CONVERGENCE ANALYSIS FOR SPLITTING OF THE ABSTRACT DIFFERENTIAL RICCATI EQUATION

ESKIL HANSEN* AND TONY STILLFJORD†

Abstract. We consider a splitting-based approximation of the abstract differential 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. In this paper we conduct a temporal error analysis and prove that the splitting method converges with the same order as the implicit Euler scheme, under the same low regularity requirements on the initial values. For a subsequent spatial discretization, the abstract setting also yields uniform temporal error bounds with respect to the spatial discretization parameter. The spatial discretizations commonly lead to large-scale problems, where the use of structural properties of the solution is essential. We therefore conclude by proving that the splitting method preserves low-rank structure in the matrix-valued case. Numerical results demonstrate the validity of the convergence analysis.

Key words. Abstract differential 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

$$\begin{aligned} \dot{P}(t) + A^*P(t) + P(t)A + P(t)^2 = Q, \quad t \in (0,T), \\ P(0) = P_0. \end{aligned} \tag{1.1}$$

This is a semi-linear operator-valued evolution equation for P, where A and Q are given 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 $\dot{x} + Ax = 0$ 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 [13] 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. [2, 10].

Previous approaches to approximate the solution of the infinite-dimensional Riccati equation (1.1) include spatial Galerkin methods [11, 16], temporal BDF and Rosenbrock methods [6] and temporal first-order splitting methods [4, 21]. While these studies show that the respective methods converge, they lack a convergence analysis which describes how quickly the convergence occurs.

It has also been noted that the solutions to the matrix-valued Riccati equation, for example arising after a spatial discretization, can often be closely approximated by a matrix-valued function of low rank. Apart from the papers [1, 14] there is to the best of our knowledge no theory for predicting precisely when such low-rank

* Centre for Mathematical Sciences, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden (eskil@maths.lth.se). The work of the first author was supported by the Swedish Research Council under grant 621-2011-5588.

† Centre for Mathematical Sciences, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden (tony@maths.lth.se).

structure exists. Nevertheless, for large-scale Riccati equations it is vital to exploit such structure, in order to avoid unfeasible computational times and memory storage requirements.

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 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, whose efficiency stems from the fact that it does not have to solve any nonlinear equations. In order to introduce our scheme, we define the operators

$$\mathcal{F}P = A^*P + PA - Q \quad \text{and} \tag{1.2}$$

$$\mathcal{G}P = P^2. \tag{1.3}$$

The two sub-problems of interest are now

$$\dot{P} + \mathcal{F}P = 0, \quad P(0) = P_0 \quad \text{and} \tag{1.4}$$

$$\dot{P} + \mathcal{G}P = 0, \quad P(0) = P_0, \tag{1.5}$$

where (1.4) is affine and (1.5) can be solved exactly. The time-stepping operator $\mathcal{S}_h$ of our splitting scheme is then given by

$$\mathcal{S}_h = (I + h\mathcal{F})^{-1}\mathrm{e}^{-h\mathcal{G}}, \tag{1.6}$$

and $\mathcal{S}_h^n P_0$ 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 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 $\mathcal{L}(X,Y)$. The (possibly infinite) Lipschitz constant of a generic nonlinear map $F : \mathcal{D}(F) \subset X \to X$ is denoted by $L[F]$. In the following, we assume that all occuring Hilbert spaces are real and separable.

With this in place, let the Hilbert space V be densely and compactly 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 $u, v \in V$*

$$|a(u,v)| \le C_1\|u\|_V\|v\|_V \quad \textit{and} \quad a(u,u) \ge C_2\|u\|_V^2.$$

The operators $A \in \mathcal{L}(V, V^*)$ and $A^* \in \mathcal{L}(V, V^*)$ 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 Ω be an open, bounded subset of $\mathbb{R}^d$ with a sufficiently regular boundary. Take $H = L^2(\Omega)$ and let V be either $H_0^1(\Omega)$, $H^1(\Omega)$ or $H_{per}^1(\Omega)$ depending on boundary conditions. Further assume that $\alpha \in C(\overline{\Omega})$ is a positive function. Then with $\lambda > 0$ (or $\lambda \geq 0$ for the Dirichlet case) and*

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

the above construction yields the diffusion operator $A = -\nabla \cdot \big(\alpha\nabla u\big) + \lambda I$.

