Iterative Solution of a Matrix Riccati Equation Arising in Stochastic Control

Chun-Hua Guo

Dedicated to Peter Lancaster on the occasion of his 70th birthday

We consider iterative methods for finding the maximal Hermitian solution of a matrix Riccati equation arising in stochastic control. Newton's method is very expensive when the size of the problem is large. A much less expensive iteration is introduced and shown to have several convergence properties similar to those of Newton's method. In ordinary situations, the convergence of the new iteration is linear while the convergence of Newton's method is quadratic. In extreme cases, the convergence of the new iteration may be sublinear while the convergence of Newton's method may be linear. We also show how the performance of Newton's method can be improved when its convergence is not quadratic.

1. Introduction

Let $\mathcal{H}$ be the linear space of all $n \times n$ Hermitian matrices over the field $\mathbb{R}$. For any $X, Y \in \mathcal{H}$, we write $X \geq Y$ (or $Y \leq X$) if $X - Y$ is positive semidefinite. For any $A \in \mathbb{C}^{n\times n}$, the spectrum of A will be denoted by $\sigma(A)$. The transpose and the conjugate transpose of A will be denoted by A^T and A^*, respectively. We denote by $\mathbb{C}_<$ (resp. $\mathbb{C}_\leq$) the set of complex numbers with negative (resp. nonpositive) real parts. A matrix A is said to be *stable* if $\sigma(A) \subset \mathbb{C}_<$. For any matrices $A, B, C \in \mathbb{C}^{n\times n}$, the pair (A, B) is *stabilizable* if $A - BK$ is stable for some $K \in \mathbb{C}^{n\times n}$. The pair (C, A) is *detectable* if (A^*, C^*) is stabilizable.

In this paper, we are concerned with the numerical solution of the matrix Riccati equation

$$\mathcal{R}(X) = A^*X + XA + \Pi(X) + C - XDX = 0, \tag{1.1}$$

where $A, C, D \in \mathbb{C}^{n\times n}$, $C^* = C$, $D^* = D$, $D \geq 0$, and Π is a *positive* linear operator from $\mathcal{H}$ into itself, i.e., $\Pi(X) \geq 0$ whenever $X \geq 0$. The Riccati function $\mathcal{R}$ is thus a mapping from $\mathcal{H}$ into itself.

Matrix Riccati equations of this type were first studied by W. M. Wonham [11]. The following result is a slight different presentation of a result in [11]. It establishes the existence of a positive semidefinite solution to (1.1) under some additional conditions.

Theorem 1.1. *If $C \geq 0$, (A, D) is stabilizable, (C, A) is detectable, and*

$$\inf_{K \in \mathcal{H}} \left\| \int_0^\infty e^{t(A-DK)^*} \Pi(I) e^{t(A-DK)} dt \right\| < 1, \tag{1.2}$$

where $\| \cdot \|$ is the spectral norm, then (1.1) *has at least one solution $\hat{X} \geq 0$ such that $A - D\hat{X}$ is stable.*

The above result was proved in [11] by using an iterative procedure, which is in fact the Newton iteration. For the Riccati function $\mathcal{R}$, the first Fréchet derivative of $\mathcal{R}$ at a matrix $X \in \mathcal{H}$ is a linear map $\mathcal{R}'_X : \mathcal{H} \to \mathcal{H}$ given by

$$\mathcal{R}'_X(H) = (A - DX)^* H + H(A - DX) + \Pi(H). \tag{1.3}$$

Also the second derivative at X, $\mathcal{R}''_X : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$, is given by

$$\mathcal{R}''_X(H_1, H_2) = -H_1 D H_2 - H_2 D H_1. \tag{1.4}$$

The Newton method for the solution of (1.1) is

$$X_{i+1} = X_i - (\mathcal{R}'_{X_i})^{-1} \mathcal{R}(X_i), \quad i = 0, 1, \ldots, \tag{1.5}$$

given that the maps $\mathcal{R}'_{X_i}$ are all invertible. In view of (1.3), the iteration (1.5) is equivalent to

$$(A - DX_i)^* X_{i+1} + X_{i+1}(A - DX_i) + \Pi(X_{i+1}) = -X_i D X_i - C, \quad i = 0, 1, \ldots. \tag{1.6}$$

Newton's method has been studied recently by T. Damm and D. Hinrichsen [2] for a rational matrix equation which includes (1.1) as a special case. We first give a definition from [2].

Definition 1.2. A matrix $X \in \mathcal{H}$ is called *stabilizing* for $\mathcal{R}$ if $\sigma(\mathcal{R}'_X) \subset \mathbb{C}_<$ and *almost stabilizing* if $\sigma(\mathcal{R}'_X) \subset \mathbb{C}_\leq$.

When $\Pi = 0$, it is readily seen that $\sigma(\mathcal{R}'_X) \subset \mathbb{C}_<$ (resp. $\mathbb{C}_\leq$) if and only if $\sigma(A - DX) \subset \mathbb{C}_<$ (resp. $\mathbb{C}_\leq$).

A solution X_+ of (1.1) is called maximal if $X_+ \geq X$ for any solution X. The maximal solution is the most desirable solution in applications. When $\Pi = 0$, the maximal solution may be found by subspace methods (see [10], for example). However, those methods are not applicable when $\Pi \neq 0$.

The following result, given in [2], is a generalization of Theorem 9.1.1 of [9]. It shows that the maximal solution of (1.1) can be found by Newton's method under mild conditions.

