CONVERGENCE ANALYSIS FOR SPLITTING OF THE ABSTRACT RICCATI EQUATION

ESKIL HANSEN∗ AND TONY STILLFJORD†

Abstract. We consider a splitting-based approximation of the abstract Riccati equation in the setting of Hilbert–Schmidt operators. The Riccati equation arises in many different areas and is important within the field of optimal control. While convergence of different methods for approximating the Riccati equation is discussed in several studies, none of them rigorously prove an order of convergence. In this paper we conduct such a convergence analysis and show that the splitting method converges with the same order as the implicit Euler scheme. As it requires no Newton iterations we expect it to be more efficient than common non-splitting methods. We also show that the splitting method preserves low-rank structure in the matrix-valued case, which is essential for large-scale problems. A numerical experiment demonstrates the validity of our theory.

Key words. Abstract Riccati equation, splitting, convergence order, low-rank approximation, Hilbert–Schmidt operators

AMS subject classifications. 65M12, 47H06, 49M30

1. Introduction. We consider the abstract Riccati equation:

\begin{equation}
\dot{P}(t) + A^*P(t) + P(t)A + P(t)^2 = Q, \quad t \in (0, T),
\end{equation}

\begin{equation}
P(0) = P_0.
\end{equation}

This is a semi-linear operator-valued evolution equation for \( P \), where \( A \) and \( Q \) are given (unbounded) linear operators. A prototypical \( A \) would be an elliptic differential operator.

The Riccati equation arises in many different areas, for example in the field of optimal control. Within this field, two important applications are linear quadratic regulator problems and stochastic filtering problems. In the former, one aims to steer the solution of \( \dot{x} = Ax \) to a desired state by adding a perturbation \( u \), the control input. Under certain quadratic constraints the solution to the Riccati equation provides a relation between the state and the optimal input. See [12] for an in-depth treatment. In stochastic filtering, one tries to find the best possible estimate of the state when it is perturbed by random noise. In this case, the solution to the Riccati equation is the covariance of the error of the optimal estimator. For more information see e.g. [1, 8].

Previous approaches to approximate the solution of the Riccati equation (1.1) include spatial Galerkin methods [10, 17], temporal BDF and Rosenbrock methods [5] and temporal first-order splitting methods [3, 21]. While these studies show that the respective methods converge, they lack a convergence analysis which describes how quickly this convergence occurs.

It has also been noted that the solutions to the matrix-valued Riccati equation, for example arising after a spatial discretization, frequently exhibit low-rank structure. Apart from the study [15] there is to the best of our knowledge no theory for predicting precisely when such features exist. Nevertheless, for large-scale Riccati equations it...
is vital to exploit such structure, in order to avoid that the computational time and memory storage requirements become unfeasible.

In light of these observations, the aim of this study is twofold. First, we aim to introduce an efficient approximation scheme which can be given a rigorous convergence order analysis in a standard abstract setting, e.g. the Hilbert–Schmidt operator framework presented by Temam [21]. Secondly, we strive to find a scheme which preserves possible low-rank structure of the solution to the Riccati equation.

To this end, we propose the usage of a (formally) first-order splitting scheme which is more effective than comparable non-splitting methods, since it does not require any Newton iterations. In order to introduce our scheme, we define the operators
\[
F P = A^* P + PA - Q \quad \text{and} \quad G P = P^2.
\]

The two sub-problems of interest are now
\[
\dot{P} + FP = 0, \quad P(0) = P_0 \quad \text{and} \quad \dot{P} + GP = 0, \quad P(0) = P_0,
\]
where (1.4) is affine and (1.5) can be solved exactly. The time-stepping operator \(S_h\) of our splitting scheme is then given by
\[
S_h = (I + hF)^{-1} e^{-hG},
\]
and \(S_h P_0\) is an approximation to \(P(nh)\).

An outline of the paper is as follows. In Section 2 we describe the abstract setting in which we treat the Riccati equation, and recall some properties of the affine and nonlinear parts of the equation. The main theorem is proved in Section 3 and shows that the splitting method and the implicit Euler scheme converge with the same order. In Section 4 we consider an implementation of the splitting method that preserves low-rank structure in the matrix-valued case, and this is applied to a Riccati equation arising from a linear quadratic regulator problem in Section 5.

