Convergence analysis for the exponential Lie splitting scheme applied to the abstract differential Riccati equation

Tony Stillfjord[a]

[a] *Centre for Mathematical Sciences, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden.*

Abstract

We consider differential Riccati equations (DREs). These equations arise in many areas and are very important within the field of optimal control. In particular, DREs provide the crucial link between the state and the optimal input in the solution of linear quadratic regulator (LQR) problems. For the approximation of the solutions to DREs we consider the recently introduced splitting methods, with the aim of proving convergence orders in the space of Hilbert–Schmidt operators. The use of this abstract setting yields stronger than usual temporal convergence results, and also implies that these are independent of a subsequent (reasonable) spatial discretization. The main result is that the exponential Lie splitting is first-order convergent, under no artificial regularity assumptions. As side-effects of the analysis, we also acquire concise proofs of the existence and positivity of the exact solutions to abstract DREs, in a more general setting than previously considered.

Keywords: Differential Riccati equation, splitting, error analysis, convergence order, Hilbert–Schmidt operators

1. Introduction

We consider the differential Riccati equation:

$$\begin{aligned} \dot{P}(t) + A^*P(t) + P(t)A + P(t)SP(t) = Q, \quad t \in (0,T), \\ P(0) = P_0. \end{aligned} \tag{1}$$

This is a semi-linear operator-valued evolution equation for P, where A, S, and Q are given linear operators. A prototypical A would be an elliptic differential operator. For brevity, we denote the composition of operators by juxtaposition; thus e.g. $P(t)A$ means $P(t) \circ A$.

Such equations arise in many areas, for example in linear quadratic regulator (LQR) problems and optimal state estimation [1, 6, 12]. The main applications that we have in mind are the LQR problems, where the goal is to minimize the functional

$$J(u) = \int_0^T (\tilde{Q}y, y)_Y + (Ru, u)_U \, \mathrm{d}t,$$

Email address: tony@maths.lth.se (Tony Stillfjord)
URL: www.maths.lu.se/~tony (Tony Stillfjord)

subject to the state and output equations

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx. \end{aligned} \tag{2}$$

Here x, u and y belong to the Hilbert spaces $\big(H, (\cdot,\cdot)_H\big)$, $\big(U, (\cdot,\cdot)_U\big)$ and $\big(Y, (\cdot,\cdot)_Y\big)$, where the latter two are frequently finite-dimensional. Under certain assumptions on the involved operators, it can be shown that the optimal input u_{opt} is given in feedback form as $u_{\text{opt}}(t) = -R^{-1}B^*P(T-t)x(t)$, where $P(t)$ solves the Riccati equation (1) with $S = BR^{-1}B^*$, $Q = C^*\tilde{Q}C$ and $P(0) = 0$, see e.g. [12].

Previous approaches to approximate the solution of the infinite-dimensional Riccati equation (1) include spatial Galerkin methods [7, 16], temporal BDF and Rosenbrock methods [4] and temporal first-order splitting methods [3, 18]. In the finite-, but still high-dimensional case, the main approach has been to employ matrix-versions of BDF or Rosenbrock methods [4, 5, 14]. While these studies show that the respective methods converge, they lack a convergence analysis which describes how quickly the convergence occurs. Recently, also first- and second-order exponential splitting schemes were considered in [17], which demonstrates convergence orders in the finite-dimensional case. The analysis is however "nonstiff", and so still has a dependence on the spatial discretization parameter.

The aim of this paper is therefore to extend the results of [17] and investigate whether a "stiff" temporal error analysis can be carried out by considering the problem in the space of Hilbert-Schmidt operators, as in [8]. We note that the approach here differs from the one presented in [8] due to the fact that the nonlinearity is no longer necessarily accretive.

To informally introduce the numerical method that we consider, let

$$\begin{aligned} \mathcal{F}P &= A^*P + PA - Q \quad \text{and} \\ \mathcal{G}P &= PSP. \end{aligned}$$

Then the time-stepping operator of the exponential Lie splitting scheme is given by

$$\mathcal{S}_h = \mathrm{e}^{-h\mathcal{F}}\mathrm{e}^{-h\mathcal{G}},$$

where $\mathrm{e}^{-t\mathcal{F}}P_0$ and $\mathrm{e}^{-t\mathcal{G}}P_0$ denote the solutions to the subproblems

$$\dot{P} + \mathcal{F}P = 0, \quad P(0) = P_0 \quad \text{and} \tag{3}$$

$$\dot{P} + \mathcal{G}P = 0, \quad P(0) = P_0. \tag{4}$$

The motivation for this splitting is that the first subproblem is affine, and while the second subproblem is nonlinear it has a simple closed-form expression for relevant initial conditions.

A brief outline of the paper is as follows. In Section 2 we set up the abstract framework for the analysis, and state the assumptions on the involved operators. The consequences of these assumptions are investigated in Section 3, and lead to the convergence analysis in Section 4. In Section 5 we give a brief motivation as to why a similar convergence analys fails for the Strang splitting. Finally, in Section 6 we demonstrate that the scheme produces positive semi-definite approximations and that also the exact solution has such structure.

