

Optimal order multigrid preconditioners for the distributed control of parabolic equations with coarsening in space and time

Andrei Drăgănescu^a and Mona Hajghassem^b

^aDepartment of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250, USA; ^b Montgomery College, Germantown, MD 20876, USA

ARTICLE HISTORY

Compiled February 19, 2022

ABSTRACT

We devise multigrid preconditioners for linear-quadratic space-time distributed parabolic optimal control problems. While our method is rooted in earlier work on elliptic control, the temporal dimension presents new challenges in terms of algorithm design and quality. Our primary focus is on the $cG(s)dG(r)$ discretizations which are based on functions that are continuous in space and discontinuous in time, but our technique is applicable to various other space-time finite element discretizations. We construct and analyse two kinds of multigrid preconditioners: the first is based on full coarsening in space and time, while the second is based on semi-coarsening in space only. Our analysis, in conjunction with numerical experiments, shows that both preconditioners are of optimal order with respect to the discretization in case of $cG(1)dG(r)$ for $r = 0, 1$, and exhibits a suboptimal behavior in time for Crank-Nicolson. We also show that, under certain conditions, the preconditioner using full space-time coarsening is more efficient than the one involving semi-coarsening in space, a phenomenon that has not been observed previously. Our numerical results confirm the theoretical findings.

KEYWORDS

Multigrid; PDE-constrained optimization; parabolic equations; space-time finite element methods

This is an Accepted Manuscript of an article published by Taylor & Francis in Optimization Methods and Software on 17 Feb 2022, available online: <http://www.tandfonline.com/doi/10.1080/10556788.2021.2022145>.

1. Introduction

The focus in this article is on developing efficient multigrid preconditioners for the optimal control of the linear parabolic equation:

$$\min_{y,u} J(y, u) = \frac{1}{2} \int_0^T \int_{\Omega} (y(x, t) - y_d(x, t))^2 dxdt + \frac{\beta}{2} \int_0^T \int_{\Omega} u(x, t)^2 dxdt \quad (1)$$

CONTACT Andrei Drăgănescu. Email: draga@umbc.edu

subject to

$$\begin{aligned}
\partial_t y - \Delta y &= u + f && \text{in } Q, \\
y(x, t) &= 0 && \text{on } \partial\Omega \times (0, T), \\
y(x, 0) &= y_0(x) && \text{in } \Omega.
\end{aligned} \tag{2}$$

Here $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) is a bounded convex domain, $Q = \Omega \times (0, T)$, $y_d \in L^2(Q)$ is the desired state, $f \in L^2(Q)$ is a given forcing term, and $y_0 \in L^2(\Omega)$. We refer to y as the state, while $u \in L^2(Q)$ is the control. Non-homogeneous Dirichlet conditions, as well as Neumann or mixed boundary conditions can also be considered, but we opted for homogeneous conditions for simplicity and clarity.

The problem (1)-(2) is a key textbook case [36, 8] in the optimal control of partial differential equations (PDEs), and perhaps the easiest in a series of problems where the controls are distributed in space and time; therefore, it forms an important stepping stone for testing new methods and ideas. Over the last two decades PDE-constrained optimization has received an increasing level of attention from the engineering and computational mathematics communities due to its wide range of applicability. In particular, the optimal control of time-dependent systems includes classes of problems like inversion of reaction-diffusion equations [6], control of fluid flows [17], with applicability to 3D and 4D variational data assimilation for weather prediction [2], and reservoir history matching in gas and oil extraction [9]. All of the aforementioned applications share a common feature: the models can lead to extremely large problems that have to be solved in a relatively short amount of time.

When speaking of solver strategies for optimal control problems with PDE constraints, one usually distinguishes between the *optimize-then-discretize* and *discretize-then-optimize* approaches. However, the more relevant question is whether to focus on solving the first order optimality conditions, i.e., the Karush-Kuhn-Tucker (KKT) system, or a reduced system obtained by eliminating the PDE-constraints from the KKT system, or even from the original optimization problem. There are pros and cons for each approach: here the KKT is a large linear system exhibiting a saddle-point structure, but it is sparse. Instead, reduced systems are positive definite, but they are expected to be dense, thus leading to the need for matrix free algorithms. Needless to say, both strategies require good preconditioners. Many, perhaps the majority of works [30, 29, 34, 24] focus on preconditioning directly the unreduced KKT system in association with Krylov solvers such as MINRES [32]. Usually these preconditioning techniques are matrix-based, and they tend to exploit special structures of the matrices involved, which is why oftentimes they are restricted to backward Euler discretizations in time. Nonetheless, the use of multigrid methods for solving optimization problems constrained by PDEs have also been used primarily on KKT systems [8] rather than reduced systems. This is due to the fact that multigrid has long been regarded as an optimal method for solving large-scale discretized PDEs, and KKT systems are essentially coupled systems of two PDEs with a similar character, namely the state equation representing the constraints and the adjoint equation. For example, for the problem (1)-(2) both the state and the adjoint equations are parabolic PDEs. For parabolic equations it has long been known [21] that semi-coarsening in space works significantly better than full coarsening in space and time, although recent work beginning with [15] shows full coarsening also to be effective if certain conditions are satisfied. Many authors [5, 16, 7, 26, 27, 25] suggest that semi-coarsening in space works better than full space-time coarsening also for the KKT systems associated with

parabolic control problems. Of course, full coarsening in space and time is preferable, when effective, due to the lower computational cost for each iteration. It should also be noted that most aforementioned works on multigrid for parabolic control use finite difference discretizations for the parabolic equations in the KKT system, some of which are specially designed for the KKT system, as are the leap-frog methods in [26, 27, 25].

Our procedure is rooted in the two-grid and multigrid methods similar to [31, 20, 11], and is certainly related to multigrid methods for integral equations of the second kind [19] and ill-posed problems [23, 22]; these differ fundamentally from multigrid methods designed for discretizations of PDEs. The approach we take diverges from the aforementioned preconditioning work on parabolic control in several relevant aspects. First, we work with the reduced system - the reduced Hessian - obtained by removing the state constraints from the discrete optimization problem resulted from the discretize-then-optimize strategy. Second, we use space-time finite element methods as opposed to finite difference discretizations. While many such finite element methods are available in principle, we focus primarily on the $cG(s)dG(r)$ method as described in [28], and we follow [3] in regarding Crank-Nicolson as a space-time finite element method. Third, our method is essentially matrix-free, in the sense that we do not take into account the structure of the large space-time matrices in the KKT system, nor do we ever form them; we only rely on standard parabolic solution methods for computing Hessian-vector multiplications. Each high-resolution Hessian-vector multiplication is very expensive, as it requires two parabolic solves; however, our goal is to show that not many such operations are necessary at high-resolution.

The technique we present has the advantage of being quite general, as it can be applied, in principle, to various space-time finite element discretizations, which form a more natural setting for defining transition operators between levels, as well as allow for more flexibility in terms of geometry and time stepping strategies. While multigrid methods of this sort have been known since their original introduction [18] to lead to a decrease in number of iterations as the resolution increases, the exact behavior for the multigrid preconditioning methods under scrutiny, and their interaction with certain discretizations is still a topic of current research.

Our main contribution is the design and partial analysis of two multigrid preconditioners for the reduced Hessian: the first preconditioner uses full coarsening in space and time (MGST), while the second one uses semi-coarsening in space only (MGSO). A third version, namely semi-coarsening in time only (MGTO) is also possible and analyzable, but is not expected to be efficient in this context. We present numerical and analytical evidence of the approximation properties of the two preconditioning strategies for three finite element discretizations: the first two are drawn from the continuous-in-space-discontinuous-in-time family $cG(s)dG(r)$ family discussed in [28], and the third is essentially a standard Crank-Nicolson discretization. Largely based on the analysis in [28], we present a full analysis of the optimality (with respect to the discretization) of the MGST preconditioner for $cG(1)dG(0)$ (piecewise constants in time), as shown in Theorem 3.1. For the MGSO preconditioner our analysis covers all discretizations of the type $cG(1)dG(r)$ with $r \geq 0$ (discontinuous piecewise polynomials in time or arbitrary degree), as shown in Theorem 3.2. We also show substantial numerical evidence for the optimality of both preconditioners for the $cG(1)dG(1)$ discretization, which is second-order in space and time, and has potentially a higher practical value. By contrast, the MGST preconditioner for the Crank-Nicolson based discretization exhibits a suboptimal behavior, namely it shows a loss of half an order in time-approximation. We show that, under certain conditions, full coarsening in space and time is more efficient than semi-coarsening in space only, which is probably

the first instance of such behavior for the optimal control of parabolic PDEs. While for cG(1)dG(0) the practical value of the method may be limited, since it requires the time-steps to be small compared to the spatial mesh size, for the cG(1)dG(1) discretization the preconditioner shows a great improvement over the unpreconditioned case under more relaxed requirements. Incidentally, the drop by only half an order in approximation for the Crank-Nicolson discretized problem still allows the MGST preconditioner to be sufficiently efficient in practice.

This paper is organized as follows. In Section 2 we introduce the formal details of the problem, the discretizations, and we review the background material needed for the analysis. Section 3 is central to this article as it contains our design and analysis of our multigrid preconditioners. In Section 4 we discuss the implementation and we showcase numerical results in support of our theoretical analysis, as well as a comparison with KKT-based strategies. We end with concluding remarks in Section 5.

2. Problem description and background

Prior to entering the detailed technical discussion we simplify our problem. Due to the linearity of the parabolic equation (2), we can write its solution as $y = \hat{y} + \tilde{y}$, where \hat{y} corresponds to $f = 0$ and $y_0 = 0$, while \tilde{y} corresponds to $u = 0$. After modifying $y_d := y_d - \tilde{y}$, the remaining problem is one where $f = 0$ and $y_0 = 0$. This remains valid through most of the paper (except Section 4.4).

We begin by discussing in detail the continuous problem in Section 2.1, followed by the discrete optimization problem in Section 2.2. The presentation of the background material follows closely [28].

2.1. The continuous problem

Let $I = (0, T)$, $\mathcal{U} = L^2(Q)$ be the control space, and $\|\cdot\|_I$ be the norm on \mathcal{U} . Furthermore, let $\|\cdot\|$ be the norm and (\cdot, \cdot) be the inner product in $L^2(\Omega)$. For an interval $(s, t) \subset \mathbb{R}$ let $(\cdot, \cdot)_{(s,t)}$ denote the inner product in $L^2((s, t), L^2(\Omega))$ (which is isometric to $L^2(\Omega \otimes (s, t))$), and let $\|\cdot\|_{(s,t)}$ be the corresponding norm. Naturally, $(\cdot, \cdot)_I$ is the inner product in $L^2(Q)$. We define the space

$$\mathcal{X} = \{v \in L^2(I, H_0^1(\Omega)) : \partial_t v \in L^2(I, H^{-1}(\Omega))\}. \quad (3)$$

Cf. [28], the weak formulation of (2) with $f = 0$ and $y_0 = 0$ reads: find $y \in \mathcal{X}$ so that

$$\begin{aligned} (\partial_t y, \varphi)_I + (\nabla y, \nabla \varphi)_I &= (u, \varphi)_I \quad \forall \varphi \in \mathcal{X}, \\ y(x, 0) &= 0. \end{aligned} \quad (4)$$

The following existence and regularity result holds (see [28], Proposition 2.1).

Proposition 2.1. *Given $u \in \mathcal{U}$, there exists a unique solution $y \in \mathcal{X}$ of (4), which also exhibits the additional regularity*

$$y \in L^2(I, H_0^1(\Omega) \cap H^2(\Omega)) \cap H^1(I, L^2(\Omega)), \quad (5)$$

and the following stability estimate holds

$$\|\partial_t y\|_I + \|\nabla^2 y\|_I \leq C\|u\|_I. \quad (6)$$

Thus we define the solution operator $\mathcal{K} \in \mathfrak{L}(\mathcal{U}, \mathcal{Y})$ (this denotes the space of bounded linear operators from \mathcal{U} to \mathcal{Y}) with the state space given by

$$\mathcal{Y} \stackrel{\text{def}}{=} L^2(I, H_0^1(\Omega) \cap H^2(\Omega)) \cap H^1(I, L^2(\Omega)) \hookrightarrow \mathcal{U}. \quad (7)$$

Since \mathcal{K} can be regarded as an operator in $\mathfrak{L}(\mathcal{U})$ (we omit the embedding operator $\mathcal{Y} \hookrightarrow \mathcal{U}$ to simplify notation), its adjoint $\mathcal{K}^* \in \mathfrak{L}(\mathcal{U})$ is defined by

$$(\mathcal{K}u, v)_I = (u, \mathcal{K}^*v)_I \quad \forall u, v \in \mathcal{U}. \quad (8)$$

It is important to note that \mathcal{K}^* satisfies a regularity estimate similar to that of \mathcal{K} .

Proposition 2.2. *Given $v \in \mathcal{U}$, the function $z = \mathcal{K}^*v$ lies in \mathcal{X} and satisfies*

$$\begin{aligned} -(\partial_t z, \varphi)_I + (\nabla z, \nabla \varphi)_I &= (v, \varphi)_I \quad \forall \varphi \in \mathcal{X}, \\ z(x, T) &= 0. \end{aligned} \quad (9)$$

Moreover, we have $z \in \mathcal{Y}$, and z also satisfies the stability estimate

$$\|\partial_t z\|_I + \|\nabla^2 z\|_I \leq C\|v\|_I. \quad (10)$$

Proof. The fact that z satisfies (9) can be found in [36]. The stability estimate (10) follows from (6) and the fact that the function $t \mapsto z(\cdot, T - t)$ satisfies (4) with $u(\cdot, t) = v(\cdot, T - t)$. \square

We turn our attention to the reduced problem. Using the solution operator, the original optimization problem (1)–(2) is equivalent to the unconstrained problem

$$\min_{u \in \mathcal{U}} \hat{J}(u) \stackrel{\text{def}}{=} J(\mathcal{K}u, u) = \frac{1}{2}\|\mathcal{K}u - y_d\|_I^2 + \frac{\beta}{2}\|u\|_I^2. \quad (11)$$

We then rewrite the reduced cost functional as

$$\hat{J}(u) = \frac{1}{2}(\mathcal{K}u - y_d, \mathcal{K}u - y_d)_I + \frac{\beta}{2}(u, u)_I = \frac{1}{2}(u, (\mathcal{K}^*\mathcal{K} + \beta\mathcal{I})u - \mathcal{K}^*y_d)_I + \frac{1}{2}(y_d, y_d)_I.$$

By taking the Fréchet derivative in the direction w and using the definition of the gradient with respect to the inner product $(\cdot, \cdot)_I$, we obtain

$$(w, \nabla \hat{J}(u))_I \stackrel{\text{def}}{=} \hat{J}(u)'(w) = (w, (\mathcal{K}^*\mathcal{K} + \beta\mathcal{I})u - \mathcal{K}^*y_d)_I,$$

where \mathcal{I} denotes the identity operator in \mathcal{U} . Since \hat{J} is quadratic and strictly convex, the necessary and sufficient optimality condition for (11) is $\nabla \hat{J}(u) = 0$, which is equivalent to the normal equations

$$\mathcal{G}u \stackrel{\text{def}}{=} (\mathcal{K}^*\mathcal{K} + \beta\mathcal{I})u = \mathcal{K}^*y_d. \quad (12)$$

It is notable that the solution \hat{u} of (12) satisfies the same regularity as the state (see [28], Proposition 2.2), namely $\hat{u} \in \mathcal{Y}$, although we do not make use of this fact. The goal of this article is to develop and analyze efficient preconditioners for discrete versions of the reduced Hessian \mathcal{G} .

2.2. The $cG(s)dG(r)$ discretization

The discretization is performed in two steps. First, the parabolic equation is discretized in time, then in space. The semidiscrete problem in time plays an important role in our multigrid analysis, as we shall see in Section 3. While we are mostly interested in the $cG(1)dG(r)$ method with $r = 0, 1$, we describe the $cG(s)dG(r)$ method in full generality.

