_1 \right\}
\end{equation}

\begin{equation}
+ 2 \sum_{j=3}^{n} \alpha_{n,j} \left( \tau_{j-2,1} + \beta_{j-1} \tau_{j-2,0} \right),
\end{equation}

and

\begin{equation}
\tau_{n,k} = \tau_{n-1,k+2} + 2\beta_n \tau_{n-1,k+1} + \alpha_n \tau_{n-2,k} + \beta_n \tau_{n-1,k}
\end{equation}

\begin{equation}
+ \alpha_n \left\{ (-)^{n} \alpha_{n-1,1} p_{k+2} + \alpha_{n-1,2} u_{k+2} + \alpha_{n-1,3} \beta_n u_{k+1} \right\}
\end{equation}

\begin{equation}
+ 2 \sum_{j=3}^{n} \alpha_{n-1,j} \tau_{j-2,k+2} + (\beta_n + \beta_{j-1}) \tau_{j-2,k+1} + \beta_n \beta_{j-1} \tau_{j-2,k} \right\}
\end{equation}

\begin{equation}
k = (0, 1, 2, \ldots, m), \quad n = (2, 3, 4, \ldots).
\end{equation}

The starting values for computation are

\begin{equation}
A_0 = y_0, \quad B_0 = 1,
\end{equation}

\begin{equation}
A_1 = y_0 + a_{1,1} x, \quad B_1 = 1 + b_{1,1} x, \quad a_{1,1} = \frac{c_2^2 - c_2 y_0}{c_1} \quad \text{and} \quad b_{1,1} = -\frac{c_2}{c_1},
\end{equation}

\begin{equation}
A_k = (y_0 b_{1,k} - a_{1,1}), \quad B_k = b_{1,k},
\end{equation}

\begin{equation}
\tau_{0,k} = y_0 q_k + y_0^2 r_k + s_k
\end{equation}

\begin{equation}
\tau_{1,k} = -a_1 p_{k+2} + y_0 q_{k+2} + y_0^2 r_{k+2} + s_{k+2} + (a_{1,1} + b_{1,1} y_0) q_{k+1}
\end{equation}

\begin{equation}
+ 2y_0 a_{1,1} r_{k+1} + 2b_{1,1} s_{k+1} + a_{1,1} b_{1,1} q_k + a_{1,1}^2 r_k + b_{1,1}^2 s_k
\end{equation}

\begin{equation}
k = 0, 1, 2, \ldots, m;
\end{equation}

and from (1.10) the values of $u_k$ are

\begin{equation}
u_k = 2y_0 q_k + 2y_0^2 r_k + 2s_k + (a_{1,1} + y_0 b_{1,1}) q_{k-1} + 2y_0 a_{1,1} r_{k-1} + 2b_{1,1} s_{k-1}
\end{equation}

\begin{equation}
k = 0, 1, 2, \ldots, m + 1.
\end{equation}

Theoretically, we can eliminate the $\tau_{n,k}$'s from the equations (1.11), (1.12) and (1.13) and obtain $\alpha_n$ and $\beta_n$ in terms of their previous values. In general, it does not seem possible to obtain closed form expressions for $\alpha_n$ and $\beta_n$. In Section III, however, we give some important examples for which this can be done.

It should be mentioned that this method of obtaining Padé approximations to functions can be applied to first order linear differential equations with polynomial coefficients by setting $R = 0$ in (1.5). This is pertinent since many important transcendental functions can be defined by an equation of this type, see [1] and [7].
II. Convergence of the Padé Approximation. Let
\[ Y(x) = \lim_{n \to \infty} y_n(x), \]
where \( y_n(x) \) is defined in (1.5). The convergence of the continued fraction (1.4) insures the convergence of \( Y(x) \). The criteria given here for the convergence of (1.4) follow from Wall [6, p. 42, 109 and 110].

Case 1.
\[ \lim_{n \to \infty} \alpha_n = 0. \]
The continued fraction (1.4) converges for \( x \) in any finite closed region which omits poles of (1.4).

Case 2.
\[ \lim_{n \to \infty} \alpha_n \neq 0. \]
We assume that the \( \alpha_n \) and \( \beta_n \) are real and that (1.4) is eventually positive definite, i.e., there is an \( N \) such that for \( n \geq N, \alpha_n < 0 \). The continued fraction (1.4) converges, except at isolated poles, for \( \text{Im}(x) \neq 0 \) if either
\[ \sum_{k=N}^{\infty} |\alpha_k|^{-1/2} = \infty, \]
or
\[ \sum_{k=N}^{\infty} |\beta_{k+1}| |\alpha_k\alpha_{k+1}|^{-1/2} = \infty. \]
Further, if
\[ \lim_{n \to \infty} \beta_n \neq 0 \quad \text{and} \quad \lim_{n \to \infty} \frac{|\alpha_n|}{\beta_n^2} \leq \frac{1}{4}, \]
then (1.4) converges in a neighborhood of the origin and any finite closed domain which omits poles of (1.4) and which excludes the negative real axis.