2. Abstract framework for the Riccati equation. We start by fixing the notation. Given a real Hilbert space \(X\), we denote its inner product by \((\cdot, \cdot)_X\) and its norm by \(\|\cdot\|_X\). The dual space of \(X\) is denoted \(X^*\), and we write the dual pairing between \(u \in X^*\) and \(v \in X\) as \(\langle u, v \rangle_{X^* \times X}\). The space of linear bounded operators from \(X\) to another Hilbert space \(Y\) is denoted by \(L(X, Y)\).

With this in place, let the Hilbert space \(V\) be densely and continuously embedded in the Hilbert space \(H\), which gives the usual Gelfand triple
\[
V \hookrightarrow H \cong H^* \hookrightarrow V^*.
\]
To define a class of suitable operators \(A\) and \(A^*\) we introduce a bilinear form \(a: V \times V \to \mathbb{R}\), satisfying the following:

**Assumption 1.** The bilinear form \(a: V \times V \to \mathbb{R}\) is bounded and coercive, i.e. there exists positive constants \(C_1, C_2\) such that for all \(u, v \in V\)
\[
|a(u, v)| \leq C_1 \|u\|_V \|v\|_V \quad \text{and} \quad a(u, u) \geq C_2 \|u\|_V^2.
\]

The operators \(A \in L(V, V^*)\) and \(A^* \in L(V, V^*)\) are then given by
\[
\langle Au, v \rangle_{V^* \times V} = a(u, v) \quad \text{and} \quad \langle A^* u, v \rangle_{V^* \times V} = a(v, u).
\]
Example 1. Let $\Omega$ be an open, bounded subset of $\mathbb{R}^d$ with a sufficiently regular boundary. Take $H = L^2(\Omega)$ and let $V$ be either $H^1(\Omega)$, $H^1_0(\Omega)$ or $H^1_{per}(\Omega)$ depending on boundary conditions. Further assume that $\alpha \in C(\overline{\Omega})$ is a positive function. Then with

$$a(u, v) = (\sqrt{\alpha} \nabla u, \sqrt{\alpha} \nabla v)_H$$

the above construction yields the anisotropic diffusion operator $A = -\nabla \cdot (\alpha \nabla u)$.

Consider now the Riccati equation (1.1). Requiring that $Q$ belongs to $\mathcal{L}(V, V^*)$ yields solutions $P(t)$ in $L(H, V) \cap L(V^*, H)$. For the analysis in this paper, we will restrict ourselves to the case when both $P(t)$ and $Q$ are self-adjoint, positive semi-definite Hilbert–Schmidt operators. This setting was for example advocated by Temam [21]. Considering the kind of applications giving rise to Riccati equations, this is a reasonable restriction. For example, in the introductory example regarding stochastic filtering, covariances are always positive semi-definite and self-adjoint.

We proceed to recap a few basic properties of these classes of operators. See e.g. [17] and [2, Sections II:3.3 and III:2.3] for a complete exposition. Let $H_i$ denote generic separable Hilbert spaces. An operator $F \in \mathcal{L}(H_1, H_2)$ is said to be Hilbert–Schmidt if

$$\sum_{k=1}^{\infty} (F e_k, F e_k)_{H_2} < \infty,$$

where $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis of $H_1$. We denote the space of all Hilbert–Schmidt operators from $H_1$ to $H_2$ by $\mathcal{HS}(H_1, H_2)$ and note that this is a Hilbert space when equipped with the inner product

$$(F, G)_{\mathcal{HS}(H_1, H_2)} = \sum_{k=1}^{\infty} (F e_k, G e_k)_{H_2}.$$  

The corresponding induced Hilbert–Schmidt norm is denoted $\|\cdot\|_{\mathcal{HS}(H_1, H_2)}$.