2.2.1. Semidiscretization in time

Partition the interval $[0, T] = \bar{I}$ into

$$\bar{I} = \{0\} \cup \bigcup_{m=1}^M I_m, \quad I_m = (t_{m-1}, t_m],$$

and let $k = \max\{k_m : 1 \leq m \leq M\}$ with $k_m = (t_m - t_{m-1})$ being the m^{th} time step. The semidiscrete state and control spaces are

$$\mathcal{Y}_k^r \stackrel{\text{def}}{=} \{v \in L^2(I, H_0^1(\Omega)) : v|_{I_m} \in \mathcal{P}_r(I_m, H_0^1(\Omega))\}, \quad (13)$$

$$\mathcal{U}_k^r \stackrel{\text{def}}{=} \{v \in L^2(I, L^2(\Omega)) : v|_{I_m} \in \mathcal{P}_r(I_m, L^2(\Omega))\}, \quad (14)$$

where $\mathcal{P}_r(I_m, V)$ is the space of polynomial functions of degree $\leq r$ with coefficients in the space V . If $r = 0$, both spaces consist of piecewise constant functions in time. Note that functions in \mathcal{Y}_k^r and \mathcal{U}_k^r are left continuous; for $v \in \mathcal{Y}_k^r$ or \mathcal{U}_k^r we denote its right-limit at t_m and its jumps, respectively, by

$$v_m^+ = \lim_{t \rightarrow 0^+} v(t_m + t), \quad [v]_m = v_m^+ - v(t_m).$$

We define the bilinear form $B : \mathcal{Y}_k^r \times \mathcal{Y}_k^r \rightarrow \mathbb{R}$ by

$$B(y, \varphi) \stackrel{\text{def}}{=} \sum_{m=1}^M (\partial_t y, \varphi)_{I_m} + (\nabla y, \nabla \varphi)_I + \sum_{m=2}^M ([y]_{m-1}, \varphi_{m-1}^+) + (y_0^+, \varphi_0^+). \quad (15)$$

The semidiscrete-in-time $dG(r)$ formulation of (4) is: find $y \in \mathcal{Y}_k^r$ so that

$$B(y, \varphi) = (u, \varphi)_I \quad \forall \varphi \in \mathcal{Y}_k^r. \quad (16)$$

It is shown in [35] that (16) has a unique solution for any $u \in \mathcal{U}$, thus defining a semidiscrete-in-time solution operator $\mathcal{K}_k : \mathcal{U} \rightarrow \mathcal{Y}_k^r$. We will be mostly interested in the restriction of \mathcal{K}_k to $\mathcal{U}_k^r \subset \mathcal{U}$. The fact that the semidiscrete state space and the test space are the same allows for a similar definition of the adjoint operator of \mathcal{K}_k .

More precisely, given $v \in \mathcal{U}$, define $z_k \in \mathcal{Y}_k^r$ to be the only solution to the problem

$$B(\varphi, z_k) = (\varphi, v)_I \quad \forall \varphi \in \mathcal{Y}_k^r. \quad (17)$$

Hence, by letting $\varphi = z_k$ in (16) and $\varphi = y_k$ in (17) we have

$$(u, z_k)_I = B(y_k, z_k) = (y_k, v)_I = (\mathcal{K}_k u, v)_I, \quad \forall u, v \in \mathcal{U},$$

which shows that $z_k = \mathcal{K}_k^* v$. It is worth noting that the Galerkin orthogonality property $B(y - y_k, \varphi) = 0$ holds for all $\varphi \in \mathcal{Y}_k^r$, even though the dG(r) semidiscretization is nonconforming, i.e., $\mathcal{Y}_k^r \not\subseteq \mathcal{Y}$. Also, the semidiscrete adjoint equation (17) represents the semidiscretization of the adjoint equation (9). We will need the following stability and approximation properties.

Proposition 2.3. *The semidiscrete solutions of the dG(r) state and adjoint equations $y_k = \mathcal{K}_k u$, $z_k = \mathcal{K}_k^* v$, together with the continuous solution $y = \mathcal{K}u$ satisfy*

$$\sum_{m=1}^M \|\partial_t y_k\|_{I_m}^2 + \|\nabla^2 y_k\|_I^2 + \sum_{m=1}^M k_m^{-1} \|[y_k]_{m-1}\|^2 \leq C \|u\|_I^2, \quad (18)$$

$$\sum_{m=1}^M \|\partial_t z_k\|_{I_m}^2 + \|\nabla^2 z_k\|_I^2 + \sum_{m=1}^M k_m^{-1} \|[z_k]_{m-1}\|^2 \leq C \|v\|_I^2, \quad (19)$$

$$\|y_k - y\|_I \leq C k^{r+1} \|\partial_t^{r+1} y\|_I. \quad (20)$$

Proof. From Theorem 4.1 in [28] we have

$$\sum_{m=1}^M \|\partial_t y_k\|_{I_m}^2 + \|\Delta y_k\|_I^2 + \sum_{m=1}^M k_m^{-1} \|[y_k]_{m-1}\|^2 \leq C \|u\|_I^2,$$

and (18) follows by applying elliptic regularity on the convex domain Ω :

$$\|\nabla^2 y_k\| \leq C \|\Delta y_k\|.$$

Corollary 4.2 in [28] together with elliptic regularity imply (19). The approximation result (20) is stated in Theorem 5.1 in [28]. \square

2.2.2. The fully discrete problem

To construct the fully cG(s)dG(r) discrete version of (4), consider a quasi-uniform mesh on Ω with mesh size h , and a standard finite element space $V^h \subset H^1(\Omega)$ consisting of continuous piecewise polynomial functions of total degree $\leq s$ on each element. Let $V_0^h = V^h \cap H_0^1(\Omega)$. We define the discrete state and control spaces by

$$\mathcal{Y}_{k,h}^{r,s} \stackrel{\text{def}}{=} \{v \in L^2(I, H_0^1(\Omega)) : v|_{I_m} \in \mathcal{P}_r(I_m, V_0^h)\} \subset \mathcal{Y}_k^r, \quad (21)$$

$$\mathcal{U}_{k,h}^{r,s} \stackrel{\text{def}}{=} \{v \in L^2(I, H^1(\Omega)) : v|_{I_m} \in \mathcal{P}_r(I_m, V^h)\} \subset \mathcal{U}_k^r. \quad (22)$$

For simplicity we discretize the controls and state space using the same meshes in space and time. A key operator for our multigrid preconditioner is the space-time

$L^2(Q)$ -projection $\pi_{k,h} : \mathcal{U} \rightarrow \mathcal{U}_{k,h}^{r,s}$ defined for $v \in \mathcal{U}$ by

$$((\mathcal{I} - \pi_{k,h})v, \varphi)_I = 0 \quad \forall \varphi \in \mathcal{U}_{k,h}^{r,s}. \quad (23)$$

The fully discrete cG(s)dG(r) formulation of (4) is: find $y \in \mathcal{Y}_{k,h}^{r,s}$ so that

$$B(y, \varphi) = (u, \varphi)_I \quad \forall \varphi \in \mathcal{Y}_{k,h}^{r,s}. \quad (24)$$

As in the semidiscrete case, the problem (24) has a unique solution and gives rise to a solution operator $\mathcal{K}_{k,h} : \mathcal{U} \rightarrow \mathcal{Y}_{k,h}^{r,s}$. We use the same notation for its restriction to the discrete control space when there is no chance of confusion. We formulate the discrete reduced optimization problem as:

$$\min_{u \in \mathcal{U}_{k,h}^{r,s}} \hat{J}(u) = \frac{1}{2} \|\mathcal{K}_{k,h}u - \pi_{k,h}y_d\|_I^2 + \frac{\beta}{2} \|u\|_I^2. \quad (25)$$

Similar to the case of the continuous optimization problem in Section 2.1, the solution of the discrete optimal control problem is again given by the discrete normal equations

$$\mathcal{G}_{k,h}u \stackrel{\text{def}}{=} (\mathcal{K}_{k,h}^* \mathcal{K}_{k,h} + \beta \mathcal{I})u = \mathcal{K}_{k,h}^* \pi_{k,h}y_d. \quad (26)$$

The adjoint $\mathcal{K}_{k,h}^*$, considered as an operator in $\mathfrak{L}(\mathcal{U}_{k,h}^{r,s})$ (since $\mathcal{Y}_{k,h}^{r,s} \subset \mathcal{U}_{k,h}^{r,s}$) is defined by the equality

$$(\mathcal{K}_{k,h}u, v)_I = (u, \mathcal{K}_{k,h}^*v)_I \quad \forall u, v \in \mathcal{U}_{k,h}^{r,s}.$$

As in the semidiscrete case, applying the adjoint $\mathcal{K}_{k,h}^*$ to $v \in \mathcal{U}_{k,h}^{r,s}$ involves solving the problem: find $z \in \mathcal{Y}_{k,h}^{r,s}$ so that

$$B(\varphi, z) = (\varphi, v)_I \quad \forall \varphi \in \mathcal{Y}_{k,h}^{r,s}. \quad (27)$$

This implies, of course, that the range of $\mathcal{K}_{k,h}^*$ lies in $\mathcal{Y}_{k,h}^{r,s} \subset \mathcal{U}_{k,h}^{r,s}$.

Proposition 2.4. *The fully discrete solution $y_{kh} = \mathcal{K}_{k,h}u$ of the cG(s)dG(r) discretized problem (24) together with the semidiscrete dG(r) solution $y_k = \mathcal{K}_k u$ and the continuous solution $y = \mathcal{K}u$ satisfy*

$$\|y_{kh} - y_k\|_I \leq Ch^{s+1} \|\nabla^{s+1} y_k\|_I, \quad (28)$$

$$\|y_{kh} - y\|_I \leq C(k^{r+1} \|\partial_t^{r+1} y\|_I + h^{s+1} \|\nabla^{s+1} y_k\|_I), \quad (29)$$

with a constant C independent of k and h .

Proof. The estimate (28) is the result in Theorem 5.5 in [28], while (29) follows from (28) and (20). \square

For the remainder of the paper we restrict our attention to the case $s = 1$, i.e., continuous piecewise linear elements in space.

Corollary 2.5. *The fully discrete solution operator $\mathcal{K}_{k,h}$ of the $cG(1)dG(0)$ discretized problem (24) together with the continuous solution operator \mathcal{K} satisfy*

$$\|(\mathcal{K}_{k,h} - \mathcal{K})u\|_I \leq C(k + h^2)\|u\|_I, \quad (30)$$

$$\|\mathcal{K}_{k,h}u\|_I \leq C\|u\|_I, \quad (31)$$

for all $u \in \mathcal{U}$, with C independent of k, h .

Proof. Given $u \in \mathcal{U}$, let $y_{kh} = \mathcal{K}_{k,h}u$, $y_k = \mathcal{K}_k u$, and $y = \mathcal{K}u$. By (29)

$$\|y_{kh} - y\|_I \leq C(k\|\partial_t y\|_I + h^2\|\nabla^2 y_k\|_I) \stackrel{(6),(18)}{\leq} C(k + h^2)\|u\|_I,$$

which implies (30). Let $\|\mathcal{K}\|_I$ be the operator-norm of \mathcal{K} as an operator in $\mathfrak{L}(\mathcal{U})$. Then

$$\|\mathcal{K}_{k,h}u\|_I \leq \|(\mathcal{K}_{k,h} - \mathcal{K})u\|_I + \|\mathcal{K}u\|_I \stackrel{(30)}{\leq} (C(k + h^2) + \|\mathcal{K}\|_I)\|u\|_I,$$

which implies (31). \square

Corollary 2.6. *For $r \geq 0$, the fully discrete solution operator $\mathcal{K}_{k,h}$ of the $cG(1)dG(r)$ discretized problem (24) together with the semidiscrete $dG(r)$ solution operator \mathcal{K}_k satisfy*

$$\|(\mathcal{K}_{k,h} - \mathcal{K}_k)u\|_I \leq Ch^2\|u\|_I, \quad (32)$$

$$\|\mathcal{K}_k u\|_I \leq C\|u\|_I, \quad (33)$$

with C independent of k, h .

Proof. Given $u \in \mathcal{U}$ define $y_{kh} = \mathcal{K}_{k,h}u$ and $y_k = \mathcal{K}_k u$. By (28)

$$\|y_{kh} - y_k\|_I \leq Ch^2\|\nabla^2 y_k\|_I \stackrel{(18)}{\leq} Ch^2\|u\|_I,$$

which translates into (32). The estimate (33) is shown in a way similar to (31). \square

While the estimates in Corollaries 2.5 and 2.6 are straightforward consequences of the existing stability and approximation estimates in [28], it is critical for our multigrid analysis that the same norm, namely the $L^2(Q)$ -norm, appears on both sides of the inequalities (30)–(33), and that they are optimal with respect to the discretization parameters k and h .

Remark 1. In case of the $cG(1)dG(1)$ discretization, an estimate similar to (30) with k^2 in place of k does not hold, in spite of the optimal estimate (29). The issue is that the latter involves the term $\|\partial_t^2 y\|_I$ which is not bounded above by $\|u\|_I$. Consequently, we cannot generalize the multigrid analysis for $cG(1)dG(0)$ to the case of $cG(1)dG(1)$.

2.3. Crank-Nicolson as a space-time finite element method

We briefly invoke a space-time finite element discretization extracted from [3]. Here, the state space is given by continuous piecewise linear functions in time, i.e.,

$$\tilde{\mathcal{Y}}_{k,h} \stackrel{\text{def}}{=} \mathcal{Y}_{k,h}^{1,1} \cap \mathcal{C}(I, V_0^h), \quad (34)$$

while the test space is $\mathcal{Y}_{k,h}^{0,1}$, that is, the same test (and function) function space used in cG(1)dG(0). The discrete variational formulation of (4) is: find $y \in \tilde{\mathcal{Y}}_{k,h}$ so that

$$\sum_{m=1}^M (\partial_t y, \varphi)_{I_m} + (\nabla y, \nabla \varphi)_I = (u, \varphi)_I, \quad \forall \varphi \in \mathcal{Y}_{k,h}^{0,1}. \quad (35)$$

The choice of the discrete control space is, in principle, independent of the discretization of the state equation; for consistency with our other choices, we opt for continuous-in-time piecewise linear controls

$$\tilde{\mathcal{U}}_{k,h} \stackrel{\text{def}}{=} \mathcal{U}_{k,h}^{1,1} \cap \mathcal{C}(I, V^h). \quad (36)$$

It is shown in [3] that (35) is equivalent to the Crank-Nicolson discretization with time-averaged data. This gives rise to a new discrete solution operator $\tilde{\mathcal{K}}_{k,h} : \tilde{\mathcal{U}}_{k,h} \rightarrow \tilde{\mathcal{Y}}_{k,h}$ which is used to replace $\mathcal{K}_{k,h}$ in (25), and the associated reduced Hessian becomes

$$\tilde{\mathcal{G}}_{k,h} \stackrel{\text{def}}{=} (\tilde{\mathcal{K}}_{k,h}^* \tilde{\mathcal{K}}_{k,h} + \beta \mathcal{I}) \in \mathfrak{L}(\tilde{\mathcal{U}}_{k,h}). \quad (37)$$

Since the state space differs from the test space, it is no longer true that $\tilde{\mathcal{K}}_{k,h}^*$ is the Crank-Nicolson discretization of \mathcal{K}^* . This issue has been addressed in [37] by modifying the discrete adjoint equation in order to improve the overall convergence rate. However, this modification does not appear to have an effect on the two-grid approximation rate of interest to us in this work and further discussed in Section 4.2.4.

3. The multigrid preconditioners

The primary goal of this article is to devise and analyze efficient multigrid preconditioners for the discrete reduced Hessian $\mathcal{G}_{k,h}$ in (26) and (37). As can be easily seen, $\text{cond}(\mathcal{G}_{k,h}) = O(\beta^{-1})$, with the bound being uniform with respect to k and h . Hence, the system (26) can be solved by unpreconditioned conjugate gradient (CG) in a mesh-independent number of iterations. That number, however, could be quite large. This is a serious problem, since applying the Hessian is very expensive, as it requires two PDE solves. Therefore, we want to construct multigrid preconditioners that significantly speed up CG, especially for large problems, i.e., when $k, h \ll 1$. Our proposed multigrid strategy originates in [11], where it was used for the initial value control of a parabolic equation rather than space-time distributed control.