2. Problem setting

Let us first fix the notation. Given a real Hilbert space X, we denote its inner product by $(\cdot,\cdot)_X$, its norm by $\|\cdot\|_X$ and its dual space by X^*. The space of linear bounded operators from X to another Hilbert space Y is denoted by $\mathscr{L}(X,Y)$. Finally, an operator F is locally Lipschitz continuous if for any ball $B_r = \{x \in X \ ; \ \|x\| \leq r\}$ there exists a local Lipschitz constant $L_r[F] < \infty$ such that

$$\|Fx - Fy\| \leq L_r[F]\|x - y\|$$

for all $x, y \in B_r$. If $L[F] = \sup_{r>0} L_r[F] < \infty$ we say that F is globally Lipschitz continuous with Lipschitz constant $L[F]$.

The following is essentially the same setting as presented in [8], but reproduced here for convenience. Let the Gelfand triple

$$V \hookrightarrow H \cong H^* \hookrightarrow V^*$$

of Hilbert spaces V and H be given. We define the class of suitable operators A and A^* by introducing 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 $u, v \in V$*

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

The operators $A \in \mathscr{L}(V, V^*)$ and $A^* \in \mathscr{L}(V, V^*)$ 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 $H^1(\Omega)$, $H_0^1(\Omega)$ 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 $A = -\nabla \cdot \big(\alpha\nabla u\big) + \lambda I$.

We will look for solutions $P(t)$ to the Riccati equation (1) in the setting of Hilbert-Schmidt operators, as advocated by e.g. Temam [18]. For completeness, we therefore now recall some basic theory about such operators, see e.g. [2, Sections II:3.3 and III:2.3] and [16, 18] for a more extensive exposition. Let H_i denote generic Hilbert spaces. An operator $F \in \mathscr{L}(H_1, H_2)$ 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

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

Further, Hilbert–Schmidt operators are invariant under composition with linear bounded operators from both the left and from the right. That is, if $F \in \mathcal{HS}(H_2, H_3)$, $G_1 \in \mathscr{L}(H_1, H_2)$ and $G_2 \in \mathscr{L}(H_3, H_4)$ then $G_2 F G_1 \in \mathcal{HS}(H_1, H_4)$ and

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

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 $\mathcal{HS}(V, H) + \mathcal{HS}(H, V^*)$. 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 $\mathscr{L}(\mathcal{V}, \mathcal{V}^*)$ and we consider the restriction $\mathcal{L} : \mathcal{D}(\mathcal{L}) \subset \mathcal{H} \to \mathcal{H}$ given by

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

Assumption 2. *The operator Q is an element of $\mathcal{H}$.*

Under Assumption 2, we can also consider the perturbed operator $\mathcal{F} : \mathcal{D}(\mathcal{F}) \subset \mathcal{H} \to \mathcal{H}$, given by

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

For the nonlinearity PSP we will consider two possibilities. First:

Assumption 3. *The operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ is given by $\mathcal{G}P = PSP$, with an operator $S \in \mathscr{L}(V^*, H) \cap \mathscr{L}(H, V)$.*

Note that the definition makes sense, as the requirement on S also means that $S \in \mathscr{L}(H, H)$. The reason for not only requiring $S \in \mathscr{L}(H, H)$ is that we will need terms of the form $\mathcal{F}\mathcal{G}P$ to be well defined. This is the case under Assumption 3, since it is easily verified that if $P \in \mathcal{V}$, then also $PSP \in \mathcal{V}$. It is also the case in the absence of S. More precisely, we have the following alternative assumption:

Assumption 4. *The operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ is given by $\mathcal{G}P = sPP$, with $s \in \mathbb{R}$.*

Given either Assumption 3 or 4, we define the full problem by

$$\mathcal{D}(\mathcal{F} + \mathcal{G}) = \mathcal{D}(\mathcal{L}) \quad \text{and} \quad (\mathcal{F} + \mathcal{G})P = \mathcal{F}P + \mathcal{G}P.$$

3. Preliminary results

The abstract setting at hand is motivated by the results in [2, 18], which essentially say that under Assumption 4 with $s > 0$, the operators $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{F}+\mathcal{G}$ are all accretive, and satisfy useful range conditions. In the situation of Assumption 3, the results for $\mathcal{F}$ are of course unchanged, and while $\mathcal{G}$ is no longer necessarily accretive[1], it is still a locally Lipschitz continuous operator. This turns out to be enough to draw similar conclusions. We start out with the following immediate extension of [2, Lemma II:3.5]:

Lemma 1. *Under Assumption 1, the operator $\mathcal{L}$ is strongly $\mathcal{V}$-coercive, i.e. there exists an $\alpha > 0$ such that*

$$(\mathcal{L}P, P)_{\mathcal{H}} \geq \alpha \|P\|_{\mathcal{V}}^2 \quad \textit{for all } P \in \mathcal{D}(\mathcal{L}).$$

As a consequence, both $\mathcal{L}$ and $\mathcal{F}$ are maximal accretive operators and the solution operators $\mathrm{e}^{-t\mathcal{L}}$ and $\mathrm{e}^{-t\mathcal{F}}$ exist and are nonexpansive.