It is clear that the Hilbert–Schmidt norm is denoted stronger than the operator norm, and in fact

$$\|F\|_{\mathcal{L}(H_1, H_2)} \leq \|F\|_{\mathcal{HS}(H_1, H_2)}.$$ 

Further, Hilbert–Schmidt operators are invariant under multiplication by linear bounded operators from both the left and from the right. That is, if $F \in \mathcal{HS}(H_2, H_3)$ and $G \in \mathcal{L}(H_1, H_2)$ then $FG \in \mathcal{HS}(H_1, H_3)$ and

$$\|FG\|_{\mathcal{HS}(H_1, H_3)} \leq \|F\|_{\mathcal{HS}(H_2, H_3)} \|G\|_{\mathcal{L}(H_1, H_2)}.$$ 

Similarly, if $F \in \mathcal{HS}(H_1, H_2)$ and $G \in \mathcal{L}(H_2, H_3)$ then $GF \in \mathcal{HS}(H_1, H_3)$ and

$$\|GF\|_{\mathcal{HS}(H_1, H_3)} \leq \|F\|_{\mathcal{HS}(H_1, H_2)} \|G\|_{\mathcal{L}(H_2, H_3)}.$$ 

Based on this, we define the spaces

$$\mathcal{V} = \mathcal{HS}(H, V) \cap \mathcal{HS}(V^*, H) \quad \text{and} \quad \mathcal{H} = \mathcal{HS}(H, H).$$

These can be shown to give rise to a new Gelfand triple

$$\mathcal{V} \hookrightarrow \mathcal{H} \cong \mathcal{H}^* \hookrightarrow \mathcal{V}^*,$$
where \( V^* \) is identified with \( \mathcal{H}S(V, H) + \mathcal{H}S(H, V^*) \) [17, 21]. If \( P \in V \) then \( A^*P \in \mathcal{H}S(H, V^*) \) and \( PA \in \mathcal{H}S(V, H) \), i.e. \( A^*P + PA \in V^* \). The operator \( P \mapsto A^*P + PA - Q \) thus belongs to \( \mathcal{L}(V, V^*) \) and we consider the related Friedrich extension \( F : \mathcal{D}(F) \subset \mathcal{H} \to \mathcal{H} \), defined by

\[
\mathcal{D}(F) = \{ P \in V : A^*P + PA - Q \in \mathcal{H} \} \quad \text{and} \quad \mathcal{F}P = A^*P + PA - Q \quad \text{for all} \ P \in \mathcal{D}(F).
\]

To simplify the notation, we also introduce the subset \( C \subset \mathcal{H} \) of self-adjoint positive semi-definite operators:

\[
\mathcal{C} = \{ P \in \mathcal{H} : P = P^* \text{ and } (Pu, u)_H \geq 0 \text{ for all } u \in H \}
\]

Finally, define the operator \( G : \mathcal{D}(G) \subset \mathcal{H} \to \mathcal{H} \) by

\[
\mathcal{D}(G) = \mathcal{C} \quad \text{and} \quad GP = P^2 \quad \text{for all} \ P \in \mathcal{D}(G).
\]

We summarize now some important properties of the operators \( F \) and \( G \) and their sum. First recall that an operator \( F : \mathcal{D}(F) \subset X \to X \) on a real Hilbert space \( X \) is accretive if

\[
(Fu - Fv, u - v)_X \geq 0
\]

for all \( u \) and \( v \) in \( \mathcal{D}(F) \).

**Lemma 2.1.** Under Assumption 1, the operators \( F \), \( G \) and \( F + G \) with \( \mathcal{D}(F + G) = \mathcal{D}(F) \cap \mathcal{C} \) are all accretive. Further, for all \( h > 0 \) we have \( \mathcal{R}(I + hF) = \mathcal{H} \), and the closed, convex set \( \mathcal{C} \) is contained in the sets \( \mathcal{R}(I + hG) \) and \( \mathcal{R}(I + h(F + G)) \).

This is proved in [2, II:3.3]. The following lemma [7, Theorem I] demonstrates the implications of these properties.

