MULTISCALE DIFFERENTIAL RICCATI EQUATIONS FOR LINEAR QUADRATIC REGULATOR PROBLEMS*

AXEL MÅLQVIST†, ANNA PERSSON†, AND TONY STILLFJORD†

Abstract. We consider approximations to the solutions of differential Riccati equations in the context of linear quadratic regulator problems, where the state equation is governed by a multiscale operator. Similarly to elliptic and parabolic problems, standard finite element discretizations perform poorly in this setting unless the grid resolves the fine-scale features of the problem. This results in unfeasible amounts of computation and high memory requirements. In this paper, we demonstrate how the localized orthogonal decomposition method may be used to acquire accurate results also for coarse discretizations, at the low cost of solving a series of small, localized elliptic problems. We prove second-order convergence (except for a logarithmic factor) in the L^2 operator norm, and first-order convergence in the corresponding energy norm. These results are both independent of the multiscale variations in the state equation. In addition, we provide a detailed derivation of the fully discrete matrix-valued equations, and show how they can be handled in a low-rank setting for large-scale computations. In connection to this, we also show how to efficiently compute the relevant operator-norm errors. Finally, our theoretical results are validated by several numerical experiments.

Key words. Multiscale, localized orthogonal decomposition, finite elements, linear quadratic regulator problems, differential Riccati equations

AMS subject classifications. 49N10, 65N12, 65N30, 93C20

1. Introduction.

In a linear quadratic regulator (LQR) problem, the state x is a model of a system whose evolution can be influenced through the input u. The goal is to drive certain measurable quantities of the system, the output y, to a given target which is typically zero. The relations between x, u and y are given by the state and output equations

$$\dot{x} = \mathcal{A}x + \mathcal{B}u, \quad x(0) = x_0, \tag{1}$$

$$y = \mathcal{C}x, \tag{2}$$

where $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are given operators. The optimal input function u^* is found by minimizing the cost functional

$$J(u) = \int_0^T (\mathcal{Q}y, y) + (\mathcal{R}u, u)\ \mathrm{d}t + (\mathcal{G}y(T), y(T))\,,$$

where $\mathcal{Q}$, $\mathcal{R}$ and $\mathcal{G}$ are given weighting factors. It can be shown (see e.g. [1, 21]) that u^* is given in feedback form as $u^*(t) = -\mathcal{R}^{-1}\mathcal{B}^*X(T-t)x(t)$, where X is the solution to an operator-valued differential Riccati equation (DRE):

$$\begin{aligned} &\dot{X}(t) = \mathcal{A}^*X(t) + X(t)\mathcal{A} + \mathcal{C}^*\mathcal{Q}\mathcal{C} - X(t)\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*X(t), \\ &X(0) = \mathcal{G}. \end{aligned} \tag{3}$$

In the case of a nonzero output target, one additional differential equation for the evolution of u^* has to be solved.

*Submitted to the editors 2017-06-13.

Funding: This work was supported by the Swedish Research Council under grant no. 2015-04964.

†Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg (axel@chalmers.se, peanna@chalmers.se, tony.stillfjord@gu.se.)

In this paper, we consider the case when the operator $\mathcal{A}$ exhibits multiscale behaviour. In particular, we consider diffusion problems where the spatial variation of the diffusion coefficient is on a fine scale compared to the computational domain. This e.g. occurs in the modeling of composite materials and flows in porous media. Numerically approximating the solutions to elliptic or parabolic equations given by such operators in the usual way is difficult, because a very fine discretization is necessary to resolve the fine-scale structure. These difficulties are exacerbated when considering DREs such as (3), as their solution essentially requires solving many parabolic equations.

A by now well established method for multiscale elliptic and parabolic problems is the localized orthogonal decomposition (LOD) [25, 15]. It is a modification of the finite element method (FEM), which incorporates some of the fine-scale structure into a coarse discretization by precomputing a series of localized fine-scale problems. Due to the localization, these are much cheaper to evaluate than the full fine-scale problem and may additionally be solved in parallel.

We note that finite elements were introduced for the approximation of optimal control problems already in the 1970's, see e.g. [27, 10, 14, 36], and the field has grown much in several different directions since then. When diffusion problems have been considered, the focus has typically been on constant or slowly varying diffusion. Recently, however, also optimal control problems of multiscale type have been considered in e.g. [11, 12, 22]. None of these consider the LOD approach, instead preferring homogenization or asymptotic expansions. Additionally, a common assumption is that the multiscale features are periodic, which is frequently not the case in applications.

The focus in this paper is on the approximation of DREs such as (3). In contrast to the forward-adjoint approach, which solves a specific optimal control problem, the DRE provides the feedback laws for all problems defined by the operators $\mathcal{A}, \mathcal{B}, \mathcal{C}$. While more expensive to solve, it can be precomputed and reused in many different situations. We refer to [8, 21] for an overview of Riccati theory, with the latter reference treating very general problems.

Our main result is that LOD-approximations to the solution of (3) with a mesh size H converge with order $H^2 \log(H^{-1})$ in the L^2 operator norm to a given accurate fine-scale FEM approximation. The convergence in the corresponding operator energy norm is shown to be of order H. We note that $H^2 \log(H^{-1})$-convergence of FEM approximations to the exact solution of (3) has previously been shown in [18], and similar results for algebraic Riccati equations can be found in [21]. (See also [30, 6] for convergence results without orders in related settings.) However, the error constants in these results depend on the multiscale variations of $\mathcal{A}$, and thus such convergence is not observed in practice. This is not the case for our present results.

For practical computations, also a temporal discretization is necessary; for this we consider a low-rank splitting scheme as introduced in [32]. Such methods decompose the DRE into its affine and nonlinear parts and approximate these separately, thereby greatly reducing the computational cost. The affine problem requires the approximation of several parabolic equations involving $\mathcal{A}$ in each time step. As the computational efficiency gain for LOD increases with the number of times the modified basis may be reused, splitting schemes are thus particularly well suited to be combined with the LOD method.

We demonstrate how to transform the FEM and LOD discretizations into matrix-valued equations, and how to implement the fully discrete methods. Even if LOD reduces the need for very fine discretizations, large 2D or 3D-problems may still yield large matrices. We therefore consider the low-rank approach, which greatly reduces

the necessary amount of computations. As a side effect, this also allows us to compute errors in the operator norms very efficiently.

A brief outline of the paper is as follows: We formalize the setting and our basic assumptions in section 2, and define the different spatial discretizations in section 3. Convergence of the LOD approximations with the appropriate order is then shown in section 4. The matrix-valued formulations of the discretized DREs and related questions are discussed in section 5, while section 6 is devoted to the temporal discretization and low-rank setting. Finally, we present several numerical experiments and their results in section 7.

2. Setting. Let $\Omega \in \mathbb{R}^d$, $d \leq 3$, be a bounded polygonal/polyhedral domain. We consider the separable Hilbert spaces $L^2(\Omega)$, $V = H_0^1(\Omega)$, U and Z, where $L^2(\Omega)$ corresponds to the state space, U is the control space and Z is the observation space. In the following, the specification of Ω will be omitted. We write $(\cdot,\cdot)$ and $\|\cdot\|$ for the inner product and norm on L^2, and denote the corresponding quantities on V, U and Z by subscripts. To define the state evolution operator $\mathcal{A}$, we assume that the inner product $a(u,v) = \int \kappa \nabla u \cdot \nabla v$ on $V \times V$ is given, with assumptions on κ given below. Then $\mathcal{A}\colon L^2 \supset \mathcal{D}(\mathcal{A}) \to L^2$ is defined by $(\mathcal{A}u, v) = -a(u,v)$ and $\mathcal{D}(\mathcal{A}) = \{u \in V \mid \mathcal{A}u \in L^2\}$.

Further, let the input operator $\mathcal{B}\colon U \to L^2$ and the output operator $\mathcal{C}\colon L^2 \to Z$ be given. We also consider the output and input weighting operators $\mathcal{Q}\colon Z \to Z$ and $\mathcal{R}\colon U \to U$ (which could be included in $\mathcal{C}$ and $\mathcal{B}$ but are typically not) and the final state weighting operator $\mathcal{G}\colon L^2 \to L^2$. By *, we denote Hilbert-adjoint operators with respect to L^2, so that e.g. $\mathcal{B}^*\colon L^2 \to U$ satisfies $(\mathcal{B}x, y) = (x, \mathcal{B}^* y)$ for all $x \in U$ and $y \in L^2$. Finally, we denote the linear bounded operators from one generic Hilbert space, Y, to another, W, by $\mathcal{L}(Y,W)$. When $W = Y$, we abbreviate $\mathcal{L}(Y) = \mathcal{L}(Y,Y)$.

In this notation, the weak form of (3) is to find $X \in \mathcal{L}(L^2)$ satisfying

$$\left(\dot{X}x, y\right) = (Xx, \mathcal{A}y) + (Xy, \mathcal{A}x) + (\mathcal{Q}\mathcal{C}x, \mathcal{C}y)_Z - \left(\mathcal{R}^{-1}\mathcal{B}^* Xx, \mathcal{B}^* Xy\right)_U, \tag{4}$$

for all $x, y \in \mathcal{D}(\mathcal{A})$.

Assumption 2.1. *The diffusion coefficient $\kappa \in L^\infty(\mathbb{R}^{d\times d})$ is symmetric and satisfies*

$$0 < \alpha := \operatorname*{ess\,inf}_{x\in\Omega} \inf_{v\in\mathbb{R}^d\setminus\{0\}} \frac{\kappa(x)v\cdot v}{v\cdot v},$$
$$\infty > \beta := \operatorname*{ess\,sup}_{x\in\Omega} \sup_{v\in\mathbb{R}^d\setminus\{0\}} \frac{\kappa(x)v\cdot v}{v\cdot v}.$$

In addition, $\mathcal{B} \in \mathcal{L}(U, L^2)$, $\mathcal{C} \in \mathcal{L}(L^2, Z)$, $\mathcal{Q} \in \mathcal{L}(Z)$, $\mathcal{R} \in \mathcal{L}(U)$ is invertible with $\mathcal{R}^{-1} \in \mathcal{L}(U)$ and $X(0) = \mathcal{G} \in \mathcal{L}(L^2)$.

The first part of Assumption 2.1 shows that a is a bounded and coercive bilinear form, which means that $\mathcal{A}$ is the generator of an analytic semigroup $\mathrm{e}^{t\mathcal{A}}\colon L^2 \to L^2$, see e.g. [34, Theorem 3.6.1]. In conjunction with the boundedness assumptions on $\mathcal{B}$, $\mathcal{C}$, $\mathcal{Q}$ and $\mathcal{R}$, this guarantees the existence and uniqueness of a solution to (4). In fact, there is even a classical solution to (3) [8, Part IV, Ch. 3], which means that the $\mathcal{A}^*X + X\mathcal{A}$ term can be extended to an operator in $\mathcal{L}(L^2)$. As a consequence, Equation (4) holds also for $x, y \in L^2$. We note that these conclusions are valid also under various weaker forms of Assumption 2.1, which additionally permit the treatment of boundary control and observation [21]. A discussion on an extension of our results to such a setting may be found in subsection 8.1.

3. Spatial discretization. We first introduce the FEM approximation of (4). To this end, we let $\mathcal{T}_h$ be a triangulation of Ω with meshwidth h and N_h internal nodes. The subspace $V_h \subset V$ denotes the space of continuous and piecewise affine functions on $\mathcal{T}_h$, and we denote the corresponding nodal basis functions by $\{\varphi_i^h\}_{i=1}^{N_h}$. This discretization is referred to as the fine, or sometimes also reference, mesh, see further subsection 3.1 below.

We also consider a coarse discretization space $V_H \subset V_h$ for $H > h$, with the corresponding family of triangulations $\{\mathcal{T}_H\}_{H>h}$, which is assumed to be quasi-uniform. For these triangulations, we let B_K be the largest ball contained in the triangle K and denote by $\gamma > 0$ the shape regularity of the mesh, defined by

$$\gamma := \max \gamma_K, \quad \gamma_K := \frac{\operatorname{diam} B_K}{\operatorname{diam} K}, \quad \forall K \in \mathcal{T}_H,\ H > h.$$

Furthermore, we let $\mathrm{Id}_H^h \colon V_H \to V_h$ denote the identity operator between these spaces, i.e. $\mathrm{Id}_H^h\, u = u$ for all $u \in V_H$. Similarly, $\mathrm{Id}_h \colon V_h \to L^2$ is the identity operator mapping into L^2 and Id_h^* is the L^2-orthogonal projection of L^2 onto V_h.