We thus get the existence of a classical solution to the affine subproblem $\dot{P} + \mathcal{F}P = 0$ if $P_0 \in \mathcal{D}(\mathcal{L})$. For the operator $\mathcal{G}$, we have:

Lemma 2. *The operator $\mathcal{G}$ is infinitely many times Fréchet differentiable, and its first two derivatives are given by*

$$\mathrm{D}\mathcal{G}[P_1]P_2 = P_1SP_2 + P_2SP_1 \quad \textit{and} \quad \mathrm{D}^2\mathcal{G}[P_1](P_2, P_3) = P_2SP_3 + P_3SP_2.$$

As a consequence, $\mathcal{G}$ is locally Lipschitz continuous with the local Lipschitz constant $L_r[\mathcal{G}] \leq 2r\|S\|_{\mathscr{L}(H,H)}$.

Proof. The differentiability of $\mathcal{G}$ follows directly from the definition after a straightforward calculation. The bound for the local Lipschitz constant follows from the differentiability, but also immediately from the identity

$$\mathcal{G}P_1 - \mathcal{G}P_2 = P_1S(P_1 - P_2) + (P_1 - P_2)SP_2.$$

□

On sufficiently short time intervals, the local Lipschitz continuity of $\mathcal{G}$ guarantees the existence of a strong solution $\mathrm{e}^{-tG}P_0$ to the nonlinear subproblem for all $P_0 \in \mathcal{H}$, see e.g. [15, Theorem 6.1.4, 6.1.6]. Further, this solution satisfies the equation

$$\mathrm{e}^{-t\mathcal{G}}P_0 = P_0 - \int_0^t \mathcal{G}\mathrm{e}^{-\tau\mathcal{G}}P_0 \mathrm{d}\tau. \tag{5}$$

Since also the mapping $P \mapsto \mathcal{G}P - Q$ is locally Lipschitz continuous, we have in fact shown the existence of a strong solution $\mathrm{e}^{-t(\mathcal{F}+\mathcal{G})}P_0$ to the full problem on a time interval $[0, T]$, where T depends on P_0. Similarly to the nonlinear subproblem, this solution satisfies the variation of constants formula

$$\mathrm{e}^{-t(\mathcal{F}+\mathcal{G})}P_0 = \mathrm{e}^{-t\mathcal{L}}P_0 + \int_0^t \mathrm{e}^{-(t-\tau)\mathcal{L}}\big(-\mathcal{G}\mathrm{e}^{-\tau(\mathcal{F}+\mathcal{G})}P_0 + Q\big)\mathrm{d}\tau. \tag{6}$$

[1] As a simple counterexample, consider $H = \mathbb{R}^2$. Then $\mathcal{H}$ consists of 2×2-matrices, and $(P_1, P_2)_{\mathcal{H}} = (P_1e_1, P_2e_1) + (P_1e_2, P_2e_2)$, where $e_1 = (1, 0)^T$ and $e_2 = (0, 1)^T$. With the positive definite matrix $S = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix}$ it is easily seen that for, e.g., $P_1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ and $P_2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ we have $(\mathcal{G}P_1 - \mathcal{G}P_2, P_1 - P_2)_{\mathcal{H}} = -1$.

Consider now terms of the form $\mathcal{L}\mathcal{G}P_0$. As the next Lemma shows, these are well defined in $\mathcal{H}$ for $P_0 \in \mathcal{D}(\mathcal{L})$, and the operator $\mathcal{L}\mathcal{G}$ thus defined satisfies a local Lipschitz condition.

Lemma 3. *Under Assumptions 1, 2 and either 3 or 4, $\mathcal{D}(\mathcal{L})$ is invariant under $\mathcal{G}$. Additionally, $\mathcal{G}$ is locally Lipschitz continuous on $\mathcal{D}(\mathcal{L})$ equipped with the graph norm.*

Proof. Let $P \in \mathcal{D}(\mathcal{L})$. We consider only Assumption 3, as the case of Assumption 4 is analogous and simpler. Thus, we have that

$$A^*PSP \in \mathcal{HS}(V^*, V^*) \subset \mathcal{HS}(H, V^*) \quad \text{and} \quad PSPA \in \mathcal{HS}(V, V) \subset \mathcal{HS}(V, H),$$

i.e. $\mathcal{L}\mathcal{G}P \in \mathcal{V}^*$. Further, by writing

$$A^*PSP + PSPA = (A^*P + PA)SP + PS(A^*P + PA) - PASP - PSA^*P,$$

we see that in fact $\mathcal{L}\mathcal{G}P \in \mathcal{H}$ since the four terms all belong to $\mathcal{H}$. To show the Lipschitz continuity, assume that P_1 and P_2 belong to $\mathcal{D}(\mathcal{L})$. Then[2]

