



## Multiscale differential riccati equations for linear quadratic regulator problems

Downloaded from: <https://research.chalmers.se>, 2025-12-10 00:27 UTC

Citation for the original published paper (version of record):

Målqvist, A., Persson, A., Stillfjord, T. (2018). Multiscale differential riccati equations for linear quadratic regulator problems. SIAM Journal of Scientific Computing, 40(4): A2406-A2426.  
<http://dx.doi.org/10.1137/17M1134500>

N.B. When citing this work, cite the original published paper.

## MULTISCALE DIFFERENTIAL RICCATI EQUATIONS FOR LINEAR QUADRATIC REGULATOR PROBLEMS\*

AXEL MÅLQVIST<sup>†</sup>, ANNA PERSSON<sup>†</sup>, AND TONY STILLFJORD<sup>†</sup>

**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

**DOI.** 10.1137/17M1134500

**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

$$\begin{aligned} (1) \quad & \dot{x} = \mathcal{A}x + \mathcal{B}u, \quad x(0) = x_0 \\ (2) \quad & y = \mathcal{C}x, \end{aligned}$$

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

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

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):

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

\*Submitted to the journal's Methods and Algorithms for Scientific Computing section June 13, 2017; accepted for publication (in revised form) May 29, 2018; published electronically August 2, 2018.

<http://www.siam.org/journals/sisc/40-4/M113450.html>

**Funding:** This work was supported by the Swedish Research Council under grant 2015-04964.

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

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

In this paper, we consider the case when the operator  $\mathcal{A}$  exhibits multiscale behavior. 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 1970s (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 the reader 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 two- or three-dimensional-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  $\Omega$  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  $\kappa$  given below. Then  $\mathcal{A}: L^2 \supset \mathcal{D}(\mathcal{A}) \rightarrow L^2$  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  $\mathcal{B}: U \rightarrow L^2$  and the output operator  $\mathcal{C}: L^2 \rightarrow Z$  be given. We also consider the output and input weighting operators  $\mathcal{Q}: Z \rightarrow Z$  and  $\mathcal{R}: U \rightarrow U$  (which could be included in  $\mathcal{C}$  and  $\mathcal{B}$  but are typically not) and the final state weighting operator  $\mathcal{G}: L^2 \rightarrow L^2$ . By  $*$ , we denote Hilbert-adjoint operators with respect to  $L^2$ , so that, e.g.,  $\mathcal{B}^*: L^2 \rightarrow U$  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

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

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  $e^{t\mathcal{A}}: L^2 \rightarrow L^2$ ; 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, Chapter 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, (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 can 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  $\Omega$  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

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

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

The semidiscretized weak form of (3) is defined by

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

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

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

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

**3.1. Localized orthogonal decomposition.** If  $\kappa$  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 behavior 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  $V_{\text{ms}} \subset V_h$  of the same dimension as the coarse space  $V_H$ . To obtain such a space, we use the 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 the reader to [25, 15], for the details.

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

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

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 Cl  ment interpolant as in [25]. Let  $V_f$  denote the kernel of  $I_H$ ,

$$V_f := \ker I_H = \{v \in V_h : I_H v = 0\},$$

and note that  $V_h$  can be decomposed as  $V_h = V_H \oplus V_f$ , meaning that every  $v_h \in V_h$  can be written as  $v_h = v_H + v_f$  with  $v_H \in V_H, v_f \in V_f$ . We now introduce the (global) correction operator  $\hat{Q}_h: V_H \rightarrow V_f$  by

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

and define the (global) multiscale space as  $\hat{V}_{\text{ms}} := \hat{R}_h V_H = V_H - \hat{Q}_h V_H$  with  $\hat{R}_h := \text{Id}_H^h - \hat{Q}_h$ . This leads to the decomposition  $V_h = \hat{V}_{\text{ms}} \oplus V_f$  with the orthogonality  $a(\hat{v}_{\text{ms}}, v_f) = 0$ ,  $\hat{v}_{\text{ms}} \in \hat{V}_{\text{ms}}, v_f \in V_f$ . 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  $\hat{V}_{\text{ms}}$  is the orthogonal complement to  $V_f$ . From the construction, it follows that  $\dim \hat{V}_{\text{ms}} = \dim V_H$ . Indeed, a basis for  $\hat{V}_{\text{ms}}$  is given by  $\{\varphi_i^H - \hat{Q}_h \varphi_i^H : i = 1, \dots, N_H\}$ .