As in earlier works, to quantify the quality of the preconditioner we use the spectral distance between symmetric positive definite operators $\mathcal{A}, \mathcal{B} \in L^2(\mathcal{H})$, defined in [11]

as

$$d_\sigma(\mathcal{A}, \mathcal{B}) = \sup_{u \in \mathcal{H} \setminus \{0\}} \left| \ln \frac{(\mathcal{A}u, u)_\mathcal{H}}{(\mathcal{B}u, u)_\mathcal{H}} \right| = \max(|\ln \rho(\mathcal{A}^{-1}\mathcal{B})|, |\ln \rho(\mathcal{B}^{-1}\mathcal{A})|), \quad (38)$$

where $(\mathcal{H}, (\cdot, \cdot))$ is a Hilbert space. In addition to being intimately connected to the condition number of the preconditioned system (as shown in (38)), this tool allows for a streamlined extension of our analysis from two-grid to multigrid [11, 4]. For initial value control of parabolic equations it was shown in [11] that $d_\sigma(\mathcal{G}_h, \mathcal{T}_h) = O(h^2/\beta)$, where \mathcal{G}_h denotes the reduced Hessian, and \mathcal{T}_h the two-grid preconditioner. We show in this paper that similar results hold for (1)-(2) under certain discretizations.

From here on we omit the superscripts r, s from the spaces, as it should be obvious from the context what polynomial degrees are being used. In this section we introduce two strategies that give rise to multigrid algorithms with similar characteristics but different qualities; each of the two strategies has advantages and disadvantages, which will be further contrasted in Section 4. We consider that, for a given space and time discretization, i.e., given a choice of k and h , the problem at hand (26) represents the finest level in a multigrid hierarchy. We assume the geometric multigrid context, in which our finest-level problem is the result of a successive refinement of grids, so that discretizations at several coarser levels are available for use. This point of view is consistent with earlier embodiments of multigrid preconditioners for optimal control problems [12, 33], as well as with the algebraic multigrid context [4], where coarser discretizations are not available, but need to be created. For the first preconditioner we coarsen in space and time (MGST), while the second only uses semi-coarsening in space only (MGSO). In Section 3.1 we present the construction of the two-grid preconditioners and state the main results, while Section 3.2 is devoted to their analysis. Finally, in Section 3.3 we describe the extension to multigrid.

3.1. The two-grid preconditioners

The description of the two-grid preconditioners is similar for all the discretizations introduced earlier, with only the spaces and transition operators (embedding and projection) being specific for each discretization; hence, we restrict the discussion to the $cG(s)dG(r)$ case.

We begin with the two-grid preconditioner that uses coarsening in space and time. The construction is valid for any degrees $s, r \geq 0$. We assume the meshes leading up to the finest are the result of uniform mesh-refinement, so that the immediately coarser *spatial* finite element space is V^{2h} (V_0^{2h} for the state, respectively), and the coarser time-step is $2k$. This leads us to consider $\mathcal{U}_{2k,2h} \subset \mathcal{U}_{k,h}$ as the immediately coarser *space-time* finite element space for the controls. For the state space there also is a coarser version $\mathcal{Y}_{2k,2h} \subset \mathcal{Y}_{k,h}$, all giving rise to a coarse solver $\mathcal{K}_{2k,2h} \in \mathfrak{L}(\mathcal{U}_{2k,2h}, \mathcal{Y}_{2k,2h})$ and a coarse Hessian $\mathcal{G}_{2k,2h} \in \mathfrak{L}(\mathcal{U}_{2k,2h})$. If $\pi_{2k,2h}$ denotes the $L^2(Q)$ -projection on $\mathcal{U}_{2k,2h}$, and the embedding $\mathcal{U}_{2k,2h} \hookrightarrow \mathcal{U}_{k,h}$ is $\mathcal{E}^{2k,2h} = \pi_{2k,2h}^*$, then we define the two-grid space-time preconditioner by

$$\mathcal{T}_{k,h} \stackrel{\text{def}}{=} \mathcal{E}^{2k,2h} \cdot \mathcal{G}_{2k,2h} \cdot \pi_{2k,2h} + \beta(\mathcal{I} - \mathcal{E}^{2k,2h} \cdot \pi_{2k,2h}) \in \mathfrak{L}(\mathcal{U}_{k,h}). \quad (39)$$

The two-grid preconditioner above is analogous to the case when controls reside in space only, as used for the initial value control of parabolic equations [11], the dis-

tributed or boundary control of elliptic equations [12], or the distributed optimal control of stationary fluid flows [14, 33]. It should be noted that the inverse of $\mathcal{T}_{k,h}$, that we actually need in practice, is given by

$$(\mathcal{T}_{k,h})^{-1} = \mathcal{E}^{2k,2h} \cdot (\mathcal{G}_{2k,2h})^{-1} \cdot \pi_{2k,2h} + \beta^{-1}(\mathcal{I} - \mathcal{E}^{2k,2h} \cdot \pi_{2k,2h}). \quad (40)$$

This fact can be verified by direct multiplication between the two expressions in (39) and (40), after taking into account that the operator $\mathcal{P} = \mathcal{E}^{2k,2h} \cdot \pi_{2k,2h}$ is a projector, i.e., $\mathcal{P}^2 = \mathcal{P}$, and that $\pi_{2k,2h} \mathcal{E}^{2k,2h} = \mathcal{I}_{\mathcal{U}_{2k,2h}}$. This can also be seen from the matrix form of the operators in Section 4.1.

Naturally, we do not expect to use the two-grid preconditioner in practice, since it involves inverting the coarser Hessian $\mathcal{G}_{2k,2h}$, but the construction of the two-grid preconditioner is the critical step towards multigrid preconditioning.

For the second preconditioner we keep the original fine-level time grid, and we coarsen in space only. That is, our coarse control space is $\mathcal{U}_{k,2h} \subset \mathcal{U}_{k,h}$, and the coarse state space is $\mathcal{Y}_{k,2h} \subset \mathcal{Y}_{k,h}$. Hence, we will use the projection in $\mathfrak{L}(\mathcal{U}_{k,h}, \mathcal{U}_{k,2h})$ denoted by $\pi_{k,2h}$, and the embedding $\mathcal{E}^{k,2h} = \pi_{k,2h}^*$. As in the previous case, we define the space-only two-grid preconditioner by

$$\mathcal{S}_{k,h} \stackrel{\text{def}}{=} \mathcal{E}^{k,2h} \cdot \mathcal{G}_{k,2h} \cdot \pi_{k,2h} + \beta(\mathcal{I} - \mathcal{E}^{k,2h} \cdot \pi_{k,2h}). \quad (41)$$

The following two theorems referring to the cG(s)dG(r) discretization will be proved in Section 3.2.

Theorem 3.1. *For $s = 1$ and $r = 0$, there exists a constant $C_{\text{ST}} > 0$ independent of k, h so that the space-time two-grid preconditioner satisfies*

$$\|(\mathcal{G}_{k,h} - \mathcal{T}_{k,h})u\|_I \leq C_{\text{ST}}(k + h^2)\|u\|_I, \quad \forall u \in \mathcal{U}_{k,h}. \quad (42)$$

Moreover, if $C_{\text{ST}}(k + h^2) \leq \beta/2$, then

$$d_\sigma(\mathcal{G}_{k,h}, \mathcal{T}_{k,h}) \leq 2C_{\text{ST}}\beta^{-1}(k + h^2). \quad (43)$$

Theorem 3.2. *For $s = 1$ and $r \geq 0$, there exists a constant $C_{\text{SO}} > 0$ independent of k, h so that the space-only two-grid preconditioner satisfies*

$$\|(\mathcal{G}_{k,h} - \mathcal{S}_{k,h})u\|_I \leq C_{\text{SO}}h^2\|u\|_I, \quad \forall u \in \mathcal{U}_{k,h}. \quad (44)$$

Moreover, if $C_{\text{SO}}h^2 \leq \beta/2$, then

$$d_\sigma(\mathcal{G}_{k,h}, \mathcal{S}_{k,h}) \leq 2C_{\text{SO}}\beta^{-1}h^2. \quad (45)$$

It is remarkable that both of the estimates above are of optimal order, the first one with respect to the space-time discretization, and the second one with respect to the space discretization. In both cases, when solving (26) using multigrid-preconditioned CG, the number of iterations is expected to decrease at the optimal rate with increasing resolution, meaning as $k, h \downarrow 0$. Naturally, due to the presence of the term k in (43), we expect $\mathcal{T}_{k,h}$ -preconditioned CG to require more iterations than the $\mathcal{S}_{k,h}$ -preconditioned CG; however, the space-time coarsening is significantly more aggressive than space-only coarsening, so each application of $\mathcal{T}_{k,h}$ will be more cost-efficient than

$\mathcal{S}_{k,h}$. We contrast the two approaches in Section 4. In addition, the numerical results in the Sections 4.2.3 and 4.2.4 strongly suggest the following two conjectures. The first conjecture refers to the cG(1)dG(1) discretization, while the second is related to the Crank-Nicolson discretization.

Conjecture 3.3. *For $s = r = 1$, there exists a constant $C_{\text{ST}} > 0$ independent of k, h so that the space-time two-grid preconditioner satisfies*

$$d_{\sigma}(\mathcal{G}_{k,h}, \mathcal{T}_{k,h}) \leq C\beta^{-1}(k^2 + h^2). \quad (46)$$

We should note that the k^2 -term in (46) is not a simple consequence of using a second-order-in-time method; such approximation is not automatically guaranteed by using other discretizations that are second-order convergent in time, as shown in Conjecture 3.4. Also, it is precisely the k^2 -term in (46) that makes the coarsening in space and time strategy more efficient than coarsening just in space.

Conjecture 3.4. *For the reduced Hessian obtained using the Crank-Nicolson discretization, there exists a constant $C > 0$ independent of k, h so that the space-time two-grid preconditioner satisfies*

$$d_{\sigma}(\tilde{\mathcal{G}}_{k,h}, \tilde{\mathcal{T}}_{k,h}) \leq C\beta^{-1}(k^{\frac{3}{2}} + h^2), \quad (47)$$

where $\tilde{\mathcal{T}}_{k,h}$ is the two-grid space-time preconditioner analogous to $\mathcal{T}_{k,h}$.

We should remark that the suboptimal approximation rate in time of $O(k^{\frac{3}{2}})$ in (47) is encountered in a slightly different context as well, namely when replacing $\tilde{\mathcal{K}}_{k,h}^*$ in (37) with the Crank-Nicolson discretization of \mathcal{K}^* . The resulting Hessian-like operator actually appears as a Schur-complement in the KKT system resulted from the first-optimize-then-discretize strategy, as in [24]. We will address this preconditioning avenue in future work.

3.2. Two-grid analysis for certain cG(s)dG(r) discretizations

Throughout this section we consider the cG(s)dG(r) discretized version of the optimal control problem with functions that are continuous piecewise linear in space ($s = 1$). The technique of proving Theorems 3.1 and 3.2 is, to some extent, similar to the case of elliptic control, in that it relies on error estimates like the ones in Corollaries 2.5 and 2.6.

3.2.1. Interpolants and their properties

We first introduce three space-time interpolants that will play essential roles in the analysis. We begin with interpolation in spatial direction. Consider the standard Lagrange finite element interpolant $I_h : H_0^1(\Omega) \cap H^2(\Omega) \rightarrow V_0^h$ (e.g., see [10]), which satisfies

$$\|y - I_h y\| \leq Ch^2 \|\nabla^2 y\|, \quad (48)$$

with C independent of h . Define the space of piecewise polynomial functions of degree $\leq r$ in $H_0^1(\Omega) \cap H^2(\Omega)$, namely

$$\tilde{\mathcal{Y}}_k^r = \{v \in L^2(I, H_0^1(\Omega) \cap H^2(\Omega)) : v|_{I_m} \in \mathcal{P}_r(I_m, H_0^1(\Omega) \cap H^2(\Omega))\} \subset \mathcal{Y}_k^r.$$

We extend the definition of I_h to $\tilde{\mathcal{Y}}_k^r$ in a pointwise (in time) manner:

$$Q_h^r : \tilde{\mathcal{Y}}_k^r \rightarrow \mathcal{Y}_{k,h}^{r,1}, \quad (Q_h^r y)(\cdot, t) \stackrel{\text{def}}{=} I_h(y(\cdot, t)). \quad (49)$$

After squaring (48) and integrating in time we obtain its space-time analogue

$$\|y - Q_h^r y\|_I \leq Ch^2 \|\nabla^2 y\|_I, \quad \forall y \in \tilde{\mathcal{Y}}_k^r. \quad (50)$$

Next we turn to interpolation in the time direction, as a first step towards space and time interpolation. This is applied to functions that are piecewise constant in time. We start with a one-dimensional result. Define $\tilde{P}_k \in \mathfrak{L}(L^2(I))$ as the projection onto the piecewise constant functions, which is given by

$$\tilde{P}_k v|_{I_m} \stackrel{\text{def}}{=} k_m^{-1} \int_{I_m} v(t) dt. \quad (51)$$

Lemma 3.5. *For $v \in H^1(I)$ the projection \tilde{P}_k satisfies the estimate*

$$\|\tilde{P}_k v - v\|_{L^2(I)} \leq k \|v'\|_{L^2(I)}. \quad (52)$$

Proof. Assume v is continuously differentiable on I_m and let $v_m = k_m^{-1} \int_{I_m} v(t) dt$. Then there exists $c_m \in I_m$ so that $v(c_m) = v_m$, and

$$v(t) - v_m = v(t) - v(c_m) = \int_{c_m}^t v'(s) ds.$$

Therefore

$$\begin{aligned} \int_{I_m} |v(t) - v_m|^2 dt &= \int_{t_{m-1}}^{t_m} \left(\int_{c_m}^t v'(s) ds \right)^2 dt \leq \int_{t_{m-1}}^{t_m} |t - c_m| \left(\int_{c_m}^t |v'(s)|^2 ds \right) dt \\ &\leq \int_{t_{m-1}}^{t_m} k_m \left(\int_{t_{m-1}}^{t_m} |v'(s)|^2 ds \right) dt = k_m^2 \int_{I_m} |v'(s)|^2 ds \end{aligned}$$

The inequality extends to $H^1(I_m)$ by density of $C^1(I_m)$ in $H^1(I_m)$. Now

$$\|\tilde{P}_k v - v\|_{L^2(I)}^2 = \sum_{m=1}^M \int_{I_m} |v(t) - v_m|^2 dt \leq \sum_{m=1}^M k_m^2 \int_{I_m} |v'(s)|^2 ds \leq k^2 \|v'\|_{L^2(I)}^2,$$

which proves (52). □

The interpolant \tilde{P}_k extends to functions in \mathcal{Y} by projecting along time. Define

$$P_k : \mathcal{Y} \rightarrow \tilde{\mathcal{Y}}_k^0, \quad P_k y(x)|_{I_m} \stackrel{\text{def}}{=} k_m^{-1} \int_{I_m} y(x, t) dt. \quad (53)$$

The time-integral in (53) is finite for a.e. $x \in \Omega$, giving rise to a piecewise constant function (with respect to time) in $L^2(I, H_0^1(\Omega) \cap H^2(\Omega))$.

Lemma 3.6. *The operator P_k satisfies the approximation and stability estimates*

$$\|y - P_k y\|_I \leq k \|\partial_t y\|_I, \quad (54)$$

$$\|\nabla^2 P_k y\|_I \leq \|\nabla^2 y\|_I. \quad (55)$$

Proof. From Lemma 3.5 we have

$$\int_I |y(x, t) - P_k y(x, t)|^2 dt \leq k^2 \int_I |\partial_t y(x, t)|^2 dt \quad \text{a.e. } x \in \Omega.$$

The estimate (54) follows from integrating above over $x \in \Omega$. Since P_k is the L^2 -projection onto the space of piecewise constant functions, we have

$$\int_I |P_k y(x, t)|^2 dt \leq \int_I |y(x, t)|^2 dt \quad \text{a.e. } x \in \Omega.$$

After integrating above over $x \in \Omega$ we obtain

$$\|P_k y\|_I^2 \leq \|y\|_I^2. \quad (56)$$

The inequality (55) now follows from (56) and the commutation $P_k(\nabla^2 y) = \nabla^2(P_k y)$ which holds for functions $y \in L^2(I, H_0^1(\Omega) \cap H^2(\Omega))$. \square

We now introduce a space-time state interpolant $\mathcal{I}_{k,h} : \mathcal{Y} \rightarrow \mathcal{Y}_{k,h}^{0,1}$ which is obtained by first interpolating in time, then in space:

$$\mathcal{I}_{k,h} y \stackrel{\text{def}}{=} Q_h^0(P_k y). \quad (57)$$

Lemma 3.7. *The interpolant $\mathcal{I}_{k,h}$ satisfies the approximation property*

$$\|y - \mathcal{I}_{k,h} y\|_I \leq k \|\partial_t y\|_I + Ch^2 \|\nabla^2 y\|_I, \quad \forall y \in \mathcal{Y}, \quad (58)$$

with C being the constant in (48), therefore independent of k, h .