$$\begin{aligned}
\mathcal{L}\big(P_1SP_1\big) - \mathcal{L}\big(P_2SP_2\big) &= \mathcal{L}\big(P_1S(P_1 - P_2) + (P_1 - P_2)SP_2\big) \\
&= \big(\mathcal{L}P_1\big)S(P_1 - P_2) + \big(\mathcal{L}(P_1 - P_2)\big)SP_2 \\
&\quad + P_1S\mathcal{L}(P_1 - P_2) + (P_1 - P_2)S\mathcal{L}P_2 \\
&\quad - P_1\big(AS + SA^*\big)(P_1 - P_2) - (P_1 - P_2)\big(AS + SA^*\big)P_2,
\end{aligned}$$

so that

$$\begin{aligned}
\|\mathcal{L}\mathcal{G}P_1 - \mathcal{L}\mathcal{G}P_2\|_{\mathcal{H}} &\leq \|P_1 - P_2\|_{\mathcal{H}}\|S\|_{\mathscr{L}(H,H)}\big(\|\mathcal{L}P_1\|_{\mathcal{H}} + \|\mathcal{L}P_2\|_{\mathcal{H}}\big) \\
&\quad + \|\mathcal{L}(P_1 - P_2)\|_{\mathcal{H}}\|S\|_{\mathscr{L}(H,H)}\big(\|P_1\|_{\mathcal{H}} + \|P_2\|_{\mathcal{H}}\big) \\
&\quad + \|A\|_{\mathscr{L}(V,V^*)}\big(\|S\|_{\mathscr{L}(H,V)} + \|S\|_{\mathscr{L}(V^*,H)}\big)\|P_1 - P_2\|_{\mathcal{V}}\big(\|P_1\|_{\mathcal{V}} + \|P_2\|_{\mathcal{V}}\big).
\end{aligned}$$

Since $\mathcal{L}$ is strictly $\mathcal{V}$-coercive by Lemma 1, we have the bound

$$\|P\|_{\mathcal{V}} \leq C\big(\|P\|_{\mathcal{H}} + \|\mathcal{L}P\|_{\mathcal{H}}\big),$$

and thus get

$$\|\mathcal{G}P_1 - \mathcal{G}P_2\|_{\mathcal{H}} + \|\mathcal{L}\mathcal{G}P_1 - \mathcal{L}\mathcal{G}P_2\|_{\mathcal{H}} \leq C\big(\|P_2 - P_1\|_{\mathcal{H}} + \|\mathcal{L}(P_2 - P_1)\|_{\mathcal{H}}\big),$$

where the constant C depends on S, $\|P_1\|_{\mathcal{D}(\mathcal{L})}$ and $\|P_2\|_{\mathcal{D}(\mathcal{L})}$.

□

Remark 1. A side-effect of Lemma 3 is that if $Q \in \mathcal{D}(\mathcal{L})$ then $\mathrm{e}^{-t(\mathcal{F}+\mathcal{G})}P_0$ is actually a *classical* solution to Equation (1) on its interval of existence [15, Theorem 6.1.7].

[2] The reader should write this calculation down in terms of A^* and A. The reason for the seemingly needlessly complicated expression is that the terms A^*P and PA are not necessarily in $\mathcal{H}$ separately, even though their sum is.

4. Convergence

In order to prove that the Lie splitting scheme is first-order convergent, we first establish its consistency.

Lemma 4. *Let Assumptions 1, 2 and either 3 or 4 be fulfilled. Then if $P_0 \in \mathcal{D}(\mathcal{L})$, we have*

$$\|\mathrm{e}^{-h\mathcal{F}}\mathrm{e}^{-h\mathcal{G}}P_0 - \mathrm{e}^{-h(\mathcal{F}+\mathcal{G})}P_0\|_{\mathcal{H}} \leq C(P_0, S)\, h^2,$$

where $C(P_0, S)$ indicates a constant depending continuously on P_0 and on S.

For the proof, we will follow the idea presented in [10]. This consists of repeated applications of the variation of constants formula, combined with Taylor expansions of G and $\mathrm{e}^{-t\mathcal{G}}$ (allowable due to Lemma 2). Finally, the local errors are identified as quadrature errors. The latter part was originally used in [13].