Theorem 1.3. *Assume that there exist a solution $\hat{X}$ to $\mathcal{R}(X) \geq 0$ and a stabilizing matrix X_0. Then the Newton sequence is well defined and, moreover, the following are true:*

1. $X_k \geq X_{k+1}, \quad X_k \geq \hat{X}, \quad \mathcal{R}(X_k) \leq 0, \quad k \geq 1.$

2. $\sigma(\mathcal{R}'_{X_k}) \subset \mathbb{C}_<, \quad k \geq 0.$

3. $\lim_{k\to\infty} X_k = X_+$ *is the maximal solution of* (1.1).

4. $\sigma(\mathcal{R}'_{X_+}) \subset \mathbb{C}_\leq$.

Remark 1.4. If (1.2) is true, then

$$\left\| \int_0^\infty e^{t(A-DX_0)^*} \Pi(I) e^{t(A-DX_0)} dt \right\| < 1$$

for some $X_0 \in \mathcal{H}$. It is noted in [2] that this X_0 is necessarily stabilizing for $\mathcal{R}$. The assumption that $C \geq 0$ and (C, A) is detectable is not needed for the above theorem. As a result, the maximal solution is not necessarily positive semidefinite. It can be seen that Theorem 1.3 is also a generalization of Theorem 3.3 of [4].

Note that the solution of the linear equation (1.6) is required in each step of the Newton iteration. The presence of the linear operator Π on the left hand side will make solving this equation very expensive when n is large. For example, if Π is given by $\Pi(H) = B^*HB$, then we will need to solve a linear matrix equation of the form

$$A^*X + XA + B^*XB = C$$

in each step of the Newton iteration. This equation is equivalent to

$$(I \otimes A^* + A^T \otimes I + B^T \otimes B^*)\mathrm{vec}\, X = \mathrm{vec}\, C,$$

where $\otimes$ is the Kronecker product and the vec operator stacks the columns of a matrix into a long vector (see [9], for example). A direct solution of this equation would require $O(n^6)$ operations. On the other hand, a matrix equation of the form

$$A^*X + XA = C$$

can be solved by the Bartels-Stewart algorithm [1] in $O(n^3)$ operations when it has a unique solution.

This observation leads us to consider the iteration

$$(1.7) \quad (A - DX_i)^* X_{i+1} + X_{i+1}(A - DX_i) = -\Pi(X_i) - X_i D X_i - C, \qquad i = 0, 1, \ldots.$$

Iteration (1.7) is obtained by replacing $\Pi(X_{i+1})$ with $\Pi(X_i)$ in iteration (1.6). A new convergence analysis will be needed for iteration (1.7).

2. Convergence of the iteration (1.7)

In this section, we will show that iteration (1.7) has several convergence properties similar to those of the Newton iteration. First, we note that iteration (1.7) can be rewritten as

$$(A - DX_i)^*(X_{i+1} - X_i) + (X_{i+1} - X_i)(A - DX_i) = -\mathcal{R}(X_i), \quad i = 0, 1, \ldots . \tag{2.1}$$

We will also need the following well known result (see [9], for example).

Lemma 2.1. *Let $A, C \in \mathbb{C}^{n \times n}$ with A stable and C Hermitian. Then the Lyapunov equation $A^*X + XA = C$ has a unique solution X (necessarily Hermitian). If $C \leq 0$, then $X \geq 0$.*

Theorem 2.2. *Assume that there exist a solution $\hat{X}$ to $\mathcal{R}(X) \geq 0$ and a Hermitian matrix X_0 such that $X_0 \geq \hat{X}$, $\mathcal{R}(X_0) \leq 0$, and $A - DX_0$ is stable. Then the iteration* (1.7) *defines a sequence $\{X_k\}$ such that*

1. $X_k \geq X_{k+1}, \quad X_k \geq \hat{X}, \quad \mathcal{R}(X_k) \leq 0, \quad k \geq 0.$
2. $\sigma(A - DX_k) \subset \mathbb{C}_<, \quad k \geq 0.$
3. $\lim_{k\to\infty} X_k = \tilde{X}$ *is a solution of* (1.1) *and* $\tilde{X} \geq \hat{X}$.
4. $\sigma(A - D\tilde{X}) \subset \mathbb{C}_\leq$.

Proof. We prove by induction that for each $i \geq 0$, X_{i+1} is uniquely determined and

$$X_i \geq X_{i+1}, \quad X_i \geq \hat{X}, \quad \mathcal{R}(X_i) \leq 0, \quad \sigma(A - DX_i) \subset \mathbb{C}_<. \tag{2.2}$$

For $i = 0$, we already have $X_0 \geq \hat{X}$, $\mathcal{R}(X_0) \leq 0$, and $\sigma(A - DX_0) \subset \mathbb{C}_<$. By (2.1) with $i = 0$ and Lemma 2.1, X_1 is uniquely determined and $X_0 \geq X_1$. We now assume that X_{k+1} is uniquely determined and (2.2) is true for $i = k$ ($k \geq 0$). By (1.7) with $i = k$,

$$\begin{aligned}
& (A - DX_k)^*(X_{k+1} - \hat{X}) + (X_{k+1} - \hat{X})(A - DX_k) \\
= \ & -\Pi(X_k) - X_kDX_k - C - A^*\hat{X} - \hat{X}A + X_kD\hat{X} + \hat{X}DX_k \\
\leq \ & -\Pi(X_k) - X_kDX_k + \Pi(\hat{X}) - \hat{X}D\hat{X} + X_kD\hat{X} + \hat{X}DX_k \\
= \ & -\Pi(X_k - \hat{X}) - (X_k - \hat{X})D(X_k - \hat{X}) \leq 0.
\end{aligned}$$