The semi-discretized weak form of (3) is defined by

$$\left(\dot{X}_h x, y\right) = (X_h x, \mathcal{A}_h y) + (X_h y, \mathcal{A}_h x) + (\mathcal{Q}\mathcal{C}_h x, \mathcal{C}_h y)_Z - \left(\mathcal{R}^{-1}\mathcal{B}_h^* X_h x, \mathcal{B}_h^* X_h y\right)_U. \tag{5}$$

for all $x, y \in V_h$ and with $X_h \colon V_h \to V_h$ satisfying $X_h(0) = \mathrm{Id}_h^*\, X(0)\, \mathrm{Id}_h$. Here, the operators $\mathcal{A}_h \colon V_h \to V_h$, $\mathcal{B}_h \colon U \to V_h$ and $\mathcal{C}_h \colon V_h \to Z$ satisfy

$$(\mathcal{A}_h x, y) = (\mathcal{A} x, y), \quad (\mathcal{B}_h u, y) = (\mathcal{B} u, y) \quad \text{and} \quad (\mathcal{C}_h x, z) = (\mathcal{C} x, z)$$

for all $x, y \in V_h$, $u \in U$ and $z \in Z$. We note that X can be proven to be self-adjoint, so we additionally require that X_h is self-adjoint.

For the coarse discretization, we have the same equation but with H instead of h. We observe that the coarse and fine operators are related in the following way:

$$\mathcal{A}_H = (\mathrm{Id}_H^h)^* \mathcal{A}_h\, \mathrm{Id}_H^h, \quad \mathcal{B}_H = (\mathrm{Id}_H^h)^* \mathcal{B}_h \quad \text{and} \quad \mathcal{C}_H = \mathcal{C}_h\, \mathrm{Id}_H^h, \tag{6}$$

and that the natural extension of X_H to a map on V_h is given by $\mathrm{Id}_H^h\, X_H (\mathrm{Id}_H^h)^*$. Here, $(\mathrm{Id}_H^h)^*$ is the L^2-orthogonal projection of V_h onto V_H.

3.1. Localized orthogonal decomposition. If κ is varying on a small scale of size $\epsilon > 0$, then the classical FEM approximation of a parabolic problem $\dot{x} = \mathcal{A}x + f$ may yield poor results, unless h is sufficiently small to resolve the fine-scale variations. That is, we typically do not observe $O(h^2)$-convergence until $h < \epsilon$, which requires infeasible amounts of computation. The same behaviour occurs for the X_h-discretizations of (4).

To this end, we assume that h is sufficiently small so that X_h is a good approximation of X. That is, $h < \epsilon$, and we refer to X_h as the reference solution. The aim is now to approximate X_h by using a multiscale space $V_{\mathrm{ms}} \subset V_h$ of the same dimension as the coarse space V_H. To obtain such a space, we use the localized orthogonal decomposition (LOD) method introduced in [25], which incorporates fine-scale information in the coarse-scale space. The construction involves the solution of several fine-scale, but localized and parallelizable, problems. We briefly summarize the procedure here and refer to [25, 15], for the details.

To define the multiscale space V_{ms}, we first introduce an interpolation operator $I_H \colon V_h \to V_H$ that fulfills

$$H^{-1}\|v - I_H v\|_{L^2(K)} + \|\nabla I_H v\|_{L^2(K)} \leq C \|\nabla v\|_{L^2(\omega_K)}, \forall v \in V_h,$$

for all triangles $K \in \mathcal{T}_H$, where $\omega_K := \cup\{\hat{K} \in \mathcal{T}_H : \hat{K} \cap K \neq \emptyset\}$. In this paper we use the weighted Clément interpolant as in [25]. Let V_{f} denote the kernel of I_H,

$$V_{\mathrm{f}} := \ker I_H = \{v \in V_h : I_H v = 0\},$$

and note that V_h can be decomposed as $V_h = V_H \oplus V_{\mathrm{f}}$, meaning that every $v_h \in V_h$ can be written as $v_h = v_H + v_{\mathrm{f}}$ with $v_H \in V_H, v_{\mathrm{f}} \in V_{\mathrm{f}}$. We now introduce the (global) correction operator $\hat{Q}_h \colon V_H \to V_{\mathrm{f}}$ by

$$a(\hat{Q}_h v, w) = a(v, w), \quad \forall w \in V_{\mathrm{f}},$$

and define the (global) multiscale space as $\hat{V}_{\mathrm{ms}} := \hat{R}_h V_H = V_H - \hat{Q}_h V_H$, with $\hat{R}_h := \mathrm{Id}_H^h - \hat{Q}_h$. This leads to the decomposition $V_h = \hat{V}_{\mathrm{ms}} \oplus V_{\mathrm{f}}$ with the orthogonality $a(\hat{v}_{\mathrm{ms}}, v_{\mathrm{f}}) = 0$, $\hat{v}_{\mathrm{ms}} \in \hat{V}_{\mathrm{ms}}$, $v_{\mathrm{f}} \in V_{\mathrm{f}}$. Note that $\hat{Q}_h$ is the orthogonal projection onto V_{f} with respect to the inner product $a(\cdot,\cdot)$, i.e. the Ritz projection onto V_{f}, and $\hat{V}_{\mathrm{ms}}$ is the orthogonal complement to V_{f}. From the construction it follows that $\dim \hat{V}_{\mathrm{ms}} = \dim V_H$. Indeed, a basis for $\hat{V}_{\mathrm{ms}}$ is given by $\{\varphi_i^H - \hat{Q}_h \varphi_i^H : i = 1, ..., N_H\}$.

In general, the corrections $\hat{Q}_h \varphi_i^H$ have global support and are expensive to compute, since they are posed in the entire fine scale space $V_{\mathrm{f}} \subseteq V_h$. To overcome this, it is observed that the corrections have exponential decay away from the i:th node of $\mathcal{T}_H$ (see [25, 15]), which motivates a truncation of the corrections. For this purpose, we define patches $\omega_k(K)$ of size k around each $K \in \mathcal{T}_H$ by the following:

$$\begin{aligned}
\omega_0(K) &:= \operatorname{int} K, \\
\omega_k(K) &:= \operatorname{int}\big(\cup\{\hat{K} \in \mathcal{T}_H : \hat{K} \cap \overline{\omega_{k-1}(K)} \neq \emptyset\}\big), \quad k = 1, 2, ...
\end{aligned}$$

Further, we define $V_{\mathrm{f}}^K := \{v \in V_{\mathrm{f}} : v(z) = 0 \text{ on } \overline{\Omega} \setminus \omega_k(K)\}$ to be the restriction of V_{f} to the patch $\omega_k(K)$. For brevity, we do not include the dependence on k in the notation. Now note that the correction operator $\hat{Q}_h$ can be written as the sum $\hat{Q}_h = \sum_{K \in \mathcal{T}_H} \hat{Q}_h^K$, where

$$a(\hat{Q}_h^K v, w) = \int_K \kappa \nabla v \cdot \nabla w, \quad \forall w \in V_{\mathrm{f}},\ v \in V_H,\ K \in \mathcal{T}_H.$$

We can now localize these computations by replacing V_{f} with V_{f}^K. Define $Q_h^K \colon V_H \to V_{\mathrm{f}}^K$ such that

$$a(Q_h^K v, w) = \int_K \kappa \nabla v \cdot \nabla w, \quad \forall w \in V_{\mathrm{f}}^K,\ v \in V_H,\ K \in \mathcal{T}_H.$$

Finally, we can define a local operator $Q_h := \sum_{K \in \mathcal{T}_H} Q_h^K$ and a localized space $V_{\mathrm{ms}} := R_h V_H = V_H - Q_h V_H$, with $R_h := \mathrm{Id}_H^h - Q_h$.

The approximation properties (and the required computational effort) of the space V_{ms} depends on the choice of k. In [15] it is proven that convergence of order H^2 is obtained if k is chosen proportional to $\log H^{-1}$. In this paper we therefore assume that $k \sim \log H^{-1}$ to avoid explicitly stating the dependence on k.

To define an LOD-approximation to the solution X_h in (5), we additionally need to introduce the identity operator $\mathrm{Id}_{\mathrm{ms}}^h : V_{\mathrm{ms}} \to V_h$, $\mathrm{Id}_{\mathrm{ms}}^h u = u$. Its L^2-adjoint is the L^2-orthogonal projection of V_h onto V_{ms}. Replacing the space V_h with V_{ms} then results in the problem to find $X_h^{\mathrm{ms}} \colon V_{\mathrm{ms}} \to V_{\mathrm{ms}}$ satisfying

$$\begin{aligned}
\left(\dot{X}_h^{\mathrm{ms}} u, v\right) &= (X_h^{\mathrm{ms}} u, \mathcal{A}_h^{\mathrm{ms}} v) + (X_h^{\mathrm{ms}} v, \mathcal{A}_h^{\mathrm{ms}} u) \\
&\quad + (\mathcal{Q}\mathcal{C}_h^{\mathrm{ms}} u, \mathcal{C}_h^{\mathrm{ms}} v)_Z - \left(\mathcal{R}^{-1}(\mathcal{B}_h^{\mathrm{ms}})^* X_h^{\mathrm{ms}} u, (\mathcal{B}_h^{\mathrm{ms}})^* X_h^{\mathrm{ms}} v\right)_U
\end{aligned} \tag{7}$$

for all $u, v \in V_{\mathrm{ms}}$ and the initial condition $X_h^{\mathrm{ms}}(0) = (\mathrm{Id}_{\mathrm{ms}}^h)^* \,\mathrm{Id}_h^* \, X(0) \,\mathrm{Id}_h \,\mathrm{Id}_{\mathrm{ms}}^h$. Here, the operators $\mathcal{A}_h^{\mathrm{ms}} \colon V_{\mathrm{ms}} \to V_{\mathrm{ms}}$, $\mathcal{B}_h^{\mathrm{ms}} \colon U \to V_{\mathrm{ms}}$ and $\mathcal{C}_h^{\mathrm{ms}} \colon V_{\mathrm{ms}} \to Z$ are given by

$$(\mathcal{A}_h^{\mathrm{ms}} v, w) = (\mathcal{A} v, w)\,, \ (\mathcal{B}_h^{\mathrm{ms}} u, w)_U = (\mathcal{B} u, w)_U \ \text{ and } \ (\mathcal{C}_h^{\mathrm{ms}} v, z)_Z = (\mathcal{C} u, z)_Z$$

for all $v, w \in V_{\mathrm{ms}}$, $u \in U$ and $z \in Z$. Similar to (6) we have

$$\mathcal{A}_h^{\mathrm{ms}} = (\mathrm{Id}_{\mathrm{ms}}^h)^* \mathcal{A}_h \,\mathrm{Id}_{\mathrm{ms}}^h, \quad \mathcal{B}_h^{\mathrm{ms}} = (\mathrm{Id}_{\mathrm{ms}}^h)^* \mathcal{B}_h \quad \text{and} \quad \mathcal{C}_h^{\mathrm{ms}} = \mathcal{C}_h \,\mathrm{Id}_{\mathrm{ms}}^h \,. \tag{8}$$

The natural V_h-extension of X_h^{ms} is given by $\mathrm{Id}_{\mathrm{ms}}^h \, X_h^{\mathrm{ms}} (\mathrm{Id}_{\mathrm{ms}}^h)^*$, similar to the X_H-case.