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  $V_f \subseteq V_h$ . 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:

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

Further, we define  $V_f^K := \{v \in V_f : v(z) = 0 \text{ on } \bar{\Omega} \setminus \omega_k(K)\}$  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

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

We can now localize these computations by replacing  $V_f$  with  $V_f^K$ . Define  $Q_h^K: V_H \rightarrow V_f^K$  such that

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

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

The approximation properties (and the required computational effort) of the space  $V_{\text{ms}}$  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  $\text{Id}_{\text{ms}}^h : V_{\text{ms}} \rightarrow V_h$ ,  $\text{Id}_{\text{ms}}^h u = u$ . Its  $L^2$ -adjoint is the  $L^2$ -orthogonal projection of  $V_h$  onto  $V_{\text{ms}}$ . Replacing the space  $V_h$  with  $V_{\text{ms}}$  then results in the problem to find  $X_h^{\text{ms}} : V_{\text{ms}} \rightarrow V_{\text{ms}}$  satisfying

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

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

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

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

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

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

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

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

and we consequently define the corrected coarse-scale operators

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

**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  $\kappa$ . We start by gathering some useful results.

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

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

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

**LEMMA 4.2.** *Under Assumption 2.1, the operators*

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

*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  $\|E_{\text{ms}}(t)\|_{\mathcal{L}(L^2)} \leq 1$ .*

By arguing as in [24] but for the (simpler) semidiscrete case, we have (choosing  $k \sim \log H$ ) the following lemma.

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

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

Here, the constant  $C$  depends on  $T$ ,  $\alpha$ ,  $\beta$ , and  $\gamma$  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  $u_{ms}(t) = e^{t\mathcal{A}_h^{ms}} P_{ms} v$  for  $v \in L^2(\Omega)$ . By introducing the Ritz projection  $R_{ms}: V_h \rightarrow V_{ms}$  satisfying  $a(R_{ms}v, w) = a(v, w)$  for all  $w \in V_{ms}, v \in V_h$ , we get (see [35, Chapter 3], [24])

$$\|u_h - u_{ms}\| \leq Ct^{-1} \sup_{s \leq t} \left\{ s^2 \|\dot{\rho}\| + s \|\rho\| + \left\| \int_0^s \rho(r) dr \right\| \right\},$$

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

$$\|u_h - u_{ms}\| \leq CH^2 t^{-1} \sup_{s \leq t} (s^2 \|\ddot{u}_h(s)\| + s \|\dot{u}_h(s)\| + \|u_h(s)\| + \|v\|).$$

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

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

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

Then we have the following lemma.

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

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

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*

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

Here, the constant  $C$  depends on  $T$ ,  $\alpha$ ,  $\beta$ ,  $\gamma$ , 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)

$$X_h(t) = e^{t\mathcal{A}_h^*} X_h(0) e^{t\mathcal{A}_h} + \int_0^t e^{(t-s)\mathcal{A}_h^*} \left( \mathcal{C}_h^* \mathcal{Q} \mathcal{C}_h - X_h(s) \mathcal{B}_h \mathcal{R}^{-1} \mathcal{B}_h^* X_h(s) \right) e^{(t-s)\mathcal{A}_h} ds$$