Therefore, $X_{k+1} \geq \hat{X}$ by Lemma 2.1. To show that $A - DX_{k+1}$ is stable, we will use an argument in [5]. Note first that, by writing $A - DX_{k+1} = A - DX_k + D(X_k - X_{k+1})$,

$$(A - DX_{k+1})^*(X_{k+1} - \hat{X}) + (X_{k+1} - \hat{X})(A - DX_{k+1})$$

$$\begin{aligned} &\leq -\Pi(X_k-\hat{X})-(X_k-\hat{X})D(X_k-\hat{X}) \\ &\quad +(X_k-X_{k+1})D(X_{k+1}-\hat{X})+(X_{k+1}-\hat{X})D(X_k-X_{k+1}) \\ (2.3) \qquad &= -\Pi(X_k-\hat{X})-(X_{k+1}-\hat{X})D(X_{k+1}-\hat{X}) \\ &\quad -(X_k-X_{k+1})D(X_k-X_{k+1}) \\ &\leq -(X_k-X_{k+1})D(X_k-X_{k+1}). \end{aligned}$$

If $A-DX_{k+1}$ is not stable, we let λ be an eigenvalue of $A-DX_{k+1}$ with $\mathrm{Re}(\lambda)\geq 0$ and $(A-DX_{k+1})x=\lambda x$ for some $x\neq 0$. Now, by (2.3),

$$2\mathrm{Re}(\lambda)x^*(X_{k+1}-\hat{X})x \leq -x^*(X_k-X_{k+1})D(X_k-X_{k+1})x.$$

Therefore, $x^*(X_k-X_{k+1})D(X_k-X_{k+1})x=0$ and thus $D(X_k-X_{k+1})x=0$. Now, $(A-DX_k)x=(A-DX_{k+1})x=\lambda x$, which is contradictory to the stability of $A-DX_k$. We have thus proved that $A-DX_{k+1}$ is stable. So, X_{k+2} is uniquely determined and

$$\begin{aligned} &(A-DX_{k+1})^*(X_{k+1}-X_{k+2})+(X_{k+1}-X_{k+2})(A-DX_{k+1}) \\ = \;& \big(A-DX_k+D(X_k-X_{k+1})\big)^*X_{k+1}+X_{k+1}\big(A-DX_k+D(X_k-X_{k+1})\big) \\ &+\Pi(X_{k+1})+X_{k+1}DX_{k+1}+C \\ = \;& -\Pi(X_k-X_{k+1})-X_kDX_k+X_{k+1}DX_{k+1} \\ &+(X_k-X_{k+1})DX_{k+1}+X_{k+1}D(X_k-X_{k+1}) \\ = \;& -\Pi(X_k-X_{k+1})-(X_k-X_{k+1})D(X_k-X_{k+1})\leq 0. \end{aligned}$$

Therefore, $X_{k+1}\geq X_{k+2}$. Since

$$(A-DX_{k+1})^*(X_{k+1}-X_{k+2})+(X_{k+1}-X_{k+2})(A-DX_{k+1})=\mathcal{R}(X_{k+1}),$$

we also get $\mathcal{R}(X_{k+1})\leq 0$. The induction process is now complete. Thus, the sequence $\{X_k\}$ is well defined, monotonically decreasing, and bounded below by $\hat{X}$. Let $\lim_{k\to\infty}X_k=\tilde{X}$. We have $\tilde{X}\geq\hat{X}$. By taking limits in (1.7), we see that $\tilde{X}$ is a solution of (1.1). Since $\sigma(A-DX_k)\subset\mathbb{C}_<$ for each k, $\sigma(A-D\tilde{X})\subset\mathbb{C}_\leq$. □

Remark 2.3. If, in addition, X_0 is an upper bound for the solution set of (1.1) (i.e., $X_0\geq X$ for all solutions of (1.1)), then $\tilde{X}$ is the maximal solution of (1.1).

To further study the convergence behaviour of iteration (1.7), we need some results from [2].

We first note that $\mathcal{H}$ is a Hilbert space with the Frobenius inner product $\langle X,Y\rangle=\mathrm{trace}(XY)$. For a linear operator $\mathcal{L}$ on $\mathcal{H}$, let $\rho(\mathcal{L})=\max\{|\lambda|:\lambda\in\sigma(\mathcal{L})\}$ denote the spectral radius, and $\beta(\mathcal{L})=\max\{\mathrm{Re}(\lambda):\lambda\in\sigma(\mathcal{L})\}$ the spectral abscissa. The identity map is denoted by I. As for matrices, $\mathcal{L}$ is called *stable* if $\sigma(\mathcal{L})\subset\mathbb{C}_<$.

Definition 2.4. A linear operator $\mathcal{L}$ on $\mathcal{H}$ is called *positive* if $\mathcal{L}(H)\geq 0$ whenever $H\geq 0$. $\mathcal{L}$ is called *inverse positive* if $\mathcal{L}^{-1}$ exists and is positive. $\mathcal{L}$ is called *resolvent positive* if the operator $\alpha I - \mathcal{L}$ is inverse positive for all sufficiently large $\alpha > 0$.