Consider now the Riccati equation (1.1). 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. [3, Sections II:3.3 and III:2.3] and [16, 21] for a complete exposition. Let H_i denote generic Hilbert spaces. An operator $F \in \mathcal{L}(H_1, H_2)$ is said to be Hilbert–Schmidt if

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

where $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis of H_1. Note that the definition is independent of the choice of the basis. 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

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

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

It is clear that the Hilbert–Schmidt norm is 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 composition with linear bounded operators from both the left and from the right. That is, if $F \in \mathcal{HS}(H_2, H_3)$, $G_1 \in \mathcal{L}(H_1, H_2)$ and $G_2 \in \mathcal{L}(H_3, H_4)$ then $G_2FG_1 \in \mathcal{HS}(H_1, H_4)$ and

$$\|G_2FG_1\|_{\mathcal{HS}(H_1,H_4)} \leq \|G_2\|_{\mathcal{L}(H_3,H_4)}\|F\|_{\mathcal{HS}(H_2,H_3)}\|G_1\|_{\mathcal{L}(H_1,H_2)}.$$

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 $\mathcal{V}^*$ is identified with $\mathcal{HS}(V,H)+\mathcal{HS}(H,V^*)$ and the inclusions are dense and continuous. If $P\in\mathcal{V}$ then $A^*P\in\mathcal{HS}(H,V^*)$ and $PA\in\mathcal{HS}(V,H)$, i.e. $A^*P+PA\in\mathcal{V}^*$. The operator $P\mapsto A^*P+PA$ thus belongs to $\mathcal{L}(\mathcal{V},\mathcal{V}^*)$ and we consider the related perturbed restriction $\mathcal{F}:\mathcal{D}(\mathcal{F})\subset\mathcal{H}\to\mathcal{H}$, defined by

$$\begin{aligned}\mathcal{D}(\mathcal{F})&=\{P\in\mathcal{V};\ A^*P+PA-Q\in\mathcal{H}\}\quad\text{and}\\ \mathcal{F}P&=A^*P+PA-Q\quad\text{for all } P\in\mathcal{D}(\mathcal{F}).\end{aligned}$$

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

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

We take the nonlinearity of the Riccati equation to be defined on this set, i.e.

$$\mathcal{G}:\mathcal{C}\to\mathcal{H}:P\mapsto P^2,$$

and let the domain of the full operator $\mathcal{F}+\mathcal{G}$ be $\mathcal{D}(\mathcal{F})\cap\mathcal{C}$.

Example 2. *In the context of Example 1, an operator $P\in\mathcal{H}$ can be identified as an integral operator of the form*

$$(Pu)(x)=\int_\Omega p(x,\xi)u(\xi)\,\mathrm{d}\xi,\quad \text{a.e. on }\Omega,$$

with the kernel $p\in L^2(\Omega\times\Omega)$ and $u\in H$. Further, for the case $V=H_0^1(\Omega)$ the space $\mathcal{V}$ can similarly be characterized by integral operators with kernels in $H_0^1(\Omega\times\Omega)$, see e.g. [16, Section 5] and [21, Example 1]. If the kernel is additionally in $H^2\cap H_0^1(\Omega\times\Omega)$, the function α is sufficiently smooth and $Q\in\mathcal{H}$, the corresponding operator P belongs to $\mathcal{D}(\mathcal{F})$. Finally, elements of the set $\mathcal{C}$ can be identified with symmetric and nonnegative kernels in $L^2(\Omega\times\Omega)$.

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

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

for all u and v in $\mathcal{D}(F)$. A direct consequence of F being accretive is that the corresponding resolvent is nonexpansive, i.e. $L[(I+hF)^{-1}]\leq 1$ for all $h>0$. Under the additional assumption that $\overline{\mathcal{D}(F)}\subset\mathcal{R}(I+hF)$ for all $h>0$ it can further be shown [9, Theorem I] that the limit

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

exists for all $u\in\overline{\mathcal{D}(F)}$, $t\geq 0$, and generates a semigroup $\{\mathrm{e}^{-tF}\}_{t\geq 0}$. For each $t\geq 0$, the nonlinear operator e^{-tF} is nonexpansive and maps $\overline{\mathcal{D}(F)}$ into itself. The continuous function $t\mapsto\mathrm{e}^{-tF}u_0$ then defines the unique (mild) solution to the abstract evolution equation $\dot u+Fu=0$, $u(0)=u_0$.