Since V_{ms} has the same dimension as V_H, there is a lower-dimensional representative for X_h^{ms}, given by $X_h^{\mathrm{ms}} = R_h X_H^{\mathrm{ms}} R_h^{-1}$. By inserting $u = R_h x$ and $v = R_h y$, with $x, y \in V_H$, in (7) we see that

$$\begin{aligned}\left(\dot{X}_H^{\mathrm{ms}} x, R_h^* R_h y\right) &= (X_H^{\mathrm{ms}} x, R_h^* \mathcal{A}_h R_h y) + (X_H^{\mathrm{ms}} y, R_h^* \mathcal{A}_h R_h x) \\ &\quad + (\mathcal{Q} \mathcal{C}_h R_h x, \mathcal{C}_h R_h y)_Z - \left(\mathcal{R}^{-1} \mathcal{B}_h^* R_h X_H^{\mathrm{ms}} x, \mathcal{B}_h^* R_h X_H^{\mathrm{ms}} y\right)_U \,,\end{aligned}$$

and we consequently define the corrected coarse-scale operators

$$\mathcal{A}_H^{\mathrm{ms}} = R_h^* \mathcal{A}_h R_h, \quad \mathcal{B}_H^{\mathrm{ms}} = R_h^* \mathcal{B}_h \quad \text{and} \quad \mathcal{C}_H^{\mathrm{ms}} = \mathcal{C}_h R_h.$$

4. Error analysis. In the following, C denotes a generic constant which may take different values at different occasions. It may depend on the problem data and the size of the domain, but is independent of h and H. Moreover, it does not depend on the multiscale variations of $\mathcal{A}$, i.e. any derivatives of κ. We start by gathering some useful results:

4.1. Preliminaries. Recall that $\mathrm{Id}_h \colon V_h \to L^2$ is the identity mapping, $P_h = \mathrm{Id}_h^* \colon L^2 \to V_h$ denotes the L^2-orthogonal projection onto V_h, and P_{ms} is the L^2-orthogonal projection onto V_{ms}. We have $P_{\mathrm{ms}} = (\mathrm{Id}_{\mathrm{ms}}^h)^* P_h$, i.e. we first project onto V_h and then onto V_{ms}. Straightforward calculations show the following:

LEMMA 4.1. *Under Assumption 2.1, it holds that* $\mathrm{Id}_h \, \mathcal{B}_h \in \mathcal{L}(U, L^2)$, $\mathcal{C}_h P_h \in \mathcal{L}(L^2, Z)$ *and* $S_h := \mathrm{Id}_h \, \mathcal{B}_h \mathcal{R}^{-1} \mathcal{B}_h^* \, \mathrm{Id}_h^* \in \mathcal{L}(L^2)$.

Further, let $\mathrm{e}^{t\mathcal{A}_h}$ denote the solution operator to the equation $\dot{u} + \mathcal{A}_h u = 0$, i.e. the semigroup generated by $\mathcal{A}_h$. Similarly, $\mathrm{e}^{t\mathcal{A}_h^{\mathrm{ms}}}$ is the semigroup generated by $\mathcal{A}_h^{\mathrm{ms}}$. Because $\mathcal{A}$ generates an analytic semigroup on L^2, these operators are analytic semigroups on V_h and V_{ms}, respectively. More specifically, we have

LEMMA 4.2. *Under Assumption 2.1, the operators*

$$E_h(t) = \mathrm{Id}_h \, \mathrm{e}^{t\mathcal{A}_h} \, \mathrm{Id}_h^* \quad \textit{and} \quad E_{ms}(t) = \mathrm{Id}_h \, \mathrm{Id}_{ms}^h \, \mathrm{e}^{t\mathcal{A}_h^{ms}} (\mathrm{Id}_{ms}^h)^* \, \mathrm{Id}_h^*,$$

are both in $\mathcal{L}(L^2)$ *for* $t \in [0, T]$, *with the uniform bounds* $\|E_h(t)\|_{\mathcal{L}(L^2)} \leq 1$ *and* $\|E_{ms}(t)\|_{\mathcal{L}(L^2)} \leq 1$.

By arguing as in [24], but for the (simpler) semi-discrete case, we have (choosing $k \sim \log H$)

LEMMA 4.3. *For* $t \in (0, T]$ *it holds that*

$$\|E_h(t) - E_{ms}(t)\|_{\mathcal{L}(L^2)} \leq C H^2 t^{-1}.$$

Here, the constant C *depends on* T, α, β, *and* γ, *but not on the multiscale variations of* $\mathcal{A}$.

Proof. We only comment briefly on the proof here. Let $u_h(t) = e^{t\mathcal{A}_h} P_h v$ and $u_{\mathrm{ms}}(t) = e^{t\mathcal{A}_h^{\mathrm{ms}}} P_{\mathrm{ms}} v$, for $v \in L^2(\Omega)$. By introducing the Ritz projection $R_{\mathrm{ms}} \colon V_h \to V_{\mathrm{ms}}$ satisfying $a(R_{\mathrm{ms}} v, w) = a(v, w)$ for all $w \in V_{\mathrm{ms}}, v \in V_h$ we get, see [35, Chapter 3] and [24],

$$\|u_h - u_{\mathrm{ms}}\| \le C t^{-1} \sup_{s \le t} \Big\{ s^2 \|\dot\rho\| + s\|\rho\| + \|\int_0^s \rho(r)\,\mathrm{d}r\| \Big\},$$

where $\rho := u_h - R_{\mathrm{ms}} u_h$. From the error bounds of R_{ms} in [25], see also [24], we get

$$\|u_h - u_{\mathrm{ms}}\| \le C H^2 t^{-1} \sup_{s \le t} \big(s^2 \|\ddot u_h(s)\| + s\|\dot u_h(s)\| + \|u_h(s)\| + \|v\| \big).$$

The regularity estimates $\|D_t^l u_h(t)\| \le C t^{-l} \|v\|$, for $t = 0, 1, 2$, [35, Lemma 2.5], completes the proof. ∎

Finally, from Lemma 4.1, we get the existence and uniqueness of solutions X_h and X_h^{ms} to the discretized DREs (5) and (7), respectively. Let us abbreviate

$$\tilde X(t) = \mathrm{Id}_h \, X_h(t) \, \mathrm{Id}_h^* \quad \text{and} \quad \tilde Y(t) = \mathrm{Id}_h \, \mathrm{Id}_{\mathrm{ms}}^h \, X_h^{\mathrm{ms}} (\mathrm{Id}_{\mathrm{ms}}^h)^* \, \mathrm{Id}_h^* \,.$$

Then we have

LEMMA 4.4. *There is a constant $C > 0$ which is independent of the multiscale variations of $\mathcal{A}$ but may depend on α and β, such that*

$$\|\tilde X(t)\|_{\mathcal{L}(L^2)} + \|\tilde Y(t)\|_{\mathcal{L}(L^2)} \le C,$$

for $t \in [0, T]$.

4.2. Error analysis. We are now ready for the main theorem of this paper:

THEOREM 4.5. *Suppose that Assumption 2.1 is fulfilled. Then for $t \in (0, T]$ it holds that*

$$\|\tilde X(t) - \tilde Y(t)\|_{\mathcal{L}(L^2)} \le C H^2 \big(\log(H^{-1}) + t^{-1} \big).$$

Here, the constant C depends on T, α, β, γ, and $\|X(0)\|_{\mathcal{L}(L^2)}$, but not on the multiscale variations of $\mathcal{A}$.

Proof. We utilize the integral form of (5). If X_h solves (5) then it satisfies

(9)
$$X_h(t) = \mathrm{e}^{t\mathcal{A}_h^*} X_h(0) \mathrm{e}^{t\mathcal{A}_h} + \int_0^t \mathrm{e}^{(t-s)\mathcal{A}_h^*} \Big(\mathcal{C}_h^* \mathcal{Q} \mathcal{C}_h - X_h(s) \mathcal{B}_h \mathcal{R}^{-1} \mathcal{B}_h^* X_h(s) \Big) \mathrm{e}^{(t-s)\mathcal{A}_h} \,\mathrm{d}s.$$

(see e.g. [8, Chapter IV:3, Proposition 2.1]). Recalling that $\mathrm{Id}_h^* \, \mathrm{Id}_h$ and $(\mathrm{Id}_{\mathrm{ms}}^h)^* \, \mathrm{Id}_{\mathrm{ms}}^h$ are the identity operators on V_h and V_{ms}, respectively, and using (8) therefore shows that

$$\begin{aligned} \tilde X(t) &= E_h(t)^* \tilde X(0) E_h(t) \\ &\quad + \int_0^t E_h(t-s)^* \Big((\mathcal{C}_h P_h)^* \mathcal{Q} \mathcal{C}_h P_h - \tilde X(s) S_h \tilde X(s) \Big) E_h(t-s) \,\mathrm{d}s, \end{aligned}$$

as well as

$$\begin{aligned} \tilde Y(t) &= E_{\mathrm{ms}}(t)^* \tilde X(0) E_{\mathrm{ms}}(t) \\ &\quad + \int_0^t E_{\mathrm{ms}}(t-s)^* \Big((\mathcal{C}_h P_h)^* \mathcal{Q} \mathcal{C}_h P_h - \tilde Y(s) S_h \tilde Y(s) \Big) E_{\mathrm{ms}}(t-s) \,\mathrm{d}s. \end{aligned}$$

(Note the $\tilde{X}(0)$ in the first term, since we suppose $X_h^{\mathrm{ms}}(0) = (\mathrm{Id}_{\mathrm{ms}}^h)^* X_h(0)\,\mathrm{Id}_{\mathrm{ms}}^h$.) Subtracting these expressions yields

$$\begin{aligned}
\tilde{X}(t) - \tilde{Y}(t) &= E_h(t)^* \tilde{X}(0)\Big(E_h(t) - E_{\mathrm{ms}}(t)\Big) + \Big(E_h(t) - E_{\mathrm{ms}}(t)\Big)^* \tilde{X}(0) E_{\mathrm{ms}}(t) \\
&\quad + \int_0^t E_h(t-s)^* (\mathcal{C}_h P_h)^* \mathcal{Q} (\mathcal{C}_h P_h)\Big(E_h(t-s) - E_{\mathrm{ms}}(t-s)\Big) \\
&\quad + \Big(E_h(t-s) - E_{\mathrm{ms}}(t-s)\Big)^* (\mathcal{C}_h P_h)^* \mathcal{Q} (\mathcal{C}_h P_h) E_{\mathrm{ms}}(t-s) \\
&\quad + \Big(E_h(t-s) - E_{\mathrm{ms}}(t-s)\Big)^* \tilde{X}(s) S_h \tilde{X}(s) E_h(t-s) \\
&\quad + E_{\mathrm{ms}}(t-s)^* \tilde{X}(s) S_h \tilde{X}(s)\Big(E_h(t-s) - E_{\mathrm{ms}}(t-s)\Big) \\
&\quad + E_{\mathrm{ms}}(t-s)^* \Big(\tilde{X}(s) - \tilde{Y}(s)\Big) S_h \tilde{X}(s) E_{\mathrm{ms}}(t-s) \\
&\quad + E_{\mathrm{ms}}(t-s)^* \tilde{Y}(s) S_h \Big(\tilde{X}(s) - \tilde{Y}(s)\Big) E_{\mathrm{ms}}(t-s)\,\mathrm{d}s \\
&=: R_1 + R_2 + \int_0^t \sum_{j=3}^{8} R_j(s)\,\mathrm{d}s,
\end{aligned}$$

so that

$$\|\tilde{X}(t) - \tilde{Y}(t)\|_{\mathcal{L}(L^2)} \le \|R_1\|_{\mathcal{L}(L^2)} + \|R_2\|_{\mathcal{L}(L^2)} + \int_0^t \Big\| \sum_{j=3}^{8} R_j(s) \Big\|_{\mathcal{L}(L^2)} \mathrm{d}s.$$

We observe that for all $G\colon L^2 \to Y$ (with a generic Hilbert space Y) it holds that $\|G\|_{\mathcal{L}(L^2,Y)} = \|G^*\|_{\mathcal{L}(Y,L^2)}$. Thus, using Lemmas 4.2 to 4.4 we get

$$\|R_1\|_{\mathcal{L}(L^2)} = \|R_2\|_{\mathcal{L}(L^2)} \le CH^2 t^{-1} \|\tilde{X}(0)\|_{\mathcal{L}(L^2)} \le CH^2 t^{-1},$$

Additionally using Lemma 4.1 shows that the last two integrands satisfy

$$\begin{aligned}
\|R_7(s) + R_8(s)\|_{\mathcal{L}(L^2)} &\le C\Big(\|\tilde{X}(s)\|_{\mathcal{L}(L^2)} + \|\tilde{Y}(s)\|_{\mathcal{L}(L^2)}\Big)\|\tilde{X}(s) - \tilde{Y}(s)\|_{\mathcal{L}(L^2)} \\
&\le C\|\tilde{X}(s) - \tilde{Y}(s)\|_{\mathcal{L}(L^2)}.
\end{aligned}$$

Due to the singularity at $s = t$ in the bound on $\|E_h(t-s) - E_{\mathrm{ms}}(t-s)\|_{\mathcal{L}(L^2)}$, we split the integrals of the remaining R_j-terms into two parts. For R_3, we find

$$\begin{aligned}
\int_0^t \|R_3(s)\|_{\mathcal{L}(L^2)}\,\mathrm{d}s &\le \int_0^{t-H^2} C\|\mathcal{C}_h P_h\|_{\mathcal{L}(L^2)}^2 H^2 (t-s)^{-1}\,\mathrm{d}s + \int_{t-H^2}^t 2\|\mathcal{C}_h P_h\|_{\mathcal{L}(L^2)}^2\,\mathrm{d}s \\
&\le CH^2\big(\log t - 2\log H\big) + CH^2 \\
&\le CH^2\big(\log(H^{-1}) + t^{-1}\big),
\end{aligned}$$

where we have used $t \le T$ for the crude estimate $\log t \le Ct^{-1}$, since a t^{-1}-term already appears in the bounds of R_1 and R_2. The same bound holds for R_4, and, by Lemma 4.1 and Lemma 4.4, also for R_5 and R_6. In conclusion, we thus have

$$\|\tilde{X}(t) - \tilde{Y}(t)\|_{\mathcal{L}(L^2)} \le CH^2\big(\log(H^{-1}) + t^{-1}\big) + C\int_0^t \|\tilde{X}(s) - \tilde{Y}(s)\|_{\mathcal{L}(L^2)}\,\mathrm{d}s,$$

which by Grönwall's lemma yields the statement of the theorem. ∎

Remark 4.6. In the common situation that $X(0) = 0$, corresponding to the case of no final state penalization, the t^{-1}-singularity disappears.