Proof. If $y \in \mathcal{Y}$ we have

$$\begin{aligned} \|y - \mathcal{I}_{k,h} y\|_I &\leq \|y - P_k y\|_I + \|P_k y - \mathcal{I}_{k,h} y\|_I \stackrel{(54)}{\leq} k \|\partial_t y\|_I + \|P_k y - Q_h^0(P_k y)\|_I \\ &\stackrel{(50)}{\leq} k \|\partial_t y\|_I + Ch^2 \|\nabla^2(P_k y)\|_I \end{aligned}$$

and the estimate (58) now follows from (55). \square

3.2.2. Space-time coarsening

In this section we consider $r = 0$ and $s = 1$. The following proposition is the key argument for the proof of Theorem 3.1.

Proposition 3.8. *The fully discrete solution operators satisfy the approximation property*

$$\|(\mathcal{K}_{k,h} - \mathcal{K}_{2k,2h}\pi_{2k,2h})u\|_I \leq C(k+h^2)\|u\|_I \quad (59)$$

with C independent of k, h .

Proof. We have

$$\begin{aligned} & \|(\mathcal{K}_{k,h} - \mathcal{K}_{2k,2h}\pi_{2k,2h})u\|_I \leq \\ & \leq \|(\mathcal{K}_{k,h} - \mathcal{K})u\|_I + \|\mathcal{K}(\mathcal{I} - \pi_{2k,2h})u\|_I + \|(\mathcal{K} - \mathcal{K}_{2k,2h})\pi_{2k,2h}u\|_I. \end{aligned}$$

The first and third terms on the right-hand side in the estimate above are bounded from above by $C(k+h^2)\|u\|_I$ due to (30). To estimate the middle term we first note that for $v \in \mathcal{U}$ we have $z = \mathcal{K}^*v \in \mathcal{Y}$, cf. Proposition 2.2. By (58) we obtain

$$\|z - \mathcal{I}_{2k,2h}z\|_I \leq 2k\|\partial_t z\|_I + 4Ch^2\|\nabla^2 z\|_I \stackrel{(10)}{\leq} C(k+h^2)\|v\|_I, \quad (60)$$

with $C > 0$ a generic constant. Then

$$\begin{aligned} \|\mathcal{K}(\mathcal{I} - \pi_{2k,2h})u\|_I &= \sup_{v \in \mathcal{U} \setminus \{0\}} \frac{(\mathcal{K}(\mathcal{I} - \pi_{2k,2h})u, v)_I}{\|v\|_I} = \sup_{v \in \mathcal{U} \setminus \{0\}} \frac{((\mathcal{I} - \pi_{2k,2h})u, \overbrace{\mathcal{K}^*v}^z)_I}{\|v\|_I} \\ &= \sup_{v \in \mathcal{U} \setminus \{0\}} \frac{((\mathcal{I} - \pi_{2k,2h})u, z - \mathcal{I}_{2k,2h}z)_I}{\|v\|_I} \\ &\stackrel{(60)}{\leq} C(k+h^2)\|(\mathcal{I} - \pi_{2k,2h})u\|_I \leq C(k+h^2)\|u\|_I, \end{aligned}$$

where we used the orthogonality $(\mathcal{I} - \pi_{2k,2h})u \perp_I \mathcal{I}_{2k,2h}z$. \square

We should note that the argument in the proof above expressed precisely the fact that $(\mathcal{I} - \pi_{2k,2h})$ is a smoother with optimal quality. We are now in the position to prove Theorem 3.1.

Proof. The first goal is to prove (42). A simple verification shows that

$$\mathcal{G}_{k,h} - \mathcal{T}_{k,h} = \mathcal{K}_{k,h}^* \mathcal{K}_{k,h} - \mathcal{E}^{2k,2h} \mathcal{K}_{2k,2h}^* \mathcal{K}_{2k,2h} \pi_{2k,2h}. \quad (61)$$

Hence,

$$\begin{aligned} & |((\mathcal{G}_{k,h} - \mathcal{T}_{k,h})u, u)_I| \\ &= \left| \|\mathcal{K}_{k,h}u\|_I^2 - \|\mathcal{K}_{2k,2h}\pi_{2k,2h}u\|_I^2 \right| \\ &= \left| \|\mathcal{K}_{k,h}u\|_I - \|\mathcal{K}_{2k,2h}\pi_{2k,2h}u\|_I \right| \cdot (\|\mathcal{K}_{k,h}u\|_I + \|\mathcal{K}_{2k,2h}\pi_{2k,2h}u\|_I) \\ &\stackrel{(31)}{\leq} C\|(\mathcal{K}_{k,h} - \mathcal{K}_{2k,2h}\pi_{2k,2h})u\|_I \cdot \|u\|_I \stackrel{(59)}{\leq} C(k+h^2)\|u\|_I^2. \end{aligned} \quad (62)$$

We denote the last constant in (62) by C_{ST} ; it is independent of k and h , but it depends on the many constants involved in the prior estimates. The estimate (42) now follows from the symmetry of the operators $\mathcal{G}_{k,h}$ and $\mathcal{T}_{k,h}$.

The inequality (43) follows from (42), as shown in the proof of Lemma 1 in [4]. For completeness we give the short argument. We use the fact that for $|x - 1| \leq 1/2$ we have $|\ln x| \leq 2|x - 1|$ (this inequality is not sharp, a fact that is not essential). Let $u \in \mathcal{U}$ be arbitrary. Then for $x = (\mathcal{T}_{k,h}u, u)_I / (\mathcal{G}_{k,h}u, u)_I$ we have

$$|x - 1| = \left| \frac{((\mathcal{T}_{k,h} - \mathcal{G}_{k,h})u, u)_I}{(\mathcal{G}_{k,h}u, u)_I} \right| \leq \beta^{-1} \|\mathcal{T}_{k,h} - \mathcal{G}_{k,h}\|_I \leq \beta^{-1} C_{\text{ST}}(k + h^2) \leq \frac{1}{2},$$

where we used that $(\mathcal{G}_{k,h}u, u) \geq \beta\|u\|^2$. Hence

$$\left| \ln \frac{(\mathcal{T}_{k,h}u, u)_I}{(\mathcal{G}_{k,h}u, u)_I} \right| = |\ln x| \leq 2|x - 1| \leq 2\beta^{-1} C_{\text{ST}}(k + h^2),$$

which proves (43). \square

3.2.3. Space-only coarsening

In this section we assume $r \geq 0$ is fixed and $s = 1$. We begin by proving the analogue of (59).

Proposition 3.9. *The fully discrete solution operators satisfy the two-level approximation property*

$$\|(\mathcal{K}_{k,h} - \mathcal{K}_{k,2h}\pi_{k,2h})u\|_I \leq Ch^2\|u\|_I \quad (63)$$

with C independent of k, h .

Proof. The proof is similar to that of proof of Proposition 3.8, except for we use as common comparison term the solution of the semidiscrete equation instead of the continuous equation. More precisely, for $u \in \mathcal{U}_{k,h}$ we have

$$\begin{aligned} \|(\mathcal{K}_{k,h} - \mathcal{K}_{k,2h}\pi_{k,2h})u\|_I &\leq \\ &\leq \|(\mathcal{K}_{k,h} - \mathcal{K}_k)u\|_I + \|\mathcal{K}_k(\mathcal{I} - \pi_{k,2h})u\|_I + \|(\mathcal{K}_k - \mathcal{K}_{k,2h})\pi_{k,2h}u\|_I. \end{aligned}$$

The first and third terms on the right-hand side in the estimate above are bounded from above by $Ch^2\|u\|_I$ due to (32). In order to estimate the middle term we first note that for $v \in \mathcal{U}$ we have $z_k = \mathcal{K}_k^*v \in \mathcal{Y}_k^r$, cf. Proposition 2.3. By (50) we obtain

$$\|z_k - Q_{2h}^r z_k\|_I \leq 4Ch^2 \|\nabla^2 z_k\|_I \stackrel{(19)}{\leq} Ch^2 \|v\|_I, \quad (64)$$

with $C > 0$ being a generic constant independent of k and h . Then

$$\begin{aligned} \|\mathcal{K}_k(\mathcal{I} - \pi_{k,2h})u\|_I &= \sup_{v \in \mathcal{U} \setminus \{0\}} \frac{(\mathcal{K}_k(\mathcal{I} - \pi_{k,2h})u, v)_I}{\|v\|_I} = \sup_{v \in \mathcal{U} \setminus \{0\}} \frac{((\mathcal{I} - \pi_{k,2h})u, \overbrace{\mathcal{K}_k^*v}^{z_k})_I}{\|v\|_I} \\ &= \sup_{v \in \mathcal{U} \setminus \{0\}} \frac{((\mathcal{I} - \pi_{k,2h})u, z_k - Q_{2h}^r z_k)_I}{\|v\|_I} \\ &\stackrel{(64)}{\leq} Ch^2 \|(\mathcal{I} - \pi_{k,2h})u\|_I \leq Ch^2 \|u\|_I, \end{aligned}$$

where we used the orthogonality $(\mathcal{I} - \pi_{k,2h})u \perp_I Q_{2h}^r z_k$. □

We return to the proof of Theorem 3.2.

Proof. As in the proof of Theorem 3.1, it is verified that

$$\mathcal{G}_{k,h} - \mathcal{S}_{k,h} = \mathcal{K}_{k,h}^* \mathcal{K}_{k,h} - \mathcal{E}^{k,2h} \mathcal{K}_{k,2h}^* \mathcal{K}_{k,2h} \pi_{k,2h}. \quad (65)$$

Therefore

$$\begin{aligned} & |((\mathcal{G}_{k,h} - \mathcal{S}_{k,h})u, u)_I| \\ &= | \|\mathcal{K}_{k,h}u\|_I - \|\mathcal{K}_{k,2h}\pi_{k,2h}u\|_I | \cdot (\|\mathcal{K}_{k,h}u\|_I + \|\mathcal{K}_{k,2h}\pi_{k,2h}u\|_I) \\ &\stackrel{(31)}{\leq} C \|(\mathcal{K}_{k,h} - \mathcal{K}_{k,2h}\pi_{k,2h})u\|_I \cdot \|u\|_I \stackrel{(63)}{\leq} Ch^2 \|u\|_I^2. \end{aligned} \quad (66)$$

We denote the last constant in (66) by C_{SO} . The rest of the proof is similar to the one of Theorem 3.1. □

3.3. The multigrid preconditioner

The transition from two-grid to multigrid has been streamlined in [11], and is similar for space-time and space-only coarsening. Two elements are key in transitioning to multigrid: the preconditioner must have a W-cycle structure, and the coarsest grid has to be sufficiently fine, i.e., may not be the coarsest grid available. In describing the transition from two- to multigrid we follow the presentation from [4]. In the interest of self-containedness, but also to show certain differences between the two preconditioners, we show the construction and analysis of their multigrid counterparts.

3.3.1. Construction of the multigrid preconditioner

The construction below is valid for all the discretizations under scrutiny, with spaces and transition operators adapted accordingly. We adopt the algebraic multigrid notation where numbering of the levels begins at the finest. The finest space is $\mathcal{U}_0 = \mathcal{U}_{k,h}$ for both space-time and space-only coarsening. If we use $\ell > 1$ levels, we will assume, for simplicity, that the finest mesh, which is associated with the finite element space V^h , has been obtained by uniform mesh refinement from the coarsest grid with mesh size $h_{\ell-1}$, leading to a hierarchy of grids with mesh-sizes $h_j = 2^j h$; time-step refinement is also assumed to be uniform, namely it is given by $k_j = 2^j k$ with $j = 0, \dots, \ell - 1$. For space-time coarsening the j^{th} control space is \mathcal{U}_{k_j, h_j} with state space \mathcal{Y}_{k_j, h_j} ; the corresponding reduced Hessian \mathcal{G}_{k_j, h_j} lies in $\mathfrak{L}(\mathcal{U}_{k_j, h_j})$. For space-only coarsening the j^{th} control and state spaces are \mathcal{U}_{k_0, h_j} and \mathcal{Y}_{k_0, h_j} , respectively, with reduced Hessian \mathcal{G}_{k_0, h_j} lying in $\mathfrak{L}(\mathcal{U}_{k_0, h_j})$.

To describe the multigrid preconditioner, let \mathcal{U}_j be either \mathcal{U}_{k_j, h_j} or \mathcal{U}_{k_0, h_j} , and \mathcal{G}_j be the corresponding reduced Hessian. Note that $\mathcal{U}_j \subset \mathcal{U}_{j-1}$ for $j = 1, \dots, \ell - 1$. Let $\pi_j : \mathcal{U}_{j-1} \rightarrow \mathcal{U}_j$ be the L^2 -projection, and $\mathcal{E}^j : \mathcal{U}_{j+1} \rightarrow \mathcal{U}_j$ the embedding. We construct two sequences of operators $\mathcal{V}_j, \mathcal{W}_j \in \mathfrak{L}(\mathcal{U}_j)$, both of which are approximating \mathcal{G}_j^{-1} . The preconditioner of interest for the operator under scrutiny \mathcal{G}_0 will be \mathcal{W}_0 . The operators are defined in Algorithm 1.

Several remarks are in order. First, note that the inversion at the coarsest level (line 2) does not have to be exact. In practice we use iterative methods to perform

Algorithm 1 The multigrid preconditioner – operator version

```

1:  $j \leftarrow \ell - 1$ 
2:  $\mathcal{W}_{\ell-1} = \mathcal{G}_{\ell-1}^{-1}$ 
3: while  $j > 0$  do
4:    $j \leftarrow j - 1$ 
5:    $\mathcal{V}_j = \mathcal{E}^j \mathcal{W}_{j+1} \pi_{j+1} + \beta^{-1} (I - \mathcal{E}^j \pi_{j+1})$ 
6:   if  $j > 0$  then
7:      $\mathcal{W}_j = 2\mathcal{V}_j - \mathcal{V}_j \mathcal{G}_j \mathcal{V}_j$ 
8:   else
9:      $\mathcal{W}_j = \mathcal{V}_j$ 
10:  end if
11: end while
12: return  $\mathcal{W}_0$ 

```

inexact solves. Second, the definition of \mathcal{V}_j (line 5) is so that

$$\mathcal{V}_j^{-1} = \mathcal{E}^j \mathcal{W}_{j+1}^{-1} \pi_{j+1} + \beta (I - \mathcal{E}^j \pi_{j+1}).$$

Therefore, if $\ell = 2$, then $\mathcal{W}_0^{-1} = \mathcal{V}_0^{-1}$ coincides with the two-level preconditioner. Third, at the intermediate levels $1 \leq j \leq \ell - 2$ (if any) we perform one Newton step (line 7) for the operator equation $\mathcal{X}^{-1} - \mathcal{G}_j = 0$ to replace the preconditioner \mathcal{V}_j with \mathcal{W}_j . Finally, none of the operators $\mathcal{V}_j, \mathcal{W}_j$ are built in practice, the algorithm is implemented matrix free, as shown in [4].

3.3.2. Analysis

The analysis is based on Theorem 2 in [4] which we include in this paper as Theorem 3.10 for convenience. The theorem makes reference to the two-grid operators discussed in Section 3.1, which we denote here generically by

$$\mathcal{T}_j = \mathcal{E}^j \mathcal{G}_{j+1} \pi_{j+1} + \beta (I - \mathcal{E}^j \pi_{j+1}). \quad (67)$$

Theorem 3.10. *Using the notation from Algorithm 1, assume there are $\mathfrak{K} > 0$ and $0 < \mathfrak{f} \leq 1$ so that*

$$d_\sigma(\mathcal{T}_j, \mathcal{G}_j) = d_j \leq \mathfrak{K} \mathfrak{f}^{\ell-1-j}, \quad j = 0, 1, \dots, \ell - 1, \quad (68)$$

with $\mathfrak{K} \leq \min(0.1, \mathfrak{f}/8)$. Then

$$d_\sigma(\mathcal{W}_0, \mathcal{G}_0^{-1}) \leq \left(\frac{1}{4} + \mathfrak{K} \right) \mathfrak{f}^{\ell-1}. \quad (69)$$

The factor \mathfrak{f} is related to the approximation rates of the two-grid preconditioners, as shown below where we apply Theorem 3.10 to the two cases studied here.