Theorem 2.5. (cf. [2]) *Let $\mathcal{L} : \mathcal{H} \to \mathcal{H}$ be resolvent positive and $\Pi : \mathcal{H} \to \mathcal{H}$ be positive. Then $\mathcal{L} + \Pi$ is also resolvent positive. Moreover, the following are equivalent.*

1. *$\mathcal{L} + \Pi$ is stable.*
2. *$-(\mathcal{L} + \Pi)$ is inverse positive.*
3. *$\mathcal{L}$ is stable and $\rho(\mathcal{L}^{-1}\Pi) < 1$.*

Theorem 2.6. (cf. [2]) *If $\mathcal{L} : \mathcal{H} \to \mathcal{H}$ is resolvent positive, then $\beta(\mathcal{L}) \in \sigma(\mathcal{L})$ and there exists a nonzero matrix $V \geq 0$ such that $\mathcal{L}(V) = \beta(\mathcal{L})V$.*

As noted in [2], if $\mathcal{L}$ is resolvent positive, then the adjoint operator $\mathcal{L}^*$ is also resolvent positive and $\beta(\mathcal{L}^*) = \beta(\mathcal{L})$.

Lemma 2.7. *For any $A \in \mathbb{C}^{n\times n}$, the linear operator $\mathcal{L} : \mathcal{H} \to \mathcal{H}$ defined by*

$$\mathcal{L}(H) = A^*H + HA$$

is resolvent positive. The adjoint operator of $\mathcal{L}$ is given by

$$\mathcal{L}^*(H) = AH + HA^*.$$

Proof. The first part of the lemma is proved in [2]. For any $U, V \in \mathcal{H}$, $\langle \mathcal{L}U, V\rangle = \mathrm{trace}(\mathcal{L}UV) = \mathrm{trace}(A^*UV) + \mathrm{trace}(UAV) = \mathrm{trace}(UVA^*) + \mathrm{trace}(UAV) = \langle U, AV + VA^*\rangle$. This proves the second part of the lemma. □

We are now ready to prove the following convergence result for iteration (1.7).

Theorem 2.8. *Assume that there exist a solution $\hat{X}$ to $\mathcal{R}(X) \geq 0$ and a Hermitian matrix X_0 such that $\mathcal{R}(X_0) \leq 0$ and $\mathcal{R}'_{X_0}$ is stable. Then the iteration* (1.7) *defines a sequence $\{X_k\}$ such that*

1. $X_k \geq X_{k+1}, \quad X_k \geq \hat{X}, \quad \mathcal{R}(X_k) \leq 0, \quad k \geq 0.$
2. $\sigma(\mathcal{R}'_{X_k}) \subset \mathbb{C}_<, \quad k \geq 0.$
3. $\lim_{k\to\infty} X_k = X_+$*, the maximal solution of* (1.1).
4. $\sigma(\mathcal{R}'_{X_+}) \subset \mathbb{C}_\leq$.

Proof. By Theorem 1.3, $X_1^N = X_0 - (\mathcal{R}'_{X_0})^{-1}\mathcal{R}(X_0) \geq \hat{X}$. Since $\mathcal{R}(X_0) \leq 0$ and $-\mathcal{R}'_{X_0}$ is inverse positive by Theorem 2.5 and Lemma 2.7, we also have $X_0 \geq X_1^N$.

Thus, $X_0 \geq \hat{X}$ is necessarily true. Since $\mathcal{R}'_{X_0}$ is stable, we know from Theorem 2.5 that the operator $\mathcal{L}$ given by

$$\mathcal{L}(H) = (A - DX_0)^* H + H(A - DX_0)$$

is also stable. Thus, $A - DX_0$ is a stable matrix. Therefore, all the conclusions of Theorem 2.2 are true. Since $\lim_{k\to\infty} X_k = \tilde{X} \geq \hat{X}$ and $\hat{X}$ can be taken to be any solution of (1.1), we have $\tilde{X} = X_+$. We have thus proved items 1 and 3 of the theorem. Since item 4 follows from item 2, we need only to prove item 2. Assume that $\mathcal{R}'_{X_k}$ is stable for some $k \geq 0$. We need to prove that $\mathcal{R}'_{X_{k+1}}$ is also stable. If $\mathcal{R}'_{X_{k+1}}$ is not stable, we know from Theorem 2.6 and the note that follows it that $(\mathcal{R}'_{X_{k+1}})^*(V) = \beta V$ for some nonzero $V \geq 0$ and some number $\beta \geq 0$. Therefore,

$$\langle V, \mathcal{R}'_{X_{k+1}}(X_{k+1} - \hat{X})\rangle = \langle \beta V, X_{k+1} - \hat{X}\rangle \geq 0.$$

On the other hand, we have by (2.3) that

$$\begin{aligned}
& \mathcal{R}'_{X_{k+1}}(X_{k+1} - \hat{X}) \\
= \quad & (A - DX_{k+1})^*(X_{k+1} - \hat{X}) + (X_{k+1} - \hat{X})(A - DX_{k+1}) + \Pi(X_{k+1} - \hat{X}) \\
\leq \quad & -\Pi(X_k - X_{k+1}) - (X_{k+1} - \hat{X})D(X_{k+1} - \hat{X}) \\
& -(X_k - X_{k+1})D(X_k - X_{k+1}) \\
\leq \quad & -(X_k - X_{k+1})D(X_k - X_{k+1}).
\end{aligned}$$

