# AN APPROXIMATE SOLUTION OF THE RICCATI MATRIX EQUATION[1]

M. S. HENRY AND F. M. STEIN

ABSTRACT. The Riccati matrix equation has been the subject of several recent papers. In the present paper, the solution to this equation is approximated by a sequence of matrices whose elements are rational functions. It is shown that the sequence converges uniformly to the solution. Furthermore, each element of the sequence is constructed from a matrix polynomial which is in a sense the best approximation to the solution of the linear system associated with the Riccati matrix equation.

1. **The Riccati matrix equation.** If $A$, $B$, $C$, and $D$ are $n\times n$ matrices whose elements are continuous functions of the real variable $x$ over the interval $[a, a+\alpha_1]$, the *Riccati matrix differential equation* is defined to be

$$(1)\qquad R[W(x)]\equiv W'(x)+W(x)A(x)+D(x)W(x)+W(x)B(x)W(x)=C(x).$$

(See [4] and the references given in [4] for a number of places where this equation arises.)

Associated with (1) is the initial condition

$$(2)\qquad W(a) = W_a, \qquad a \leqq x \leqq a+\alpha_1,$$

where $W_a$ is an $n\times n$ nonsingular matrix. Although the system is known to have a solution under the conditions given [1], the solution in closed form may be difficult or impossible to find. Consequently we consider the possibility of approximating this solution in some sense. Since the differential equation (1) is nonlinear, we examine the linear system associated with (1) and (2) in the following discussion.

2. **The associated linear system.** For any $r\times r$ matrix $H(x)$ whose elements are continuous functions for $a\leqq x\leqq b$, we define the *absolute value function* to be

$$(3)\qquad \|H(x)\| = \sum_{i,j=1}^{r} |h_{ij}|.$$

Presented to the Society, April 19, 1968 under the title *A best approximate solution of the Riccati matrix equation*; received by the editors February 3, 1969.

*AMS Subject Classifications.* Primary 4115, 4117, 4140; Secondary 3535.

*Key Words and Phrases.* Riccati matrix equation, associated Riccati system, best approximation, matrix polynomial, bounded linear operation, Banach algebra, rational function, rate of convergence.

[1] This work was partially supported by AFOSR Grant 185-67.

From (3) it follows that

$$\|H(x)\|_m = \max_{a \leqq x \leqq b} \|H(x)\| \tag{4}$$

defines a norm for matrices of this type.

The $2n \times 2n$ matrix integral equation

$$L[\phi(x)] \equiv \phi(x) - \int_a^x K(t)\phi(t)dt = Y_a, \tag{5}$$

where

$$K(x) = \begin{bmatrix} A(x) & B(x) \\ C(x) & -D(x) \end{bmatrix} \quad \text{and} \quad Y_a = \begin{bmatrix} W_a^{-1} & 0 \\ I & W_a^{-1} \end{bmatrix}, \tag{6}$$

is defined to be the *associated Riccati system.* If

$$Y(x) = \begin{bmatrix} Y_{11}(x) & Y_{12}(x) \\ Y_{21}(x) & Y_{22}(x) \end{bmatrix} \tag{7}$$

is the unique solution of (5) in the generalized rectangle

$$\|\phi - Y_a\| \leqq c, \qquad a \leqq x \leqq a + \alpha_2 = b, \qquad \alpha_2 \leqq \alpha_1, \tag{8}$$

such that $Y_{11}(x)$ has an inverse for $a \leqq x \leqq b$ (it will, for example, if $c < 1/\|W_a\|$), then it can be shown (see [4]) that

$$W(x) = Y_{21}(x) Y_{11}^{-1}(x) \tag{9}$$

is the unique solution of (1) and (2) in the generalized rectangle

$$\|W - W_a\| \leqq \|Y_{21}\|_m \|Y_{11}^{-1}\|_m + \|W_a\|, \qquad a \leqq x \leqq b. \tag{10}$$