First we focus on the space-time coarsening case for $r = 0$ and $s = 1$. Let $\mathcal{W}_{k,h}^{\text{ST}}$ be the multigrid preconditioner (i.e., \mathcal{W}_0) associated with the space-time coarsening strategy. Then $\mathcal{T}_j = \mathcal{T}_{k_j, h_j}$. Cf. Theorem 3.1 we have

$$d_\sigma(\mathcal{G}_{k_j, h_j}, \mathcal{T}_{k_j, h_j}) \leq d_j \stackrel{\text{def}}{=} 2C_{\text{ST}} \beta^{-1} (k_j + h_j^2), \quad (70)$$

provided $C_{\text{ST}}(k_j + h_j^2) \leq \beta/2$. Note that $k_j = (1/2)^{\ell-1-j}k_{\ell-1}$, $h_j = (1/2)^{\ell-1-j}h_{\ell-1}$. Therefore,

$$d_j = (1/2)^{\ell-1-j} 2C_{\text{ST}}\beta^{-1}(k_{\ell-1} + (1/2)^{\ell-1-j}h_{\ell-1}^2) \leq \mathfrak{K}_{\text{ST}}(1/2)^{\ell-1-j}, \quad (71)$$

with

$$\mathfrak{K}_{\text{ST}} \stackrel{\text{def}}{=} 2C_{\text{ST}}\beta^{-1}(k_{\ell-1} + h_{\ell-1}^2). \quad (72)$$

We can now apply Theorem 3.10 with $\mathfrak{f} = 1/2$ to get the following.

Corollary 3.11. *Assume $r = 0$ and $s = 1$. If $\mathfrak{K}_{\text{ST}} < 1/16$, i.e.,*

$$(k_{\ell-1} + h_{\ell-1}^2) \leq \beta C_{\text{ST}}^{-1}/32, \quad (73)$$

then

$$d_{\sigma}(\mathcal{W}_{k,h}^{\text{ST}}, \mathcal{G}_{k,h}^{-1}) \leq \left(\frac{1}{4} + \mathfrak{K}_{\text{ST}}\right) \left(\frac{1}{2}\right)^{\ell-1}. \quad (74)$$

Condition (73) states that the coarsest-level mesh-size and time-step have to be sufficiently small relative to β . If (73) is satisfied, then the quality of the multigrid preconditioner increases with the number of levels, since the preconditioners $\mathcal{W}_{k,h}^{\text{ST}}$ approximates $\mathcal{G}_{k,h}^{-1}$ increasingly well.

We turn our attention to the case of space-only coarsening for $r \geq 0$ and $s = 1$. Denote by $\mathcal{W}_{k,h}^{\text{SO}}$ the multigrid preconditioner (i.e., \mathcal{W}_0) associated with the space-only coarsening strategy. Then $\mathcal{T}_j = \mathcal{T}_{k_0,h_j}$. Cf. Theorem 3.2 we have

$$d_{\sigma}(\mathcal{G}_{k_0,h_j}, \mathcal{T}_{k_0,h_j}) \leq \hat{d}_j \stackrel{\text{def}}{=} 2C_{\text{SO}}\beta^{-1}h_j^2, \quad (75)$$

provided $C_{\text{SO}}h_j^2 \leq \beta/2$. Again,

$$\hat{d}_j = (1/4)^{\ell-1-j} 2C_{\text{SO}}\beta^{-1}h_{\ell-1}^2 \leq \mathfrak{K}_{\text{SO}}(1/4)^{\ell-1-j}, \quad (76)$$

with

$$\mathfrak{K}_{\text{SO}} \stackrel{\text{def}}{=} 2C_{\text{SO}}\beta^{-1}h_{\ell-1}^2. \quad (77)$$

We can now apply Theorem 3.10 with $\mathfrak{f} = 1/4$ to get the following.

Corollary 3.12. *Assume $r \geq 0$ and $s = 1$. If $\mathfrak{K}_{\text{SO}} < 1/16$, i.e.,*

$$h_{\ell-1}^2 \leq \beta C_{\text{SO}}^{-1}/32, \quad (78)$$

then

$$d_{\sigma}(\mathcal{W}_{k,h}^{\text{SO}}, \mathcal{G}_{k_0,h}^{-1}) \leq \left(\frac{1}{4} + \mathfrak{K}_{\text{SO}}\right) \left(\frac{1}{4}\right)^{\ell-1}. \quad (79)$$

As with space-time coarsening, condition (78) restricts the base-level mesh size, but much less than condition (73) restricts the size of the coarsest time step. Also, the rate of $(1/4)^{\ell-1}$ shows that we expect the space-only strategy to require approximately half the number of iterations needed when using the space-time coarsening. However, in terms of complexity, space-time coarsening is more advantageous than space-only, as we see next.

Finally we should remark that, if Conjecture 3.3 is true, then the space-time coarsening strategy for the cG(1)dG(1) discretization gives rise to a multigrid preconditioner with $\mathfrak{f} = 1/4$, and we expect a behavior as in Corollary 3.12. Similarly, if Conjecture 3.4 holds, then the space-time coarsening strategy for the Crank-Nicolson discretization gives rise to a multigrid preconditioner with $\mathfrak{f} = 1/2\sqrt{2}$.

3.3.3. A formula for complexity

To compute the complexity of applying the multigrid preconditioners we follow the strategy in [11]. Let $P_j = \dim(\mathcal{U}_j)$. We assume that at the intermediate levels the cost of projecting, embedding, and other algebraic manipulations of vectors – all amounting to a small multiple of P_j – is negligible compared to that of applying the Hessian, which amounts to two PDE-solves. Recall that at the finest level we do not apply the Hessian inside the preconditioner. We also assume that at the coarsest level we solve the system using unpreconditioned conjugate gradient, which requires a number of iterations N_{cg} that is bounded with respect to the discretization (verified in practice), but dependent on β and tolerances. Furthermore, the cost $G(j)$ of applying the Hessian at level j satisfies $G(j+1) \approx \mathfrak{g}G(j)$, $j = 0, 1, \dots, \ell-2$, with $0 < \mathfrak{g} < 1/2$. Hence, due to the W-cycle structure of the preconditioner, we have the following recursion:

$$W(j) = \begin{cases} (O(P_{\ell-1}) + G(\ell-1))N_{\text{cg}} & \approx G(\ell-1)N_{\text{cg}} & \text{if } j = \ell-1, \\ O(P_j) + G(j) + 2W(j+1) & \approx G(j) + 2W(j+1) & \text{if } 0 < j < \ell-1, \\ O(P_0) + W(1) & & \text{if } j = 0. \end{cases}$$

We used $x \approx y$ to denote the existence of two constants $c_1 < 1 < c_2$ independent of ℓ and relatively close to 1 so that $c_1 y \leq x \leq c_2 y$. The solution of the above recursion is

$$W(0) \approx G(0) \cdot \mathfrak{g} \left(\frac{1 - (2\mathfrak{g})^{\ell-2}}{1 - (2\mathfrak{g})} + (2\mathfrak{g})^{\ell-2} N_{\text{cg}} \right) \stackrel{\text{def}}{=} \mathfrak{g}_{\text{eff}} G(0). \quad (80)$$

Furthermore, let $\widehat{N}_j = \dim(V^{h_j})$ and $M_j = T/k_j$ be the number of time steps at level j . We assume the parabolic equation is solved iteratively (see next section) with an efficient classical multigrid solver at each time-step; hence, the cost of a parabolic solve at level $j = 0$ is $G(0) \approx C_{\text{hess}} M_0 \widehat{N}_0 \approx C_{\text{hess}} P_0$. Hence we assume there exists a constant independent of the discretization parameters (otherwise potentially large) so that $G(j) \approx C_{\text{hess}} P_j$.

We apply the above complexity estimation to the two scenarios. For space-time coarsening with d spatial dimensions we have $P_j \approx M_j \widehat{N}_j \approx 2^{d+1} M_{j+1} \widehat{N}_{j+1} \approx 2^{d+1} P_{j+1}$. For space-only coarsening we have $P_j \approx M_0 \widehat{N}_j \approx 2^d M_0 \widehat{N}_{j+1} \approx 2^d P_{j+1}$. Hence we can use in (80) the following factors:

$$\mathfrak{g}^{\text{ST}} = 2^{-(d+1)}, \quad \mathfrak{g}^{\text{SO}} = 2^{-d}. \quad (81)$$

4. Implementation and numerical results

In order for the preconditioner to behave as predicted, it is essential to preserve the problem formulation as described, i.e., not replace mass matrices with spectrally equivalent diagonal matrices, projections with restrictions, etc. To facilitate reproducibility of our results we present in Section 4.1 more details on the discrete formulation and implementation. Then we show two kinds of numerical results. In Section 4.2 we verify directly the convergence rates shown or conjectures in Section 3. Since this requires the construction of the space-time matrices \mathbf{G} and \mathbf{T} described in Section 4.1, these experiments are restricted to one spatial dimension. The second set of numerical results are given in Section 4.3, and they show the results of actual optimization solves in two spatial dimensions. The numerical results in Section 4.4 show a comparison and a discussion with other type of KKT-based preconditioning techniques.

4.1. Implementation for the $cG(1)dG(0)$ case

We restrict the discussion of the implementation to the two-level case for the $cG(1)dG(0)$ preconditioner. Let φ_i , $i = 1, \dots, N$ be the nodal basis functions of the finest space V_0^h , with φ_i , $i = N + 1, \dots, \hat{N}$ being the boundary basis functions in V^h . Denote by $\mathbf{A} \in \mathcal{M}_N$ the associated spatial stiffness matrix $\mathbf{A}_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$, $1 \leq i, j \leq N$. Let $\mathbf{M}^u \in \mathcal{M}_{\hat{N}}$ be the control mass matrix, i.e., $\mathbf{M}^u_{ij} = (\varphi_j, \varphi_i)$, $1 \leq i, j \leq \hat{N}$, and

$$\mathbf{M}^{yu} = \mathbf{M}^u(1 : N, :) \in \mathcal{M}_{N \times \hat{N}}, \quad \mathbf{M}^y = \mathbf{M}^u(1 : N, 1 : N) \in \mathcal{M}_{N \times N}$$

be the control-to-state mass matrix, and the state mass matrix, respectively. Let ψ_m , $m = 1, \dots, M$ be the basis functions for the time-domain, that is ψ_m is the indicator function of I_m . The nodal basis functions for $\mathcal{Y}_{k,h}$ are $\Phi_{m,i} = \psi_m \otimes \varphi_i$, $1 \leq m \leq M$, $1 \leq i \leq N$, and the basis functions for $\mathcal{U}_{k,h}$ are $\Phi_{m,i} = \psi_m \otimes \varphi_i$, $1 \leq m \leq M$, $1 \leq i \leq \hat{N}$. If the vector $\mathbf{y}^m \in \mathbb{R}^N$ represents $y_m = y|_{I_m}$, i.e.,

$$y = \sum_{m=1}^M y_m = \sum_{m=1}^M \sum_{i=1}^N \mathbf{y}_i^m \Phi_{m,i}, \quad (82)$$

then we represent the state y by the vector $\mathbf{y}^T = [(\mathbf{y}^1)^T, \dots, (\mathbf{y}^M)^T] \in \mathbb{R}^{MN}$. Similarly, we represent the control as $\mathbf{u}^T = [(\mathbf{u}^1)^T, \dots, (\mathbf{u}^M)^T] \in \mathbb{R}^{M\hat{N}}$. The fully discrete $cG(1)dG(0)$ matrix form of (24) and solution operator are given by

$$\hat{\mathbf{A}}\mathbf{y} = \hat{\mathbf{M}}^{yu}\mathbf{u}, \quad \mathbf{K} \stackrel{\text{def}}{=} (\hat{\mathbf{A}})^{-1}\hat{\mathbf{M}}^{yu}. \quad (83)$$

where

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{M}^y + k_1\mathbf{A} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ -\mathbf{M}^y & \mathbf{M}^y + k_2\mathbf{A} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & -\mathbf{M}^y & \mathbf{M}^y + k_3\mathbf{A} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{M}^y + k_n\mathbf{A} \end{bmatrix} \quad (84)$$

and $\hat{\mathbf{M}}^{\mathbf{y}\mathbf{u}}$ is a block diagonal (non-square) matrix with m^{th} diagonal block equal to $k_m \mathbf{M}^{\mathbf{y}\mathbf{u}}$. For the discrete optimal control problem we also need the state mass matrix $\hat{\mathbf{M}}^{\mathbf{y}}$ and control mass matrix $\hat{\mathbf{M}}^{\mathbf{u}}$ in space-time, defined to be a block diagonal (square) matrix with m^{th} diagonal block equal to $k_m \mathbf{M}^{\mathbf{y}}$, and $k_m \mathbf{M}^{\mathbf{u}}$, respectively. Note that if y is expressed as (82), then $\|y\|_I^2 = \mathbf{y}^T \hat{\mathbf{M}}^{\mathbf{y}} \mathbf{y}$; similarly, if \mathbf{u} represents a control $u \in \mathcal{U}_{k,h}$, then $\|u\|_I^2 = \mathbf{u}^T \hat{\mathbf{M}}^{\mathbf{u}} \mathbf{u}$. In matrix form the discrete reduced optimization problem (25) reads:

$$\min_{\mathbf{u} \in \mathbb{R}^{M\hat{N}}} \hat{J}(\mathbf{u}) = \frac{1}{2} (\mathbf{K}\mathbf{u} - \mathbf{y}_d)^T \hat{\mathbf{M}}^{\mathbf{y}} (\mathbf{K}\mathbf{u} - \mathbf{y}_d) + \frac{\beta}{2} \mathbf{u}^T \mathbf{M}^{\mathbf{u}} \mathbf{u}. \quad (85)$$

The familiar form of the linear system equation defining the solution of (85) is

$$\mathbf{H}\mathbf{u} \stackrel{\text{def}}{=} (\mathbf{K}^T \hat{\mathbf{M}}^{\mathbf{y}} \mathbf{K} + \beta \hat{\mathbf{M}}^{\mathbf{u}}) \mathbf{u} = \mathbf{K}^T \hat{\mathbf{M}}^{\mathbf{y}\mathbf{u}} \mathbf{y}_d. \quad (86)$$

Please note that the Hessian-vector multiplication is performed by simply solving the two parabolic equations – forward and adjoint – using appropriate methods (see (91) below), and is the single most expensive operation in the solution process. Note that the preconditioners **are not designed** for the matrix \mathbf{H} , but rather for

$$\mathbf{G} = (\hat{\mathbf{M}}^{\mathbf{u}})^{-1} \mathbf{H} = (\mathbf{M}^{\mathbf{u}})^{-1} \mathbf{K}^T \hat{\mathbf{M}}^{\mathbf{y}} \mathbf{K} + \beta \mathbf{I} = \mathbf{K}^* \mathbf{K} + \beta \mathbf{I}, \quad (87)$$

where $\mathbf{K}^* = (\hat{\mathbf{M}}^{\mathbf{u}})^{-1} \mathbf{K}^T \hat{\mathbf{M}}^{\mathbf{y}}$. The reason for the notation lies in the fact that \mathbf{K}^* is the matrix representation in the nodal basis of $\mathcal{K}_{k,h}^*$, because for any $y \in \mathcal{Y}$ with vector representation \mathbf{y} , and $u \in \mathcal{U}$ with vector representation \mathbf{u} we have

$$\mathbf{y}^T \hat{\mathbf{M}}^{\mathbf{y}} \mathbf{K} \mathbf{u} = (\mathcal{K}_{k,h} u, y)_I = (u, \mathcal{K}_{k,h}^* y)_I = (\mathbf{K}^* \mathbf{y})^T \hat{\mathbf{M}}^{\mathbf{u}} \mathbf{u} = \mathbf{y}^T (\mathbf{K}^*)^T \hat{\mathbf{M}}^{\mathbf{u}} \mathbf{u}. \quad (88)$$

So \mathbf{K}^* is the adjoint of \mathbf{K} with respect to the $L^2(Q)$ -induced inner-product on $\mathbb{R}^{M\hat{N}}$.