**Lemma 2.2.** Let \( F : \mathcal{D}(F) \subset X \to X \) be an accretive operator that satisfies

\[
\overline{\mathcal{D}(F)} \subset \mathcal{R}(I + hF)
\]

for all \( h > 0 \). Then the limit

\[
e^{-tF}u = \lim_{n \to \infty} (I + t/nF)^{-n}u
\]

exists for all \( u \in \overline{\mathcal{D}(F)} \), \( t \geq 0 \), and defines a (nonlinear) semigroup.

The proof is based on the fact that the implicit Euler scheme generates a Cauchy sequence satisfying

\[
\|(I + t/nF)^{-n}u - (I + t/mF)^{-m}u\|_X \leq C(1/n - 1/m)^{1/2}\|Fu\|_X,
\]

as shown in [7, Equation 1.10].

The mapping \( t \mapsto e^{-tF}u_0 \) is called a mild solution to the abstract evolution equation \( \dot{u} + Fu = 0 \), \( u(0) = u_0 \), and it is a continuous function of \( t \). It follows from Lemma 2.1 and Theorem 2.2 that there is a mild solution to the Riccati equation (1.1) if \( Q \in \mathcal{H} \), as well as to the sub-problems (1.4) and (1.5), all with suitable initial conditions.

In the general case, a strong solution coincides with \( e^{-tF}u_0 \), if, for example, the range condition is strengthened to \( \text{conv} \overline{\mathcal{D}(F)} \subset \mathcal{R}(I + hF) \) for all \( h > 0 \), see e.g. [6, Corollary 2.1]. This is certainly true in our case, as the domains of \( G \) and \( F + G \) are contained in the convex set \( C \). Proof of existence of a strong solution and further regularity results can be found in [2, Theorem III:2.8].
3. Convergence analysis. We consider now the approximation of (1.1) by the splitting scheme (1.6). We prove convergence in four steps. The first step is a related result for the implicit Euler method given by the time-stepping operator $R_h$:

\[ R_h = (I + h(F + G))^{-1}. \]  

**Lemma 3.1.** Let Assumption 1 be valid, assume that $P_0 \in D(F) \cap C$ and let $Q \in C$. Then the implicit Euler scheme given by $R_h$ converges to the solution of the Riccati equation (1.1). More specifically,

\[ \| e^{-nh(F+G)}P_0 - R_h^n P_0 \|_{\mathcal{H}} \leq C h^q \| (F + G) P_0 \|_{\mathcal{H}}, \quad 0 \leq nh \leq T, \]

with $q \in [1/2, 1]$. The constant $C$ depends on $T$, but not on $n$ or $h$ separately.

**Proof.** This follows by combining Lemma 2.1 with Lemma 2.2 together with the bound (2.1). \qed

We note here that this order of convergence is optimal in the general nonlinear case as illustrated in [18, Example 3].

Next, we consider just the nonlinear part (1.5) of the full problem.

**Lemma 3.2.** The solution to (1.5) is given by

\[ e^{-tG} P_0 = (I + tP_0)^{-1} P_0, \]

where $(I + tP_0)^{-1} P_0$ denotes the composition of two operators in $\mathcal{L}(H, H)$. Furthermore, the function $t \mapsto e^{-tG} P_0$ belongs to $C_\infty(0, T; \mathcal{H})$.

**Proof.** Assume that

\[ P(t) = (I + tP_0)^{-1} P_0. \]

As $P_0$ is an element of $C$ it is both self-adjoint and compact, since all Hilbert–Schmidt operators are compact. By the Hilbert–Schmidt spectral theorem [20, Theorem VI.16], one therefore has the representation

\[ (I + tP_0)^{-1} v = \sum_{k=1}^{\infty} \frac{1}{1 + t \lambda_k} (v, e_k) e_k, \]

where $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis for $H$ and $\{\lambda_k\}_{k=1}^{\infty}$ are the eigenvalues of $P_0$. Since $P_0$ belongs to $C$ it is positive semi-definite and one has that $\lambda_k \geq 0$ for all $k \geq 1$. Hence,

\[ \|(I + tP_0)^{-1}\|_{\mathcal{L}(H, H)} \leq 1, \]