Remark 4.7. We note that a bound of the same form has been shown in [18] for the FEM error. However, the error constant then depends on the variations in κ, and one does not observe the given convergence order until $H < \epsilon$.

Similar to the parabolic case, the error bound becomes less singular near $t = 0$ if we measure in the V-norm. To prove this we need the following, slightly stronger, assumptions on the operators (cf. Assumption 2.1):

ASSUMPTION 4.8. *In addition to Assumption 2.1, $\mathcal{B} \in \mathcal{L}(U, V)$, $\mathcal{C} \in \mathcal{L}(V, Z)$, and $X(0) = \mathcal{G} \in \mathcal{L}(V)$. Moreover, we assume that the mesh $\mathcal{T}_h$ is of a form such that P_h is stable in the V-norm.*

Remark 4.9. In particular, quasi-uniform meshes satisfy Assumption 4.8. We refer to [2] for a discussion on more general permissible meshes.

THEOREM 4.10. *Suppose that Assumption 4.8 is fulfilled. For $t \in (0, T]$ it holds that*

$$\|\tilde{X}(t) - \tilde{Y}(t)\|_{\mathcal{L}(V)} \leq CHt^{-1/2}.$$

Here, the constant C depends on T, α, β, γ, and $\|X(0)\|_{\mathcal{L}(V)}$, but not on the multiscale variations of $\mathcal{A}$.

Proof. We start by noting that $\|\mathrm{Id}_h\|_{\mathcal{L}(V_h,V)} \leq 1$. Furthermore, since P_h is stable in the V-norm, the following bound holds

$$\|P_h\|_{\mathcal{L}(V,V_h)} = \sup_{v\in V} \frac{\|P_h v\|_V}{\|v\|_V} \leq \sup_{v\in V} \frac{C\|v\|_V}{\|v\|_V} \leq C.$$

Now, note that if the initial data $v \in V$ then we may instead of Lemma 4.3 prove the following, less singular, error bound

$$\|E_h(t) - E_{\mathrm{ms}}(t)\|_{\mathcal{L}(V)} \leq CHt^{-1/2}.$$

In addition, parabolic regularity gives the bounds $\|E_h(t)\|_{\mathcal{L}(V)}, \|E_{\mathrm{ms}}(t)\|_{\mathcal{L}(V)} \leq C$.

Note that $\|\mathcal{C}_h\|_{\mathcal{L}(V_h,Z)} \leq \|\mathcal{C}\|_{\mathcal{L}(V,Z)}$, so from Assumption 4.8 it follows that

$$\begin{aligned} &\|(\mathcal{C}_h P_h)^* \mathcal{Q}(\mathcal{C}_h P_h)\|_{\mathcal{L}(V)} \\ &\qquad \leq \|P_h^*\|_{\mathcal{L}(V_h,V)} \|\mathcal{C}_h^*\|_{\mathcal{L}(Z,V_h)} \|\mathcal{Q}\|_{\mathcal{L}(Z)} \|\mathcal{C}_h\|_{\mathcal{L}(V_h,Z)} \|P_h\|_{\mathcal{L}(V,V_h)} \leq C. \end{aligned} \tag{10}$$

Similarly, $\|\mathcal{B}_h\|_{\mathcal{L}(U,V_h)} \leq \|\mathcal{B}\|_{\mathcal{L}(U,V)}$, and we have

$$\begin{aligned} \|S_h\|_{\mathcal{L}(V)} &\leq \|\mathrm{Id}_h\|_{\mathcal{L}(V_h,V)} \|\mathcal{B}_h\|_{\mathcal{L}(U,V_h)} \|\mathcal{R}^{-1}\|_{\mathcal{L}(U)} \|\mathcal{B}_h^*\|_{\mathcal{L}(V_h,U)} \|\mathrm{Id}_h^*\|_{\mathcal{L}(V,V_h)} \\ &\leq C. \end{aligned} \tag{11}$$

As in the proof of Theorem 4.5 we can write the difference $\tilde{X}(t) - \tilde{Y}(t)$ as a sum of eight terms so that

$$\|\tilde{X}(t) - \tilde{Y}(t)\|_{\mathcal{L}(V)} \leq \|R_1\|_{\mathcal{L}(V)} + \|R_2\|_{\mathcal{L}(V)} + \int_0^t \sum_{j=3}^{8} \|R_j\|_{\mathcal{L}(V)} \, \mathrm{d}s.$$

For R_1 we have

$$\begin{aligned}\|R_1\|_{\mathcal{L}(V)} &\leq \|E_h(t)^*\|_{\mathcal{L}(V)}\|\tilde{X}(0)\|_{\mathcal{L}(V)}\|E_h(t) - E_{\mathrm{ms}}(t)\|_{\mathcal{L}(V)} \\ &\leq CHt^{-1/2}\|X(0)\|_{\mathcal{L}(V)} \leq CHt^{-1/2},\end{aligned}$$

and similarly we prove $\|R_2\|_{\mathcal{L}(V)} \leq CHt^{-1/2}$, where we have used that

$$\|X_h(0)\|_{\mathcal{L}(V)} = \|\mathrm{Id}_h^*\, X(0)\, \mathrm{Id}_h\|_{\mathcal{L}(V)} \leq C\|X(0)\|_{\mathcal{L}(V)},$$

which is bounded due to Assumption 4.8. Using the bounds (10) and (11) we get

$$\int_0^t \sum_{j=3}^{6} \|R_j\|_{\mathcal{L}(V)}\,\mathrm{d}s \leq \int_0^t CH(t-s)^{-1/2}\,\mathrm{d}s \leq CHt^{1/2},$$

and

$$\int_0^t \|R_7\|_{\mathcal{L}(V)} + \|R_8\|_{\mathcal{L}(V)}\,\mathrm{d}s \leq \int_0^t C\|\tilde{X}(s) - \tilde{Y}(s)\|_{\mathcal{L}(V)}\,\mathrm{d}s.$$

By applying Grönwall's lemma we obtain the desired error bound. ∎

5. Matrix-valued formulation. To perform actual computations, we write the finite-dimensional equations on matrix form by expressing the equations in the FEM or LOD bases. To this end, let the function $x \in V_h$ and the operator $X_h \colon V_h \to V_h$ have the vector and matrix representations $\boldsymbol{x} \in \mathbb{R}^{N_h}$ and $\tilde{\boldsymbol{X}}^h \in \mathbb{R}^{N_h \times N_h}$, i.e.

$$x = \sum_{j=1}^{N_h} \boldsymbol{x}_j \varphi_j^h \quad \text{and} \quad X_h x = \sum_{i,j=1}^{N_h} \tilde{\boldsymbol{X}}_{i,j}^h \boldsymbol{x}_j \varphi_i^h \tag{12}$$

Since exactly the same results hold for V_H upon replacing h by H, we frequently omit the h sub- and superscripts in the following manipulations. They will be reinstated later when we compare different discretizations. The coordinates satisfy

$$\boldsymbol{M}\boldsymbol{x} = \big((x,\varphi_i)\big)_{i=1}^N \quad \text{and} \quad \boldsymbol{M}\tilde{\boldsymbol{X}} = \big((X_h\varphi_j,\varphi_i)\big)_{i,j=1}^N,$$

where $\boldsymbol{M}$ denotes the (symmetric) mass matrix, $\boldsymbol{M}_{i,j} = (\varphi_j,\varphi_i)$. Unfortunately, we will not recover the usual form of the matrix-valued DRE when working in these coordinates. Therefore, we perform the change of variables

$$\boldsymbol{X}\boldsymbol{M} = \tilde{\boldsymbol{X}}.$$

Coincidentally, this means that we actually have

$$X_h x = \sum_{i,j=1}^{N} \boldsymbol{X}_{i,j}\,(x,\varphi_j)\,\varphi_i. \tag{13}$$

Equation (5) is equivalent to

$$\begin{aligned}\left(\dot{X}_h\varphi_i,\varphi_j\right) &= (X_h\varphi_i,\mathcal{A}_h\varphi_j) + (X_h\varphi_j,\mathcal{A}_h\varphi_i) \\ &\quad + (\mathcal{Q}\mathcal{C}_h\varphi_i,\mathcal{C}_h\varphi_j)_Z - \left(\mathcal{R}^{-1}\mathcal{B}_h^*X_h\varphi_i,\mathcal{B}_h^*X_h\varphi_j\right)_U\end{aligned} \tag{14}$$

for $1 \le i, j \le N$, and since $X_h\varphi_i = \sum_{k=1}^N (\boldsymbol{X}\boldsymbol{M})_{k,i}\varphi_k$, the first term becomes

$$\sum_{k=1}^N (\dot{\boldsymbol{X}}\boldsymbol{M})_{k,i}\boldsymbol{M}_{j,k} = (\boldsymbol{M}\dot{\boldsymbol{X}}\boldsymbol{M})_{j,i}$$

Likewise, with the (negative) stiffness matrix $\boldsymbol{A}_{i,j} = (\mathcal{A}\varphi_j, \varphi_i)$, the second and third terms become

$$\sum_{k=1}^N (\boldsymbol{X}\boldsymbol{M})_{k,i}\boldsymbol{A}_{k,j} + \sum_{k=1}^N (\boldsymbol{X}\boldsymbol{M})_{k,j}\boldsymbol{A}_{k,i} = (\boldsymbol{A}^T\boldsymbol{X}\boldsymbol{M})_{j,i} + (\boldsymbol{M}\boldsymbol{X}\boldsymbol{A})_{j,i},$$

due to the symmetry of $\boldsymbol{M}$ and $\boldsymbol{X}$. (Recall that we search for a self-adjoint operator X_h.) Finally, the last two terms can be written

$$(\boldsymbol{C}^T\boldsymbol{Q}\boldsymbol{C})_{j,i} \quad \text{and} \quad \left(\boldsymbol{M}\boldsymbol{X}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{M}\right)_{j,i},$$

where $\boldsymbol{B}_{i,j} = \left(\mathcal{B}\varphi_j^U, \varphi_i\right)$, $\boldsymbol{Q}_{i,j} = \left(\mathcal{Q}\varphi_j^Z, \varphi_i^Z\right)$, $\boldsymbol{R}_{i,j} = \left(\mathcal{R}\varphi_j^U, \varphi_i^U\right)$, $\boldsymbol{C}_{i,j} = \left(\mathcal{C}\varphi_j, \varphi_i^Z\right)$ and $\{\varphi_i^U\}$, $\{\varphi_i^Z\}$ denote orthonormal bases for U and Z, respectively. Summarizing, we can write the equation on matrix form as

$$\boldsymbol{M}\dot{\boldsymbol{X}}\boldsymbol{M} = \boldsymbol{M}\boldsymbol{X}\boldsymbol{A} + \boldsymbol{A}^T\boldsymbol{X}\boldsymbol{M} + \boldsymbol{C}^T\boldsymbol{Q}\boldsymbol{C} - \boldsymbol{M}\boldsymbol{X}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{M}. \tag{15}$$