We turn to the matrix form of the two-grid preconditioner. For either space-time or space-only coarsening assume that the coarse level matrices $\hat{\mathbf{A}}_c, \hat{\mathbf{M}}_c^{\mathbf{y}}, \hat{\mathbf{M}}_c^{\mathbf{y}\mathbf{u}}, \hat{\mathbf{M}}_c^{\mathbf{u}}, \mathbf{K}_c$ have been built together with the coarse-to-fine embedding matrix \mathbf{E} . The fine-to-coarse projection matrix (representing either $\pi_{2k,2h}$ or $\pi_{k,2h}$) is given by

$$\mathbf{\Pi} = \mathbf{E}^* = (\hat{\mathbf{M}}_c^{\mathbf{u}})^{-1} \mathbf{E}^T \hat{\mathbf{M}}^{\mathbf{u}}. \quad (89)$$

Therefore, the two-grid matrix preconditioner for \mathbf{G} , as the matrix representation of $\mathcal{T}_{k,h}$ or $\mathcal{S}_{k,h}$ is

$$\mathbf{T} = \mathbf{E} \mathbf{G}_c \mathbf{\Pi} + \beta (\mathbf{I} - \mathbf{E} \mathbf{\Pi}) = \mathbf{E} ((\hat{\mathbf{M}}_c^{\mathbf{u}})^{-1} \mathbf{K}_c^T \hat{\mathbf{M}}_c^{\mathbf{y}} \mathbf{K}_c + \beta \mathbf{I}_c) \mathbf{\Pi} + \beta (\mathbf{I} - \mathbf{E} \mathbf{\Pi}). \quad (90)$$

We should make a remark with respect to symmetry, which plays a critical role in the choice of a Krylov solver. The matrix \mathbf{H} in (86) is obviously symmetric, in the sense that $\mathbf{H} = \mathbf{H}^T$. However, the matrix \mathbf{G} in (87), representing the reduced Hessian, is symmetric with respect to the $L^2(Q)$ -induced inner product, i.e., it satisfies $\mathbf{G}^T \hat{\mathbf{M}}^{\mathbf{u}} = \hat{\mathbf{M}}^{\mathbf{u}} \mathbf{G}$, so it is not expected to be symmetric in a standard sense. Therefore, when using an off-the-shelf preconditioned CG implementation, one has to solve (86) instead of (87), case in which the preconditioner for \mathbf{H} is $\hat{\mathbf{M}}^{\mathbf{u}} \mathbf{T}$, not just \mathbf{T} from (90).

Finally, we note that most of the solution process of (86) is normally performed matrix free. This means that only the matrices $\mathbf{A}, \mathbf{M}^y, \mathbf{M}^{y\mathbf{u}}, \mathbf{M}^{\mathbf{u}}$ involving spatial finite elements are built, and all the internal solves are approximative. This includes solving the PDE (applying \mathbf{K}), which involves inverting the lower block-triangular matrix $\hat{\mathbf{A}}$. Therefore, the action $\mathbf{y} = \mathbf{K}\mathbf{u}$ is replaced by the block forward substitution

$$(\mathbf{M}^y + k_m \mathbf{A})\mathbf{y}^m = \mathbf{M}^y \mathbf{y}^{m-1} + k_m \mathbf{M}^{y\mathbf{u}} \mathbf{u}^m, \quad m = 1, \dots, M, \quad (91)$$

where we have set $\mathbf{y}^0 = \mathbf{0}$. A similar process – based on block backward substitution – is used to compute the application of $\mathbf{z} = \mathbf{K}^* \mathbf{v}$. The systems in (91) can be solved inexactly, or using direct methods, depending on resources. For applying the space-time projection $\mathbf{\Pi}$ on the coarser grid we should note that, due to the fact that controls are discontinuous in time, the projection is localized in time, hence only the spatial mass matrix $\mathbf{M}^{\mathbf{u}}$ needs to be inverted. As shown in [38, 4], mass matrices can be inverted easily using simple iterative methods.

4.2. Direct numerical verification of the two-grid approximation

In order to verify directly the approximation rates in Theorems 3.1 and 3.2 as well as Conjectures 3.3 and 3.4 we consider the one-dimensional case ($d = 1$) with $\Omega = [0, 1]$, $\beta = 1$; varying β does not alter the powers of k and h significantly, as shown in similar studies [1]. We show approximation results for the two-grid preconditioner using space-time and space-only coarsening for the cG(1)dG(0) discretization, and for space-time coarsening using cG(1)dG(1) finite elements; the latter results are particularly relevant, since they confirm Conjecture 3.3; we also show results for the space-time preconditioner associated with the Crank-Nicolson discretization to confirm Conjecture 3.4. For all these cases we consider a hierarchy of spaces $\mathcal{U}_0 \supset \mathcal{U}_1 \supset \dots \supset \mathcal{U}_\ell$, and we actually build the reduced Hessian matrices \mathbf{G}_j and the two-grid preconditioners \mathbf{T}_j , as shown in Section 4.1, as well as their cG(1)dG(1) and Crank-Nicolson-based counterparts. With the matrices being constructed, the spectral distances estimated in (43), (45), (46), and (47) can be computed directly using the formula

$$d_j = \{\max |\ln \lambda| : \lambda \in \sigma(\mathbf{G}_j, \mathbf{T}_j)\}. \quad (92)$$

where $\sigma(\mathbf{A}, \mathbf{B})$ is the set of generalized eigenvalues of \mathbf{A} and \mathbf{B} . Then we study the change of d_j by varying mesh sizes and time steps.

4.2.1. Space-time coarsening for cG(1)dG(0)

We consider the case $T = 1$ for this section. The easiest way to verify the power of k in (43) is to examine cases where h is relatively small, so that $h^2 \ll k$. We take $k_0 = 1/16$, $h_0 = 1/128$, and $k_j = 2^{-j} k_0$ for $j = 0, \dots, 4$, and we consider $\mathcal{U}_j = \mathcal{U}_{k_j, h_0}$. The results of computing d_j , reported in Table 1, suggest that the ratios d_j/d_{j+1} approach 2 as k_j gets smaller (as long as $h_0^2 \ll k_j$), showing consistency with the power of k being 1 in (43).

Verifying that the power of h is 2 in (43) is more challenging, since it requires matrix-based computations with $k \ll h^2$, which proved to be impractical due to memory constraints. Instead, we produced 49 datapoints reported in Table 2, meaning $d_{kh} = d_\sigma(\mathcal{G}_{k,h}, \mathcal{T}_{k,h})$ with $k = 2^{-s}$, $h = 2^{-t}$, $3 \leq s \leq 12$, $3 \leq t \leq 9$, and $s + t \leq 15$; the last constraint is due both to memory limitations and execution time. At first

Table 1. Spectral distance for $d_j = d_\sigma(\mathcal{G}_{k_j, h_0}, \mathcal{T}_{k_j, h_0})$ with $h_0 = 1/128$. Parameters are $T = 1, \beta = 1$.

k^{-1}	1/16	1/32	1/64	1/128	1/256
d_j	1.6296e-02	9.4806e-03	5.2892e-03	2.8258e-04	1.4677e-04
d_j/d_{j+1}	1.7189	1.7924	1.8717	1.9253	–

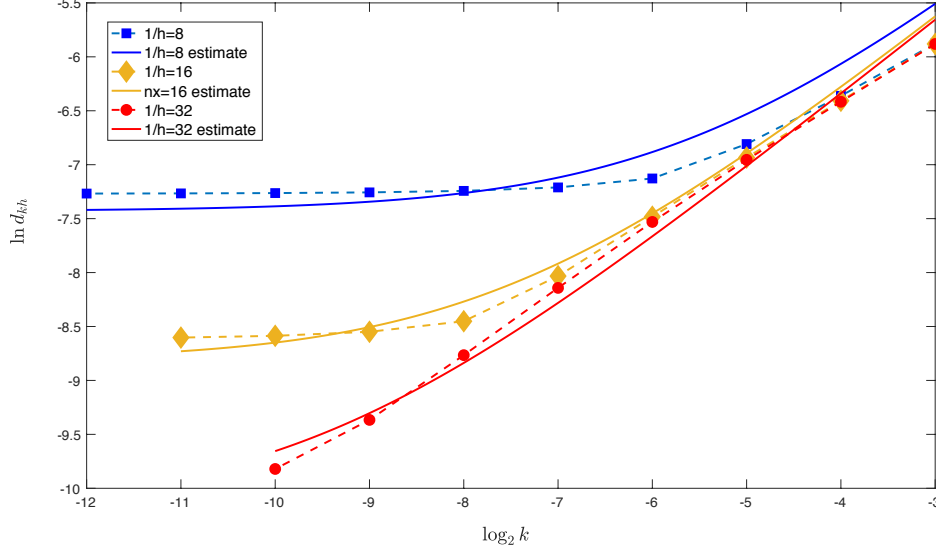


Figure 1. Comparison of actual $\ln d_{kh}$ with $\ln(c_1 k + c_2 h^2)$.

glance, the numbers in Table 2 do not appear to bring new information. However, we performed a nonlinear least-squares matching using the data in Table 2 between $\ln d_{kh}$ and $\ln(c_1 k + c_2 h^\gamma)$, thus treating the power γ as unknown. The optimal result is $c_1 \approx 0.0277$, $c_2 \approx 0.0368$, $\gamma \approx 1.9864$, which strongly supports the optimality of the estimate (43).

In addition, we performed a non-linear least squares matching between $\ln d_{kh}$ and $\ln(c_1 k + c_2 h^2)$, and we obtain the values $c_1 \approx 0.0277$ and $c_2 \approx 0.0379$. In Figure 1 we show how the resulting estimate fits with the numbers in the first three columns in Table 2. Note that we used the entire data in Table 2 to find the best fit for c_1, c_2 , not just the first three columns. The graphs in Figure 1 suggest that the match is quite reasonable.

Table 2. Spectral distance $d_{kh} = d_\sigma(\mathcal{G}_{k,h}, \mathcal{T}_{k,h})$ for $k = 2^{-s}, h = 2^{-t}, 3 \leq s \leq 12, 3 \leq t \leq 9$, and $s + t \leq 15$. Parameters are $T = 1, \beta = 1$.

$k^{-1} \setminus h^{-1}$	8	16	32	64	128	256	512
8	2.80e-03	2.79e-03	2.79e-03	2.79e-03	2.79e-03	2.79e-03	2.79e-03
16	1.73e-03	1.65e-03	1.63e-03	1.63e-03	1.63e-03	1.63e-03	1.63e-03
32	1.10e-03	9.75e-04	9.54e-04	9.49e-04	9.48e-04	9.48e-04	9.48e-04
64	8.02e-04	5.63e-04	5.36e-04	5.30e-04	5.29e-04	5.29e-04	5.28e-04
128	7.39e-04	3.25e-04	2.91e-04	2.84e-04	2.83e-04	2.82e-04	
256	7.15e-04	2.14e-04	1.56e-04	1.48e-04	1.47e-04		
512	7.05e-04	1.94e-04	8.56e-05	7.69e-05			
1024	7.01e-04	1.87e-04	5.43e-05				
2048	6.99e-04	1.84e-04					
4096	6.98e-04						

4.2.2. Space-only coarsening for $cG(1)dG(0)$

Verifying the estimate (45) for the case of space-only coarsening is less challenging, since only one parameter is involved. However, here we are also interested in the robustness of the constant in front of h^2 with respect to k and even to the end time T , the latter not being addressed by the theory. In Table 3 we compute $d_{kh} = d_\sigma(\mathcal{G}_{k,h}, \mathcal{T}_{k,h})$ for $k = 2^{-s}$, $h = 2^{-t}$ with $3 \leq s \leq 5$, $3 \leq t \leq 8$, $T = 1, 2, 4$, and $\beta = 1$, and we report the ratios d_{kh}/h^2 . Note that number of time steps for each computation is T/k ; hence, for $T = 2$, $k^{-1} = 32$ indicates 64 time steps.

For each T the ratios under scrutiny increase very slowly with $k \downarrow 0$, supporting the fact that C_{SO} in (45) is uniform with respect to k . Moreover, the increase in $\max_{kh} d_{kh}/h^2$ is very small with respect to T , from ≈ 0.0459 at $T = 1$ to ≈ 0.0534 for $T = 4$, suggesting the growth is at most logarithmic with T .

Table 3. Spectral distance $d_{kh} = d_\sigma(\mathcal{G}_{k,h}, \mathcal{T}_{k,h})$ and ratios d_{kh}/h^2 for $k = 2^{-s}$, $h = 2^{-t}$, $3 \leq s \leq 5$, $3 \leq t \leq 8$, $T = 1, 2, 4$, $\beta = 1$.

$k^{-1} \setminus h^{-1}$	8	16	32	64	128	256
$T = 1$						
8	6.4816e-04 0.041482	1.6818e-04 0.043053	4.2430e-05 0.043448	1.0632e-05 0.043547	2.6594e-06 0.043572	6.6495e-07 0.043578
16	6.6863e-04 0.042792	1.7355e-04 0.044429	4.3789e-05 0.04484	1.0972e-05 0.044943	2.7447e-06 0.044969	6.8627e-07 0.044976
32	6.8156e-04 0.04362	1.7695e-04 0.045299	4.4649e-05 0.045721	1.1188e-05 0.045826	2.7986e-06 0.045853	6.9976e-07 0.045859
$T = 2$						
8	7.4481e-04 0.047668	1.9419e-04 0.049712	4.9053e-05 0.05023	1.2295e-05 0.05036	3.0757e-06 0.050392	7.6905e-07 0.0504
16	7.5468e-04 0.0483	1.9682e-04 0.050385	4.9720e-05 0.050913	1.2462e-05 0.051046	3.1176e-06 0.051079	7.7953e-07 0.051088
32	7.6021e-04 0.048653	1.9829e-04 0.050762	5.0094e-05 0.051296	1.2556e-05 0.05143	3.1411e-06 0.051464	7.8539e-07 0.051472
$T = 4$						
8	7.8109e-04 0.04999	2.0402e-04 0.052229	5.1561e-05 0.052798	1.2925e-05 0.052941	3.2335e-06 0.052977	8.0850e-07 0.052986
16	7.8440e-04 0.050202	2.0491e-04 0.052457	5.1787e-05 0.053029	1.2982e-05 0.053173	3.2476e-06 0.053209	8.1205e-07 0.053218
32	7.8615e-04 0.050313	2.0538e-04 0.052576	5.1906e-05 0.053151	1.3012e-05 0.053295	3.2551e-06 0.053332	8.1391e-07 0.053341

4.2.3. Space-time coarsening for $cG(1)dG(1)$

In this section we repeat the procedure from Section 4.2.1 that allows us to “guess” the correct powers in Conjecture 3.3. Again, we compute $d_{kh} = d_\sigma(\mathcal{G}_{k,h}, \mathcal{T}_{k,h})$ for $k = 2^{-s}$, $h = 2^{-t}$, $4 \leq s, t \leq 9$, and $s + t \leq 14$. The results are included in Table 5. To assess the decrease rate with respect to k , we first fix $h = 1/128$, and we compute the ratios $d_{kh}/d_{\frac{k}{2}h}$ in Table 4. The results suggest that for sufficiently small h we have $d_{kh} = O(k^2)$ (compared with $d_{kh} = O(k)$ in the $cG(1)dG(0)$ case). Furthermore, we perform a nonlinear least squares minimization using the entire data in Table 5 to fit the expression $\ln(c_1 k^\delta + c_2 h^\gamma)$ with $\ln d_{kh}$, with c_1, c_2, δ, γ treated as unknowns. The optimal parameters are $c_1 \approx 0.2413$, $\delta \approx 2.0042$, $c_2 \approx 0.06388$, $\gamma \approx 2.1603$. These results strongly suggest that

$$d_{kh} = O(k^2 + h^2), \quad (93)$$

a result that is optimal with respect to both parameters. (There is no reason to believe that the power of h would be greater than 2.)