Lemma 2.1. *Under Assumption 1, the operators $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{F}+\mathcal{G}$ are all accretive. If $Q\in\mathcal{C}$ then the nonexpansive resolvents $(I+h\mathcal{F})^{-1}$, $(I+h\mathcal{G})^{-1}$ and $(I+h(\mathcal{F}+\mathcal{G}))^{-1}$ all map $\mathcal{C}$ into $\mathcal{C}$.*

This follows by minor modifications of the proofs in [3, II:3.3, III:2.3]. Thus the discussion above yields that with suitable P_0 and Q there exists a solution $\mathrm{e}^{-t(\mathcal{F}+\mathcal{G})}P_0$ to the Riccati equation (1.1), as well as a solution $\mathrm{e}^{-t\mathcal{G}}P_0$ to the subproblem (1.5). Furthermore, the splitting scheme $\mathcal{S}_h$ (1.6) is well-defined as a mapping from $\mathcal{C}$ to $\mathcal{C}$.

3. Convergence analysis. We now consider the approximation of the solution to (1.1) by the splitting scheme (1.6), with the aim of proving a convergence order. The main challenge is the lack of higher-order time regularity of the solution, which prohibits the standard ODE-type consistency argument. However, the existence proof [9, Theorem I] of a mild solution $\mathrm{e}^{-tF}u_0$ is based on the bound

$$\|(I + t/nF)^{-n}u_0 - (I + t/mF)^{-m}u_0\|_X \leq C(1/n - 1/m)^{1/2}\|Fu_0\|_X, \tag{3.1}$$

which yields the remarkable "byproduct" that the implicit Euler scheme converges with at least an order of $q = 1/2$. As illustrated in [17, Example 3], this convergence order is optimal in the general accretive case, though higher orders $q > 1/2$ can be observed if the vector field F possesses more structure (or if X is finite-dimensional). For the abstract Riccati equation, a single implicit Euler step is given by

$$\mathcal{R}_h = \big(I + h(\mathcal{F} + \mathcal{G})\big)^{-1}, \tag{3.2}$$

and the approximation converges as follows:

LEMMA 3.1. *If Assumption 1 is valid, $P_0 \in \mathcal{D}(\mathcal{F}) \cap \mathcal{C}$ and $Q \in \mathcal{C}$, then*

$$\|\mathrm{e}^{-nh(\mathcal{F}+\mathcal{G})}P_0 - \mathcal{R}_h^n P_0\|_{\mathcal{H}} \leq Ch^q\|(\mathcal{F} + \mathcal{G})P_0\|_{\mathcal{H}}, \quad 0 \leq nh \leq T,$$

for a fixed parameter $q \geq 1/2$. Parameter values $q > 1/2$ may be obtained under extra structural assumptions on $\mathcal{F} + \mathcal{G}$. The constant C depends on T, but not on n or h separately.

We now compare the proposed splitting method (1.6) with the implicit Euler scheme, instead of the exact solution. This approach avoids the need for higher differentiability of the exact solution and allows us to derive a general consistency concept by purely algebraic manipulations of the time-stepping operators. In order to demonstrate this, we state the following theorem for a broader class of splitting schemes, given by the time-stepping operators

$$\mathcal{S}_h = (I + h\mathcal{F})^{-1}\mathcal{T}_{h\mathcal{G}}. \tag{3.3}$$

THEOREM 3.2. *Let Assumption 1 be valid, $P_0 \in \mathcal{D}(\mathcal{F}) \cap \mathcal{C}$ and $Q \in \mathcal{C}$. Furthermore, assume that the operator $\mathcal{T}_{h\mathcal{G}}$ maps $\mathcal{C}$ into itself, satisfies $L[\mathcal{T}_{h\mathcal{G}}] \leq 1$ and fulfills the consistency bound*

$$\|(I - h\mathcal{G}\mathcal{R}_h)\mathcal{R}_h^j P_0 - \mathcal{T}_{h\mathcal{G}}\mathcal{R}_h^j P_0\|_{\mathcal{H}} \leq Ch^{1+p}, \quad j = 0, \ldots, n, \tag{3.4}$$

for a given $p > 0$. Then the splitting scheme (3.3) *converges to the solution of the Riccati equation* (1.1)*. More specifically,*

$$\|\mathrm{e}^{-nh(\mathcal{F}+\mathcal{G})}P_0 - \mathcal{S}_h^n P_0\|_{\mathcal{H}} \leq C(h^p + h^q), \quad 0 \leq nh \leq T,$$

where q is the convergence order of the implicit Euler scheme.