3. **The best approximation to $L[Y(x)]$.** For our approximating functions we choose matrix polynomials of the class

$$P_k(x) = \sum_{i=0}^{k} x^i \sum_{m,j=1}^{2n} c_{mj}^i E_{mj}, \tag{11}$$

where the $E_{mj}$ are $2n \times 2n$ matrices having one as the $(m, j)$ element and zeros elsewhere. The coefficients $c_{mj}^i$ of the matrix polynomials $P_k(x)$ are required to be such that $P_k(a) = Y_a$. Since $L$ is a linear operator it is not difficult to show that

$$\min_{[c_{mj}^i]} \|L[Y(x)] - L[P_k(x)]\|_m \tag{12}$$

is attained for some matrix polynomial $P_k^*(x)$ for fixed $k$, the maximum degree of the polynomial elements of $P_k(x)$. That is, $L[P_k^*(x)]$

is the *best approximation* to $L[Y(x)]$ for fixed $k$.

Since $P_k^*(x)$ is the minimizing matrix polynomial of degree $k$ in (12), it follows by use of the Weierstrass theorem on polynomial approximation [2] that

$$\lim_{k\to\infty} \|L[Y(x)] - L[P_k^*(x)]\|_m = 0. \tag{13}$$

Therefore, since $L$ is a one-to-one continuous operator of a Banach space onto itself, (13) implies that

$$\lim_{k\to\infty} \|L^{-1}\|_1 \|L[Y(x)] - L[P_k^*(x)]\|_m = 0,$$

where $\|L^{-1}\|_1$ is the norm of the bounded linear operator $L^{-1}$, see [5]. Consequently, since

$$\|L^{-1}\|_1 \|L[Y(x)] - L[P_k^*(x)]\|_m \geqq \|L^{-1}\{L[Y(x)] - L[P_k^*(x)]\}\|_m$$

we have that

$$\lim_{k\to\infty} \|Y(x) - P_k^*(x)\|_m = 0. \tag{14}$$

4. **An approximation to** $W(x)$. We write the minimizing matrix polynomial of degree $k$ as

$$P_k^*(x) = \begin{bmatrix} P_{11}^k(x) & P_{12}^k(x) \\ P_{21}^k(x) & P_{22}^k(x) \end{bmatrix}, \tag{15}$$

where $P_{ij}^k(x)$ is an $n\times n$ matrix polynomial. Then from (9) we arrive at the following theorem.

THEOREM. *If* $W(x)$ *is the unique solution of* (1) *and* (2) *in the generalized rectangle* (10), *then*

$$\lim_{k\to\infty} \|W(x) - P_{21}^k(x)[P_{11}^k(x)]^{-1}\|_m = 0, \tag{i}$$

*and*

$$P_{21}^k(a)[P_{11}^k(a)]^{-1} = W_a, \qquad k \geqq k_0. \tag{ii}$$

PROOF. We first show that if $k$ is sufficiently large, say $k \geqq k_0$, then $P_{11}^k(x)$ has an inverse. If no $k_0$ exists such that $P_{11}^k(x)$ has an inverse, then there exists a subsequence of the sequence $\{P_{11}^k(x)\}$, say $\{P_{11}^{k_j}(x)\}$, such that for each $j$ the matrix polynomial $P_{11}^{k_j}(x)$ is singular. Thus, since

$$\lim_{j\to\infty} \|Y_{11}(x) - P_{11}^{k_j}(x)\|_m = 0$$

by (14), it would follow that $Y_{11}(x)$ must also be singular, a contradiction. Hence, $P_{11}^k(x)$ has an inverse for sufficiently large $k$.