Therefore,

$$\langle V, (X_k - X_{k+1})D(X_k - X_{k+1})\rangle = 0.$$

So, $\operatorname{trace}\big(V^{1/2}(X_k - X_{k+1})D^{1/2}D^{1/2}(X_k - X_{k+1})V^{1/2}\big) = 0$. It follows that $D^{1/2}(X_k - X_{k+1})V^{1/2} = 0$ and thus $D(X_k - X_{k+1})V = 0$. Now, by Lemma 2.7,

$$\begin{aligned}
(\mathcal{R}'_{X_k})^*(V) &= (A - DX_k)V + V(A - DX_k)^* + \Pi^*(V) \\
&= (\mathcal{R}'_{X_{k+1}})^*(V) + D(X_{k+1} - X_k)V + V(X_{k+1} - X_k)D \\
&= (\mathcal{R}'_{X_{k+1}})^*(V) = \beta V,
\end{aligned}$$

which is contradictory to the stability of $\mathcal{R}'_{X_k}$. □

We will now make a comparison between Theorem 1.3 and Theorem 2.8. Note first that we need to assume $\mathcal{R}(X_0) \leq 0$ in Theorem 2.8. The Newton iteration does not need this assumption and $\mathcal{R}(X_1) \leq 0$ is necessarily true. The conclusions in Theorem 1.3 and Theorem 2.8 are almost the same. The only difference is that the first conclusion is generally not true for $k = 0$ in Theorem 1.3, since $\mathcal{R}(X_0) \leq 0$ is not assumed there. But that conclusion will be true for $k = 0$ if we also assume $\mathcal{R}(X_0) \leq 0$ in Theorem 1.3. If it is difficult to choose an X_0 with $\mathcal{R}'_{X_0}$ stable *and* $\mathcal{R}(X_0) \leq 0$, we may get such an X_0 by applying one Newton iteration on a Hermitian matrix Y_0 stabilizing for $\mathcal{R}$.

From the above discussions, the following conclusions can be made.

- Under the conditions of Theorem 2.8, the four conclusions in the theorem would remain valid if the sequence $\{X_k\}_{k=1}^{\infty}$ were obtained by using Newton's method and iteration (1.7) in an arbitrary combination.

- Under the conditions of Theorem 1.3, the four conclusions in the theorem would remain valid if, after X_1 has been obtained by Newton's method, the sequence $\{X_k\}_{k=2}^{\infty}$ were obtained by using Newton's method and iteration (1.7) in an arbitrary combination.

Before we can determine a good combination of the Newton iteration and the iteration (1.7), we need to have some idea about the convergence rates of these two iterations.

3. Convergence rates of the two iterations

We start with a result on the convergence rate of the Newton iteration.

Theorem 3.1. *If $\mathcal{R}'_{X_+}$ is stable in Theorem* 1.3*, then the convergence of Newton's method is quadratic.*

The above result was proved in [2]. It also follows directly from Theorem 1.3 and a result on the local quadratic convergence of Newton's method in general Banach spaces (see [7], for example).

For iteration (1.7), linear convergence can be guaranteed when $\mathcal{R}'_{X_+}$ is stable. This will be a consequence of the following general result.

Theorem 3.2. (cf. [8, p. 21]) *Let T be a (nonlinear) operator from a Banach space E into itself and $x^* \in E$ be a solution of $x = Tx$. If T is Fréchet differentiable at x^* with $\rho(T'_{x^*}) < 1$, then the iterates $x_{n+1} = Tx_n$ $(n = 0, 1, \ldots)$ converge to x^*, provided that x_0 is sufficiently close to x^*. Moreover, for any $\epsilon > 0$,*

$$\|x_n - x^*\| \le c(x_0; \epsilon)\big(\rho(T'_{x^*}) + \epsilon\big)^n,$$

where $\|\cdot\|$ is the norm in E and $c(x_0; \epsilon)$ is a constant independent of n.

Theorem 3.3. *Let the sequence $\{X_k\}$ be as in Theorem* 2.8*. If $\mathcal{R}'_{X_+}$ is stable, then*

$$\limsup_{k\to\infty} \sqrt[k]{\|X_k - X_+\|} \le \rho\big((\mathcal{L}_{X_+})^{-1}\Pi\big) < 1,$$

where $\|\cdot\|$ is any matrix norm and the operator $\mathcal{L}_X$ is defined by

$$\mathcal{L}_X(H) = (A - DX)^* H + H(A - DX).$$

Proof. The iteration (1.7) can be written as $X_{k+1} = G(X_k)$ with

$$G(X) = (\mathcal{L}_X)^{-1}(-\Pi(X) - XDX - C).$$

It can easily be shown that

$$G(X_+ + H) - G(X_+) = -(\mathcal{L}_{X_+})^{-1}\Pi(H) + o(H),$$

where $o(H)$ denotes some matrix $W(H)$ with $\lim_{\|H\|\to 0} \frac{\|W(H)\|}{\|H\|} = 0$. Therefore, the Fréchet derivative of G at the matrix X_+ is $G'_{X_+} = -(\mathcal{L}_{X_+})^{-1}\Pi$. Since $\mathcal{R}'_{X_+}$ is stable, we have $\rho\big((\mathcal{L}_{X_+})^{-1}\Pi\big) < 1$ by Theorem 2.5. Therefore,

$$\limsup_{k\to\infty} \sqrt[k]{\|X_k - X_+\|} \le \rho\big((\mathcal{L}_{X_+})^{-1}\Pi\big) < 1$$

by Theorems 2.8 and 3.2. □

Therefore, when $\mathcal{R}'_{X_+}$ has no eigenvalues on the imaginary axis, the convergence of iteration(1.7) is linear while the convergence of the Newton iteration is quadratic. Next we will examine the convergence rates of the two iterations when $\mathcal{R}'_{X_+}$ has some eigenvalues on the imaginary axis.