Proof. Due to the convergence result of Lemma 3.1, it is enough to prove that

$$\|\mathcal{R}_h^n P_0 - \mathcal{S}_h^n P_0\|_{\mathcal{H}} \leq Ch^p.$$

By Lemma 2.1 and the stability assumption $L[\mathcal{T}_{h\mathcal{G}}] \leq 1$, one obtains that

$$\begin{aligned}\|\mathcal{R}_h^n P_0 - \mathcal{S}_h^n P_0\|_{\mathcal{H}} &= \sum_{j=1}^{n} \|\mathcal{S}_h^{n-j}\mathcal{R}_h^j P_0 - \mathcal{S}_h^{n-j+1}\mathcal{R}_h^{j-1} P_0\|_{\mathcal{H}} \\ &\leq \sum_{j=1}^{n} L[\mathcal{S}_h]^{n-j} L[(I+h\mathcal{F})^{-1}] \|(I+h\mathcal{F})\mathcal{R}_h^j P_0 - \mathcal{T}_{h\mathcal{G}}\mathcal{R}_h^{j-1} P_0\|_{\mathcal{H}} \\ &\leq \sum_{j=1}^{n} \|(I-h\mathcal{G}\mathcal{R}_h)\mathcal{R}_h^{j-1} P_0 - \mathcal{T}_{h\mathcal{G}}\mathcal{R}_h^{j-1} P_0\|_{\mathcal{H}}.\end{aligned}$$

Employing the consistency bound then yields the desired convergence order. ∎

Note that the proof also holds for the less stringent stability conditions that the Lipschitz constants of $\mathcal{T}_{h\mathcal{G}}$ and $(I+h\mathcal{F})^{-1}$ are bounded by $1+Ch$ instead of 1, though this yields a minor step size restriction.

The operator $\mathcal{T}_{h\mathcal{G}}$ has to be selected with care depending on the problem at hand in order to ensure stability, consistency and efficiency. In the Riccati case we can compute the solution to (1.5) explicitly, and we therefore choose

$$\mathcal{T}_{h\mathcal{G}} = \mathrm{e}^{-h\mathcal{G}}.$$

Its nonexpansivity follows directly from Lemma 2.1, and in order to prove the consistency (3.4) we first prove that $\mathcal{G}$ generates a smooth flow.

LEMMA 3.3. *The nonlinear semigroup generated by $\mathcal{G}$ is given by*

$$\mathrm{e}^{-t\mathcal{G}} P_0 = (I+tP_0)^{-1} P_0,$$

where $P_0 \in \mathcal{C}$ and $(I+tP_0)^{-1}P_0$ denotes the composition of two operators in $\mathcal{L}(H,H)$. Furthermore, $P : t \mapsto \mathrm{e}^{-t\mathcal{G}} P_0$ is in $C^\infty([0,T];\mathcal{H})$ and $\mathrm{d}^n/\mathrm{d}t^n P(t) = (-1)^n n! P(t)^{n+1}$.

Proof. Assume that

$$P(t) = (I+tP_0)^{-1} P_0.$$

As P_0 is an element of $\mathcal{C}$ it is both self-adjoint and compact, since all Hilbert–Schmidt operators are compact. By the Hilbert–Schmidt spectral theorem [15, 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, consisting of eigenvectors of P_0 with corresponding eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$. Since P_0 is positive semi-definite, $\lambda_k \geq 0$ for all $k \geq 1$. Hence,