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



$$\begin{aligned}\tilde{X}(t) &= E_h(t)^* \tilde{X}(0) E_h(t) \\ &\quad + \int_0^t E_h(t-s)^* \left( (C_h P_h)^* \mathcal{Q} C_h P_h - \tilde{X}(s) S_h \tilde{X}(s) \right) E_h(t-s) \, ds\end{aligned}$$

as well as

$$\begin{aligned}\tilde{Y}(t) &= E_{\text{ms}}(t)^* \tilde{X}(0) E_{\text{ms}}(t) \\ &\quad + \int_0^t E_{\text{ms}}(t-s)^* \left( (C_h P_h)^* \mathcal{Q} C_h P_h - \tilde{Y}(s) S_h \tilde{Y}(s) \right) E_{\text{ms}}(t-s) \, ds.\end{aligned}$$

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

$$\begin{aligned}\tilde{X}(t) - \tilde{Y}(t) &= E_h(t)^* \tilde{X}(0) \left( E_h(t) - E_{\text{ms}}(t) \right) + \left( E_h(t) - E_{\text{ms}}(t) \right)^* \tilde{X}(0) E_{\text{ms}}(t) \\ &\quad + \int_0^t E_h(t-s)^* (C_h P_h)^* \mathcal{Q} (C_h P_h) \left( E_h(t-s) - E_{\text{ms}}(t-s) \right) \\ &\quad + \left( E_h(t-s) - E_{\text{ms}}(t-s) \right)^* (C_h P_h)^* \mathcal{Q} (C_h P_h) E_{\text{ms}}(t-s) \\ &\quad + \left( E_h(t-s) - E_{\text{ms}}(t-s) \right)^* \tilde{X}(s) S_h \tilde{X}(s) E_h(t-s) \\ &\quad + E_{\text{ms}}(t-s)^* \tilde{X}(s) S_h \tilde{X}(s) \left( E_h(t-s) - E_{\text{ms}}(t-s) \right) \\ &\quad + E_{\text{ms}}(t-s)^* \left( \tilde{X}(s) - \tilde{Y}(s) \right) S_h \tilde{X}(s) E_{\text{ms}}(t-s) \\ &\quad + E_{\text{ms}}(t-s)^* \tilde{Y}(s) S_h \left( \tilde{X}(s) - \tilde{Y}(s) \right) E_{\text{ms}}(t-s) \, ds \\ &=: R_1 + R_2 + \int_0^t \sum_{j=3}^8 R_j(s) \, ds,\end{aligned}$$

so that

$$\|\tilde{X}(t) - \tilde{Y}(t)\|_{\mathcal{L}(L^2)} \leq \|R_1\|_{\mathcal{L}(L^2)} + \|R_2\|_{\mathcal{L}(L^2)} + \int_0^t \left\| \sum_{j=3}^8 R_j(s) \right\|_{\mathcal{L}(L^2)} \, ds.$$

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

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

Additionally, using Lemma 4.1 shows that the last two integrands satisfy

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

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

$$\begin{aligned}\int_0^t \|R_3(s)\|_{\mathcal{L}(L^2)} \, ds &\leq \int_0^{t-H^2} C \|C_h P_h\|_{\mathcal{L}(L^2)}^2 H^2 (t-s)^{-1} \, ds + \int_{t-H^2}^t 2 \|C_h P_h\|_{\mathcal{L}(L^2)}^2 \, ds \\ &\leq CH^2 (\log t - 2 \log H) + CH^2 \\ &\leq CH^2 (\log(H^{-1}) + t^{-1}),\end{aligned}$$

where we have used  $t \leq T$  for the crude estimate  $\log t \leq Ct^{-1}$  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 Lemmas 4.1 and 4.4, also for  $R_5$  and  $R_6$ . In conclusion, we thus have

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

which by Grönwall's lemma yields the statement of the theorem.  $\square$

*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  $\kappa$ , 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 the reader 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  $T$ ,  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\|X(0)\|_{\mathcal{L}(V)}$  but not on the multiscale variations of  $\mathcal{A}$ .