Similar to the relations between the fine and coarse operators (6), it is easily shown that their matrix representations satisfy

$$\boldsymbol{A}_H = (\boldsymbol{I}_H^h)^T\boldsymbol{A}_h\boldsymbol{I}_H^h, \quad \boldsymbol{B}_H = (\boldsymbol{I}_H^h)^T\boldsymbol{B}_h, \quad \boldsymbol{C}_H = \boldsymbol{C}_h\boldsymbol{I}_H^h \quad \text{and} \quad \boldsymbol{M}_H = (\boldsymbol{I}_H^h)^T\boldsymbol{M}_h\boldsymbol{I}_H^h,$$

where $\boldsymbol{I}_H^h \in \mathbb{R}^{N_h \times N_H}$ is the prolongation matrix that satisfies $\boldsymbol{I}_H^h\boldsymbol{x}^H = \boldsymbol{x}^h$ if $x = \sum_{j=1}^{N_H} \boldsymbol{x}_j^H\varphi_j^H$ and $\mathrm{Id}_H^h\, x = \sum_{j=1}^{N_h} \boldsymbol{x}_j^h\varphi_j^h$. By expressing the φ^H functions in terms of φ^h, it can be seen that $(\boldsymbol{I}_H^h)_{i,j} = \varphi_j^H(z_i)$, where z_i is the i:th node of $\mathcal{T}_h$. Thus the coarse systems are easily constructed when the fine system is known. Note, however, that the matrix representation of $(\mathrm{Id}_H^h)^*$ is not $(\boldsymbol{I}_H^h)^T$ but $\boldsymbol{M}_H^{-1}(\boldsymbol{I}_H^h)^T\boldsymbol{M}_h$.

For the LOD case, we let $\boldsymbol{Q}_h$ and $\boldsymbol{R}_h = \boldsymbol{I}_H^h - \boldsymbol{Q}_h$ be the matrix representations of Q_h and R_h, respectively. To compute them efficiently, we follow [13]. Then

$$X_H^{\mathrm{ms}}x = \sum_{i=1}^{N_H} (\boldsymbol{X}_H^{\mathrm{ms}}\boldsymbol{M}_{\mathrm{ms}}\boldsymbol{x})_i\varphi_i^H,$$

where $\boldsymbol{X}_H^{\mathrm{ms}}$ is symmetric and satisfies

$$\begin{aligned}\boldsymbol{M}_{\mathrm{ms}}\dot{\boldsymbol{X}}_H^{\mathrm{ms}}\boldsymbol{M}_{\mathrm{ms}} = {}& \boldsymbol{M}_{\mathrm{ms}}\boldsymbol{X}_H^{\mathrm{ms}}\boldsymbol{A}_{\mathrm{ms}} + \boldsymbol{A}_{\mathrm{ms}}^T\boldsymbol{X}_H^{\mathrm{ms}}\boldsymbol{M}_{\mathrm{ms}} \\ & + \boldsymbol{C}_{\mathrm{ms}}^T\boldsymbol{Q}\boldsymbol{C}_{\mathrm{ms}} - \boldsymbol{M}_{\mathrm{ms}}\boldsymbol{X}_H^{\mathrm{ms}}\boldsymbol{B}_{\mathrm{ms}}\boldsymbol{R}^{-1}\boldsymbol{B}_{\mathrm{ms}}^T\boldsymbol{X}_H^{\mathrm{ms}}\boldsymbol{M}_{\mathrm{ms}},\end{aligned}$$

with the matrices

$$\boldsymbol{A}_{\mathrm{ms}} = \boldsymbol{R}_h^T\boldsymbol{A}_h\boldsymbol{R}_h, \quad \boldsymbol{B}_{\mathrm{ms}} = \boldsymbol{R}_h^T\boldsymbol{B}_h, \quad \boldsymbol{C}_{\mathrm{ms}} = \boldsymbol{C}_h\boldsymbol{R}_h \quad \text{and} \quad \boldsymbol{M}_{\mathrm{ms}} = \boldsymbol{R}_h^T\boldsymbol{M}_h\boldsymbol{R}_h.$$

Finally, we note that if $u \in V_h$, $w \in V_H$ and $(\mathrm{Id}_{\mathrm{ms}}^h)^*u = R_h w$, then in coordinates we have $\boldsymbol{w} = \boldsymbol{M}_{\mathrm{ms}}^{-1}\boldsymbol{R}_h^T\boldsymbol{M}_h\boldsymbol{u}$. This means that the matrix representation of $\mathrm{Id}_{\mathrm{ms}}^h\, X_h^{\mathrm{ms}}(\mathrm{Id}_{\mathrm{ms}}^h)^*$ is $\boldsymbol{R}_h\boldsymbol{X}_H^{\mathrm{ms}}\boldsymbol{R}_h^T\boldsymbol{M}_h$.

5.1. Error computation. We measure the quality of different approximations as the $\mathcal{L}(L^2)$-normed distance to a reference approximation at the final time T. In order to find a matrix representation for this, we first observe that since $\|P_h x\| \leq \|x\|$, we have

$$\|\mathrm{Id}_h\, X_h P_h\|_{\mathcal{L}(L^2)} = \sup_{\substack{x\in L^2\\ x\neq 0}} \frac{\|X_h P_h x\|}{\|x\|} \leq \sup_{\substack{x\in L^2\\ x\neq 0}} \frac{\|X_h P_h x\|}{\|P_h x\|} = \sup_{\substack{x\in V_h\\ x\neq 0}} \frac{\|X_h x\|}{\|x\|}.$$

But $P_h x = x$ for $x \in V_h$, so since $V_h \subset L^2$ we also get

$$\|\mathrm{Id}_h\, X_h P_h\|_{\mathcal{L}(L^2)} \geq \sup_{\substack{x\in V_h\\ x\neq 0}} \frac{\|X_h P_h x\|}{\|x\|} = \sup_{\substack{x\in V_h\\ x\neq 0}} \frac{\|X_h x\|}{\|x\|}.$$

To compute the $\mathcal{L}(L^2)$-norm it is thus enough to test with $x = \sum_{i=1}^{N_h} \boldsymbol{x}_i \varphi_i^h \in V_h$. Again omitting the h sub- and superscripts, we have that $(x, x) = \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}$, and similarly

$$\begin{aligned}(X_h x, X_h x) &= \sum_{i,j,k,l=1}^{N} (\boldsymbol{X}\boldsymbol{M})_{i,j} \boldsymbol{x}_j (\boldsymbol{X}\boldsymbol{M})_{k,l} \boldsymbol{x}_l \,(\varphi_i, \varphi_k) \\ &= \boldsymbol{x}^T \boldsymbol{M}^T \boldsymbol{X}^T \boldsymbol{M} \boldsymbol{X} \boldsymbol{M} \boldsymbol{x}.\end{aligned}$$

Since $\boldsymbol{M}$ is symmetric positive definite, we may do a Cholesky factorization $\boldsymbol{M} = \boldsymbol{L}_{\boldsymbol{M}} \boldsymbol{L}_{\boldsymbol{M}}^T$, and the change of variables $\boldsymbol{y} = \boldsymbol{L}_{\boldsymbol{M}}^T \boldsymbol{x}$ yields

$$\|\mathrm{Id}_h\, X_h P_h\|_{\mathcal{L}(L^2)} = \sup_{\substack{\boldsymbol{y}\in\mathbb{R}^N\\ \boldsymbol{y}\neq 0}} \frac{\left(\boldsymbol{y}^T \boldsymbol{L}_{\boldsymbol{M}}^T \boldsymbol{X} \boldsymbol{L}_{\boldsymbol{M}} \boldsymbol{L}_{\boldsymbol{M}}^T \boldsymbol{X} \boldsymbol{L}_{\boldsymbol{M}} \boldsymbol{y}\right)^{1/2}}{\left(\boldsymbol{y}^T \boldsymbol{y}\right)^{1/2}} = \|\boldsymbol{L}_{\boldsymbol{M}}^T \boldsymbol{X} \boldsymbol{L}_{\boldsymbol{M}}\|_{\mathbb{R}^{N\times N}},$$

where $\|\cdot\|_{\mathbb{R}^{N\times N}}$ denotes the standard spectral matrix norm. Recalling the matrix representation $\boldsymbol{I}_H^h$ of Id_H^h, we now get that

$$\|\mathrm{Id}_h\, X_h P_h - \mathrm{Id}_H\, X_H P_H\|_{\mathcal{L}(L^2)} = \|\boldsymbol{L}_{\boldsymbol{M}}^T (\boldsymbol{X}_h - \boldsymbol{I}_H^h \boldsymbol{X}_H (\boldsymbol{I}_H^h)^T) \boldsymbol{L}_{\boldsymbol{M}}\|_{\mathbb{R}^{N\times N}}.$$

The LOD error is completely analogous, using instead $\boldsymbol{R}_h$ and $\boldsymbol{X}_H^{\mathrm{ms}}$.

A similar approach also allows us to compute $\mathcal{L}(V)$-errors. Let $\boldsymbol{A} = \boldsymbol{L}_{\boldsymbol{A}} \boldsymbol{L}_{\boldsymbol{A}}^T$ be a Cholesky factorization of $\boldsymbol{A}$. Then

$$\|\mathrm{Id}_h\, X_h P_h - \mathrm{Id}_H\, X_H P_H\|_{\mathcal{L}(V)} \leq \|P_h\|_{\mathcal{L}(V,V_h)} \|\boldsymbol{L}_{\boldsymbol{A}}^T (\boldsymbol{X}_h - \boldsymbol{I}_H^h \boldsymbol{X}_H (\boldsymbol{I}_H^h)^T) \boldsymbol{M} \boldsymbol{L}_{\boldsymbol{A}}^{-T}\|_{\mathbb{R}^{N\times N}}.$$

We also get that $\|\mathrm{Id}_h\, X_h P_h - \mathrm{Id}_H\, X_H P_H\|_{\mathcal{L}(V)}$ is bounded from below by $\|\boldsymbol{L}_{\boldsymbol{A}}^T (\boldsymbol{X}_h - \boldsymbol{I}_H^h \boldsymbol{X}_H (\boldsymbol{I}_H^h)^T) \boldsymbol{M} \boldsymbol{L}_{\boldsymbol{A}}^{-T}\|_{\mathbb{R}^{N\times N}}$, i.e. the latter quantity can be thought of as an equivalent norm. Since $\boldsymbol{L}_{\boldsymbol{A}}$ is triangular, the extra cost required for the computation of $\boldsymbol{L}_{\boldsymbol{A}}^{-T}$ is negligible. If the low-rank formulation is used (see subsection 6.1), only a small number of linear equation systems involving $\boldsymbol{L}_{\boldsymbol{A}}$ needs to be solved, reducing the cost even further.

