Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Computational and Applied Mathematics 173 (2005) 169–198
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

A time-domain decomposition iterative method for the solution of distributed linear quadratic optimal control problems

Matthias Heinkenschloss*,[1]

Department of Computational and Applied Mathematics, Rice University, MS 134, Houston, TX 77005-1892, USA

Received 23 October 2003; received in revised form 4 March 2004

Abstract

We study a class of time-domain decomposition-based methods for the numerical solution of large-scale linear quadratic optimal control problems. Our methods are based on a multiple shooting reformulation of the linear quadratic optimal control problem as a discrete-time optimal control (DTOC) problem. The optimality conditions for this DTOC problem lead to a linear block tridiagonal system. The diagonal blocks are invertible and are related to the original linear quadratic optimal control problem restricted to smaller time-subintervals. This motivates the application of block Gauss–Seidel (GS)-type methods for the solution of the block tridiagonal systems. Numerical experiments show that the spectral radii of the block GS iteration matrices are larger than one for typical applications, but that the eigenvalues of the iteration matrices decay to zero fast. Hence, while the GS method is not expected to convergence for typical applications, it can be effective as a preconditioner for Krylov-subspace methods. This is confirmed by our numerical tests.

A byproduct of this research is the insight that certain instantaneous control techniques can be viewed as the application of one step of the forward block GS method applied to the DTOC optimality system.

Keywords: Linear quadratic optimal control problems; Instantaneous control; Suboptimal control; Multiple shooting; Discrete-time optimal control problem; Gauss–Seidel method; Krylov subspace methods

1. Introduction

This paper is concerned with the numerical solution of linear quadratic optimal control problems governed by parabolic partial differential equations (PDEs). Such problems are difficult to solve in practice because of the large storage requirements arising out of the strong coupling in space

* Tel.: +7133485176; fax: +7133485318.
E-mail address: heinken@rice.edu (M. Heinkenschloss).
[1] Supported by NSF Grants DMS-0075731 and ACI-0121360.

0377-0427/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2004.03.005

and time of state (solution of the governing PDE), the so-called adjoint, and the control. Several techniques, including snapshot techniques for storage management and instantaneous control for the computation of suboptimal solutions, both of which will be discussed in more detail in Section 1.2, have been proposed to address this difficulty. The goal of this paper is to present another approach that is related to storage management and instantaneous control techniques.

Our approach is based on time-domain decomposition. More precisely, we will use a multiple shooting approach to equivalently reformulate our problem as a so-called discrete-time optimal control (DTOC) problem. For our target problems the operators/matrices arising in the multiple shooting reformulation of the problem cannot be computed explicitly; only operator/matrix-vector products can be formed. Therefore, existing multiple shooting implementations for systems of ODEs or 'small' PDEs (see, e.g., [1,6,8,9,17,28,29,32,36]), which rely on matrix factorizations are not applicable. Our approach is matrix free. It is based on the observation that the optimality conditions for DTOC problem are a block tridiagonal system. For each block operator/matrix-vector products can be formed. These operations are related to PDE and adjoint PDE solves on time-subdomains. Moreover, the diagonal blocks are invertible and the application of the inverse of a diagonal block is equivalent to solving a linear quadratic optimal control problem on a time-subdomain. This time-subdomain linear quadratic optimal control problem is essentially a smaller copy of the original linear quadratic optimal control problem. The block tridiagonal structure motivates the application of block Gauss–Seidel (GS) methods. Numerical experiments show that the spectral radii of the block GS iteration matrices are larger than one for typical applications, but that the eigenvalues of the iteration matrices decay to zero fast. Hence, while the GS method is not expected to converge for typical applications, we propose to use them as preconditioners in Krylov-subspace methods. We will demonstrate that this approach is effective for a class of problems. Moreover, we will show connections between the block GS iteration and existing approaches.

1.1. Problem formulation

To make the introductory discussion more concrete, we consider problems of the form

$$\min J(u) = \frac{1}{2}\int_0^T \|u(t)\|_U^2 \,\mathrm{d}t + \frac{\alpha_1}{2}\int_0^T \|C(t)y(t) - z_1(t)\|_Z^2 \,\mathrm{d}t + \frac{\alpha_2}{2}\|C_T y(T) - z_2\|_{Z_T}^2, \tag{1}$$

where $y = y(u)$ is the solution of a PDE, the state equation, which is abstractly written as

$$\frac{\partial}{\partial t} y(t) + A(t)y(t) = B(t)u(t) + f(t), \quad t \in (0,T), \tag{2a}$$