*Proof.* We start by noting that  $\|\text{Id}_h\|_{\mathcal{L}(V_h, V)} \leq 1$ . 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:

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

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

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

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

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

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

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

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

For  $R_1$ , we have

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

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

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

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

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

and

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

By applying Grönwall's lemma, we obtain the desired error bound.  $\square$

**5. Matrix-valued formulation.** To perform actual computations, we write the finite-dimensional equations in 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  $X_h: V_h \rightarrow V_h$  have the vector and matrix representations  $\mathbf{x} \in \mathbb{R}^{N_h}$  and  $\tilde{\mathbf{X}}^h \in \mathbb{R}^{N_h \times N_h}$ , i.e.,

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

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

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

where  $\mathbf{M}$  denotes the (symmetric) mass matrix,  $\mathbf{M}_{i,j} = (\varphi_j, \varphi_i)$ . 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

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

Coincidentally, this means that we actually have

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

Equation (5) is equivalent to

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

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

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

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

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

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

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

where  $\mathbf{B}_{i,j} = (\mathcal{B} \varphi_j^U, \varphi_i)$ ,  $\mathbf{Q}_{i,j} = (\mathcal{Q} \varphi_j^Z, \varphi_i^Z)$ ,  $\mathbf{R}_{i,j} = (\mathcal{R} \varphi_j^U, \varphi_i^U)$ ,  $\mathbf{C}_{i,j} = (\mathcal{C} \varphi_j, \varphi_i^Z)$ , 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

$$(15) \quad \mathbf{M} \dot{\mathbf{X}} \mathbf{M} = \mathbf{M} \mathbf{X} \mathbf{A} + \mathbf{A}^T \mathbf{X} \mathbf{M} + \mathbf{C}^T \mathbf{Q} \mathbf{C} - \mathbf{M} \mathbf{X} \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{X} \mathbf{M}.$$

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

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

where  $\mathbf{I}_H^h \in \mathbb{R}^{N_h \times N_H}$  is the prolongation matrix that satisfies  $\mathbf{I}_H^h \mathbf{x}^H = \mathbf{x}^h$  if  $x = \sum_{j=1}^{N_H} \mathbf{x}_j^H \varphi_j^H$  and  $\text{Id}_H^h x = \sum_{j=1}^{N_h} \mathbf{x}_j^h \varphi_j^h$ . By expressing the  $\varphi^H$  functions in terms of  $\varphi^h$ , it can be seen that  $(\mathbf{I}_H^h)_{i,j} = \varphi_j^H(z_i)$ , 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  $(\text{Id}_H^h)^*$  is not  $(\mathbf{I}_H^h)^T$  but  $\mathbf{M}_H^{-1} (\mathbf{I}_H^h)^T \mathbf{M}_h$ .

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

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

where  $\mathbf{X}_H^{\text{ms}}$  is symmetric and satisfies

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

with the matrices

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

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

**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

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

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

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

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

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

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

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

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

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

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

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

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

We also get that  $\|\text{Id}_h X_h P_h - \text{Id}_H X_H P_H\|_{\mathcal{L}(V)}$  is bounded from below by  $\|\mathbf{L}_A^T (\mathbf{X}_h - \mathbf{I}_H^h \mathbf{X}_H (\mathbf{I}_H^h)^T) \mathbf{M} \mathbf{L}_A^{-T}\|_{\mathbb{R}^{N \times N}}$ ; i.e., the latter quantity can be thought of as an equivalent norm. Since  $\mathbf{L}_A$  is triangular, the extra cost required for the computation of  $\mathbf{L}_A^{-T}$  is negligible. If the low-rank formulation is used (see subsection 6.1), only a small number of linear equation systems involving  $\mathbf{L}_A$  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  $e^{t\mathbf{A}^T}$ , i.e., essentially solving a parabolic problem. Let  $\tau$  denote a fixed time step, and let  $t_j = j\tau$ ,  $j = 0, \dots, N_t$ , be the time discretization of the interval  $[0, T]$ . We split (15) into two parts,  $\dot{\mathbf{X}} = \mathcal{F} \mathbf{X} + \mathcal{G} \mathbf{X}$ , where