It should be noted that if (1.4) converges and \( x_j^{(n)}, j = 1, 2, \ldots, n \) are the zeros of \( B_n(x) \) then \( \lim_{n \to \infty} x_j^{(n)} = x_j \) where \( x_j \) is a pole of \( y(x) \). This offers a constructive method of obtaining the poles (and zeros) of \( y(x) \). See the third example in Section III.

III. Examples and Applications. Here we consider some important special cases of (1.1) and, for a number of these, deduce closed form expressions for \( \alpha_n \) and \( \beta_n \) defined in (1.6).

1. Let \( u = \Phi(a; b; x) \) where \( \Phi(a; b; x) \) is the confluent hypergeometric function, see [8, p. 248]. Then \( y = u'/u \) satisfies the Ricatti equation
\[ xy' + (b - x)y + xy^2 - a = 0, \quad y_0 = y(0) = \frac{a}{b}. \]

Following the development in Section I, we find:
\[ A_0 = y_0, \quad B_0 = 1, \]
\[ A_1 = y_0 + a_{1,1} x, \quad B_1 = 1 + b_{1,1} x, \]
\[ a_{1,1} = \frac{a(a + 1)}{b(b + 1)(b + 2)}, \quad b_{1,1} = \frac{2a - b}{b(b + 2)}, \]
\[ \alpha_1 = \frac{a(a - b)}{b^2(b + 1)}, \quad \beta_1 = b_{1,1}, \]
\[ \tau_{n,0} = 0, \quad n = 0, 1, 2, \ldots, \]
\[ \tau_{0,1} = (b + 1)\alpha_1 , \quad \tau_{1,1} = \frac{a(a + 1)(a - b)(a - b - 1)}{b^2(b + 1)^2(b + 2)^2}, \]
\[ u_0 = 0, \quad u_1 = (b + 2)\alpha_1 , \quad u_2 = b\alpha_1 \beta_1 , \]
and the equations defining \( \alpha_n, \beta_n \) and \( \tau_{n,1} \) are
\[ \alpha_n = -\tau_{n-1,1} \left\{ (\cdot)^n \alpha_{n-1,1} + \alpha_{n-1,2} u_1 + 2 \sum_{j=3}^{n} \alpha_{n-1,j} \tau_{j-2,1} \right\}, \]
\[ \beta_n = -\left\{ \alpha_{n,2} u_2 + 2 \sum_{j=3}^{n} \alpha_{n,j} \beta_{j-1} \tau_{j-2,1} \right\} \left\{ 2\tau_{n-1,1} + \alpha_{n,2} u_1 + 2 \sum_{j=3}^{n} \alpha_{n,j} \tau_{j-2,1} \right\}, \]
\[ \tau_{n,1} = \alpha_n \tau_{n-2,1} + \beta_n^2 \tau_{n-1,1} + \alpha_n \beta_n \tau_{n-1,1} + 2\beta_n \sum_{j=3}^{n} \alpha_{n,j} \beta_{j-1} \tau_{j-2,1} . \]
We can eliminate \( \tau_{n,1} \) from the equations (3.3) and show by mathematical induction that
\[ \alpha_n = \frac{(a + n)(b - a + n)}{(b + 2n - 1)(b + 2n + 1)(b + 2n + 1)}, \]
\[ \beta_n = \frac{2a - b}{(b + 2n - 2)(b + 2n)}, \]
and the approximants converge to a function which is convergent except possibly at isolated poles.