for all $t \geq 0$. This implies that

\[
\|P(t + h) - P(t)\|_{\mathcal{H}} = \| (I + (t + h)P_0)^{-1} - (I + tP_0)^{-1} (I + (t + h)P_0) (I + tP_0)^{-1} P_0 \|_{\mathcal{H}} \\
= \| - h (I + (t + h)P_0)^{-1} P_0 (I + tP_0)^{-1} P_0 \|_{\mathcal{H}} \\
\leq h \| (I + (t + h)P_0)^{-1} \|_{\mathcal{L}(H, H)} \| P_0 \|_{\mathcal{L}(H, H)} \| (I + tP_0)^{-1} \|_{\mathcal{L}(H, H)} \| P_0 \|_{\mathcal{H}} \\
\leq h \| P_0 \|_{\mathcal{H}}^2,
\]

which tends to zero as $h$ tends to zero. Hence, $t \mapsto P(t)$ is continuous in $\mathcal{H}$. By the same construction we obtain that

\[
\lim_{h \to 0} \| (P(t + h) - P(t))/h + P(t)P(t) \|_{\mathcal{H}} = 0,
\]
i.e. \( t \mapsto P(t) \) is continuously differentiable and satisfies the equation (1.5). By application of the chain rule,

\[
\dot{P} = -\frac{d}{dt}P^2 = -\dot{P}P - PP = 2P^3,
\]

and by repeating this procedure we see that all higher derivatives can be rewritten as powers of \( P \). Since \( P \) is continuous, this observation proves the claim that \( t \mapsto e^{-tG}P_0 \) belongs to \( C^\infty(0; T; \mathcal{H}) \).

We also require the following lemma, to which we include the short proof for convenience:

**Lemma 3.3.** Let \( X \) be a real Hilbert space. Suppose \( F : \mathcal{D}(F) \subset X \to X \) is an accretive operator. Then the resolvent \((I + hF)^{-1}\) is non-expansive on its domain for all \( h > 0 \), i.e.

\[
\| (I + hF)^{-1}u - (I + hF)^{-1}v \|_X \leq \| u - v \|_X \quad \text{for all } u, v \in \mathcal{R}(I + hF).
\]

**Proof.** Since \( F \) is accretive, for \( u, v \in \mathcal{D}(F) \) we have

\[
\| (I + hF)u - (I + hF)v \|_X^2 = \| u - v \|_X^2 + 2h(Fu - Fv, u - v)_H + h^2\| Fu - Fv \|_X^2 \geq \| u - v \|_X^2.
\]

If instead \( u, v \in \mathcal{R}(I + hF) \) then \((I + hF)^{-1}u \) and \((I + hF)^{-1}v \) are both in \( \mathcal{D}(F) \). Replacing \( u \) and \( v \) by these elements in the above inequality proves the lemma. \( \square \)

**Theorem 3.4.** Let Assumption 1 be valid, assume that \( P_0 \in \mathcal{D}(F) \cap C \) and let \( Q \in C \). Then the splitting scheme given by \( S_h \) (1.6) converges to the solution of the Riccati equation (1.1). More specifically,

\[
\| e^{-nh(F + \mathcal{G})}P_0 - S_h^n P_0 \|_H \leq C(h + h^q), \quad 0 \leq nh \leq T,
\]

where \( q \in [1/2, 1] \) is the convergence order of the implicit Euler scheme, as in Lemma 3.1. The constant \( C \) depends on \( T \), \( \| (F + \mathcal{G})P_0 \|_H \) and \( \| P_0 \|_H \), but not on \( n \) or \( h \) separately.

**Proof.** From Lemma 3.1 and the inequality

\[
\| e^{-nh(F + \mathcal{G})}P_0 - S_h^n P_0 \|_H \leq \| e^{-nh(F + \mathcal{G})}P_0 - R_h^n P_0 \|_H + \| R_h^n P_0 - S_h^n P_0 \|_H,
\]

it follows that we only need to consider the distance between the splitting approximation and the implicit Euler approximation.