$$y(0) = y_0. \tag{2b}$$

The functions u and y are called control and state, respectively. Problem (1), (2) is posed in a Hilbert space setting which will be made precise in Section 2. One may view $A(t)$, $B(t)$, $C(t)$, C_T to be large sparse matrices, where the square matrix $A(t)$ is obtained from the spatial discretization of an elliptic PDE. We assume that the state equation (2) admits a unique solution y for every control u and that the optimal control problem (1), (2) has a solution. This is guaranteed under the conditions stated in Section 1.1. The solution of (1), (2) is characterized by the optimality conditions

which consist of the state equations (2), the adjoint equations

$$-\frac{\partial}{\partial t}\,p(t) + A(t)^* p(t) = \alpha_1 C(t)^*(C(t)y(t) - z_1(t)), \quad t \in (0,T), \tag{3a}$$

$$p(T) = \alpha_2 C_T^*(C_T y(T) - z_2) \tag{3b}$$

and the equation

$$u(t) + B(t)^* p(t) = 0, \quad t \in (0,T). \tag{4}$$

Here M^* denotes the adjoint of the operator M, which is the transpose if M is a real matrix. The left-hand side of (4) is the gradient of the objective function J. The optimality system (2), (3), (4) reveals an important structure. While algorithms for the solution of (2) often involve marching in time, starting from an initial condition, the optimization results in a much stronger coupling in time. State y information from all times feeds into the adjoint equation (3), which has to be solved backward in time. The adjoints p determine the controls u (see Eq. (4)), which feed back into the state equation (2). This coupling in time makes the practical solution of these very large-scale optimization problems challenging.

1.2. Background

Snapshot techniques, instantaneous control approaches, and reduced basis techniques are used to cope with the storage and computing time demands for solving distributed optimal control problems. We review the first two techniques only, because they are related to our approach.

The snapshot technique has been applied in [20] in the context of reverse mode automatic differentiation and in, e.g., [4,34] to implement gradient-based optimization methods for optimal control and data assimilation, respectively. Roughly speaking, snapshot techniques only keep a copy of the state at certain times $0 = T_0 < T_1 < \cdots < T_{N_t}$. During the adjoint p computation they recompute y for times $t \in (T_i, T_{i+1})$ by resolving (2) on (T_i, T_{i+1}). In [4,20,34] more refined storage techniques are used to balance storage requirements and the computing time for state recomputations. Like the snapshot technique, our approach only requires permanent storage for states y and adjoints p at times $0 = T_0 < T_1 < \cdots < T_{N_t} = T$. Storage for $y(t)$ and $p(t)$, $t \in (T_i, T_{i+1})$ needs to be provided temporarily for subproblem computations and this storage could be shared by different subproblem solvers. In [4,34] storage management techniques are used within gradient computations to cope with the large storage requirements; storage management techniques do not alter the gradient based optimization algorithms for the solution of (1). The time-domain decomposition approach discussed in this paper, effects the optimization algorithm and it may be viewed as an approach to integrate (simple) storage management techniques into the optimization, rather than using it only as a black-box for gradient $\nabla J(u)$ computations.

In their simplest form, instantaneous control strategies split the optimal control problem (1), (2) into a set of smaller problems of the same type. The smaller problems are obtained by restricting the original optimal control problem to time intervals (T_i, T_{i+1}), where $0 = T_0 < T_1 < \cdots < T_{N_t} = T$. These problems are then solved sequentially to obtain a 'suboptimal' control. (At this point we use 'suboptimal' loosely.) The computation of suboptimal controls proceeds as follows. Suppose suboptimal controls $\hat{u}_i$ and corresponding states $\hat{y}_i$ have been computed on the subintervals