Table 4. Spectral distance between space-time preconditioner and Hessian using cG(1)dG(1) elements; $d_j = d_\sigma(\mathcal{G}_{k_j, h}, \mathcal{T}_{k_j, h})$ with $h = 1/128$, $k_j = 2^{-4-j}$, $j = 0, 1, 2, 3$.

k^{-1}	1/16	1/32	1/64	1/128
d_j	9.3497e-04	2.6418e-04	6.8177e-05	1.7179e-05
d_j/d_{j+1}	3.5391	3.8750	3.9685	–

Table 5. Spectral distance between space-time preconditioner and Hessian using cG(1)dG(1) elements; $d_{kh} = d_\sigma(\mathcal{G}_{k, h}, \mathcal{T}_{k, h})$ for the cG(1)dG(0) discretization $k = 2^{-s}$, $h = 2^{-t}$, $4 \leq s, t \leq 9$, and $s + t \leq 14$.

$k^{-1} \setminus h^{-1}$	16	32	64	128	256	512
16	9.3510e-04	9.3499e-04	9.3497e-04	9.3497e-04	9.3497e-04	9.3496e-04
32	2.6423e-04	2.6419e-04	2.6419e-04	2.6418e-04	2.6418e-04	2.6418e-04
64	1.8107e-04	6.8178e-05	6.8177e-05	6.8177e-05	6.8176e-05	
128	1.8094e-04	4.5689e-05	1.7179e-05	1.7179e-05		
256	1.8093e-04	4.5660e-05	1.1448e-05			
512	1.8093e-04	4.5658e-05				

4.2.4. Space-time coarsening for Crank-Nicolson

We repeat the procedure from Sections 4.2.1 and 4.2.3 for the space-time preconditioner associated with the Crank-Nicolson discretization by computing $d_{kh} = d_\sigma(\tilde{\mathcal{G}}_{k, h}, \tilde{\mathcal{T}}_{k, h})$ for $k = 2^{-s}$, $h = 2^{-t}$, $3 \leq s \leq 11$, $3 \leq t \leq 9$, and $s + t \leq 15$; the results are reported in Table 4.2.4 with $\beta = 1$, $T = 1$.

Table 6. Spectral distance $d_{kh} = d_\sigma(\tilde{\mathcal{G}}_{k, h}, \tilde{\mathcal{T}}_{k, h})$ for the Crank-Nicolson discretized Hessian, $k = 2^{-s}$, $h = 2^{-t}$, $3 \leq s \leq 11$, $3 \leq t \leq 9$, and $s + t \leq 15$; parameters are $T = 1$, $\beta = 1$.

8	8	16	32	64	128	256	512
8	2.9618e-03	2.9672e-03	2.9689e-03	2.9694e-03	2.9695e-03	2.9695e-03	2.9695e-03
16	1.3825e-03	1.3668e-03	1.3644e-03	1.3638e-03	1.3637e-03	1.3636e-03	1.3636e-03
32	7.1331e-04	5.1207e-04	5.0501e-04	5.0356e-04	5.0321e-04	5.0313e-04	5.0311e-04
64	6.9811e-04	2.0185e-04	1.8014e-04	1.7811e-04	1.7766e-04	1.7755e-04	1.7752e-04
128	6.9690e-04	1.8178e-04	6.5641e-05	6.2169e-05	6.1633e-05	6.1511e-05	
256	6.9676e-04	1.8103e-04	4.6126e-05	2.2098e-05	2.1359e-05		
512	6.9674e-04	1.8095e-04	4.5709e-05	1.1720e-05			
1024	6.9674e-04	1.8094e-04	4.5664e-05				
2048	6.9674e-04	1.8093e-04					

In this case the best fit of the expression $\ln(c_1 k^\delta + c_2 h^\gamma)$ with $\ln d_{kh}$, with c_1, c_2, δ, γ treated as unknowns yields the optimal parameters $c_1 \approx 0.0675$, $\delta \approx 1.4608$, $c_2 \approx 0.0553$, $\gamma \approx 2.1402$. In addition, if we fix $\gamma = 2$ in the fitting process, then the optimal parameters are $c_1 \approx 0.0704$, $\delta \approx 1.4756$, $c_2 \approx 0.0384$. These results suggest

$$d_{kh} = O(k^{\frac{3}{2}} + h^2), \quad (94)$$

thus supporting Conjecture 3.4.

To conclude this section, we should add that in our parameter matching process we used higher accuracies for the actual numbers representing spectral distances than we are able to reproduce in the tables above.

4.3. Linear solves

The second set of experiments are actual solves in two spatial dimensions. The domain is $\Omega = [0, 1]^2$ and end time is $T = 2$. The desired state and optimal control are $y_d(x_1, x_2, t) = t \sin(\pi x_1) x_2^2 (1 - x_2)$, $u_d = (\partial_t - \Delta) y_d$, that is, $u_d = \sin(\pi x_1) x_2^2 (1 - x_2) + \pi^2 t \sin(\pi x_1) x_2^2 (1 - x_2) - t \sin(\pi x_1) (2 - 6x_2)$. A discretized y_d is introduced into the

optimal control problem, and u_d is only used for verification of convergence. We solve the normal equations (26) in three ways. **(i)** unpreconditioned CG; **(ii)** preconditioned CG with space-time multigrid preconditioning; **(iii)** preconditioned CG with space-only multigrid preconditioning. For preconditioned CG we use up to four levels. At the coarsest level we solve the system (invert the coarsest Hessian) using unpreconditioned CG. We solve the parabolic equation inside the Hessian $\mathcal{G}_{k,h}$ iteratively with direct, prefactored direct solves at each time step. For cG(1)dG(0), the process is described by (91). Mass-matrices are also inverted using direct solves. The tolerances are set at 10^{-8} for CG and 10^{-10} for the coarsest Hessian, as we are not focused in this work on the effects of inexact solves. The regularization parameter is $\beta = 10^{-5}$, a number chosen so that we can observe the transition of the space-time coarsening strategy starting as less efficient, then becoming more efficient as the number of time-steps increases. All computations were performed using MATLAB on a system with two eight-core 2.9 GHz Intel Xeon E5-2690 CPUs and 256 GB memory.

4.3.1. Linear solves for the cG(1)dG(0) discretization

We run the code with mesh sizes $h = 1/64, 1/128, 1/256$, and time steps $k = 2/128, 2/256, \dots, 2/4096$. Note that the dimension of the control space is $M \cdot (n+1)^2$, where $M = 2/k$ is the number of time steps and $n = 1/h$ is the partition size in each direction. The largest problem we were able to solve had $2048 \cdot 257^2 = 135,268,352$ control variables. We report in Table 7 the number of iterations and the wall-clock times in seconds. We first note that unpreconditioned CG converges in a number of iterations that is fairly stable for all the cases, namely between 75 and 79 iterations. At the same time, the number of iterations for the space-only multigrid preconditioned CG is remarkably stable, as predicted by the theory, decreasing from 4 iterations when $h = 1/64$ to 3 when $h = 1/128, 1/256$, independently of k . The dependence on k for the case of space-time preconditioning is also consistent with theory. The two-level preconditioner shows a remarkable drop in number of iterations as k decreases, from 29 at $k = 2/128$ down to 6 at $k = 2/4096$. However, the multigrid version of the space-time preconditioner only works when the coarsest level is sufficiently fine: the three-grid preconditioner becomes useful only when $k \leq 2/2048$, and the four-grid preconditioner works only when $k \leq 2/4096$. As with other instances of this multigrid strategy, the wall-clock savings only become effective at the higher-resolutions. For space-only preconditioning the four-grid preconditioned CG runs 7-9 times faster than unpreconditioned CG at $h = 1/256$. However, when the four-grid preconditioner becomes available for use within the space-time coarsening framework (due the the base case being sufficiently fine), it ultimately becomes more efficient than the space-only version, in spite of the higher number of iterations, as seen in the last row in Table 7.

It is worth exploring the value of the complexity estimates from Section 3.3.3 in light of our wall-clock measurements. We focus on the cases with $k = 2/4096$, where we were able to run both strategies with $\ell = 4$ levels. We adopt $N_{\text{cg}} = 80$. For the space-time coarsening case we obtain $\mathbf{g}_{\text{eff}}^{\text{ST}} \approx 0.78$, whereas for space-only coarsening $\mathbf{g}_{\text{eff}}^{\text{SO}} \approx 5.375$. Taking into account that one PCG iteration essentially requires one Hessian-vector multiplication and one application of the preconditioner, for $h_0 = 1/64$ with $\ell = 4$ levels, the ratio of the estimated costs is

$$\frac{6 \text{ (SO)PCG iterations}}{4 \text{ (ST)PCG iterations}} = \frac{6(1 + 0.78)}{4(1 + 5.375)} \approx 0.42,$$

which compares surprisingly well to the ratio of the actual wall-clock times of 0.48. However, the complexity estimates are only orientative, since they assume a linear growth with spatial problem size in the cost of solving the parabolic equation, which is not the case for our implementation due to the use of direct methods.

Table 7. Iteration counts for cG(1)dG(0) discretized unpreconditioned CG (1 level), space-time preconditioned CG (ST), and space-only preconditioned CG (SO). Parameters are $T = 2$, $\beta = 10^{-5}$. Wall-clock times are shown in seconds.

$M \setminus h^{-1}$	# lev.	64		128		256	
		MGST	MGSO	MGST	MGSO	MGST	MGSO
128	1	75 (156)		75 (788)		75 (5019)	
	2	29 (386)	4 (129)	29 (1591)	3 (544)	29 (9390)	3 (3210)
	3	> 50	4 (55)	> 50	3 (226)	> 50	3 (1119)
	4	> 50	4 (39)	> 50	3 (116)	> 50	3 (556)
256	1	76 (322)		77 (1523)		77 (8418)	
	2	19 (492)	4 (307)	19 (2652)	3 (1324)	19 (14175)	3 (6143)
	3	> 50	4 (146)	> 50	3 (484)	> 50	3 (2407)
	4	> 50	4 (85)	> 50	3 (255)	> 50	3 (1230)
512	1	78 (673)		78 (4162)		78 (18669)	
	2	13 (804)	4 (614)	13 (3741)	3 (2490)	13 (21596)	3 (13469)
	3	> 50	4 (244)	> 50	3 (1017)	> 50	3 (4824)
	4	> 50	4 (161)	> 50	3 (542)	> 50	3 (2582)
1024	1	77 (1252)		78 (6694)		78 (37631)	
	2	9 (1311)	4 (1274)	9 (6651)	3 (5381)	9 (36375)	3 (30257)
	3	> 50	4 (539)	> 50	3 (2154)	> 50	3 (10331)
	4	> 50	4 (333)	> 50	3 (1189)	> 50	3 (5507)
2048	1	78 (2624)		78 (13296)		78 (77848)	
	2	7 (2193)	4 (2588)	7 (10951)	3 (10674)	7 (68917)	3 (54434)
	3	8 (677)	4 (1169)	8 (3626)	3 (4430)	8 (18619)	3 (20142)
	4	> 50	4 (679)	> 50	3 (2324)	> 50	3 (11989)
4096	1	79 (4977)		78 (28361)		-	
	2	6 (3888)	4 (5519)	6 (19480)	3 (20246)	-	-
	3	6 (1193)	4 (2423)	6 (6232)	3 (9463)	-	-
	4	6 (715)	4 (1484)	6 (3442)	3 (4698)	-	-

4.3.2. Linear solves for the cG(1)dG(1) discretization

For the optimization runs in Table 8 we use the same parameters and domain as for the cG(1)dG(0) discretization. We remark that unpreconditioned CG solves the problem in a number of iterations ranging between 77 and 80, which is very similar to the cG(1)dG(0) discretization; this is a clear sign that the difficulty of the problem (measured as the number of relevant eigenvalues of the Hessian) has not changed with discretization. Due to the optimal rate in k of (93) we can showcase the benefits of coarsening in space and time using larger time steps and mesh-sizes, namely we show $h = 1/32, 1/64, 1/128$, and $k = 2/64, \dots, 2/2048$. The benefits of having k^2 in (93) are two-fold. First, the coarsest grid can be chosen to be coarser than for the equivalent cG(1)dG(0) case; for cG(1)dG(1) we can have the coarsest step size to be $2/1024 = 1/512$, while for cG(1)dG(0) it is $2/2048 = 1/1024$. Second, the number of multigrid iterations with space-time coarsening drops significantly faster with decreasing k , so that at resolution $k = 2/2048$ and $h = 1/128$ space-time coarsening and space-only coarsening both result in 3 multigrid iterations. We should point out that, even though the estimate (93) is balanced in k and h , the (theoretical) constant multiplying h^2 may be smaller than the constant multiplying k^2 , as shown in the direct spectral distance computations in Section 4.2.3. We also show the wall-clock times for each of the runs. As expected, for relatively large time-steps the space-only multigrid is more efficient; however, as the time steps continue to decrease, space-time coarsening is faster by more than a factor of two than space-only coarsening, and by more than a factor of

ten than unpreconditioned CG. The improvement over CG measured in iteration count is much stronger, naturally.

Table 8. Iteration counts for for cG(1)dG(1) discretized unpreconditioned CG (1 level), space-time preconditioned CG (ST), and space-only preconditioned CG (SO). Parameters are $T = 2$, $\beta = 10^{-5}$. Wall-clock times are shown in seconds.

$M \setminus h^{-1}$		32		64		128	
	# lev.	MGST	MGSO	MGST	MGSO	MGST	MGSO
64	1	77 (41)		77 (266)		77 (1690)	
	2	28 (114)	5 (50)	28 (630)	4 (208)	28 (3191)	3 (1172)
	3	> 30	5 (22)	> 30	4 (92)	> 30	3 (356)
	4	> 30	> 30	> 30	4 (54)	> 30	3 (197)
128	1	76 (87)		79 (602)		77 (2750)	
	2	18 (174)	5 (117)	18 (757)	4 (525)	17 (4562)	3 (2103)
	3	> 30	5 (44)	> 30	4 (209)	> 30	3 (752)
	4	> 30	> 30	> 30	4 (99)	> 30	3 (438)
256	1	77 (172)		78 (1201)		78 (5473)	
	2	11 (218)	5 (245)	11 (1098)	4 (895)	11 (6351)	3 (4773)
	3	> 30	5 (97)	> 30	4 (400)	> 30	3 (1563)
	4	> 30	> 30	> 30	4 (215)	> 30	3 (911)
512	1	78 (357)		77 (2202)		79 (11291)	
	2	7 (307)	5 (474)	7 (1510)	4 (2053)	7 (9148)	3 (9662)
	3	> 30	5 (195)	> 30	4 (889)	> 30	3 (3181)
	4	> 30	> 30	> 30	4 (432)	> 30	3 (1814)
1024	1	78 (842)		78 (4737)		80 (22606)	
	2	5 (545)	5 (987)	4 (2085)	4 (3775)	4 (12707)	3 (19455)
	3	6 (177)	5 (407)	5 (795)	4 (1745)	5 (3798)	3 (6410)
	4	> 30	> 30	> 30	4 (876)	> 30	3 (3619)
2048	1	77 (1583)		79 (9609)		78 (43804)	
	2	5 (1045)	5 (2034)	4 (3998)	4 (7923)	3 (21217)	3 (46351)
	3	5 (299)	5 (922)	4 (1279)	4 (3361)	3 (6065)	3 (13732)
	4	> 30	> 30	4 (819)	4 (1880)	3 (3859)	3 (7741)

4.4. Comparison with Crank-Nicolson discretized optimal control problem and all-at-once methods.

In this section we conduct a side-by-side comparison of our MGST and MGSO preconditioners for the cG(1)dG(1) and Crank-Nicolson-based discretized optimal control problem with an all-at-once preconditioning strategy developed in the recent work [24]; the latter refers to a Crank-Nicolson-based first-optimize-then-discretize strategy; according to its authors, it is one of the very few attempts at preconditioning second order-based discretized KKT system using the all-at-once approach. Hence, the common ground for the numerical results included in this section is that they all use second-order schemes in space and time, even though they lead ultimately to discrete systems are not equivalent. In addition, the comparison with the results in [24] is mostly based on number of iterations and trends, since we do not have access to the optimized codes that were used.