By Lemma 2.1 and Lemma 3.3 the operators \((I + hF)^{-1}\) and \( e^{-h\mathcal{G}} \) are non-expansive, which implies that so is \( S_h \). Hence,

\[
\| R_h^n P_0 - S_h^n P_0 \|_H = \sum_{j=1}^n \| S_h^{n-j}R_h^j P_0 - S_h^{n-j+1}R_h^j P_0 \|_H \leq \sum_{j=1}^n \| R_h^j P_0 - S_h R_h^{j-1} P_0 \|_H \leq \sum_{j=1}^n \| (I + hF) R_h^j P_0 - e^{-h\mathcal{G}} R_h^{j-1} P_0 \|_H.
\]
We now consider the terms of the sum separately. From Lemma 3.2 it follows that for any \( Z \in \mathcal{C} \) we can make the expansion
\[
e^{-h\mathcal{G}} Z = Z - h\mathcal{G}Z + h^2 R,
\]
where
\[
R = \int_0^1 (1-t) \frac{d^2}{dt^2} e^{-t\mathcal{G}} Z dt = \int_0^1 2(1-t)((I+tZ)^{-1}Z)^3 dt
\]
is bounded in \( \mathcal{H} \) by \( 2\|Z\|_H^3 \). Hence,
\[
\|(I+h\mathcal{F})\mathcal{R}_h Z - e^{-h\mathcal{G}} Z\|_H = \|Z - h\mathcal{G}\mathcal{R}_h Z - (Z - h\mathcal{G}Z + h^2 R)\|_H
\leq h\|Z^2 - (\mathcal{R}_h Z)^2\|_H + 2h^2\|Z\|_H^3
= h\|(\mathcal{R}_h Z - Z)\mathcal{R}_h Z + Z(\mathcal{R}_h Z - Z)\|_H + 2h^2\|Z\|_H^3
\leq h\|(\mathcal{R}_h Z - Z)\|_H (\|\mathcal{R}_h Z\|_H + \|Z\|_H) + 2h^2\|Z\|_H^3.
\]
Again by Lemma 2.1 and Lemma 3.3, the operator \( \mathcal{R}_h \) is non-expansive, so with \( Z = \mathcal{R}_h^{j-1}P_0 \) we see that
\[
\|\mathcal{R}_h^j P_0 - \mathcal{R}_h^{j-1}P_0\|_H \leq \|\mathcal{R}_h P_0 - P_0\|_H
= \|\mathcal{R}_h P_0 - \mathcal{R}_h (I + h(\mathcal{F} + \mathcal{G}))P_0\|_H
\leq \|P_0 - (I + h(\mathcal{F} + \mathcal{G}))P_0\|_H
\leq h\|(\mathcal{F} + \mathcal{G})P_0\|_H,
\]
and we also have
\[
\|\mathcal{R}_h^j P_0\|_H \leq \|P_0\|_H + \sum_{k=1}^i \|\mathcal{R}_h^k P_0 - \mathcal{R}_h^{k-1}P_0\|_H \leq \|P_0\|_H + ih\|(\mathcal{F} + \mathcal{G})P_0\|_H.
\]
Thus, \( \|\mathcal{R}_h^j P_0\|_H \) and \( \|\mathcal{R}_h^{j-1}P_0\|_H \) are both bounded by a constant depending on \( \|P_0\|_H \), \( \|(\mathcal{F} + \mathcal{G})P_0\|_H \) and \( T \). This implies that
\[
\|(I+h\mathcal{F})\mathcal{R}_h^j P_0 - e^{-h\mathcal{G}}\mathcal{R}_h^{j-1}P_0\|_H \leq Ch^2,
\]
where the constant \( C \) depends on the same quantities. Inserting this into the telescopic sum (3.2) yields the desired bound. \( \Box \)

It should be noted that one could apply a spatial discretization to the abstract equation and analyze the resulting matrix-valued differential equation to obtain convergence results. However, the usual analysis based on Taylor expansions leads to error bounds that depend on the discretization parameters, and when the discretization is refined these bounds may tend to infinity. This is not the case for the above results, which yield uniform error bounds with respect to the spatial discretization parameter.