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

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

$$\begin{aligned}\|P(t+h) - P(t)\|_{\mathcal{H}} &= \big\| \big(I + (t+h)P_0\big)^{-1} \Big[(I+tP_0) - \big(I+(t+h)P_0\big) \Big] (I+tP_0)^{-1} P_0) \big\|_{\mathcal{H}} \\ &= \big\| -h \big(I+(t+h)P_0\big)^{-1} P_0 (I+tP_0)^{-1} P_0 \big\|_{\mathcal{H}} \\ &\leq h \|(I+(t+h)P_0)^{-1}\|_{\mathcal{L}(H,H)} \|P_0\|_{\mathcal{L}(H,H)} \|(I+tP_0)^{-1}\|_{\mathcal{L}(H,H)} \|P_0\|_{\mathcal{H}} \\ &\leq h \|P_0\|_{\mathcal{H}}^2,\end{aligned}$$

and $t \mapsto P(t)$ is therefore continuous in $\mathcal{H}$. By the same construction we obtain that

$$\lim_{h\to 0} \|(P(t+h) - P(t))/h + P(t)^2\|_{\mathcal{H}} = 0,$$

i.e. $t \mapsto P(t)$ is continuously differentiable and satisfies the equation (1.5). By application of the chain rule we see that we can express higher derivatives of P as compositions of P with itself. Since P is continuous, this observation proves the claim that $t \mapsto \mathrm{e}^{-t\mathcal{G}}P_0$ belongs to $C^\infty([0,T];\mathcal{H})$. ∎

The smoothness of $\mathrm{e}^{-t\mathcal{G}}$ and the Banach algebra setting of Hilbert–Schmidt operators now yields the consistency (3.4) of the splitting scheme:

LEMMA 3.4. *Let Assumption 1 be valid, $P_0 \in \mathcal{D}(\mathcal{F}) \cap \mathcal{C}$ and $Q \in \mathcal{C}$. Then*

$$\|(I - h\mathcal{G}\mathcal{R}_h)\mathcal{R}_h^j P_0 - \mathrm{e}^{-h\mathcal{G}}\mathcal{R}_h^j P_0\|_{\mathcal{H}} \le Ch^2,$$

for $j = 0, \ldots, n$. The constant C depends on $T \ge nh$, $\|P_0\|_{\mathcal{H}}$ and $\|(\mathcal{F}+\mathcal{G})P_0\|_{\mathcal{H}}$, but not on n or h separately.

Proof. From Lemma 3.3 it follows that for any $Z \in \mathcal{C}$ we can make the expansion

$$\mathrm{e}^{-h\mathcal{G}}Z = Z - h\mathcal{G}Z + h^2 R,$$

where the rest term is given by

$$R = \int_0^1 (1-t)\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathrm{e}^{-t\mathcal{G}}Z\,\mathrm{d}t = \int_0^1 2(1-t)\big((I+tZ)^{-1}Z\big)^3\mathrm{d}t$$

and is bounded in $\mathcal{H}$ by $2\|Z\|_{\mathcal{H}}^3$. Hence,

$$\begin{aligned}
\|(I - h\mathcal{G}\mathcal{R}_h)Z - \mathrm{e}^{-h\mathcal{G}}Z\|_{\mathcal{H}} &= \|Z - h\mathcal{G}\mathcal{R}_h Z - (Z - h\mathcal{G}Z + h^2R)\|_{\mathcal{H}} \\
&\le h\|Z^2 - (\mathcal{R}_h Z)^2\|_{\mathcal{H}} + 2h^2\|Z\|_{\mathcal{H}}^3 \\
&= h\|(\mathcal{R}_h Z - Z)\mathcal{R}_h Z + Z(\mathcal{R}_h Z - Z)\|_{\mathcal{H}} + 2h^2\|Z\|_{\mathcal{H}}^3 \\
&\le h\|(\mathcal{R}_h Z - Z)\|_{\mathcal{H}}\big(\|\mathcal{R}_h Z\|_{\mathcal{H}} + \|Z\|_{\mathcal{H}}\big) + 2h^2\|Z\|_{\mathcal{H}}^3.
\end{aligned}$$

Since the operator $\mathcal{R}_h$ is nonexpansive, setting $Z = \mathcal{R}_h^j P_0$ yields

$$\begin{aligned}
\|\mathcal{R}_h^{j+1}P_0 - \mathcal{R}_h^j P_0\|_{\mathcal{H}} &\le \|\mathcal{R}_h P_0 - \mathcal{R}_h\big(I + h(\mathcal{F}+\mathcal{G})\big)P_0\|_{\mathcal{H}} \\
&\le \|P_0 - \big(I + h(\mathcal{F}+\mathcal{G})\big)P_0\|_{\mathcal{H}} \\
&\le h\|(\mathcal{F}+\mathcal{G})P_0\|_{\mathcal{H}},
\end{aligned}$$

and we also have that

$$\|\mathcal{R}_h^i P_0\|_{\mathcal{H}} \le \|P_0\|_{\mathcal{H}} + \sum_{k=1}^{i}\|\mathcal{R}_h^k P_0 - \mathcal{R}_h^{k-1}P_0\|_{\mathcal{H}} \le \|P_0\|_{\mathcal{H}} + ih\|(\mathcal{F}+\mathcal{G})P_0\|_{\mathcal{H}}.$$