The example is extracted from Section 5.1 in [24], and includes a non-zero initial condition, so we refer back to the our formulation (1)-(2), with $\Omega = (-1, 1)^2$, $f = 0$, and end time $T = 2$; the desired state and initial value are

$$\begin{aligned}
 y_d(x, t) &= 1 + \left[\left(\frac{2}{\pi^2 \beta} + \frac{\pi^2}{2} \right) e^T + \left(1 - \frac{2}{(2 + \pi^2) \beta} - \frac{\pi^2}{2} \right) e^t \right] s(x), \\
 y_0(x) &= 1 + \left(\frac{2}{\pi^2 \beta} e^T - \frac{2}{(2 + \pi^2) \beta} \right) s(x),
 \end{aligned}$$

where $s(x) = s(x_1, x_2) = \cos\left(\frac{\pi x_1}{2}\right) \cos\left(\frac{\pi x_2}{2}\right)$. The analytic solution is given by

$$\begin{aligned} y(x, t) &= 1 + \left(\frac{2}{\pi^2 \beta} e^T - \frac{2}{(2 + \pi^2) \beta} e^t \right) s(x), \\ u(x, t) &= \beta^{-1} (e^T - e^t) s(x). \end{aligned}$$

As in [24], we consider uniform grids in space with $h = 2^{-j}$, $j = 2, 3, \dots, 7$; however, while in [24] the time step is $k = h$, we let $k = h/2$ in order for the relative errors in u to be in the same ballpark as the first-optimize-then-discretize example in [24]. At the same time, to achieve the asymptotic convergence rate of $O(k^2 + h^2)$ for cG(1)dG(1) (as shown in [28]), we set the tolerance to 10^{-10} , compared to the tolerance of 10^{-6} in [24]. Of course, in the all-at-once approach, the tolerance controls the residual of the adjoint equation as well as the state equation, whereas in our reduced approach it only forces convergence of the systems in the control variables.

For each of the aforementioned discretizations we solve the parabolic control problem for $\beta = 10^{-2}, 10^{-3}, 10^{-4}$, using both the MGST the MGSO preconditioning, as well as unpreconditioned CG, and we record the number of iterations and run-times in Table 9, the last column of which includes the number of preconditioned MINRES iterations from [24]. For both MGST and MGSO we use as base case for the spatial grid $h_{base} = 1/4$, while the coarsest time-grid for MGST varies by case, due to the limitations of the two-grid approximation. For example, for the MGST preconditioned Crank-Nicolson discretized problem with $\beta = 10^{-3}$ and $h = 1/32, k = 1/64$ we chose $k_{base} = 1/32$ as base case for the time-grid, with $h_{base} = 1/4$ for the coarsest spatial grid; the corresponding entry (7 iterations over 23 seconds) corresponds to the use of four grids, with the parameters given by $(k, h) \in \{(1/64, 1/32), (1/32, 1/16), (1/32, 1/8), (1/32, 1/4)\}$. So, for this particular case, space-time coarsening is performed only at the first coarsening step, followed by two more steps of space-only coarsening.

With respect to the cG(1)dG(1) and Crank-Nicolson solves, the findings in Table 9 are consistent with the ones in Sections 4.3.1 and 4.3.2: for each β , the number of iterations decreases with increasing resolution, and space-time coarsening becomes slightly more efficient in terms of runtime than space-only, provided the time step is sufficiently small, while space-only coarsening remains a safe choice. It is not surprising that for the same choice of (k, h) , Crank-Nicolson is faster than cG(1)dG(1), since it has about half as many degrees of freedom, but it lags behind in approximation.

We believe that a truthful comparison between reduced methods and all-at-once preconditioning methods of KKT systems deserves a separate study, perhaps co-authored by researchers from both camps. However, we wish to add a few remarks based on our limited experience and knowledge regarding the latter, and on the results from [24] reported in Table 9. In terms of number of iterations the preconditioned MINRES appears to be reasonably stable with respect to $\beta \downarrow 0$, and increasing very slowly with increasing resolution. In absolute terms, the number of MGST and MGST iterations is significantly smaller than for preconditioned MINRES, at least for higher resolutions. Of course, the cost of matrix-vector multiplications for reduced systems is much higher than for MINRES, where the matrix is sparse; however, the preconditioner in [24] also requires two imprecise parabolic solves, meaning block forward and backward substitutions, with the linear solves at each time-step being replaced by a number of AMG V-cycles, leaving the most expensive work to be performed inside the preconditioner. In lieu of a definite conclusion – which would require bipartisan efforts, we would say

that perhaps our method is expected to be more efficient than the all-at once KKT-based approaches when $h^2 \ll \beta$ (or $h^2 + k^2 \ll \beta$), while the latter would be preferable when the opposite holds.

Table 9. Iteration counts for cG(1)dG(1) vs. Crank-Nicolson vs. MINRES (MR) with $k = h/2$; wall-clock times (seconds) are in parantheses.

h^{-1}	cG(1)dG(1)				Crank-Nicolson				MR	
	$\ err_u\ $	MGST	MGSO	CG	$\ err_u\ $	MGST	MGSO	CG		
$\beta = 10^{-2}$										
4	1.16e-02	–	–	13 (0.06)	2.22e-02	–	–	13 (0.08)	17	
8	2.84e-03	7 (0.66)	5 (0.74)	14 (0.41)	9.20e-03	–	5 (0.83)	14 (0.26)	20	
16	7.11e-04	5 (2.68)	4 (3.77)	14 (5.42)	3.73e-03	6 (2.62)	4 (2.75)	15 (2.25)	20	
32	1.79e-04	4 (27)	3 (29)	14 (67)	1.44e-03	5 (11)	3 (12)	15 (15)	22	
64	4.49e-05	3 (241)	3 (352)	14 (1017)	5.36e-04	4 (72)	3 (106)	15 (172)	24	
128	1.12e-05	3 (2800)	3 (2968)	14 (9102)	1.95e-04	4 (935)	3 (818)	16 (2549)	n/a	
$\beta = 10^{-3}$										
4	3.01e-02	–	–	23 (0.11)	1.59e-02	–	–	25 (0.15)	18	
8	7.08e-03	–	6 (1.47)	24 (0.69)	5.84e-03	–	6 (2.03)	32 (0.56)	20	
16	1.60e-03	8 (6.60)	5 (6.91)	26 (7.95)	2.60e-03	–	5 (6.37)	34 (3.60)	21	
32	3.87e-04	6 (40)	4 (50)	26 (120)	1.15e-03	7 (23)	4 (23)	35 (28)	25	
64	9.64e-05	4 (313)	4 (364)	27 (1427)	4.74e-04	7 (150)	3 (120)	36 (415)	28	
128	2.41e-05	3 (2796)	3 (2845)	27 (16809)	1.82e-04	5 (1051)	5 (1672)	33 (6304)	n/a	
$\beta = 10^{-4}$										
4	5.21e-02	–	–	29 (0.11)	1.38e-02	–	–	37 (0.24)	16	
8	1.81e-02	–	9 (3.67)	41 (1.14)	3.88e-03	–	9 (7.72)	72 (1.27)	19	
16	4.44e-03	–	6 (13)	49 (16)	1.42e-03	–	6 (19)	80 (11)	22	
32	8.92e-04	10 (86)	5 (81)	55 (268)	6.89e-04	–	5 (61)	83 (76)	24	
64	1.91e-04	8 (733)	6 (554)	59 (3541)	3.40e-04	8 (321)	3 (245)	89 (1128)	28	
128	4.53e-05	5 (4465)	5 (4483)	60 (36572)	1.51e-04	6 (1645)	3 (1407)	89 (15322)	n/a	

5. Conclusions and future work

We have constructed and analyzed two multigrid preconditioners for the reduced Hessian arising in the space-time distributed control of a parabolic equation. As with other instances of the strategy, the solution process is increasingly efficient at higher resolution, resulting in a number of iterations that decreases with resolution, and in significant wall-clock savings. While the space-only coarsening strategy is a safe choice, space-time coarsening can be more efficient, especially for the cG(1)dG(1) discretization, where time steps only have to decrease proportionally with mesh size to achieve such a feat.

Several generalizations are possible and desirable. When adding control constraints and discretizing the controls using piecewise constants control in space and time, one expects an approximation rate of $O(k + h)$, as in the elliptic case [12] in connection to semismooth Newton methods or as in [13] for use with interior point methods. Changing the parabolic constraint to a semilinear one is expected to be a smooth transition from a computational standpoint, although more challenging from an analytical one. Proving Conjectures 3.3 and 3.4 (or variants) remains a high priority task; however, it requires a novel approach and a priori estimates in perhaps weaker norms. The question remains whether the space-time multigrid preconditioner shows an additional improvement when using higher order methods in time such as cG(r)dG(s), with $r \geq 2$. Parallelizability of the method is another critical aspect for the problem under scrutiny in this paper, but it needs to be considered in the wider context of parallelization of solvers for time-dependent PDEs.

Funding

This material is based upon work supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Award Number DE-SC0005455, and by the National Science Foundation under award DMS-1913201.

References

- [1] H. ANTIL, A. DRĂGĂNESCU, AND K. GREEN, *A note on multigrid preconditioning for fractional pde-constrained optimization problems*, Results in Applied Mathematics, 9 (2021), p. 100133.
- [2] M. ASCH, M. BOCQUET, AND M. NODET, *Data assimilation*, vol. 11 of Fundamentals of Algorithms, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016. Methods, algorithms, and applications.
- [3] A. K. AZIZ AND P. MONK, *Continuous finite elements in space and time for the heat equation*, Math. Comp., 52 (1989), pp. 255–274.
- [4] A. T. BARKER AND A. DRĂGĂNESCU, *Algebraic multigrid preconditioning of the Hessian in PDE-constrained optimization*, Numerical Linear Algebra with Applications, (2020).
- [5] A. BORZÌ, *Multigrid methods for parabolic distributed optimal control problems*, J. Comput. Appl. Math., 157 (2003), pp. 365–382.
- [6] A. BORZÌ, *Space-time multigrid methods for solving unsteady optimal control problems*, in Real-time PDE-constrained optimization, SIAM, 2007, pp. 97–113.
- [7] A. BORZÌ AND S. GONZÁLEZ ANDRADE, *Second-order approximation and fast multigrid solution of parabolic bilinear optimization problems*, Adv. Comput. Math., 41 (2015), pp. 457–488.
- [8] A. BORZÌ AND V. SCHULZ, *Computational optimization of systems governed by partial differential equations*, vol. 8 of Computational Science & Engineering, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2012.
- [9] J. BRANDMAN, H. DENLI, AND D. TRENEV, *Introduction to pde-constrained optimization in the oil and gas industry*, in Frontiers in PDE-Constrained Optimization, Springer, 2018, pp. 171–203.
- [10] S. C. BRENNER AND L. R. SCOTT, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.
- [11] A. DRĂGĂNESCU AND T. F. DUPONT, *Optimal order multilevel preconditioners for regularized ill-posed problems*, Math. Comp., 77 (2008), pp. 2001–2038.
- [12] A. DRĂGĂNESCU AND J. SARASWAT, *Optimal-order preconditioners for linear systems arising in the semismooth Newton solution of a class of control-constrained problems*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 1038–1070.
- [13] A. DRĂGĂNESCU AND C. PETRA, *Multigrid preconditioning of linear systems for interior point methods applied to a class of box-constrained optimal control problems*, SIAM J. Numer. Anal., 50 (2012), pp. 328–353.
- [14] A. DRĂGĂNESCU AND A. M. SOANE, *Multigrid solution of a distributed optimal control problem constrained by the Stokes equations*, Appl. Math. Comput., 219 (2013), pp. 5622–5634.
- [15] M. J. GANDER AND M. NEUMÜLLER, *Analysis of a new space-time parallel multigrid algorithm for parabolic problems*, SIAM J. Sci. Comput., 38 (2016),

- pp. A2173–A2208.
- [16] S. GONZÁLEZ ANDRADE AND A. BORZÌ, *Multigrid second-order accurate solution of parabolic control-constrained problems*, *Comput. Optim. Appl.*, 51 (2012), pp. 835–866.
 - [17] M. D. GUNZBURGER, *Perspectives in flow control and optimization*, vol. 5 of *Advances in Design and Control*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003.
 - [18] W. HACKBUSCH, *On the fast solving of parabolic boundary control problems*, *SIAM J. Control Optim.*, 17 (1979), pp. 231–244.
 - [19] ———, *Integral equations*, vol. 120 of *International Series of Numerical Mathematics*, Birkhäuser Verlag, Basel, 1995. Theory and numerical treatment, Translated and revised by the author from the 1989 German original.
 - [20] M. HANKE AND C. R. VOGEL, *Two-level preconditioners for regularized inverse problems. I. Theory*, *Numer. Math.*, 83 (1999), pp. 385–402.
 - [21] G. HORTON AND S. VANDEWALLE, *A space-time multigrid method for parabolic partial differential equations*, *SIAM J. Sci. Comput.*, 16 (1995), pp. 848–864.
 - [22] B. KALTENBACHER, *V-cycle convergence of some multigrid methods for ill-posed problems*, *Math. Comp.*, 72 (2003), pp. 1711–1730 (electronic).
 - [23] J. T. KING, *Multilevel algorithms for ill-posed problems*, *Numer. Math.*, 61 (1992), pp. 311–334.
 - [24] S. LEVEQUE AND J. W. PEARSON, *Fast iterative solver for the optimal control of time-dependent PDEs with Crank-Nicolson discretization in time*, arXiv preprint arXiv:2007.08410, (2020).
 - [25] B. LI, J. LIU, AND M. XIAO, *A new multigrid method for unconstrained parabolic optimal control problems*, *J. Comput. Appl. Math.*, 326 (2017), pp. 358–373.
 - [26] J. LIU AND M. XIAO, *A leapfrog multigrid algorithm for the optimal control of parabolic PDEs with Robin boundary conditions*, *J. Comput. Appl. Math.*, 307 (2016), pp. 216–234.
 - [27] ———, *A leapfrog semi-smooth Newton-multigrid method for semilinear parabolic optimal control problems*, *Comput. Optim. Appl.*, 63 (2016), pp. 69–95.
 - [28] D. MEIDNER AND B. VEXLER, *A priori error estimates for space-time finite element discretization of parabolic optimal control problems. I. Problems without control constraints*, *SIAM J. Control Optim.*, 47 (2008), pp. 1150–1177.
 - [29] J. W. PEARSON, M. STOLL, AND A. J. WATHEN, *Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems*, *SIAM J. Matrix Anal. Appl.*, 33 (2012), pp. 1126–1152.
 - [30] T. REES, H. S. DOLLAR, AND A. J. WATHEN, *Optimal solvers for PDE-constrained optimization*, *SIAM J. Sci. Comput.*, 32 (2010), pp. 271–298.
 - [31] A. RIEDER, *A wavelet multilevel method for ill-posed problems stabilized by Tikhonov regularization*, *Numer. Math.*, 75 (1997), pp. 501–522.
 - [32] Y. SAAD, *Iterative methods for sparse linear systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, second ed., 2003.
 - [33] A. M. SOANE AND A. DRĂGĂNESCU, *Multigrid preconditioners for the Newton-Krylov method in the optimal control of the stationary Navier-Stokes equations*, *SIAM J. Numer. Anal.*, 57 (2019), pp. 1494–1523.
 - [34] M. STOLL AND A. WATHEN, *All-at-once solution of time-dependent Stokes control*, *J. Comput. Phys.*, 232 (2013), pp. 498–515.
 - [35] V. THOMÉE, *Galerkin finite element methods for parabolic problems*, vol. 25 of *Springer Series in Computational Mathematics*, Springer-Verlag, Berlin, second ed., 2006.

- [36] F. TRÖLTZSCH, *Optimal control of partial differential equations*, vol. 112 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.
- [37] N. VON DANIELS, M. HINZE, AND M. VIERLING, *Crank-Nicolson time stepping and variational discretization of control-constrained parabolic optimal control problems*, SIAM J. Control Optim., 53 (2015), pp. 1182–1198.
- [38] A. WATHEN AND T. REES, *Chebyshev semi-iteration in preconditioning for problems including the mass matrix*, Electron. Trans. Numer. Anal., 34 (2008/09), pp. 125–135.