We now assume that $k \geqq k_0$, and hence $P_{11}^k(x)$ has an inverse. We consider the norm of the difference,

$$\begin{aligned}
&\|Y_{21}Y_{11}^{-1} - P_{21}^k(P_{11}^k)^{-1}\|_m \\
&\quad = \|Y_{21}Y_{11}^{-1} - P_{21}^kY_{11}^{-1} + P_{21}^kY_{11}^{-1} - P_{21}^k(P_{11}^k)^{-1}\|_m \\
&\quad \leqq \|Y_{11}^{-1}\|_m\|Y_{21} - P_{21}^k\|_m + \|P_{21}^k\|_m\|(P_{11}^k)^{-1} - Y_{11}^{-1}\|_m \\
&\quad \leqq \|Y_{11}^{-1}\|_m\|Y_{21} - P_{21}^k\|_m + 2\|P_{21}^k\|_m\|Y_{11}^{-1}\|_m^2\|P_{11}^k - Y_{11}\|_m.
\end{aligned} \tag{16}$$

The last inequality in (16), which is given without proof, follows from the fact that the linear space of $n \times n$ matrices whose elements are continuous functions over $[a, b]$ is a Banach algebra [5]. Thus by the use of the submatrices of (14) in (16), result (i) of the theorem is obtained.

The properties of a Banach algebra used are the following:

If $G \subseteq N_n$ ($N_n$ is a Banach algebra) is the set of elements in $N_n$ which have inverses, and if $\{A^{(k)}(x)\} \subseteq G$, and $A(x) \in G$, and if

$$\lim_{k\to\infty} \|A^{(k)}(x) - A(x)\|_m = 0,$$

then by a series of lemmas in [5, pp. 305–307], it is shown that if $T[A(x)] = A^{-1}(x)$ and $T[A^{(k)}(x)] = [A^{(k)}(x)]^{-1}$, then

$$\|T[A(x)] - T[A^{(k)}(x)]\|_m \leqq 2\|A^{(k)}(x) - A(x)\|_m\|A^{-1}(x)\|_m^2.$$

This result is used in the inequality in (16) with $A(x) = Y_{11}(x)$ and $A^{(k)}(x) = P_{11}^k(x)$.

Since we require that $P_k(a) = Y_a$, then by examining the $n \times n$ submatrices we have for $k \geqq k_0$ that

$$P_{21}^k(a)[P_{11}^k(a)]^{-1} = I(W_a^{-1})^{-1} = W_a,$$

and conclusion (ii) of the theorem is satisfied.

5. **Conclusion.** We observe that from the matrix polynomials $\{P_k^*(x)\}$ which minimize (12) we obtain a sequence of matrices of *rational functions* which converge uniformly to the solution of (1) and (2). Since the solution $Y(x)$ of (5) is best approximated without being known, the solution $W(x)$ of (1) and (2) can thus be approximated without being known.

Although the discussion has not been concerned with the numerical aspect of the problem, it might be observed that, since $L$ is a linear operator, the minimizing matrix polynomial $P_k^*(x)$, with $k$ fixed, can be obtained in a manner analogous to that used, say, in finding the best approximation to a continuous function using Chebyshev polynomials.

Also it is not given in the paper but it can be shown that the rate of convergence in (i) of the theorem is of the order $O(1/k)$ uniformly in $x$. Furthermore if $Y(x)$ possesses $p$ continuous derivatives, then the rate of convergence is of the order $O(1/k^p)$.

## References

1. E. A. Coddington and N. Levinson, *Theory of ordinary differential equations*, McGraw-Hill, New York, 1965.

2. P. J. Davis, *Interpolation and approximation*, Blaisdell, Waltham, Mass., 1963. MR 28 #393.

3. D. Jackson, *The theory of approximation*, Amer. Math. Soc. Colloq. Publ., vol. XI, Amer. Math. Soc., Providence, R. I., 1930.

4. W. T. Reid, *Riccati matrix differential equations and non-oscillation criteria for associated linear differential systems*, Pacific J. Math. **13** (1963), 665–685. MR 27 #4991.

5. G. F. Simmons, *Introduction to topology and modern analysis*, McGraw-Hill, New York, 1963. MR 26 #4145.

Montana State University, Bozeman, Montana 59715 and
Colorado State University, Fort Collins, Colorado 80521