Proof. We first consider the exact solution to the full problem (1). Since

$$\mathrm{e}^{-h\mathcal{F}}P_0 = \mathrm{e}^{-h\mathcal{L}}P_0 + \int_0^h \mathrm{e}^{-(h-\tau)\mathcal{L}}Q\,\mathrm{d}\tau,$$

the variation of constants formula (6) can also be written

$$\begin{aligned}\mathrm{e}^{-h(\mathcal{F}+\mathcal{G})}P_0 &= \mathrm{e}^{-h\mathcal{F}}P_0 - \int_0^h \mathrm{e}^{-(h-\tau)\mathcal{L}}\mathcal{G}\mathrm{e}^{-\tau(\mathcal{F}+\mathcal{G})}P_0\,\mathrm{d}\tau \\ &= \mathrm{e}^{-h\mathcal{F}}P_0 - hU(h),\end{aligned}$$

where

$$U(h) = \int_0^1 \mathrm{e}^{-h(1-\tau)\mathcal{L}}\mathcal{G}\mathrm{e}^{-h\tau(\mathcal{F}+\mathcal{G})}P_0\,\mathrm{d}\tau$$

is bounded in $\mathcal{H}$. This follows from the local Lipschitz continuity of $\mathcal{G}$, as $\mathrm{e}^{-h\mathcal{L}}$ is nonexpansive and $\mathrm{e}^{-h\tau(\mathcal{F}+\mathcal{G})}P_0$ is continuous. A repeated application of the formula for $\mathrm{e}^{-h(\mathcal{F}+\mathcal{G})}$, followed by a first-order Taylor expansion of $\mathcal{G}$ around $\mathrm{e}^{-\tau\mathcal{F}}P_0$, yields

$$\begin{aligned}\mathrm{e}^{-h(\mathcal{F}+\mathcal{G})}P_0 &= \mathrm{e}^{-h\mathcal{F}}P_0 - \int_0^h \mathrm{e}^{-(h-\tau)\mathcal{L}}\mathcal{G}\big(\mathrm{e}^{-\tau\mathcal{F}}P_0 - \tau U(\tau)\big)\,\mathrm{d}\tau \\ &= \mathrm{e}^{-h\mathcal{F}}P_0 - \int_0^h \mathrm{e}^{-(h-\tau)\mathcal{F}}\mathcal{G}\mathrm{e}^{-\tau\mathcal{F}}P_0\,\mathrm{d}\tau + h^2 R_{\mathcal{T}},\end{aligned}$$

where

$$h^2 R_{\mathcal{T}} = \int_0^h \tau \int_0^1 \mathrm{e}^{-(h-\tau)\mathcal{L}}\,\mathrm{D}\mathcal{G}[\mathrm{e}^{-\tau\mathcal{F}}P_0 - \sigma\tau U(\tau)]U(\tau)\,\mathrm{d}\sigma\,\mathrm{d}\tau + \int_0^h \int_0^{h-\tau} \mathrm{e}^{(h-\tau-\sigma)\mathcal{L}}\,Q\,\mathrm{d}\sigma\,\mathrm{d}\tau.$$

As $U(h)$ is bounded, so is the rest-term $R_{\mathcal{T}}$.

Now consider the Lie splitting, $\mathcal{S}_h = \mathrm{e}^{-h\mathcal{G}}\mathrm{e}^{-h\mathcal{F}}$. A Taylor-expansion of $\mathrm{e}^{-h\mathcal{G}}$ around $\mathrm{e}^{-h\mathcal{F}}P_0$ yields

$$\mathcal{S}_h P_0 = \mathrm{e}^{-h\mathcal{F}}P_0 - h\mathcal{G}\mathrm{e}^{-h\mathcal{F}}P_0 + h^2 R_{\mathcal{S}},$$

where the rest term

$$R_{\mathcal{S}} = \int_0^1 (1-\tau)\mathrm{D}\mathcal{G}[\mathrm{e}^{-\tau h\mathcal{G}}\mathrm{e}^{-h\mathcal{F}}P_0]\mathcal{G}\mathrm{e}^{-\tau h\mathcal{G}}\mathrm{e}^{-h\mathcal{F}}P_0\,\mathrm{d}\tau,$$

is again bounded in $\mathcal{H}$ due to the local Lipschitz continuity of $\mathcal{G}$. By introducing the function $\phi : [0,h] \to \mathcal{H}$ given by

$$\phi(\tau) = \mathrm{e}^{-(h-\tau)\mathcal{F}}\mathcal{G}\mathrm{e}^{-\tau\mathcal{F}}P_0,$$

and collecting terms, we can thus express the local error as

$$\mathcal{S}_h P_0 - \mathrm{e}^{-h(\mathcal{F}+\mathcal{G})}P_0 = \int_0^h \phi(\tau)\,\mathrm{d}\tau - h\phi(h) + h^2(R_{\mathcal{S}} - R_{\mathcal{T}}).$$

But the first two terms in this expression constitute the local error of a first-order quadrature rule applied to ϕ. In particular, we have

$$\int_0^h \phi(\tau)\,\mathrm{d}\tau - h\phi(h) = \int_0^h \int_0^1 \tau\phi'(\sigma\tau) - h\phi'(\sigma h)\mathrm{d}\sigma\mathrm{d}\tau,$$

where $\phi' : [0,h] \to \mathcal{H}$ is given by

$$\begin{aligned}\phi'(\tau) = {} & \mathcal{F}\mathrm{e}^{-(h-\tau)\mathcal{F}}\mathcal{G}\mathrm{e}^{-\tau\mathcal{F}}P_0 - \mathrm{e}^{-(h-\tau)\mathcal{F}}\mathrm{D}\mathcal{G}[\mathrm{e}^{-\tau\mathcal{F}}P_0]\mathcal{F}\mathrm{e}^{-\tau\mathcal{F}}P_0 \\ & + \int_0^{h-\tau} \mathrm{e}^{(h-\tau-\sigma)\mathcal{L}}\,Q\,\mathrm{d}\sigma.\end{aligned}$$