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

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

$$\mathbf{X}^{j+1} = e^{\frac{\tau}{2}\mathcal{F}} e^{\tau\mathcal{G}} e^{\frac{\tau}{2}\mathcal{F}} \mathbf{X}^j.$$

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

$$(16) \quad e^{t\mathcal{F}} \mathbf{X} = e^{t\mathbf{M}^{-T}\mathbf{A}^T} \mathbf{X} e^{t\mathbf{A}\mathbf{M}^{-1}} + \int_0^t e^{s\mathbf{M}^{-T}\mathbf{A}^T} \mathbf{M}^{-T} \mathbf{C}^T \mathbf{Q} \mathbf{C} \mathbf{M}^{-1} e^{s\mathbf{A}\mathbf{M}^{-1}} ds,$$

$$(17) \quad e^{t\mathcal{G}} \mathbf{X} = (\mathbf{I} + t\mathbf{X}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T)^{-1} \mathbf{X},$$

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  $\mathbf{X}$  has low rank. This is generally true for LQR problems and dramatically reduces the computational cost. In that case, we may factorize  $\mathbf{X} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ , where  $\mathbf{L} \in \mathbb{R}^{N_h \times r}$  and  $\mathbf{D} \in \mathbb{R}^{r \times r}$  with the rank  $r \ll N_h$ . Also  $e^{\tau\mathcal{F}} \mathbf{X}$  and  $e^{\tau\mathcal{G}} \mathbf{X}$ , and thus the iterates  $\mathbf{X}_j$ , may then be factorized in such a way. After a reformulation,  $e^{\tau\mathcal{G}} \mathbf{X}$  is very cheap to compute, and the computation of  $e^{\tau\mathcal{F}} \mathbf{X}$  reduces to an evaluation of  $e^{\tau\mathbf{M}^{-T}\mathbf{A}^T} \mathbf{L}$  (plus preliminary, similar work for the integral term). The latter operation is equivalent to solving  $\mathbf{M}\dot{x} = \mathbf{A}^T x$ ,  $x(0) = \mathbf{L}$ , and the matrix  $\mathbf{M}$  is thus never explicitly inverted. For further details, we refer the reader 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  $\mathbf{X}_h = \mathbf{L}_h \mathbf{D}_h \mathbf{L}_h^T$  and  $\mathbf{X}_H = \mathbf{L}_H \mathbf{D}_H \mathbf{L}_H^T$  with  $\mathbf{L}_h \in \mathbb{R}^{N_h \times r_h}$  and  $\mathbf{L}_H \in \mathbb{R}^{N_H \times r_H}$  with  $r_h, r_H \ll N_h$ , and let  $\mathbf{M}_h = \mathbf{L}_M \mathbf{L}_M^T$  be a Cholesky factorization. By setting

$$\mathbf{V} = [\mathbf{L}_M^T \mathbf{L}_h \quad \mathbf{L}_M^T \mathbf{I}_H^h \mathbf{L}_H] \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} \mathbf{D}_h & 0 \\ 0 & -\mathbf{D}_H \end{bmatrix},$$

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

$$\mathbf{L}_M^T (\mathbf{X}_h - \mathbf{I}_H^h \mathbf{X}_H (\mathbf{I}_H^h)^T) \mathbf{L}_M = \mathbf{V} \mathbf{D} \mathbf{V}^T.$$