6. Temporal discretization. We discretize the matrix-valued DREs in time by means of a low-rank splitting scheme, since the basic operation in such methods is the application of $\mathrm{e}^{t\boldsymbol{A}^T}$, i.e. essentially solving a parabolic problem. Let τ denote a fixed time step, and let $t_j = j\tau$, $j = 0, \ldots, N_t$, be the time discretization of the

interval $[0,T]$. We split Equation (15) into two parts, $\dot{\boldsymbol{X}} = \mathscr{F}\boldsymbol{X} + \mathscr{G}\boldsymbol{X}$, where

$$\mathscr{F}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{M}^{-1} + \boldsymbol{M}^{-1}\boldsymbol{A}^T\boldsymbol{X} + \boldsymbol{M}^{-1}\boldsymbol{C}^T\boldsymbol{Q}\boldsymbol{C}\boldsymbol{M}^{-1} \quad \text{and} \quad \mathscr{G}\boldsymbol{X} = -\boldsymbol{X}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^T\boldsymbol{X}.$$

Then the Strang splitting approximation at time t_j is given by $\boldsymbol{X}^j$, with $\boldsymbol{X}^0 = \boldsymbol{X}(0)$ and

$$\boldsymbol{X}^{j+1} = \mathrm{e}^{\frac{\tau}{2}\mathscr{F}}\mathrm{e}^{\tau\mathscr{G}}\mathrm{e}^{\frac{\tau}{2}\mathscr{F}}\boldsymbol{X}^j.$$

Here, the solution operators $\mathrm{e}^{t\mathscr{F}}$ and $\mathrm{e}^{t\mathscr{G}}$ satisfy

$$\mathrm{e}^{t\mathscr{F}}\boldsymbol{X} = \mathrm{e}^{t\boldsymbol{M}^{-T}\boldsymbol{A}^T}\boldsymbol{X}\mathrm{e}^{t\boldsymbol{A}\boldsymbol{M}^{-1}} + \int_0^t \mathrm{e}^{s\boldsymbol{M}^{-T}\boldsymbol{A}^T}\boldsymbol{M}^{-T}\boldsymbol{C}^T\boldsymbol{Q}\boldsymbol{C}\boldsymbol{M}^{-1}\mathrm{e}^{s\boldsymbol{A}\boldsymbol{M}^{-1}}\,\mathrm{d}s, \tag{16}$$

$$\mathrm{e}^{t\mathscr{G}}\boldsymbol{X} = \left(I + t\boldsymbol{X}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B}^T\right)^{-1}\boldsymbol{X}, \tag{17}$$

where the first equality is apparent from the integral formulation (9), while the second is easily verified by differentiation.

The low-rank version of the method relies on the assumption that the solution $\boldsymbol{X}$ has low rank. This is general true for LQR problems and dramatically reduces the computational cost. In that case, we may factorize $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$, where $\boldsymbol{L} \in \mathbb{R}^{N_h \times r}$ and $\boldsymbol{D} \in \mathbb{R}^{r\times r}$ with the rank $r \ll N_h$. Also $\mathrm{e}^{\tau\mathscr{F}}\boldsymbol{X}$ and $\mathrm{e}^{\tau\mathscr{G}}\boldsymbol{X}$, and thus the iterates $\boldsymbol{X}_j$, may then be factorized in such a way. After a reformulation, $\mathrm{e}^{\tau\mathscr{G}}\boldsymbol{X}$ is very cheap to compute, and the computation of $\mathrm{e}^{\tau\mathscr{F}}\boldsymbol{X}$ reduces to an evaluation of $\mathrm{e}^{\tau\boldsymbol{M}^{-T}\boldsymbol{A}^T}\boldsymbol{L}$ (plus preliminary, similar work for the integral term). The latter operation is equivalent to solving $\boldsymbol{M}\dot{x} = \boldsymbol{A}^T x$, $x(0) = \boldsymbol{L}$, and the matrix $\boldsymbol{M}$ is thus never explicitly inverted. For further details, we refer to [32, 33].

6.1. Low-rank errors. Also the error computations outlined in subsection 5.1 benefit from being formulated in a low-rank setting. Assume that $\boldsymbol{X}_h = \boldsymbol{L}_h\boldsymbol{D}_h\boldsymbol{L}_h^T$ and $\boldsymbol{X}_H = \boldsymbol{L}_H\boldsymbol{D}_H\boldsymbol{L}_H^T$ with $\boldsymbol{L}_h \in \mathbb{R}^{N_h\times r_h}$ and $\boldsymbol{L}_H \in \mathbb{R}^{N_H\times r_H}$ with $r_h, r_H \ll N_h$, and let $\boldsymbol{M}_h = \boldsymbol{L}_{\boldsymbol{M}}\boldsymbol{L}_{\boldsymbol{M}}^T$ be a Cholesky factorization. By setting

$$\boldsymbol{V} = \begin{bmatrix}\boldsymbol{L}_{\boldsymbol{M}}^T\boldsymbol{L}_h & \boldsymbol{L}_{\boldsymbol{M}}^T\boldsymbol{I}_H^h\boldsymbol{L}_H\end{bmatrix} \quad \text{and} \quad \boldsymbol{D} = \begin{bmatrix}\boldsymbol{D}_h & 0\\ 0 & -\boldsymbol{D}_H\end{bmatrix}$$

we see that $\boldsymbol{V} \in \mathbb{R}^{N_h\times(r_h+r_H)}$, $\boldsymbol{D} \in \mathbb{R}^{(r_h+r_H)\times(r_h+r_H)}$ and it follows that

$$\boldsymbol{L}_{\boldsymbol{M}}^T(\boldsymbol{X}_h - \boldsymbol{I}_H^h\boldsymbol{X}_H(\boldsymbol{I}_H^h)^T)\boldsymbol{L}_{\boldsymbol{M}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T.$$

Since $\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$ is not necessarily an eigenvalue decomposition, we cannot immediately determine the norm by inspection. However, performing a QR-factorization $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{R}$ is cheap if the number of columns is low, and $\boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^T \in \mathbb{R}^{(r_h+r_H)\times(r_h+r_H)}$ can also be diagonalized cheaply. (This is precisely the LDL^T column compression procedure which is applied in each time step.) We acquire $\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T = (\boldsymbol{Q}\boldsymbol{W})\tilde{\boldsymbol{D}}(\boldsymbol{Q}\boldsymbol{W})^T$, for some $\boldsymbol{W}$, where $\|\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T\| = |\tilde{\boldsymbol{D}}_{1,1}|$.

For errors in the $\mathcal{L}(V)$-norm, we do not get a symmetric matrix as above. But if $\boldsymbol{A} = \boldsymbol{L}_{\boldsymbol{A}}\boldsymbol{L}_{\boldsymbol{A}}^T$ we can still write

$$\boldsymbol{L}_{\boldsymbol{A}}^T(\boldsymbol{X}_h - \boldsymbol{I}_H^h\boldsymbol{X}_H(\boldsymbol{I}_H^h)^T)\boldsymbol{M}\boldsymbol{L}_{\boldsymbol{A}}^{-T} = \boldsymbol{G}_1\boldsymbol{D}\boldsymbol{G}_2^T,$$

with the same $\boldsymbol{D}$, and with

$$\boldsymbol{G}_1 = \begin{bmatrix}\boldsymbol{L}_{\boldsymbol{A}}^T\boldsymbol{L}_h & \boldsymbol{L}_{\boldsymbol{A}}^T\boldsymbol{I}_H^h\boldsymbol{L}_H\end{bmatrix} \quad \text{and} \quad \boldsymbol{G}_2 = \begin{bmatrix}\boldsymbol{L}_{\boldsymbol{A}}^{-1}\boldsymbol{M}\boldsymbol{L}_h & \boldsymbol{L}_{\boldsymbol{A}}^{-1}\boldsymbol{M}\boldsymbol{I}_H^h\boldsymbol{L}_H.\end{bmatrix}$$

We can cheaply QR-factorize both $\boldsymbol{G}_1 = \boldsymbol{U}\boldsymbol{R}_1$ and $\boldsymbol{G}_2 = \boldsymbol{V}\boldsymbol{R}_2$; this means that

$$\|\boldsymbol{L}_{\boldsymbol{A}}^T(\boldsymbol{X}_h - \boldsymbol{I}_H^h \boldsymbol{X}_H (\boldsymbol{I}_H^h)^T)\boldsymbol{M}\boldsymbol{L}_{\boldsymbol{A}}^{-T}\|_{\mathbb{R}^{N_h \times N_h}} = \|\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T\|_{\mathbb{R}^{N_h \times N_h}} = \|\boldsymbol{S}\|_{\mathbb{R}^{N_h \times N_h}},$$

where $\boldsymbol{S} = \boldsymbol{R}_1 \boldsymbol{D} \boldsymbol{R}_2^T$ is a small matrix.

7. Numerical experiments. We have performed a number of numerical experiments in order to verify our a priori error bounds for the LOD discretizations, and to demonstrate their efficiency in comparison to the classical FEM.

In all experiments, we compute the relevant matrices for both FEM and LOD by using efficient code written by Fredrik Hellman and Daniel Elfverson[1]. These pre-solve computations were run on a Intel® Core™ i5-4690 processor. We note that the localized elliptic fine-scale problems were not solved in parallel. Doing so would further improve the performance of LOD.

For approximating the solutions to the DREs, we employ in all cases the low-rank Strang splitting scheme (as described in section 6) with $N_t = 256$ time steps. This ensures that the temporal error is small compared to the spatial error, which is our interest here. Our implementation utilizes the DREsplit[2] library. These computations were performed on resources at Chalmers Centre for Computational Science and Engineering (C3SE) provided by the Swedish National Infrastructure for Computing (SNIC). Each simulation used a single Intel® Xeon® E5-2650 v3 processor.

The multiscale diffusion coefficients κ considered in the numerical examples are of two distinct types. In Examples 1, 2, and 4 we consider a piecewise constant coefficient, generated randomly with no spatial correlation, that varies on a fine scale, see Figure 1. In Examples 3 and 5 κ takes two values. One value in the background and one in the thin channels, see Figure 6. This is a common setup for reinforced (composite) materials. Both these cases are challenging for the finite element method.

7.1. Example 1. In this first example, we consider diffusion on the unit square. More specifically, we take $\Omega = [0,1]^2$ and set $\mathcal{A}x = \nabla \cdot (\kappa \nabla x)$ with Dirichlet boundary conditions. Here, κ is piecewise constant on a square grid of size 2^{-7} and taking randomly chosen values in $[10^{-3}, 1]$; see Figure 1 for an illustration. We consider 3 independent inputs and define the input operator $\mathcal{B}$ as the sum

$$\mathcal{B}u = \sum_{j=1}^{3} \mathcal{B}_j u_j, \quad \text{where} \quad (\mathcal{B}_j u)(\xi_1, \xi_2) = \begin{cases} u, & \frac{j}{4} \leq \xi_1, \xi_2 \leq \frac{j}{4} + \frac{1}{8} \\ 0, & \text{otherwise} \end{cases}.$$

Thus we can control the system on three small squares. As the output operator we take the mean, i.e. $\mathcal{C}x = \int_\Omega x$. We choose $\mathcal{Q}$ and $\mathcal{R}$ to be the identity operators and take $\mathcal{G} = X(0) = 0$.

For the discretization in space, we start with a coarse mesh containing 8 triangles, and then refine this 6 times, giving meshes with 2^{3+2j} triangles, for $j = 0, \ldots, 6$. One additional refinement provides the reference grid with $2^{17} = 131072$ triangles. This results in matrices $\boldsymbol{A}_j \in \mathbb{R}^{n \times n}$, $\boldsymbol{B}_j \in \mathbb{R}^{n \times 3}$ and $\boldsymbol{C}_j \in \mathbb{R}^{1 \times n}$, $j = 0, \ldots, 7$, with $n = 1, 9, 49, 225, 961, 3969, 16129, 65025$ (since we only consider the interior nodes).

The approximations are compared only at the final time, in the $\mathcal{L}(L^2)$- and $\mathcal{L}(V)$-norms as outlined in subsection 5.1, and the computed errors are shown in Figure 2. We see that the classical FEM initially struggles due to not resolving the multiscale

[1] Available on request from Fredrik Hellman, fredrik.hellman@it.uu.se.

[2] Available on request from Tony Stillfjord, tony.stillfjord@gu.se, or from http://www.tonystillfjord.net.

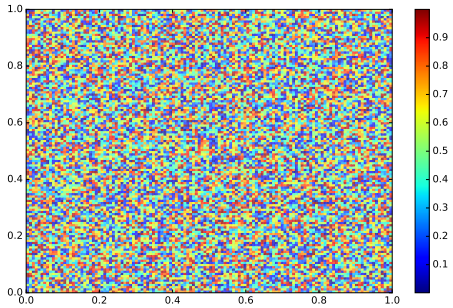

FIG. 1. *The diffusion coefficient used in Example 1, plotted over the domain* Ω. *(This figure is in color in the electronic version of the article.)*

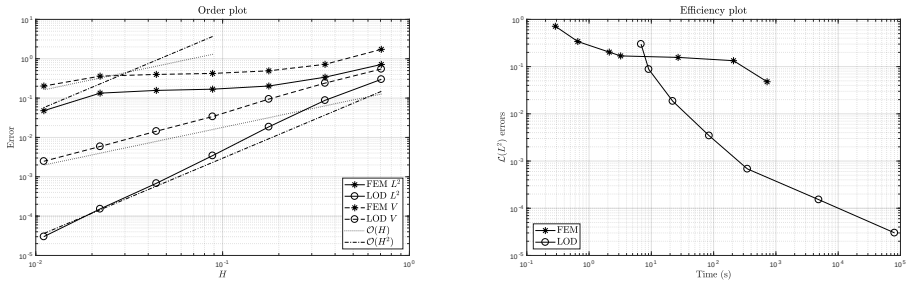

FIG. 2. *Left: The* $\mathcal{L}(L^2)$*- and* $\mathcal{L}(V)$*-norm errors of the approximations computed in Example 1, plotted against the meshwidth. Right: The* $\mathcal{L}(L^2)$*-norm errors plotted against the computation time.*

coefficient properly, but converges with order 2 when the mesh becomes fine enough. The LOD approach converges with order 2 also for the coarse meshes, and additionally results in approximations that are about one order of magnitude more accurate. The plot to the right shows the errors against the actual computation time, including the time spent on constructing the LOD bases. As can be seen, this extra effort is low enough that except for the most inaccurate cases it is always worthwhile to use the LOD approach.