This expression is well defined in view of Lemma 3, and as $\mathrm{e}^{t\mathcal{L}}$ is nonexpansive by Lemma 1 we additionally see that $\phi'(\tau)$ is uniformly bounded on $[0,h]$. Thus the quadrature error, and therefore also the local error of the Lie splitting, is $\mathcal{O}(h^2)$, as desired. The continuous dependence of the error on P_0 follows directly, as $\mathrm{e}^{-t\mathcal{F}}$ and $\mathrm{e}^{-t\mathcal{G}}$ are both continuous semigroups.

□

We are now able to prove that the Lie splitting scheme, given by $\mathcal{S}_h = \mathrm{e}^{-h\mathcal{F}}\mathrm{e}^{-h\mathcal{G}}$, converges with order 1.

Theorem 1. *Let Assumptions 1, 2 and either 3 or 4 be fulfilled. Then with $0 \le nh \le T$ and h sufficiently small we have for $P_0 \in \mathcal{D}(\mathcal{L})$ that*

$$\|\mathcal{S}_h^n P_0 - \mathrm{e}^{-nh(\mathcal{F}+\mathcal{G})}P_0\|_{\mathcal{H}} \le Ch,$$

where the constant C depends on S, P_0 and T, but not on h or n separately.

Proof. Denote $P(t) = \mathrm{e}^{-t(\mathcal{F}+\mathcal{G})}P_0$ and define $t_k = kh$ for $k = 0, 1, \ldots$. By the continuity of P, we can define the finite number

$$r_- = \max_{0\le t\le T} \|P(t)\|_{\mathcal{H}}.$$

Further take $r > r_-$ and note that for any $U \in B_r$ there is a solution $\mathrm{e}^{-t\mathcal{G}}U$ to the nonlinear subproblem, for at least those t such that $t < 1/(\|S\|_{\mathscr{L}(H,H)}\|U\|_{\mathcal{H}})$. This follows directly from solving the differential inequality

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\mathrm{e}^{-t\mathcal{G}}U\|_{\mathcal{H}} \le \|S\|_{\mathscr{L}(H,H)}\|\mathrm{e}^{-t\mathcal{G}}U\|_{\mathcal{H}}^2.$$

Thus by Lemma 1 we can choose $r_+ > r$ and h sufficiently small such that $\mathcal{S}_h U \in B_{r_+}$ for all $U \in B_r$. By the representation (5) it then follows that $\mathcal{S}_h$ is Lipschitz continuous on B_r with Lipschitz constant $\mathrm{e}^{hL_{r_+}[G]}$.

Clearly, $\mathcal{S}_h^0 P_0 = P_0 \in B_r$. Suppose additionally that $\mathcal{S}_h^k P_0 \in B_r$ for $j = 0, \ldots, n$. Then by Lemma 4,

$$\begin{aligned}\|\mathcal{S}_h^{n+1}P_0 - P(t_{n+1})\|_{\mathcal{H}} &\le \|\mathcal{S}_h P(t_n) - \mathrm{e}^{-h(\mathcal{F}+\mathcal{G})}P(t_n)\|_{\mathcal{H}} + \|\mathcal{S}_h\mathcal{S}_h^n P_0 - \mathcal{S}_h P(t_n)\|_{\mathcal{H}}\\ &\le h^2 C(P(t_n), S) + \mathrm{e}^{2hr\|S\|_{\mathscr{L}(H,H)}}\|\mathcal{S}_h^n P_0 - P(t_n)\|_{\mathcal{H}}.\end{aligned}$$

Since $\|\mathcal{S}_h^0 P_0 - P(t_0)\|_{\mathcal{H}} = 0$, solving this recursion yields

$$\begin{aligned}\|\mathcal{S}_h^{n+1}P_0 - P(t_{n+1})\|_{\mathcal{H}} &\le h^2 \sum_{k=1}^{n} C(P(t_k), S)\mathrm{e}^{2h(n-k)r\|S\|_{\mathscr{L}(H,H)}}\\ &\le hC'\mathrm{e}^{2rT\|S\|_{\mathscr{L}(H,H)}},\end{aligned}$$

where $C' = \sup_{t\in[0,T]} C(P(t_n), S)T < \infty$ by the continuous dependence of $C(P_0, S)$ on P_0. But as $P(t_{n+1}) \in B_{r_-}$, this guarantees that for small enough h, also $\mathcal{S}_h^{n+1}P_0 \in B_r$. This step size restriction can be decreased by choosing a larger r, with the drawback of simultaneously increasing the error constants. The theorem finally follows by induction over n. □

Remark 2. We note that the approach used for the proof of Lemma 4 works equally well for the scheme $\mathrm{e}^{-hF}\mathrm{e}^{-hG}$, and hence Theorem 1 applies also for this method.