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

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

$$\mathbf{L}_A^T (\mathbf{X}_h - \mathbf{I}_H^h \mathbf{X}_H (\mathbf{I}_H^h)^T) \mathbf{M} \mathbf{L}_A^{-T} = \mathbf{G}_1 \mathbf{D} \mathbf{G}_2^T$$

with the same  $\mathbf{D}$  and with

$$\mathbf{G}_1 = [\mathbf{L}_A^T \mathbf{L}_h \quad \mathbf{L}_A^T \mathbf{I}_H^h \mathbf{L}_H] \quad \text{and} \quad \mathbf{G}_2 = [\mathbf{L}_A^{-1} \mathbf{M} \mathbf{L}_h \quad \mathbf{L}_A^{-1} \mathbf{M} \mathbf{I}_H^h \mathbf{L}_H].$$

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

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

where  $\mathbf{S} = \mathbf{R}_1 \mathbf{D} \mathbf{R}_2^T$  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.<sup>1</sup> 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<sup>2</sup> library. These computations were performed on resources at the Chalmers Centre for Computational Science and Engineering provided by the Swedish National Infrastructure for Computing. Each simulation used a single Intel Xeon E5-2650 v3 processor.

The multiscale diffusion coefficients  $\kappa$  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,  $\kappa$  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 FEM.

**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,  $\kappa$  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 three 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 eight triangles and then refine this six times, giving meshes with  $2^{3+2j}$  triangles for  $j = 0, \dots, 6$ . One additional refinement provides the reference grid with  $2^{17} = 131072$  triangles. This results in matrices  $\mathbf{A}_j \in \mathbb{R}^{n \times n}$ ,  $\mathbf{B}_j \in \mathbb{R}^{n \times 3}$ , and  $\mathbf{C}_j \in \mathbb{R}^{1 \times n}$ ,  $j = 0, \dots, 7$ , 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 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.

<sup>1</sup>Available on request from Fredrik Hellman, fredrik.hellman@it.uu.se.

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

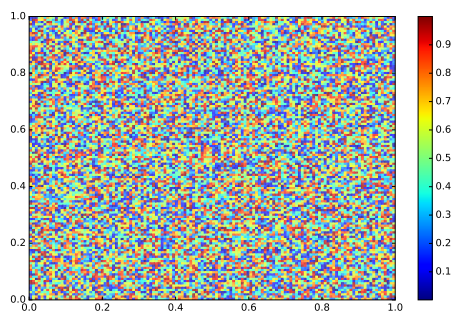


FIG. 1. The diffusion coefficient used in Example 1, plotted over the domain  $\Omega$ . (This figure is in color in the electronic version of the article.)

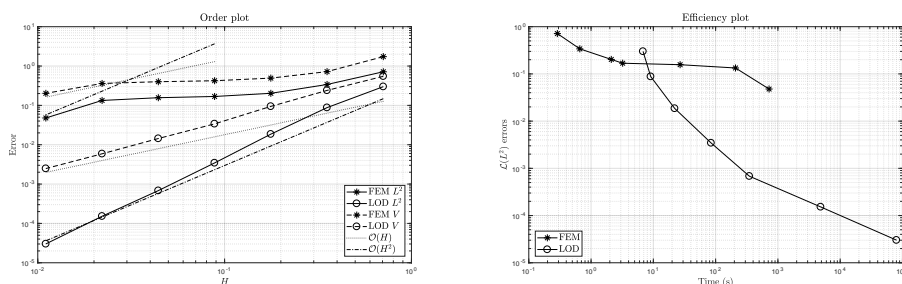


FIG. 2. Left: The  $\mathcal{L}(L^2)$ - and  $\mathcal{L}(V)$ -norm errors of the approximations computed in Example 1, plotted against the meshwidth. Right: The  $\mathcal{L}(L^2)$ -norm errors plotted against the computation time.

**7.2. Example 2.** Here, we consider an L-shaped domain  $\Omega$ , where  $[0.5, 1] \times [0.5, 0.5]$  has been removed from the unit square. The diffusion coefficient  $\kappa$  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 set up 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,  $\kappa$  takes the constant value 1 everywhere, except for in seven 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.