In the case of $\Pi = 0$, the two iterations are identical and $\mathcal{R}'_{X_+}$ has eigenvalues on the imaginary axis if and only if $A - DX_+$ has eigenvalues on the imaginary axis. A convergence rate analysis has been given in [6] when all the eigenvalues of $A - DX_+$ on the imaginary axis are semisimple (i.e., all elementary divisors associated with these eigenvalues are linear). If $A - DX_+$ has non-semisimple eigenvalues on the imaginary axis, the convergence rate analysis remains an open problem.

In general, when $\mathcal{R}'_{X_+}$ has eigenvalues on the imaginary axis, we know from Theorem 2.6 that 0 must be one of these eigenvalues. Therefore, $\mathcal{R}'_{X_+}$ is not invertible. The convergence of Newton's method is typically linear in this case (see [3], for example). If $\mathcal{L}_{X_+}$ is invertible, then $\rho\big((\mathcal{L}_{X_+})^{-1}\Pi\big) = 1$ and the convergence of iteration (1.7) is expected to be sublinear in view of Theorem 3.3. Here is one example.

Example 3.4. For the Riccati equation (1.1) with $n = 1$ and

$$A = \frac{1}{2}, \quad C = -1, \quad D = 1, \quad \Pi(X) = X,$$

it is clear that $X_+ = 1$ is the unique solution and the conditions in Theorems 1.3 and 2.8 are satisfied for any $X_0 > 1$. The iteration (1.7) is given by

$$X_{k+1} = X_k - \frac{(X_k - 1)^2}{2X_k - 1}.$$

So
$$\frac{X_{k+1}-1}{X_k-1}=1-\frac{X_k-1}{2X_k-1}.$$
Therefore, $\lim_{k\to\infty}\frac{X_{k+1}-1}{X_k-1}=1$, i.e., the convergence is sublinear. The Newton iteration is given by
$$X_{k+1}=X_k-\frac{1}{2}(X_k-1).$$
So
$$\frac{X_{k+1}-1}{X_k-1}=\frac{1}{2}.$$
Thus, the Newton iteration converges to X_+ linearly with rate $\frac{1}{2}$.

If both $\mathcal{R}'_{X_+}$ and $\mathcal{L}_{X_+}$ are singular (this is not very likely when $\Pi\neq 0$), then the convergence of the iteration (1.7) may be linear. However, the rate of convergence may be very close to 1, as the following example shows:

Example 3.5. Consider the Riccati equation (1.1) with $n=2$ and
$$A=\begin{pmatrix}0&0\\0&-1\end{pmatrix},\quad C=0,\quad D=\begin{pmatrix}1&0\\0&0\end{pmatrix},$$
$$\Pi(X)=\begin{pmatrix}0&0\\0&\epsilon\end{pmatrix}X\begin{pmatrix}0&0\\0&\epsilon\end{pmatrix},$$
where $0\le\epsilon<\sqrt{2}$. The conditions in Theorems 1.3 and 2.8 are satisfied for $X_0=I$. For iteration (1.7), the iterates are
$$X_k=\begin{pmatrix}\left(\frac{1}{2}\right)^k&0\\0&\left(\frac{\epsilon^2}{2}\right)^k\end{pmatrix},\quad k=1,2,\ldots.$$
The convergence to the maximal solution $X_+=0$ is thus linear with rate $\max\{\frac{1}{2},\frac{\epsilon^2}{2}\}$. For the Newton iteration, the iterates are
$$X_k=\begin{pmatrix}\left(\frac{1}{2}\right)^k&0\\0&0\end{pmatrix},\quad k=1,2,\ldots.$$
The convergence to the maximal solution is thus linear with rate $\frac{1}{2}$.

In summary, when $\mathcal{R}'_{X_+}$ is invertible, the convergence of iteration (1.7) is linear and the convergence of Newton's method is quadratic; when $\mathcal{R}'_{X_+}$ is not invertible, the convergence of iteration (1.7) is typically sublinear and the convergence of Newton's method is typically linear. Therefore, we should start with the much less expensive iteration (1.7) (as long as an initial guess X_0 satisfying the conditions of Theorem 2.8 is available) and switch to the Newton iteration at a later stage if the convergence of iteration (1.7) is detected to be too slow or if a very high precision is required of the approximate maximal solution. Note that *one* step of Newton iteration may be needed to find an X_0 for use with iteration (1.7), from a Hermitian matrix Y_0 such that $\mathcal{R}'_{Y_0}$ is stable. The matrix X_0 so obtained may be far away from the maximal solution X_+ and the Newton iteration could take many steps before fast convergence sets in. This makes the use of the iteration (1.7) (after one Newton iteration) particularly important.

4. Improvement of Newton's method in the singular case

From the above discussions, we can see that the Newton iteration is most useful when the convergence of iteration (1.7) is too slow, particularly when the convergence of iteration (1.7) is sublinear. However, when the convergence of iteration (1.7) is sublinear, $\mathcal{R}'_{X_+}$ is singular and the convergence of Newton's method is typically linear. Linear convergence alone is not satisfactory since the method requires a lot of computational work in each iteration. As in [6], we will show that a simple modification can improve the performance of Newton's method significantly in many cases.

We let $\mathcal{N}$ be the null space of $\mathcal{R}'_{X_+}$ and $\mathcal{M}$ be its orthogonal complement in $\mathcal{H}$. Let $P_\mathcal{N}$ and $P_\mathcal{M}$ be the orthogonal projections onto $\mathcal{N}$ and $\mathcal{M}$, respectively.