5. Higher-order analysis

We would like to perform a similar analysis for higher-order schemes, e.g. the second-order Strang splitting given by the time stepping operator

$$\mathcal{S}_h = \mathrm{e}^{-h/2\mathcal{G}}\mathrm{e}^{-h\mathcal{F}}\mathrm{e}^{-h/2\mathcal{G}}.$$

Using the same approach fails, however, due to the fact that $\mathcal{D}\left(\mathcal{L}^2\right)$ is only invariant under $\mathcal{G}$ under very restrictive conditions on S. For $\mathcal{D}(\mathcal{L})$, the problematic term is $P(AS + SA^*)P$. By the properties of S, this belongs to $\mathcal{H}$. For $\mathcal{D}\left(\mathcal{L}^2\right)$, employing the same argument unavoidably leads to the term

$$P\big(A(AS + SA^*) + (AS + SA^*)A^*\big)P,$$

or variations thereof. To ensure that this belongs to $\mathcal{H}$, and in fact that the middle expression is well defined at all, we need to assume that $AS+SA^* \in \mathscr{L}(V^*, H) \cap \mathscr{L}(H, V)$, similar to assuming $P \in \mathcal{V}$. However, with the typical $A = -\Delta$, this excludes e.g. the identity operator on $\mathcal{V}$, and essentially requires S to be more regularising than Δ^{-1}. In the LQR applications, it excludes all simple input operators of the form $B : \mathbb{R}^m \to V$ with $Bu = \sum_{k=1}^m u_k v_k$ for $v_k \in V$. It is therefore an unreasonably strict assumption. However, as mentioned in the introduction, the Hilbert-Schmidt setting is stronger than usual, and we expect that higher-order convergence could be shown in a weaker setting.

6. Preservation of positivity

In the traditional matrix-valued setting of the Riccati equation, it can be shown that the solution to (1) is self-adjoint and positive semi-definite if the same holds for Q, S and P_0. We therefore introduce the closed and convex cone $\mathcal{C}$ of such operators,

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

and make the following additional assumption:

Assumption 5. *Let Assumptions 1 and 2 be satisfied, with $Q \in \mathcal{C}$. Further, either let Assumption 3 be satisfied with $S \in \mathcal{C}$ or let Assumption 4 be satisfied with $s \geq 0$.*

There are several approaches for showing that the solution $\mathrm{e}^{-t(\mathcal{F}+\mathcal{G})}P_0$ belongs to $\mathcal{C}$ under Assumption 5 in the matrix-valued case. These all fail in the current operator-valued setting, for various reasons. For the case $\mathcal{G}P = P^2$, a proof is given in [2, Section III:2.3], but this depends heavily on the accretiveness of $\mathcal{G}$. Since this is potentially lost under Assumption 3, there is no straightforward modification in this case. However, a rather concise proof can be constructed by employing the splitting approximation.

First consider the following characterization of the solution to the nonlinear subproblem:

Lemma 5. *Under Assumption 5, the solution $\mathrm{e}^{-t\mathcal{G}}P_0$ to the problem (4) with $P_0 \in \mathcal{H} \cap \mathcal{C}$ on a compact interval $t \in [0, T]$ is given by*

$$\mathrm{e}^{-t\mathcal{G}}P_0 = (I + tP_0S)^{-1}P_0.$$

This holds in the case of Assumption 3. Under Assumption 4 we have instead $\mathrm{e}^{-t\mathcal{G}}P_0 = (I + tsP_0)^{-1}P_0$.

Proof. We consider only the case of Assumption 3, as the proof for Assumption 4 is virtually the same, but more simple. We first note that for any $U, V \in \mathcal{C}$, we have

$$\|(I + UV)^{-1}\|_{\mathscr{L}(H,H)} \leq 1 + \|U\|_{\mathscr{L}(H,H)}\|V\|_{\mathscr{L}(H,H)}$$

see e.g. [11, Lemma 2A.1]. Thus the function $P(t)$ given by

$$t \mapsto (I + tP_0S)^{-1}P_0$$

is well defined for all $t \geq 0$, and maps into $\mathcal{H}$. Further, we have