(T_i, T_{i+1}), $i = 1, \ldots, j-1$. Then the optimal control problem (1), (2) is restricted to $[T_j, T_{j+1}]$ and the initial condition is replaced by $y(T_j) = \hat{y}_{j-1}(T_j)$. If $j < N_t - 1$, the last term in the objective function (1) is dropped. An optimization procedure is applied to compute an approximation of the optimal control $\hat{u}_j$ for this problem together with the corresponding state $\hat{y}_j$. The suboptimal control $\hat{u}$ for the original problem (1), (2) is defined by connecting the piecewise controls, $\hat{u}_{[T_i,T_{i+1}]} = \hat{u}_i$, $i = 0, \ldots, N_t$. The instantaneous control strategies found in the literature differ in the way the partition $0 = T_0 < T_1 < \cdots < T_{N_t} = T$ is chosen, in the objective function chosen for the subproblems, in the optimization method applied to the subproblems, and in the truncation criteria applied in these optimization methods. Typically, instantaneous control strategies use a moving horizon, i.e., the final time T is not fixed a priori, and they stop moving the horizon T if a certain objective is met at T. Furthermore, they often use model predictive control approaches, i.e., in the j step they optimize over a time interval larger than $[T_j, T_{j+1}]$, but advance control $\hat{u}_j$ and state $\hat{y}_j$ only on $[T_j, T_{j+1}]$. Finally, they may modify the objective function used for the optimization over $[T_j, T_{j+1}]$. For the control of Navier–Stokes flow and related problems, instantaneous control techniques have been used in, e.g., [7,12–15,21–24,27,31,37]. For example, in [14,15,22,23] the partition $0 = T_0 < T_1 < \cdots < T_N = T$ is identical to the time discretization used for the numerical time integration for (2a). In [7,12,13] the size $\Delta T = T_n - T_{n-1}$ of time intervals for the instantaneous control is bigger than the step sizes δt used in the numerical time integration. In all papers [7,12–15,22,23] the final time T is not fixed, but moved until an overall objective such as drag reduction at T is met. In most cases, theoretical investigations of the instantaneous control strategies are missing. The papers [21,23,24] provide some theoretical foundation of instantaneous control techniques. In [21,23], it is shown that, under certain assumptions, their instantaneous control technique can be interpreted as a suboptimal closed-loop controller and that the closed-loop dynamical system is stable under appropriate conditions. In particular, for this result to be valid, $T_j - T_{j-1}$ has to be sufficiently small. Their theory assumes $C = I$ and $C_T = 0$ in (1). In [24] an infinite time horizon control problem for Navier–Stokes flow with a tracking type objective function and distributed control is considered. It is shown that, under suitable assumptions, the error between the velocities corresponding to the suboptimal controls and the desired velocities decays exponentially in time. Again, one of the assumptions is that $T_j - T_{j-1}$ is sufficiently small. In [7,12,13], which investigate boundary control of turbulent flows using direct numerical simulation (DNS) and large eddy simulation (LES), respectively, the partitioning $0 = T_0 < T_1 < \cdots < T_{N_t} = T$ is independent of the time integration. In these papers it was found that larger $T_j - T_{j-1}$ led to stronger decreases in drag or turbulent kinetic energy at time T, the respective objective functions in the optimization. Thus, there is still a significant gap between theoretical results and numerical observation. Our present work originated from the desire to obtain a better understanding of instantaneous control techniques and to narrow the gap between theory and numerics. See Section 4.2.4 for further discussions.

Finally, we mention that for optimal control problems governed by the wave equation, a different time-domain decomposition method has been proposed in [26]. Since the wave equation is a second order in time equation the time-domain decomposition method in [26] uses transmission conditions at time-subintervals that involve linear combinations of state and adjoint information at T_i's as well as time derivatives of states and adjoints at T_i's. It is not clear whether the method in [26] can be extended to general first order in time problems studied in this paper. However, it is noteworthy that like the approach proposed in this paper, the iteration in [26] requires the solution of optimal control problems restricted to time subdomains. The paper [26] proves convergence in

the infinite-dimensional setting, but no convergence rates are established and no numerical results are given.

1.3. Notation

Throughout this paper, we use Hilbert space notation which is suitable for linear quadratic distributed optimal control problems. In particular, given Hilbert spaces X, Y, $\mathscr{L}(X,Y)$ denotes the space of bounded linear operators from X to Y and $\mathscr{L}(X) = \mathscr{L}(X,X)$. The norm in the Hilbert space X is denoted by $\|\cdot\|_X$. The dual of a Hilbert space X is denoted by X^*. The duality pairing between the Hilbert space X and its dual X^* is denoted by $\langle\cdot,\cdot\rangle_{X^*\times X}$. Given an operator $M \in \mathscr{L}(X,Y)$ its adjoint is denoted by M^*.

Readers not familiar with this abstract setting may view $A(t)$, $B(t)$, $C(t)$, C_T to be large sparse matrices, where the square matrix $A(t)$ is obtained from the spatial discretization of an elliptic PDE. In this case $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$ and $M \in \mathscr{L}(X,Y)$ is a matrix $M \in \mathbb{R}^{m\times n}$, and $M^* = M^T$. Moreover, the duality pairing $\langle y, x\rangle_{X^*\times X}$ is simply $y^{\mathrm{T}}x$.