Theorem 4.1. *Let the sequence $\{X_k\}$ be as in Theorem 1.3 and, for any fixed $\theta > 0$, let*

$$Q = \{k : \ \|P_\mathcal{M}(X_k - X_+)\| > \theta\|P_\mathcal{N}(X_k - X_+)\|\}.$$

Then there is a constant $c > 0$ such that $\|X_k - X_+\| \le c\|X_{k-1} - X_+\|^2$ for all sufficiently large $k \in Q$.

Proof. Let $\tilde{X}_k = X_k - X_+$. Using Taylor's Theorem with (1.4) and the fact that $\mathcal{R}'_{X_+}(P_\mathcal{N}\tilde{X}_k) = 0$,

$$(4.1)\,\mathcal{R}(X_k) = \mathcal{R}(X_+) + \mathcal{R}'_{X_+}(\tilde{X}_k) + \frac{1}{2}\mathcal{R}''_{X_+}(\tilde{X}_k, \tilde{X}_k) = \mathcal{R}'_{X_+}(P_\mathcal{M}\tilde{X}_k) - \tilde{X}_k D\tilde{X}_k.$$

For $k \in Q$, we have $\|\tilde{X}_k\| \le \|P_\mathcal{M}\tilde{X}_k\| + \|P_\mathcal{N}\tilde{X}_k\| \le \left(\theta^{-1} + 1\right)\|P_\mathcal{M}\tilde{X}_k\|$. Since $\|\mathcal{R}'_{X_+}(P_\mathcal{M}\tilde{X}_k)\| \ge c_1\|P_\mathcal{M}\tilde{X}_k\|$ for some constant $c_1 > 0$, we have by (4.1)

$$(4.2)\quad \|\mathcal{R}(X_k)\| \ge c_1\|P_\mathcal{M}\tilde{X}_k\| - c_2\|\tilde{X}_k\|^2 \ge \left(c_1(\theta^{-1}+1)^{-1} - c_2\|\tilde{X}_k\|\right)\|\tilde{X}_k\|.$$

On the other hand, we have by (1.6)

$$(A - DX_{k-1})^* X_k + X_k(A - DX_{k-1}) + \Pi(X_k) = -X_{k-1}DX_{k-1} - C,$$

and obviously,

$$(A - DX_+)^* X_+ + X_+(A - DX_+) + \Pi(X_+) = -X_+DX_+ - C.$$

By subtraction, we obtain after some manipulations

$$(A - DX_{k-1})^* \tilde{X}_k + \tilde{X}_k(A - DX_{k-1}) + \Pi(\tilde{X}_k) = -\tilde{X}_{k-1} D \tilde{X}_{k-1}.$$

Writing $X_+ = X_{k-1} - \tilde{X}_{k-1}$ in (4.1) and using the last equation it is found that

$$\begin{aligned}\mathcal{R}(X_k) &= \big((A - DX_{k-1}) + D\tilde{X}_{k-1}\big)^* \tilde{X}_k + \tilde{X}_k\big((A - DX_{k-1}) + D\tilde{X}_{k-1}\big) \\ &\quad + \Pi(\tilde{X}_k) - \tilde{X}_k D \tilde{X}_k \\ &= -\tilde{X}_{k-1} D \tilde{X}_{k-1} + \tilde{X}_{k-1} D \tilde{X}_k + \tilde{X}_k D \tilde{X}_{k-1} - \tilde{X}_k D \tilde{X}_k.\end{aligned}$$

Thus,

$$\|\mathcal{R}(X_k)\| \le c_3 \|\tilde{X}_k\|^2 + c_4 \|\tilde{X}_k\| \|\tilde{X}_{k-1}\| + c_5 \|\tilde{X}_{k-1}\|^2. \tag{4.3}$$

In view of (4.2) and the fact that $X_k \neq X_+$ for any k, we have

$$c_1(\theta^{-1} + 1)^{-1} - c_2 \|\tilde{X}_k\| \le c_3 \|\tilde{X}_k\| + c_4 \|\tilde{X}_{k-1}\| + c_5 \|\tilde{X}_{k-1}\|^2 / \|\tilde{X}_k\|.$$

Since $\tilde{X}_k \to 0$ by Theorem 1.3, $\|\tilde{X}_k\| \le c \|\tilde{X}_{k-1}\|^2$ for all sufficiently large $k \in Q$.
□

Corollary 4.2. *Assume that, for given $\theta > 0$,*

$$\|P_{\mathcal{M}}(X_k - X_+)\| > \theta \|P_{\mathcal{N}}(X_k - X_+)\|$$

for all k large enough. Then $X_k \to X_+$ quadratically.

From the corollary, we see that the error will be dominated by the null space component at some stage if the convergence of Newton's method is not quadratic (no examples of quadratic convergence for Newton's method in the singular case have been found for the Riccati equation). We will now examine what will happen if the error is precisely in the null space.

Theorem 4.3. *Let the sequence $\{X_k\}$ be as in Theorem 1.3. If $\mathcal{R}'_{X_+}$ is singular and $X_k - X_+ \in \mathcal{N}$, then*

1. $X_{k+1} - X_+ = \frac{1}{2}(X_k - X_+)$.
2. $X_+ = X_k - 2(\mathcal{R}'_{X_k})^{-1} \mathcal{R}(X_k)$.