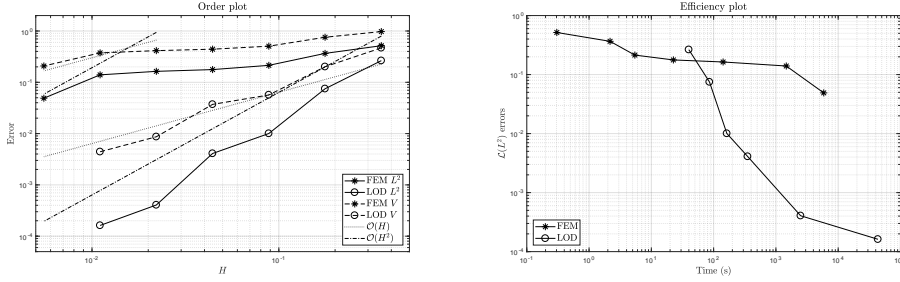


FIG. 3. Left: The  $\mathcal{L}(L^2)$ - and  $\mathcal{L}(V)$ -norm errors of the approximations computed in Example 2, plotted against the meshwidth. Right: The  $\mathcal{L}(L^2)$ -norm errors plotted against the computation time.

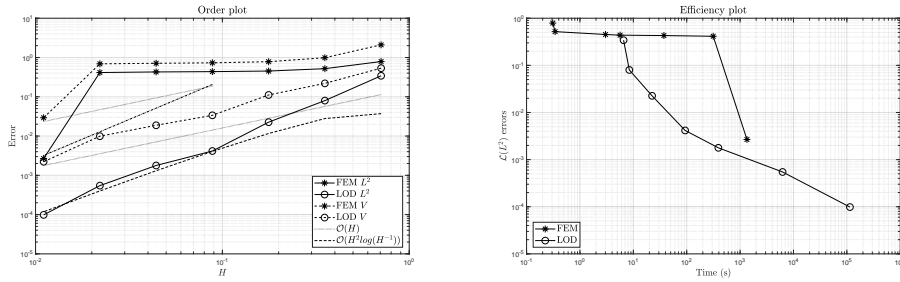


FIG. 4. Left: The  $\mathcal{L}(L^2)$ - and  $\mathcal{L}(V)$ -norm errors of the approximations computed in Example 3, plotted against the meshwidth. Right: The  $\mathcal{L}(L^2)$ -norm errors plotted against the computation time.

**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  $\Gamma_D$  and impose homogeneous Dirichlet boundary conditions there. The left and right edges we denote  $\Gamma_1$  and  $\Gamma_2$ , 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

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

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  $\Gamma_1, \Gamma_2$ , 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

$$B_{j,i}^h = \int_{\Gamma_i} \Psi \varphi_j^h.$$

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

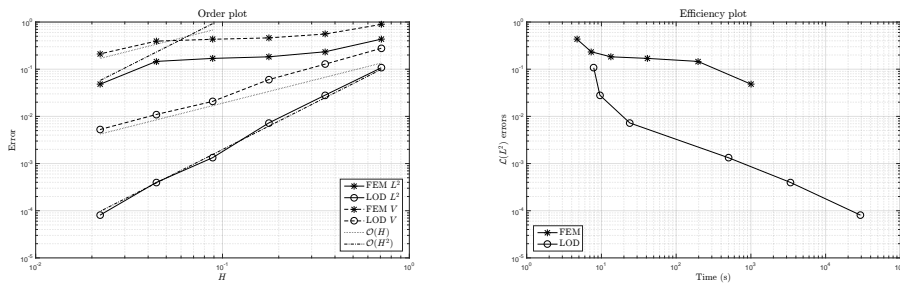


FIG. 5. Left: The  $\mathcal{L}(L^2)$ - and  $\mathcal{L}(V)$ -norm errors of the approximations computed in Example 5, plotted against the meshwidth. Right: The  $\mathcal{L}(L^2)$ -norm errors plotted against the computation time.