4. Implementation and preservation of low rank. We consider now the implementation of the splitting method (1.6). In the case when \( A \) is an elliptic partial differential operator, a straightforward discretization of Equation (1.1) would quickly lead to huge equation systems. Consider for example the linear quadratic regulator example given in the introduction. Assuming that the state \( x(t) \) is a function defined
on a subset Ω of R\(^d\) and using finite differences to discretize it with \(n\) points in each dimension leads to a solution with \(n^d\) elements. Already with \(d = 3\), representing the solution as a dense vector requires an inordinate amount of memory even for moderate values of \(n\). In our case, however, the solution is the operator \(P(t)\), which if discretized in the same way would require a matrix with \(n^{2d}\) elements. Except for in the uninteresting cases, on current computer architectures this is unfeasible.

However, as stated in the introduction, the solutions to the matrix-valued Riccati equation frequently exhibit low-rank behaviour. Throughout the rest of this section we assume that a spatial discretization has been made, so that the abstract Riccati equation becomes a matrix-valued Riccati equation. That is, now \(H = R\) for some integer \(n > 0\) and \(P(t)\) is an element of \(R^{n \times n}\). By a low-rank representation we mean that \(P(t) = zz^T\) where \(z \in R^{n \times m}\), with \(m \ll n\). We first show that the discretized version of

\[
e^{-hG}P = (I + hP)^{-1}P
\]

preserves such low-rank structure.

**Lemma 4.1.** Assume that the matrix \(P\) satisfies \(P = zz^T\) where \(z \in R^{n \times m}\). Then for all \(h > 0\) it holds that

\[
(I + hP)^{-1}P = ww^T,
\]

where \(w \in R^{n \times m}\).

**Proof.** We will employ a special case of the Woodbury matrix inversion formula which states that for matrices \(Y\) and \(Z\) of appropriate dimensions one has

\[
(I + YZ)^{-1} = I - Y(I + ZY)^{-1}Z.
\]

This can be easily verified by simply multiplying from the left and from the right by \(I + YZ\). Denote now by \(I_k\) the identity matrix in \(R^{k \times k}\). Taking \(Y = hz\) and \(Z = z^T\) we see that

\[
(I_n + hzz^T)^{-1}zz^T = zz^T - hz(I_m + z^Thz)^{-1}z^Tzz^T
\]

\[
= z(I_m - (I_m + hz^Tz)^{-1}hz^Tz)z^T
\]

\[
= z(I_m - I_m + (I_m + hz^Tz)^{-1}z^T)
\]

\[
= z(I_m + hz^Tz)^{-1}z^T.
\]

Since \(z^Tz\) is a positive semi-definite matrix, one obtains that the matrix \(I_m + hz^Tz\) is positive definite for any \(h > 0\). Hence, it can be Cholesky factorized as

\[
I_m + hz^Tz = LL^T,
\]

where \(L\) is a lower-triangular invertible matrix. This means that

\[
(I_n + hzz^T)^{-1}zz^T = zL^{-T}L^{-1}z^T = (zL^{-T})(zL^{-T})^T = ww^T,
\]

where \(wL^T = z\). 

The method described in the proof of Lemma 4.1 immediately suggests an efficient algorithm to compute the low-rank factor \(w\) of \((I + hP)^{-1}P\), which only involves operations on, and with, small \(m \times m\) matrices.
The second part of the splitting scheme (1.6) is the computation of the action of \((I + hF)^{-1}\) on \(e^{-hG}P\). Assume that \((I + hF)S = P\). This means that \(S + hA^*S + hSA - hQ = P\). But this is equivalent to the Lyapunov equation

\[(I + 2hA)^*S + S(I + 2hA) = 2P + 2hQ.\]

If both \(P\) and \(Q\) have low rank, then so has the right-hand side, and we can therefore apply standard low-rank solvers for computing \((I + hF)^{-1}P\). We refer to [4, 11, 16] for methods based on the alternating direction iteration (ADI) and to [9, 19] for methods based on Krylov subspaces.

The main point here is that the standard method to solve the algebraic Riccati equation is to employ Newton (Kleinman) iteration, which requires the solution of a Lyapunov equation in every step. Here, we only need to solve one Lyapunov equation instead of several, which is expected to substantially increase the efficiency of the method. Note that, as the nonlinear flow preserves the rank exactly, the only possible increase in rank is due to the approximation related to the affine part of the vector field.