This implies that

$$\|(I - h\mathcal{G}\mathcal{R}_h)\mathcal{R}_h^j P_0 - \mathrm{e}^{-h\mathcal{G}}\mathcal{R}_h^j P_0\|_{\mathcal{H}} \le Ch^2,$$

where the constant C depends on T, $\|P_0\|_{\mathcal{H}}$ and $\|(\mathcal{F}+\mathcal{G})P_0\|_{\mathcal{H}}$. ∎

In conclusion, we obtain the following convergence result for the splitting scheme:

COROLLARY 3.5. *If Assumption 1 is valid, $P_0 \in \mathcal{D}(\mathcal{F}) \cap \mathcal{C}$ and $Q \in \mathcal{C}$, then the splitting approximation $\mathcal{S}_h^n P_0$, with $\mathcal{S}_h = (I + h\mathcal{F})^{-1}\mathrm{e}^{-h\mathcal{G}}$, converges to the (mild) solution of the abstract Riccati equation* (1.1)*. More precisely,*

$$\|\mathrm{e}^{-nh(\mathcal{F}+\mathcal{G})}P_0 - \mathcal{S}_h^n P_0\|_{\mathcal{H}} \leq C(h + h^q), \quad 0 \leq nh \leq T,$$

where $q \geq 1/2$ is the convergence order of the implicit Euler scheme. The constant C depends on T, $\|P_0\|_{\mathcal{H}}$, and $\|(\mathcal{F}+\mathcal{G})P_0\|_{\mathcal{H}}$, but not on n or h separately.

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 finally consider 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 $\mathbb{R}^d$ and using finite differences to discretize it with n points in each dimension leads to a solution with n^d elements. Representing this solution as a dense vector requires an inordinate amount of memory already with $d = 3$ and 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 differential 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 differential Riccati equation becomes a matrix-valued differential Riccati equation. That is, now $H = \mathbb{R}^n$ for some integer $n > 0$ and $P(t)$ is an element of $\mathbb{R}^{n\times n}$. By a low-rank approximation we mean that $P(t) \approx zz^T$ where $z \in \mathbb{R}^{n\times m}$, with $m \ll n$. We first show that the discretized version of

$$\mathrm{e}^{-h\mathcal{G}}P = (I + hP)^{-1}P$$

preserves such low-rank structure.

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

$$(I + hP)^{-1}P = ww^T,$$

where $w \in \mathbb{R}^{n\times m}$.

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 $\mathbb{R}^{k\times k}$. Taking $Y = hz$ and $Z = z^T$

we see that

$$\begin{aligned}(I_n + hzz^T)^{-1}zz^T &= zz^T - hz(I_m + z^Thz)^{-1}z^Tzz^T \\ &= z(I_m - (I_m + hz^Tz)^{-1}hz^Tz)z^T \\ &= z(I_m - I_m + (I_m + hz^Tz)^{-1})z^T \\ &= z(I_m + hz^Tz)^{-1}z^T.\end{aligned}$$

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. In the more general quadratic case of $\mathcal{G}P = PBR^{-1}B^TP$, the solution becomes

$$(I_n + hzz^TBR^{-1}B^T)^{-1}zz^T = z(I_m + hz^TBR^{-1}B^Tz)^{-1}z^T,$$

which can be computed as efficiently as in the previous case if B has much fewer columns than rows, or if $BR^{-1}B^T$ has a low-rank factorization.