7.2. Example 2. Here, we consider an L-shaped domain Ω, where $[0.5, 1] \times [0.5, 0.5]$ has been removed from the unit square. The diffusion coefficient κ is piecewise constant on a square grid of size 2^{-7} and taking random values in $[10^{-3}, 1]$. We use one control input, given by the characteristic function of the square $[0.65, 0.85]^2$, and one output, the mean over the square $[0.15, 0.35]^2$. The meshes are setup as in the previous example, but now with $n = 5, 33, 161, 705, 2945, 12033, 48641$ interior nodes ($n = 195585$ for the reference solution). The time discretization and other parameters are the same as in the previous example.

The results are shown in Figure 3. Due to the reentrant corner the errors behave more erratically than in the previous example, but LOD is still clearly first- and second-order convergent in contrast to standard FEM, which performs very poorly. We also observe that LOD is more efficient in all but the coarsest cases.

7.3. Example 3. We again consider the setting of Example 1, but replace the diffusivity constant. Here, κ takes the constant value 1 everywhere, except for in 7

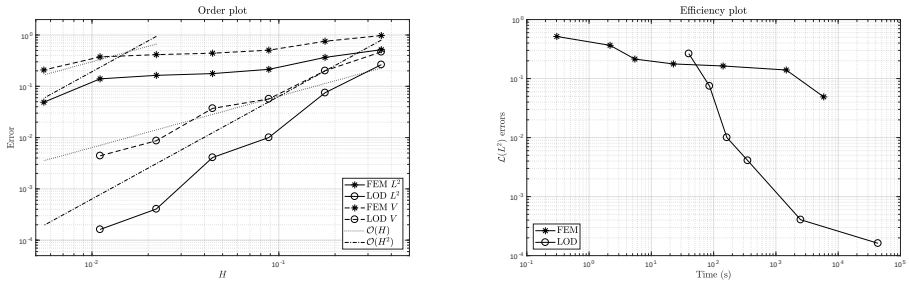

FIG. 3. *Left: The $\mathcal{L}(L^2)$- and $\mathcal{L}(V)$-norm errors of the approximations computed in Example 2, plotted against the meshwidth. Right: The $\mathcal{L}(L^2)$-norm errors plotted against the computation time.*

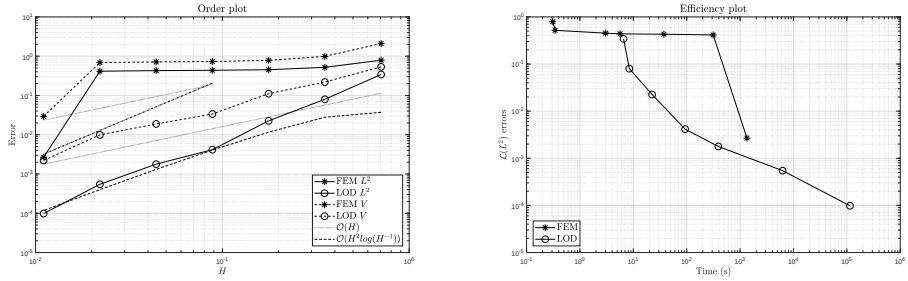

FIG. 4. *Left: The $\mathcal{L}(L^2)$- and $\mathcal{L}(V)$-norm errors of the approximations computed in Example 3, plotted against the meshwidth. Right: The $\mathcal{L}(L^2)$-norm errors plotted against the computation time.*

horizontal stripes where it is 10^{-2}. The stripes are centered around the heights $j/8$, $j = 1, \dots, 7$, and have a width of 2^{-7}.

The results are shown in Figure 4. This time, the detrimental effect on the FEM discretization is even more pronounced, with almost no convergence until the thin stripes can be resolved. The LOD approximations are once again more accurate for all H. We note that the $\mathcal{L}(L^2)$-error is not quite $\mathcal{O}(H^2)$ in this case, but rather close to $\mathcal{O}(H^2 \log H^{-1})$ as predicted by Theorem 4.5. Like in the previous example, computing the LOD bases is cheap enough that the LOD approach is more efficient in all but the least accurate cases.

7.4. Example 4. In this example, we deviate from the basic setting described in section 4 by considering a boundary control application. All parameters except for the boundary conditions and the input operator are the same as in Example 1. We call the union of the top and bottom edges of the unit square Γ_D and impose homogeneous Dirichlet boundary conditions there. The left and right edges we denote Γ_1 and Γ_2, respectively, and there we impose nonhomogeneous Neumann boundary conditions. In particular, with the outward-pointing normal denoted by n, we consider functions x satisfying

$$\kappa \nabla x \cdot n = \Psi u_i \quad \text{on } \Gamma_i.$$

Here, u_1 and u_2 are the two control inputs, and

$$\Psi \colon s \mapsto \begin{cases} 2s, & 0 \leq s \leq 1/2, \\ 2(1-s), & 1/2 < s \leq 1, \end{cases}$$

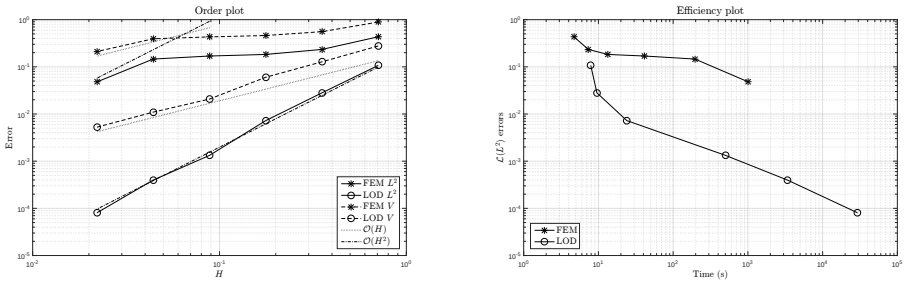

FIG. 5. *Left: The $\mathcal{L}(L^2)$- and $\mathcal{L}(V)$-norm errors of the approximations computed in Example 4, plotted against the meshwidth. Right: The $\mathcal{L}(L^2)$-norm errors plotted against the computation time.*

is a fixed function. The operator $\mathcal{A}$ now corresponds to $x \mapsto \nabla \cdot (\kappa \nabla x)$ on the space $\{x \in H^1(\Omega) \mid x_{|\Gamma_D} = 0\}$ with no conditions imposed on Γ_1, Γ_2, while the (unbounded) operator $\mathcal{B}$ implements the Neumann boundary conditions. We refrain from elaborating further on this here, and simply note that the FEM matrix representation becomes

$$\boldsymbol{B}^h_{j,i} = \int_{\Gamma_i} \Psi \varphi^h_j.$$

Fur further details on the proper abstract framework, see e.g. [21] and subsection 8.1.

Since Assumption 2.1 is no longer satisfied, we may not apply Theorem 4.5. However, the results plotted in Figure 5 are similar to the results in previous examples. Again, the LOD approximations are more efficient except for the very coarsest meshes. This indicates that our theory could be extended also to the case of unbounded operators $\mathcal{B}$ and $\mathcal{C}$.

7.5. Example 5. As a final experiment, we consider another boundary control application. The domain is formed like a lying U, see Figure 6. The thickness of each of the "handles" is $1/6$, the total horizontal extent 1 and the vertical extent $4/6$. Inside the domain are three evenly spaced stripes with a diameter of 0.0052. As previously, we consider $\mathcal{A}u = \nabla \cdot (\kappa \nabla u)$ where $\kappa = 10^{-2}$ everywhere except for in the stripes where instead $\kappa = 1$. We use homogeneous Neumann boundary conditions over the whole boundary, except for the two vertical sections on the left. On the top-most vertical part, Γ_1, we impose a nonhomogeneous Neumann condition $\kappa \nabla x \cdot n = \Psi u$ with Ψ having the same hat-shaped form as in Example 4. On the bottom vertical part, Γ_2, we impose a homogeneous Dirichlet condition. These correspond to an insulated edge, a controllable heat input and a heat sink, respectively. The operator $\mathcal{B}$ is again given by $u \mapsto u \int_{\Gamma_1} \Psi \varphi$, and as output we take the mean of the temperature over the domain; $\mathcal{C}x = \int_\Omega x$. The meshes in this example have $n = 28, 84, 280, 1008, 3808, 14784$ interior nodes, respectively, while the reference solution uses $n = 58240$.

The results are plotted in Figure 7, where we can once again observe error behaviour consistent with the bounds given in Theorem 4.5.

Remark 7.1. In all the experiments, we have chosen the fine-scale structure of the multiscale coefficient such that the reference FEM solution can resolve it, since otherwise we can not properly compute the respective errors. Decreasing the size of the fine-scale features even further would mean that the FEM convergence is further delayed, while we may still compute accurate LOD approximations. In such a case,

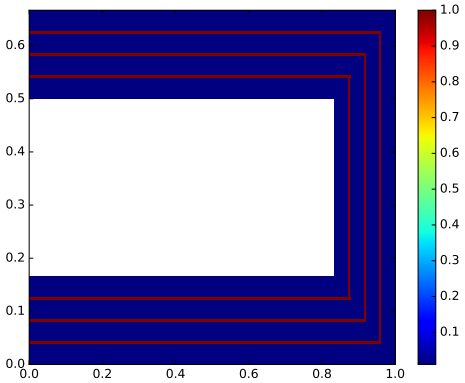

FIG. 6. *The diffusion coefficient used in Example 5, plotted over the domain* Ω. *(This figure is in color in the electronic version of the article.)*

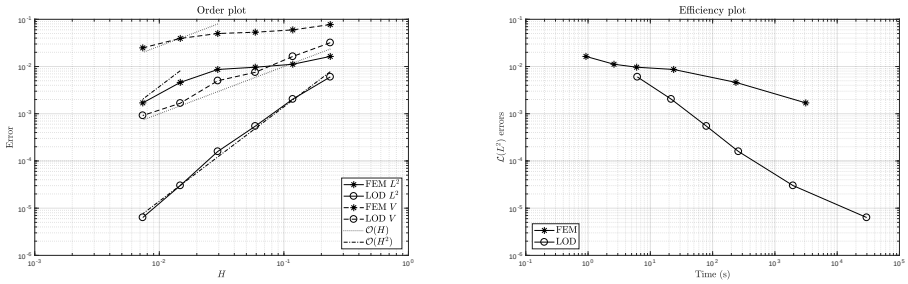

FIG. 7. *Left: The* $\mathcal{L}(L^2)$*- and* $\mathcal{L}(V)$*-norm errors of the approximations computed in Example 5, plotted against the meshwidth. Right: The* $\mathcal{L}(L^2)$*-norm errors plotted against the computation time.*

the efficiency of LOD in comparison to FEM is further (greatly) improved.

Remark 7.2. We note that the finest discretizations of the demonstrated numerical experiments are representative of large-scale DRE problems. While there is of course no strict limit, the authors would at the time of writing classify large-scale as problems of size $n \geq 10^4$. Due to the matrix-valued nature of the equations, this is naively equivalent to solving a vector-valued differential equation of size 10^8. While we do not employ a naive method, the number of unknowns still number in the millions. For readers more familiar with the theory of algebraic Riccati equations (AREs), i.e. the stationary counterparts of DREs, we note that one time step for a DRE solver is roughly equivalent to the solution of one corresponding ARE. The computational effort for solving a DRE is therefore usually at least two orders of magnitude higher, and large-scale in the ARE setting is therefore larger; starting rather at around $n = 10^5$. We also note that the approximations were here computed on the equivalent of a modern desktop computer. With the increasing availability of parallelization on clusters or GPUs, we expect to see a shift towards even larger problems in the near future. However, for multiscale problems such as these, it is still critical to employ LOD.