2. Mathematical setting

In this section, we specify our abstract formulation (1), (2), which closely follows [30]. We let H, V, U be Hilbert spaces with $V \hookrightarrow H$, V dense in H. We identify H^* with H. The control space $\mathscr{U}$ and the state space $\mathscr{Y}$ are given by

$$\mathscr{U} = L^2(0,T;U) \quad \text{and} \quad \mathscr{Y} = \left\{ v \mid v \in L^2(0,T;V), \frac{\partial}{\partial t} y \in L^2(0,T;V^*) \right\},$$

respectively. We assume that

$$A(t) \in \mathscr{L}(V,V^*), \quad t \in [0,T]$$

is a family of continuous linear operators such that

$$\forall v, w \in V, \quad t \mapsto \langle A(t)v, w\rangle_{V^*\times V} \text{ is measurable on } (0,T)$$

and there exist $c, \nu > 0$, $\lambda \geqslant 0$ such that for all $t \in [0,T]$ and for all $v, w \in V$,

$$\langle A(t)v, w\rangle_{V^*\times V} \leqslant c\|v\|_V\|w\|_V,$$

$$\langle A(t)v, v\rangle_{V^*\times V} + \lambda\|v\|_H^2 \geqslant \nu\|v\|_V^2.$$

Furthermore, we assume that $B(t) \in \mathscr{L}(U,V^*)$ depends continuously on $t \in [0,T]$ and that $f \in L^2(0,T;V^*)$, $y_0 \in H$. The state equation (2) admits a unique solution $y \in \mathscr{Y}$ which depends continuously on the initial condition and on the right-hand side (see, e.g., [3, Chapter 2]; [30, pp. 102, 103]; [40, Section 23.7]).

To specify the objective function (1), we let Z, Z_T be Hilbert spaces and we assume that $C(t) \in \mathscr{L}(V,Z)$ depends continuously on $t \in [0,T]$, $C_T \in \mathscr{L}(H,Z_T)$, $z_1 \in L^2(0,T;Z)$, and $z_2 \in Z_T$. Since $\mathscr{Y} \subset C([0,T];H)$ (see, e.g., [30, p. 102] or [40, Section 23.6]), the objective function (1) is well defined

for $u \in \mathscr{U}$ and $y \in \mathscr{Y}$. With the assumption

$$\alpha_1, \alpha_2 \geqslant 0,$$

the optimal control problem (1), (2) admits a unique solution u (see [30, Chapter III]). The solution is characterized by the optimality conditions (2)–(4).

3. Temporal decomposition of the problem

We use a multiple shooting approach to reformulate the optimization problem (1), (2). We select a partition

$$0 = T_0 < T_1 < \cdots < T_{N_t} = T$$

of $[0, T]$ and we introduce the auxiliary variables

$$\bar{y}_i \in H, \quad n = 0, \ldots, N_t,$$

where

$$\bar{y}_0 \stackrel{\text{def}}{=} y_0.$$

Moreover, we set

$$\bar{u}_i = u|_{(T_i, T_{i+1})} \in L^2(T_i, T_{i+1}; U), \quad i = 0, \ldots, N_t - 1.$$

The state equations

$$\frac{\partial}{\partial t} y_i + A(t) y_i = B(t)\bar{u}_i + f, \quad t \in (T_i, T_{i+1}), \tag{5a}$$

$$y_i(T_i) = \bar{y}_i, \tag{5b}$$

$i = 0, \ldots, N_t - 1$, together with the continuity conditions

$$\bar{y}_{i+1} = y_i(T_{i+1}), \quad i = 0, \ldots, N_t - 1 \tag{6}$$

are equivalent to the original state equation (2). If y solves (2), then $\bar{y}_i = y(T_i)$, $y_i = y|_{\Omega \times (T_i, T_{i+1})}$, $i = 0, \ldots, N_t - 1$, solves (5), (6) and vice versa. In this case,

$$\begin{aligned} &\frac{1}{2} \int_0^T \|u(t)\|_U^2 \, \mathrm{d}t + \frac{\alpha_1}{2} \int_0^T \|Cy(t) - z_1(t)\|_Z^2 \, \mathrm{d}t + \frac{\alpha_2}{2} \|C_T y(T) - z_2\|_{Z_T}^2 \\ &\quad = \sum_{i=0}^{N_t - 1} \left\{ \frac{1}{2} \int_{T_i}^{T_{i+1}} \|u_i(t)\|_U^2 \, \mathrm{d}t + \frac{\alpha_1}{2} \int_{T_i}^{T_{i+1}} \|Cy_i(t) - z_1(t)\|_Z^2 \, \mathrm{d}t \right\} \\ &\qquad + \frac{\alpha_2}{2} \|C_T y_{N_t - 1}(T_{N_t}) - z_2\|_{Z_T}^2. \end{aligned} \tag{7}$$