In order to fully implement the splitting scheme (1.6), we also need to consider the action of $(I + h\mathcal{F})^{-1}$ on $\mathrm{e}^{-h\mathcal{G}}P$. Assume therefore that $(I + h\mathcal{F})S = P$. This means that $S + hA^*S + hSA - hQ = P$. But this is equivalent to the Lyapunov equation

$$(I + 2hA)^*S + S(I + 2hA) = 2P + 2hQ.$$

There are many methods for solving Lyapunov equations where the right-hand side is of low rank. The recent surveys [8, 19] discuss the state-of-the-art for solvers based on the ADI iteration as well as Krylov-related projection methods. In our case, given the factorizations $P = zz^T$ and $Q = Q_fQ_f^T$ we see that the matrix $w = (\sqrt{2}z, \sqrt{2h}Q_f)$ is a low-rank factor of the right-hand side. As there might be some linear dependence between the columns in z and Q_f, we recommend applying a column compression technique to compute an approximative low-rank factor $\tilde{w}$ with fewer columns. See e.g. [18, Section 4.4.1], where an approach based on the rank-revealing QR decomposition (RRQR) is described.

To summarize, we present the full procedure as pseudo-code in Algorithm 1.

Algorithm 1 Computing $\mathcal{S}_hP$

Input: Low-rank factors z and Q_f such that $P = zz^T$ and $Q = Q_fQ_f^T$

1. Cholesky factorize $I + hz^Tz =: LL^T$
2. Solve $wL^T = z$
3. Form $\tilde{x} = (\sqrt{2}w, \sqrt{2h}Q_f)$
4. Column-compress $x \approx \tilde{x}$ by e.g. RRQR
5. Low-rank solve the Lyapunov equation $(I + 2hA)^*S + S(I + 2hA) = xx^T$ for $S = yy^T$ by e.g. an ADI method

Output: y

We reiterate that the proposed method essentially requires only the low-rank solution of one Lyapunov equation per step, indicating that its efficiency is on par with the best alternative solvers based on solving Lyapunov equations, see e.g. [6, 7]. Demonstrating this, as well as comparing the efficiency to that of projection methods, e.g. [12, 20], is out of the scope of this paper and will be investigated elsewhere.

5. Numerical examples. Consider a linear quadratic regulator problem, where the goal is to minimize the functional

$$J(x,u) = \int_0^T \|Cx - x_d\|^2 + \|u\|^2 \,\mathrm{d}t$$

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 [13, Chapter III.4] that the optimal control strategy is an affine mapping, $u(t) = -P(T-t)x(t) - r(T-t)$, where P and r satisfies

$$\begin{aligned} \dot{P} + PA + A^*P + P^2 = C^*C, \quad P(0) = 0, \\ \dot{r} + A^*r + Pr = -C^*x_d, \quad r(0) = 0. \end{aligned} \tag{5.1}$$

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, $\alpha(x) = 2 + \cos 2\pi x$ and $\lambda = 1$. To define C, we choose first the real trigonometric orthonormal basis for H: $\{1\} \cup \{e_k\}_{k=1}^\infty \cup \{f_k\}_{k=1}^\infty$, 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

$$C\Big(a_0 + \sum_{k=0}^{\infty} a_k e_k + b_k f_k\Big) = a_0 + \sum_{k=0}^{m} a_k e_k + b_k f_k,$$

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, and as $P_0 = 0$ clearly belongs to $\mathcal{D}(\mathcal{F})$, the assumptions in Corollary 3.5 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

$$Q_M = (E^{\mathrm{T}} zz^T E)^T (E^T zz^T E) = E^T zz^T zz^T E = E^T zz^T E,$$

we can also low-rank factorize $Q_M = ww^T$ with $w = E^T z$. 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 Corollary 3.5. The errors are measured in the Frobenius norm $\|\cdot\|_{Fro}$ 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 $(I + h\mathcal{F})^{-1}$ we have used a modified version of LyaPack 1.8 [5] 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.