8. Generalizations and future work. In this section we provide some notes on possible extensions of our theory and draw connections to related problems and methods.

8.1. Boundary control. Boundary control applications such as Example 4 occur frequently within the field of optimal control. Then either the input or output operator (or both) acts on the boundary of the computational domain. In order to put such problems into the semigroup framework, one has to allow for unbounded operators $\mathcal{B}$ and $\mathcal{C}$ [21]. Clearly, our convergence analysis is no longer valid in that case, since we can no longer guarantee that $S_h \in \mathcal{L}(L^2)$ or that $\mathcal{C}_h P_h \in \mathcal{L}(L^2, Z)$. However, it is typically assumed that $\mathcal{B}$ and $\mathcal{C}$ are not *too* unbounded. More specifically, if we suppose that $(-\mathcal{A})^{-\beta}\mathcal{B} \in \mathcal{L}(U, L^2)$ and $\mathcal{C}(-\mathcal{A})^{-\gamma} \in \mathcal{L}(L^2, Z)$, where $0 \leq \beta + \gamma < 1$, we cover a large class of applications. Here, $(-\mathcal{A})^{-\alpha}$, denotes fractional powers of $\mathcal{A}$ which exist due to Assumption 2.1. They give rise to the spaces $X_{-\alpha} \supset L^2$ as the completions of L^2 in the norm $\|x\|_{-\alpha} = \|(-\mathcal{A})^{-\alpha}x\|$. When $\gamma = 0$ we then have that $S_h \in \mathcal{L}(X_{-\beta})$, and by properly extending also the other involved operators to $X_{-\beta}$ we may follow the line of proof of Theorem 4.5 and show convergence in $\mathcal{L}(X_{-\beta})$.

Obviously, this is a sub-optimal estimation, as $\|\cdot\|_{-\beta}$ is a weaker norm than $\|\cdot\|_{L^2}$ for $\beta > 0$. However, from [21, Theorem 1.2.1.1] we have that $\tilde{X}(t)S_h\tilde{X}(t)$ is actually bounded in L^2, at least away from $t = 0$. It therefore seems likely that one could use similar ideas to prove that the same holds for $\tilde{Y}(t)S_h\tilde{Y}(t)$, in which case we would have convergence in $\mathcal{L}(L^2)$. Unfortunately, the theory required for such estimations is rather extensive, and we expect it to be even more so for the LOD approximations. We therefore leave such questions as future work.

8.2. Systems of equations and applications in multiphysics. In this paper we consider problems where the evolution operator $\mathcal{A}$ in the state equation defines an inner product of the form $a(u, v) = \int \kappa \nabla u \cdot \nabla v$. However, many interesting applications requires coupled systems to be modeled appropriately, for instance, multiphysical features such as thermoelasticity [9], which describes temperature and displacement in a material. Another example is the singularly perturbed systems [17, 29], which appear when modeling, for instance, fluid catalytic crackers. These are ill-conditioned problems due to a significantly larger time derivative for one (or more) of the equations.

The LOD method has successfully been applied to thermoelasticity and poroelasticity problems, see [23]. With more complicated models, the computational gain in using a coarse representation of the underlying partial differential equation is even greater. Analysis of such problems should be considered in the future.

8.3. Other time discretizations. It should also be noted that the LOD approach could be used with other time discretizations of the DRE. We have here chosen the Strang splitting scheme due to it being familiar to one of the authors and because an efficient implementation was readily available. However, there are also other types of splitting schemes [33, 28]. Additionally, one might instead consider e.g. BDF and Rosenbrock methods [6, 5, 20], projection-based methods [16] or even peer methods [19]. These depend on solving linear equation systems rather than computing the solutions to parabolic problems, and the error analysis approach would thus differ. However, bounds similar to that given in Lemma 4.3 naturally exist also for stationary problems [26].

8.4. Algebraic Riccati equations. The latter fact is even more relevant if one considers algebraic Riccati equations (AREs). These are the stationary counterpoints

to the time-dependent DREs and arise when the final time T in the cost functional goes to infinity. In this case, splitting does not apply, but we may still apply LOD to the equation to reduce its complexity. Then any method for AREs may be applied to solve this smaller problem, such as Newton-Kleinman ADI [4], rational Krylov subspace methods [31], or RADI [3]. See also [7] for a survey. Clearly, for each of these cases one would have to perform an error analysis such as the one provided in this paper.

Acknowledgments. We are grateful to Fredrik Hellman for his assistance with the code for computing the LOD bases.

REFERENCES

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank, *Matrix Riccati equations*, Systems & Control: Foundations & Applications, Birkhäuser, Basel, 2003, https://doi.org/10.1007/978-3-0348-8081-7.

[2] R. E. Bank and H. Yserentant, *On the H^1-stability of the L_2-projection onto finite element spaces*, Numer. Math., 126 (2014), pp. 361–381, https://doi.org/10.1007/s00211-013-0562-4.

[3] P. Benner, Z. Bujanović, P. Kürschner, and J. Saak, *RADI: a low-rank ADI-type algorithm for large scale algebraic Riccati equations*, Numer. Math., 138 (2018), pp. 301–330, https://doi.org/10.1007/s00211-017-0907-5.

[4] P. Benner, J.-R. Li, and T. Penzl, *Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems*, Numer. Linear Algebra Appl., 15 (2008), pp. 755–777, https://doi.org/10.1002/nla.622.

[5] P. Benner and H. Mena, *Rosenbrock methods for solving Riccati differential equations*, IEEE Trans. Automat. Control, 58 (2013), pp. 2950–2956, https://doi.org/10.1109/TAC.2013.2258495.

[6] P. Benner and H. Mena, *Numerical solution of the infinite-dimensional LQR problem and the associated Riccati differential equations*, J. Numer. Math., 26 (2018), pp. 1–20, https://doi.org/10.1515/jnma-2016-1039.

[7] P. Benner and 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, https://doi.org/10.1002/gamm.201310003.

[8] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and control of infinite dimensional systems*, Systems & Control: Foundations & Applications, Birkhäuser Boston, Inc., Boston, MA, second ed., 2007, https://doi.org/10.1007/978-0-8176-4581-6.

[9] M. A. Biot, *Thermoelasticity and irreversible thermodynamics*, J. Appl. Phys., 27 (1956), pp. 240–253.

[10] W. E. Bosarge, Jr., O. G. Johnson, and C. L. Smith, *A direct method approximation to the linear parabolic regulator problem over multivariate spline bases*, SIAM J. Numer. Anal., 10 (1973), pp. 35–49, https://doi.org/10.1137/0710006.

[11] L. Cao, J. Liu, W. Allegretto, and Y. Lin, *A multiscale approach for optimal control problems of linear parabolic equations*, SIAM J. Control Optim., 50 (2012), pp. 3269–3291, https://doi.org/10.1137/110828800.

[12] Y. Chen, Y. Huang, W. Liu, and N. Yan, *A mixed multiscale finite element method for convex optimal control problems with oscillating coefficients*, Comput. Math. Appl., 70 (2015), pp. 297–313, https://doi.org/10.1016/j.camwa.2015.03.020.

[13] C. Engwer, P. Henning, A. Målqvist, and D. Peterseim, *Efficient implementation of the Localized Orthogonal Decomposition method*, ArXiv e-prints, (2016), https://arxiv.org/abs/1602.01658. https://arxiv.org/abs/1602.01658.

[14] R. S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47, https://doi.org/10.1016/0022-247X(73)90022-X.

[15] P. Henning and A. Målqvist, *Localized orthogonal decomposition techniques for boundary value problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1609–A1634, https://doi.org/10.1137/130933198.

[16] A. Koskela and H. Mena, *Analysis of Krylov Subspace Approximation to Large Scale Differential Riccati Equations*, ArXiv e-prints, (2017), https://arxiv.org/abs/1705.07507,

https://arxiv.org/abs/1705.07507.

[17] S. Koskie, C. Coumarbatch, and Z. Gajic, *Exact slow-fast decomposition of the singularly perturbed matrix differential Riccati equation*, Appl. Math. Comput., 216 (2010), pp. 1401–1411, https://doi.org/10.1016/j.amc.2010.02.040.

[18] M. Kroller and K. Kunisch, *Convergence rates for the feedback operators arising in the linear quadratic regulator problem governed by parabolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1350–1385, https://doi.org/10.1137/0728071.

[19] N. Lang, *Numerical Methods for Large-Scale Linear Time-Varying Control Systems and related Differential Matrix Equations*, dissertation, Technische Universität Chemnitz, Chemnitz, Germany, June 2017.

[20] N. Lang, H. Mena, and J. Saak, *On the benefits of the LDL^T factorization for large-scale differential matrix equation solvers*, Linear Algebra Appl., 480 (2015), pp. 44–71, https://doi.org/10.1016/j.laa.2015.04.006.

[21] I. Lasiecka and R. Triggiani, *Control theory for partial differential equations: continuous and approximation theories. I*, vol. 74 of Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 2000. Abstract parabolic systems.

[22] J. Li, *A multiscale finite element method for optimal control problems governed by the elliptic homogenization equations*, Comput. Math. Appl., 60 (2010), pp. 390–398, https://doi.org/10.1016/j.camwa.2010.04.017.

[23] A. Målqvist and A. Persson, *A generalized finite element method for linear thermoelasticity*, ESAIM Math. Model. Numer. Anal., 51 (2017), pp. 1145–1171, https://doi.org/10.1051/m2an/2016054.

[24] A. Målqvist and A. Persson, *Multiscale techniques for parabolic equations*, Numer. Math., 138 (2018), pp. 191–217, https://doi.org/10.1007/s00211-017-0905-7.

[25] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, Math. Comp., 83 (2014), pp. 2583–2603.

[26] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, Math. Comp., 83 (2014), pp. 2583–2603, https://doi.org/10.1090/S0025-5718-2014-02868-8.

[27] R. S. McKnight and W. E. Bosarge, Jr., *The Ritz-Galerkin procedure for parabolic control problems*, SIAM J. Control Optim., 11 (1973), pp. 510–524.

[28] H. Mena, A. Ostermann, L.-M. Pfurtscheller, and C. Piazzola, *Numerical low-rank approximation of matrix differential equations*, J. Comput. Appl. Math., (2018), https://arxiv.org/abs/1705.10175. Accepted for publication.

[29] H. Mukaidani, H. Xu, and K. Mizukami, *Numerical algorithm for solving cross-coupled algebraic Riccati equations of singularly perturbed systems*, in Advances in dynamic games, vol. 7 of Ann. Internat. Soc. Dynam. Games, Birkhäuser Boston, Boston, MA, 2005, pp. 545–570, https://doi.org/10.1007/0-8176-4429-6_29.

[30] 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, https://doi.org/10.1016/0022-247X(91)90035-X.

[31] V. Simoncini, D. B. Szyld, and M. Monsalve, *On two numerical methods for the solution of large-scale algebraic Riccati equations*, IMA J. Numer. Anal., 34 (2014), pp. 904–920, https://doi.org/10.1093/imanum/drt015.

[32] T. Stillfjord, *Low-rank second-order splitting of large-scale differential Riccati equations*, IEEE Trans. Automat. Control, 60 (2015), pp. 2791–2796, https://doi.org/10.1109/TAC.2015.2398889.

[33] T. Stillfjord, *Adaptive high-order splitting schemes for large-scale differential Riccati equations*, Numer. Algorithms, (2017), https://doi.org/10.1007/s11075-017-0416-8.

[34] H. Tanabe, *Equations of evolution*, vol. 6 of Monographs and Studies in Mathematics, Pitman (Advanced Publishing Program), Boston, Mass.-London, 1979. Translated from the Japanese by N. Mugibayashi and H. Haneda.

[35] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[36] R. Winther, *Error estimates for a Galerkin approximation of a parabolic control problem*, Ann. Mat. Pura Appl. (4), 117 (1978), pp. 173–206, https://doi.org/10.1007/BF02417890.