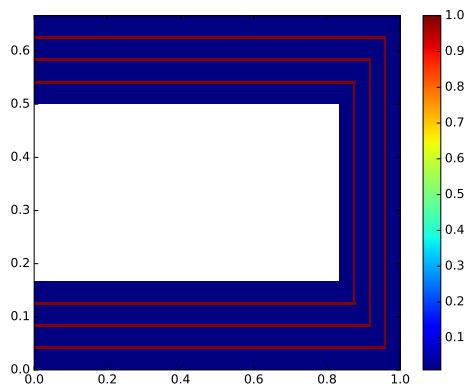


FIG. 6. The diffusion coefficient used in Example 5, plotted over the domain  $\Omega$ . (This figure is in color in the electronic version of the article.)

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 topmost vertical part,  $\Gamma_1$ , we impose a nonhomogeneous Neumann condition  $\kappa \nabla x \cdot n = \Psi u$  with  $\Psi$  having the same hat-shaped form as in Example 4. On the bottom vertical part,  $\Gamma_2$ , 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  $n = 58240$ .

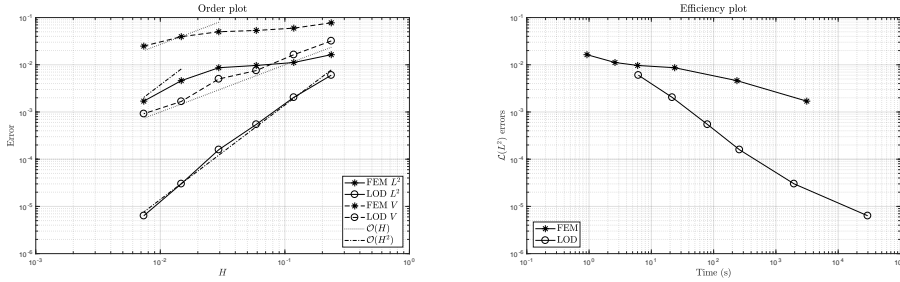


FIG. 7. Left: The  $\mathcal{L}(L^2)$ - and  $\mathcal{L}(V)$ -norm errors of the approximations computed in Example 5, plotted against the meshwidth. Right: The  $\mathcal{L}(L^2)$ -norm errors plotted against the computation time.

The results are plotted in Figure 7, where we can once again observe error behavior 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 cannot 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, 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  $C_h P_h \in \mathcal{L}(L^2, Z)$ . 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 suboptimal 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 because it is 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 AREs. These are the stationary counterparts 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 RAD [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 Found. Appl., 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*, 2nd ed., Systems Control Found. Appl., Birkhäuser, Boston, MA, 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*, preprint, arXiv:1602.01658, 2016.
- [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](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*, preprint, arXiv:1705.07507, 2017.
- [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*, Ph.D. dissertation, Technische Universität Chemnitz, Chemnitz, Germany, 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*, Encyclopedia Math. Appl. 74, Cambridge University Press, Cambridge, 2000.

- [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. PFURTSCHALLER, AND C. PIAZZOLA, *Numerical low-rank approximation of matrix differential equations*, J. Comput. Appl. Math., 340 (2018), pp. 602–614, <https://doi.org/10.1016/j.cam.2018.01.035>.
- [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, Ann. Internat. Soc. Dynam. Games 7, Birkhäuser, Boston, MA, 2005, pp. 545–570, [https://doi.org/10.1007/0-8176-4429-6\\_29](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](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*, Monogr. Stud. Math. 6, Pitman (Advanced Publishing Program), Boston, MA, 1979. Translated from the Japanese by N. Mugibayashi and H. Haneda.
- [35] V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer Ser. Comput. Math., Springer-Verlag, Berlin, 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>.