5. Numerical examples. As an example, we study a linear quadratic regulator problem, where the goal is to minimize the functional

\[J(x, u) = \int_0^T \|Cx - x_d\|^2 + \|u\|^2 \, dt\]

subject to the state equation

\[\dot{x} = Ax + u.\]

The variable \(x\) is the state, \(x_d\) is the observation of the desired state, \(u\) is the control input and \(C\) is the observation operator. We require \(C\) to be a Hilbert–Schmidt operator. It can be proved [12] that the optimal control strategy is an affine mapping, \(u(t) = -P(T-t)x(t) - r(T-t)\), where \(P\) and \(r\) satisfies

\[
\begin{aligned}
\dot{P} + PA + A^*P + P^2 &= C^*C, \quad P(0) = 0, \\
\dot{r} + A^*r + Pr &= -C^*x_d, \quad r(0) = 0.
\end{aligned}
\]

The first of these equations is of the form (1.1), with \(Q = C^*C\), and we will approximate its solution numerically.

We choose to work in the setting of Example 1, with \(\Omega = (0, 1)\), periodic boundary conditions and \(\alpha(x) = 2 + \cos 2\pi x\). To define \(C\), we choose first the real trigonometric orthonormal basis for \(H\): \(\{e_k\}_{k=0}^\infty \cup \{f_k\}_{k=0}^\infty\), where

\[e_k(x) = \sqrt{2} \cos(2\pi k x) \quad \text{and} \quad f_k(x) = \sqrt{2} \sin(2\pi k x).\]

Then we set

\[C\left(\sum_{k=0}^\infty a_k e_k + b_k f_k\right) = \sum_{k=0}^m a_k e_k + b_k f_k,\]

for a small \(m\), i.e. we simply truncate the sum. Then \(C\) is clearly Hilbert–Schmidt and it can be thought of as representing measuring equipment that can only measure low-frequency signals. The product \(Q = C^*C\) is also Hilbert–Schmidt, so the assumptions in Theorem 3.4 are fulfilled.
We discretize the problem by standard second-order finite differences and $2M + 1$ nodes in space, where we take $M = 500$. The discretization of $C$ also has a natural low-rank factorization, $zz^T$, where $z$ is a matrix of dimension $(2M + 1) \times (2m + 1)$. In order to work in the same basis we instead consider $E^T zz^T E$, where $E$ denotes the orthogonal transformation matrix between the two different bases. Let $Q_M$ be the discretization of $Q$. Since

$$Q_M = (E^T zz^T E)^T (E^T zz^T E) = E^T zz^T zz^T E = E^T z^T z E,$$

we can also low-rank factorize $Q_M = ww^T$ with $w = (zE)^T$. For this experiment we choose $m = 3$, which yields a matrix of low rank.

Figure 5.1 (left) shows that the splitting method (1.6) converges with order $q = 1$ when applied to the problem described above. This result agrees with Theorem 3.4. The errors are measured in the Frobenius norm $\| \cdot \|_{F,0}$ scaled by $1/(2M + 1)$, which is the discretized analogue of the Hilbert–Schmidt norm. To solve the Lyapunov equations involved in computing the action of $(I + hF)^{-1}$ we have used a modified version of LyaPack 1.8 [13] with a normalized residual tolerance of $10^{-6}$ in the ADI iterations. Finally, the right plot in Figure 5.1 demonstrates that the rank of the approximation stays low throughout the integration.

**Fig. 5.1.** Left: The normalized errors $\| S^n_{h}P_0 - P_{ref}\|_{F,0}/\|P_{ref}\|_{F,0}$ when approximating the solution to (5.1) for different $h = 1/N$ with $N = 2, 4, \ldots, 512$. The reference solution $P_{ref}$ was also computed by the splitting method, albeit with a finer temporal step size of $h = 1/2048$. The spatial discretization has $2M + 1 = 1001$ nodes. Right: The rank of $S^n_{h}P_0$ for $n = 1, 2, \ldots, 512$, with $h = 1/512$.

**REFERENCES**


SPLITTING THE ABSTRACT RICCATI EQUATION