2. Let \( u = J_v(z) \) where \( J_v(z) \) is the Bessel function of the first kind. Set \( y = zu/u \), and \( x = z^2 \), then \( y(x) \) satisfies
\[ 2xy' + y^2 + x - v^2 = 0, \quad y_0 = y(0) = v. \]
The development is similar to the preceding examples and we list the results:
\[ \alpha_1 = \frac{1}{2(1 + v)} \quad \text{and} \quad \beta_1 = \frac{-1}{4(1 + v)(2 + v)}, \]
After eliminating \( \tau_{n,1} \) from the equations defining \( \alpha_n \) and \( \beta_n \), we can show by mathematical induction that
\[ \alpha_n = \frac{-1}{16(2n - 3 + v)(2n - 2 + v)(2n - 1 + v)} \quad \text{and} \]
\[ \beta_n = \frac{-1}{2(2n + v)(2n - 2 + v)}. \]
Again we have convergence in any finite closed domain which excludes the poles of $Y(x)$. Since $I_n(x) = e^{-i\pi/2}J_n(ze^{i\pi/2})$ where $I_n(z)$ is the modified Bessel function of the first kind, we also have the Padé approximations to $zI_n'(z)/I_n(z)$.

3. Let $u = K_n(1/x)$ where $K_n(x)$ is the Bessel function of the second kind. Then $y = u'(x)/u(x)$ satisfies the equation

$$-x^2y' + xy + y^2 - (1 + \nu^2x^2) = 0, \quad y_0 = y(0) = -1.$$  

We have

$$A_0 = -1, \quad B_0 = 1$$

$$A_1 = -1 + a_{1,1}x, \quad B_1 = 1 + b_{1,1}x,$$

$$a_{1,1} = \nu^2 - \frac{3}{2}, \quad b_{1,1} = \frac{1}{2} - \nu^2,$$

$$\alpha_1 = \frac{1}{2}, \quad \beta_1 = \frac{1}{4} - \nu^2,$$

$$\tau_{n,0} = 0, \quad n = 1, 2, \ldots,$$

$$\tau_{n,1} = -1, \quad \tau_{1,1} = -(\frac{1}{4} + \nu^2)\beta_1,$$

$$\tau_{n,2} = -\nu^2, \quad \tau_{1,2} = -\nu^2\beta_1^2,$$

$$u_0 = 0, \quad u_1 = u_2 = -1, \quad \text{and} \quad u_3 = -2\nu^2\beta_1.$$  

The equations defining $\alpha_n$, $\beta_n$ and $\tau_{n,j}$ are

$$\alpha_n = -\tau_{n-1,1} \left\{-\alpha_{n-1,2} + 2 \sum_{j=3}^{n} \alpha_{n-1,j} \tau_{j-2,1} \right\},$$

$$\beta_n = -\frac{1}{\tau_{n-1,1}} \left\{\tau_{n-2,2} - \alpha_{n,2}[(-)^n \alpha_1 + 1] + 2 \sum_{j=3}^{n} \alpha_{n,j} \tau_{j-2,2} + \beta_j \tau_{j-2,2} \right\},$$

$$\tau_{n,1} = 2\beta_n \tau_{n-1,2} + \alpha_n \tau_{n-2,1} + \beta_n \tau_{n-1,1} + \alpha_n \tau_{n-1,2} - \alpha_n \beta_n$$

$$+ 2 \sum_{j=3}^{n} \alpha_{n,j} \left[\beta_n + \beta_{j-1} + \beta_n \beta_{j-1} \tau_{j-2,2} + \beta_n \beta_{j-1} \tau_{j-2,2}\right],$$

$$\tau_{n,2} = \alpha_n \tau_{n-2,2} + \beta_n \tau_{n-1,2} + \alpha_n \beta_n \tau_{n-1,2} + 2\beta_n \sum_{j=3}^{n} \alpha_{n,j} \beta_{j-1} \tau_{j-2,2}.$$  

We can combine the expressions in (3.10) to get

$$\alpha_n = \alpha_{n-2} + \beta_{n-1}(\beta_{n-1} - n + 1) - \beta_{n-2}(\beta_{n-2} - n + 3)$$

and

$$\beta_n = n - \frac{1}{2} - \beta_{n-1} + \frac{\alpha_{n-1}}{\alpha_n} \left\{\beta_{n-1} + \beta_{n-2} - n + \frac{5}{2} - \frac{\beta_{n-1}^2}{2\alpha_{n-1}}\right\}, \quad n = 4, 5, \ldots.$$  

For $\nu = 0$ these reduce to

$$\alpha_n = -\frac{n^2}{4} + \frac{n}{2} - \frac{3}{16}$$

and

$$\beta_n = n - \frac{1}{2},$$

so that for $\nu = 0$ we have convergence for $x$ not on the negative real axis, and in a
neighborhood of the origin. Convergence for \( v \) arbitrary has not yet been shown. However, we do have

\[
\frac{K'_v(z)}{K_v(z)} = -\frac{v}{z} - \frac{z}{(v - 1) - \frac{z}{K'_{v-1}(z)}},
\]

so that if we have a convergent expression for a given \( v \), we have a convergent expansion for \( v + 1 \), although the latter is not the Padé.