Proof. By Taylor's Theorem,

$$\mathcal{R}'_{X_k}(X_k - X_+) = \mathcal{R}'_{X_+}(X_k - X_+) + \mathcal{R}''_{X_+}(X_k - X_+, X_k - X_+).$$

Since $\mathcal{R}(X_+) = 0$ and $\mathcal{R}'_{X_+}(X_k - X_+) = 0$, we may also write

$$\begin{aligned}&\mathcal{R}'_{X_k}(X_k - X_+) \\ =\ & 2\{\mathcal{R}(X_+) + \mathcal{R}'_{X_+}(X_k - X_+) + \frac{1}{2}\mathcal{R}''_{X_+}(X_k - X_+, X_k - X_+)\} \\ =\ & 2\mathcal{R}(X_k).\end{aligned}$$

The second part of the theorem follows immediately. The first part follows easily from (1.5) and the second part. □

From this result, we know that it is possible to get a better approximation to the maximal solution by using a double Newton step when X_k approaches X_+ slowly but the error $X_k - X_+$ is rapidly dominated by its null space component.

The following example illustrates this point.

Example 4.4. Consider the Riccati equation (1.1) with $n = 2$,

$$A = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}, \quad C = \begin{pmatrix} -2 & -4 \\ -4 & -3 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

and

$$\Pi(X) = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} X \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}.$$

It can be verified that the maximal solution of the equation is

$$X_+ = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

The conditions in Theorem 2.8 are satisfied for $X_0 = 10I$. For this example, $\rho\big((\mathcal{L}_{X_+})^{-1}\Pi\big) = 1$ and the convergence of iteration (1.7) is indeed sublinear. After 10000 iterations, we get an approximate maximal solution X_{10000} with

$$X_{10000} - X_+ = \begin{pmatrix} 0 & 2.9596 \times 10^{-16} \\ 2.9596 \times 10^{-16} & 4.0049 \times 10^{-4} \end{pmatrix}.$$

We could have switched to Newton's method much earlier. For example, after 40 iterations, we get X_{40} with

$$X_{40} - X_+ = \begin{pmatrix} 5.0493 \times 10^{-12} & 1.4938 \times 10^{-8} \\ 1.4938 \times 10^{-8} & 1.1968 \times 10^{-1} \end{pmatrix}.$$

If we use $X_0^N = X_{40}$ as the initial guess for the Newton iteration, we get X_{20}^N after 20 iterations with

$$X_{20}^N - X_+ = \begin{pmatrix} 0 & 0 \\ 0 & 1.1516 \times 10^{-7} \end{pmatrix}.$$

However, the double Newton step can be used to great advantage. For example, we can get X_2^N after only two Newton iterations with

$$X_2^N - X_+ = \begin{pmatrix} 0 & 2.2547 \times 10^{-11} \\ 2.2547 \times 10^{-11} & 2.9921 \times 10^{-2} \end{pmatrix}$$

and apply a double Newton step on X_2^N to get X_3^{DN} with

$$X_3^{DN} - X_+ = \begin{pmatrix} 1.7764 \times 10^{-15} & -2.1889 \times 10^{-11} \\ -2.1889 \times 10^{-11} & 6.5890 \times 10^{-11} \end{pmatrix}.$$

Note that $\mathcal{N} = \{\mathrm{diag}(0, a) : a \in \mathbb{R}\}$ for this example.

Acknowledgements

This work was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada. The author also wishes to thank Tobias Damm for sending him the reference [2].

References

[1] Bartels, R. H., Stewart, G. W., Solution of the matrix equation $AX + XB = C$, *Comm. ACM* **15** (1972), 820-826.

[2] Damm, T., Hinrichsen, D., Newton's method for a rational matrix equation occurring in stochastic control, Preprint, 1999.

[3] Decker, D. W., Keller, H. B., Kelley, C. T., Convergence rates for Newton's method at singular points, *SIAM J. Numer. Anal.* **20** (1983), 296-314.

[4] Freiling, G., Jank, G., Existence and comparison theorems for algebraic Riccati equations and Riccati differential and difference equations, *J. Dynam. Control Systems* **2** (1996), 529-547.

[5] Gohberg, I., Lancaster, P., Rodman, L., On Hermitian solutions of the symmetric algebraic Riccati equation, *SIAM J. Control Optimization* **24** (1986), 1323-1334.

[6] Guo, C.-H., Lancaster, P., Analysis and modification of Newton's method for algebraic Riccati equations, *Math. Comp.* **67** (1998), 1089-1105.

[7] Kantorovich, L. V., Akilov, G. P., *Functional Analysis in Normed Spaces*, Pergamon, New York, 1964.

[8] Krasnoselskii, M. A., Vainikko, G. M., Zabreiko, P. P., Rutitskii, Ya. B., Stetsenko, V. Ya., *Approximate Solution of Operator Equations*, Wolters-Noordhoff Publishing, Groningen, 1972.

[9] Lancaster, P., Rodman, L., *Algebraic Riccati Equations*, Clarendon Press, Oxford, 1995.

[10] Mehrmann, V. L., *The Autonomous Linear Quadratic Control Problem*, Lecture Notes in Control and Information Sciences, Vol. 163, Springer Verlag, Berlin, 1991.

[11] Wonham, W. M., On a matrix Riccati equation of stochastic control, *SIAM J. Control* **6** (1968), 681-697.

Department of Mathematics and Statistics
University of Regina
Regina, SK S4S 0A2
Canada

1991 Mathematics Subject Classification. Primary 15A24, 93B40; Secondary 47B60, 65H10

Received