$$\begin{aligned}
&\|(I+(t+h)P_0S)^{-1}P_0 - (I+tP_0S)^{-1}P_0\|_{\mathcal{H}} = \\
&\qquad = \big\|\big(I+(t+h)P_0S\big)^{-1}\Big[(I+tP_0S) - \big(I+(t+h)P_0S\big)\Big](I+tP_0S)^{-1}P_0)\big\|_{\mathcal{H}} \\
&\qquad = \big\| - h\big(I+(t+h)P_0S\big)^{-1}P_0S(I+tP_0S)^{-1}P_0\big\|_{\mathcal{H}} \\
&\qquad \leq h\|(I+(t+h)P_0S)^{-1}\|_{\mathscr{L}(H,H)}\|P_0S\|_{\mathscr{L}(H,H)}\|(I+tP_0S)^{-1}\|_{\mathscr{L}(H,H)}\|P_0S\|_{\mathcal{H}} \\
&\qquad \leq h\big(1+(t+h)\|P_0\|_{\mathscr{L}(H,H)}\|S\|_{\mathscr{L}(H,H)}\big)\big(1+t\|P_0\|_{\mathscr{L}(H,H)}\|S\|_{\mathscr{L}(H,H)}\big)\|P_0\|_{\mathcal{H}}^2 \\
&\qquad \leq h\,C(\|P_0\|_{\mathcal{H}}, \|S\|_{\mathcal{H}}),
\end{aligned}$$

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

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

The function $t \mapsto P(t)$ is thus continuously differentiable and satisfies the equation (4), which means that it must be equal to $\mathrm{e}^{-t\mathcal{G}}P_0$.

□

Corollary 1. *Let Assumption 5 be valid and $T > 0$. Then with $0 \leq nh \leq T$ and h sufficiently small, both $\mathcal{S}_h^n P_0$ and $\mathrm{e}^{-nh(\mathcal{F}+\mathcal{G})}P_0$ belong to $\mathcal{C}$.*

Proof. We have that $\mathcal{H}\cap\mathcal{C}$ is invariant under $\mathrm{e}^{-t\mathcal{L}}$ [2, Lemma II:3.5], from which it follows by the variation of constants formula that it is also invariant under $\mathrm{e}^{-t\mathcal{F}}$. Consider next the operator $\mathrm{e}^{-t\mathcal{G}}$. For all $P_0 \in \mathcal{C}$, there exists $P_0^{1/2} \in \mathcal{C}$ such that $P_0 = P_0^{1/2}P_0^{1/2}$. By [11, Lemma 2A.1], we have

$$(I+tP_0S)^{-1} = I - tP_0^{1/2}(I+tP_0^{1/2}SP_0^{1/2})^{-1}P_0^{1/2}S,$$

which means that

$$\begin{aligned}
\mathrm{e}^{-t\mathcal{G}}P_0 &= (I+tP_0S)^{-1}P_0 \\
&= P_0^{1/2}P_0^{1/2} - P_0^{1/2}\Big((I+tP_0^{1/2}SP_0^{1/2})^{-1}tP_0^{1/2}SP_0^{1/2}\Big)P_0^{1/2} \\
&= P_0^{1/2}(I+tP_0^{1/2}SP_0^{1/2})^{-1}P_0^{1/2}.
\end{aligned}$$

But this expression is clearly self-adjoint and positive semi-definite, i.e. $\mathrm{e}^{-t\mathcal{G}}P_0 \in \mathcal{C}$. Thus the splitting approximation $\mathcal{S}_h^n P_0$ belongs to $\mathcal{C}$ for all $P_0 \in \mathcal{C}$. Now by Theorem 1, $\mathrm{e}^{-nh(\mathcal{F}+\mathcal{G})}P_0$ is the limit of a sequence of approximations all belonging to the closed set $\mathcal{C}$, and thus also belongs to $\mathcal{C}$.

□

Acknowledgement. The author would like to thank Eskil Hansen for valuable input during the preparation of this work.

References

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

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

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

[4] 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).

[5] P. Benner, H. Mena, *Rosenbrock methods for solving Riccati differential equations*, IEEE Trans. Automat. Control, 58(11), (2013), pp. 2950–2956.

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

[7] 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.

[8] E. Hansen, T. Stillfjord, *Convergence analysis for splitting of the abstract Riccati equation*, accepted for publication in SIAM J. Numer. Anal. (2014).

[9] E. Hansen, A. Ostermann, *Dimension splitting for evolution equations*, Numer. Math., 108 (2008), pp. 557–570.

[10] E. Hansen, F. Kramer, A. Ostermann, *A second-order positivity preserving scheme for semilinear parabolic problems*, Appl. Numer. Math. 62 (2012), pp. 1428–1435.

[11] I. Lasiecka, R. Triggiani, *Control theory for partial differential equations: Continuous and approximation theories*, vol. 1., Cambridge University Press, Cambridge, 2000.

[12] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.

[13] T. Jahnke, C. Lubich, *Error bounds for exponential operator splittings*, BIT 40(2) 2000, pp. 735–744.

[14] H. Mena, *Numerical Solution of Differential Riccati Equations Arising in Optimal Control Problems for Parabolic Partial Differential Equations*, Ph.D. dissertation, Escuela Politécnica Nacional, Quito, Ecuador, 2012.

[15] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Springer, New York, 1983.

[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] T. Stillfjord, *Low-rank second-order splitting of large-scale differential Riccati equations*, preprint (2014).

[18] R. Temam, *Sur l'équation de Riccati associée à des opérateurs non bornés, en dimension infinie*, J. Funct. Anal., 7 (1971), pp. 85–115.