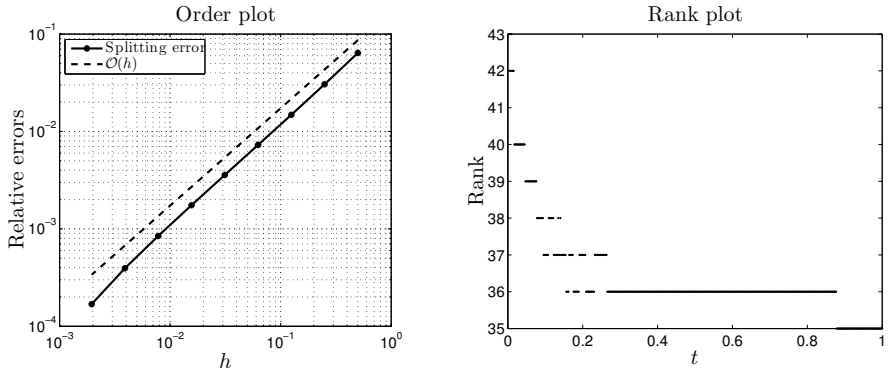

FIG. 5.1. *Left: The relative errors* $\|\mathcal{S}_h^n P_0 - P_{ref}\|_{Fro}/\|P_{ref}\|_{Fro}$ *when approximating the solution to* (5.1) *for different* $h = 1/N$ *with* $N = 2, 4, \ldots, 512$. *The reference solution* P_{ref} *was also computed by the splitting method, albeit with a finer temporal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

REFERENCES

[1] A. C. Antoulas, D. C. Sorensen, Y. Zhou, *On the decay rate of Hankel singular values and related issues*, Syst. Control. Lett., 46(5) (2002), pp. 323–342.

[2] A. V. Balakrishnan, *Applied Functional Analysis*, Springer, New York, 1976.

[3] V. Barbu, *Nonlinear Semigroups And Differential Equations In Banach Spaces*, Noordhoff, Leyden, 1976.

[4] V. Barbu, M. Iannelli, *Approximating some non-linear equations by a fractional step scheme*, Differ. Integral Equ., 6(1) (1993), pp. 15–26.

[5] P. Benner, V. Mehrmann, H. Mena, T. Penzl, J. Saak, *LyaPack 1.8*, http://www.mpi-magdeburg.mpg.de/mpcsc/software/mess.php, accessed on 2013-08-30.

[6] P. Benner, H. Mena, *Numerical solution of the infinite-dimensional LQR-Problem and the associated differential Riccati equations*, preprint, Max Planck Institute Magdeburg, MPIMD/12-13 (2012).

[7] P. Benner, J. Saak, *A Galerkin-Newton-ADI method for solving large-scale algebraic Riccati equations*, Preprint SPP1253-090, DFG Priority Programme 1253 Optimization with Partial Differential Equations (2010).

[8] P. Benner, J. Saak, *Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: A state of the art survey*, GAMM-Mitt., 36 (2013), pp. 32–52.

[9] M. G. Crandall, T. M. Liggett, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Am. J. Math., 93(2) (1971), pp. 265–298.

[10] R. Curtain, A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, in Lecture Notes in Control and Information Sciences, Vol. 8, Springer, Berlin, 1978.

[11] A. Germani, L. Jetto, M. Piccioni, *Galerkin approximation for optimal linear filtering of infinite dimensional linear systems*, SIAM J. Control Optim., 26(6) (1988), pp. 1287–1305.

[12] M. Heyouni, K. Jbilou *An extended block Arnoldi algorithm for large-scale solutions to the continuous-time algebraic Riccati equation*, Electron. T. Numer. Ana., 33 (2009), pp. 53-62

[13] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.
[14] T. Penzl, *Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case*, Syst. Control Lett., 40(2) (2000), pp. 139–144.
[15] M. Reed, B. Simon, *Functional Analysis I*, Academic Press, New York, 1980.
[16] I. G. Rosen, *Convergence of Galerkin approximations for operator Riccati equations—a nonlinear evolution equation approach*, J. Math. Anal. Appl., 155 (1991), pp. 226–248.
[17] J. Rulla, *Error analysis for implicit approximations to solutions to Cauchy problems*, SIAM J. Numer. Anal., 33(1) (1996), pp. 68–87.
[18] J. Saak, *Efficient numerical solution of large scale algebraic matrix equations in PDE control and model order reduction*, Ph.D. thesis, Faculty of Mathematics, Chemnitz University of Technology, Chemnitz, 2009.
[19] V. Simoncini, *Computational methods for linear matrix equations*, preprint, Università di Bologna (2013).
[20] V. Simoncini, D. B. Szyld, M. Monsalve, *On two numerical methods for the solution of large-scale algebraic Riccati equations*, IMA J. Numer. Anal., doi: 10.1093/imanum/drt015 (2013).
[21] R. Temam, *Sur l'équation de Riccati associée à des opérateurs non bornés, en dimension infinie*, J. Funct. Anal., 7 (1971), pp. 85–115.