For an important application of the Padé approximations to \( K'_v(x)/K_v(x) \), we consider the inversion of the Laplace integral

\[
I = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{K'_v(p)}{pK'_v(p)} e^{pt} \, dp,
\]

which is discussed in [9] and [10] in connection with the problem of supersonic flow past quasi-cylindrical bodies of almost circular cross section.

In the latter reference an approximation for \( I \) is achieved by using rational approximations for \( K_v(p) \). These rational approximations are not of the Padé type.

One method of evaluating \( I \) in (3.16) is by applying the calculus of residues which requires a knowledge of the zeros of \( K'_v(x) \). Here the Padé approximations to \( K'_v(x)/K_v(x) \) can be used to great advantage. A table is included which compares the approximate zeros obtained in this way to the zeros of \( K'_v(x) \) given in [9].

**Tables of Zeros of \( K'_v(x) \) and \( A_n(1/x) \)**

<table>
<thead>
<tr>
<th>True Zeros</th>
<th>Approximate Zeros</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n = 6 )</td>
</tr>
</tbody>
</table>

\[ \nu = 1 \]

\[
-0.64355 \pm i0.50118 \quad -0.62888 \pm i0.50755 \quad -0.64082 \pm i0.49114 \quad -0.64886 \pm i0.49965
\]

\[ \nu = 2 \]

\[
-0.83455 \pm i1.43444 \quad -0.83479 \pm i1.43502 \quad -0.83455 \pm i1.43437 \quad -0.83457 \pm i1.43442
\]

\[ \nu = 6 \]

\[
-1.23832 \pm i5.23662 \quad -1.23832 \pm i5.23631 \quad -1.23832 \pm i5.23606 \quad -1.23832 \pm i5.23661
\]

\[
-3.10823 \pm i3.10944 \quad -3.11913 \pm i3.18820 \quad -3.10747 \pm i3.10875 \quad -3.10798 \pm i3.10908
\]

\[
-3.83045 \pm i1.31040 \quad -3.26840 \pm i1.33025 \quad -3.88014 \pm i1.26068 \quad -3.87739 \pm i1.29243
\]

\[ \nu = 10 \]

\[
-1.47973 \pm i9.10691 \quad -1.50168 \pm i9.10913 \quad -1.47946 \pm i9.10680 \quad -1.47975 \pm i9.10692
\]

\[
-4.01755 \pm 6.76252 \quad -3.09425 \pm 7.31312 \quad -4.17663 \pm 6.86179 \quad -4.01704 \pm 6.76220
\]

\[
-5.34531 \pm 4.85738 \quad -2.13322 \pm 3.14663 \quad -4.26387 \pm 3.29352 \quad -5.38090 \pm 4.84802
\]

\[
-6.13751 \pm 3.05017 \quad -3.70852 \pm 1.99284 \quad -6.01367 \pm 3.38991 \quad -5.75923 \pm 1.24091
\]

4. We now apply the results of Section I to a well-known first order linear differential equation which was treated from different points of view by Luke [1] and Laguerre [7], and obtain some results found in these papers.
Let \( y = \, _2F_1(1, b + 1; c + 1; x) \). Here, \( _2F_1(a, b; c; x) \) is the hypergeometric function, see [8, Chapter 2]. The equation satisfied by \( y \) is

\[
(3.17) \quad x(1 - x) y' + [c - (b + 1)x] y - c = 0, \quad y_0 = y(0) = 1.
\]

and after eliminating \( \tau_{n,1} \) from the equations defining \( \alpha_n \) and \( \beta_n \) we can show by induction that

\[
\begin{align*}
\alpha_n &= -\frac{(n - 1)(n + c - b - 2)(n + c - 1)(n + b)}{(2n + c - 3)(2n + c - 2)(2n + c - 1)} \quad \text{and} \\
\beta_n &= -\frac{2n(n + c - 1) + bc}{(2n + c)(2n - 2 + c)}.
\end{align*}
\]

Convergence is obtained for \( x \) not on the negative real axis and also in a neighborhood of the origin.

IV. Acknowledgment. This paper covers research initiated by the Directorate of Systems Dynamic Analysis, Systems Engineering Group (RTD), United States Air Force, Wright-Patterson Air Force Base, Ohio, under Contract No. AF 33(657)-10652.

The author acknowledges the valuable advice and direction given him by Mr. Yudell L. Luke and Mr. Jet Wimp.

Midwest Research Institute
Kansas City, Missouri


License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use