Remark 1. In reformulation (7) of the objective function, we have expressed the terminal observation $y(T)$ by $y_{N_t - 1}(T_{N_t})$. Alternatively, we could have used the continuity condition $\bar{y}_{N_t} = y_{N_t - 1}(T_{N_t})$ and express the terminal observation $\|C_T y(T) - z_2\|_{Z_T}^2$ in (1) as $\|C_T \bar{y}_{N_t} - z_2\|_{Z_T}^2$. The following discussions

remain valid if this alternative formulation of the objective function is chosen. However, in our numerical experiments (7) was never inferior to the alternative formulation and often significantly better.

It is clear that the solution of the differential equation (5) is a function of $\bar{y}_i$, $\bar{u}_i$ and f. Therefore, the continuity conditions (6) and the objective function (7) can be viewed as functions of $\bar{u}_i$, $i = 0, \ldots, N_t - 1$, and $\bar{y}_i$, $i = 0, \ldots, N_t$. This will be formalized in the following. We define

$$\mathscr{U}_i = L^2(T_i, T_{i+1}; U) \quad \text{and} \quad \mathscr{Y}_i = \left\{ y \mid y \in L^2(T_i, T_{i+1}; V), \frac{\partial}{\partial t} y \in L^2(T_i, T_{i+1}; V^*) \right\}.$$

To express the continuity condition (6) in terms of $\bar{y}_i, \bar{u}_i$, we define

$$\bar{A}_i \in \mathscr{L}(H), \quad \bar{B}_i \in \mathscr{L}(\mathscr{U}_i, H), \quad \bar{b}_i \in H \quad i = 0, \ldots, N_t - 1$$

and $\bar{b}_i \in H$, $i = 0, \ldots, N_t - 1$, as follows:

$$\bar{A}_i \bar{y}_i = y_i^y(T_{i+1}), \quad \bar{B}_i \bar{u}_i = y_i^u(T_{i+1}), \quad \bar{b}_i = y_i^f(T_{i+1}), \tag{8}$$

where y_i^y is the solution of (5) with $\bar{u}_i = 0$ and $f = 0$, y_i^u is the solution of (5) with $\bar{y}_i = 0$ and $f = 0$, and y_i^f is the solution of (5) with $\bar{y}_i = 0$ and $\bar{u}_i = 0$. Using (8) the continuity conditions (6) can be written as

$$\bar{y}_{i+1} = \bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i, \quad i = 0, \ldots, N_t - 1. \tag{9}$$

To express the objective function (7) in terms of $\bar{y}_i, \bar{u}_i$, we define

$$\bar{E}_i \in \mathscr{L}(H, \mathscr{Y}_i), \quad \bar{F}_i \in \mathscr{L}(\mathscr{U}_i, \mathscr{Y}_i), \quad \bar{f}_i \in \mathscr{Y}_i, \quad i = 0, \ldots, N_t - 2$$

and

$$\bar{E}_i \in \mathscr{L}(H, \mathscr{Y}_i \times H), \quad \bar{F}_i \in \mathscr{L}(\mathscr{U}_i, \mathscr{Y}_i \times H), \quad \bar{f}_i \in \mathscr{Y}_i \times H, \quad i = N_t - 1,$$

as follows. For $i = 0, \ldots, N_t - 1$, let y_i^y be the solution of (5) with $\bar{u}_i = 0$ and $f = 0$, let y_i^u be the solution of (5) with $\bar{y}_i = 0$ and $f = 0$, and let y_i^f be the solution of (5) with $\bar{y}_i = 0$ and $\bar{u}_i = 0$. We set

$$\bar{E}_i \bar{y}_i = y_i^y, \quad \bar{F}_i \bar{u}_i = y_i^u, \quad \bar{f}_i = y_i^f, \quad i = 0, \ldots, N_t - 2 \tag{10}$$

and

$$\bar{E}_i \bar{y}_i = \begin{pmatrix} y_i^y \\ y_i^y(T) \end{pmatrix}, \quad \bar{F}_i \bar{u}_i = \begin{pmatrix} y_i^u \\ y_i^u(T) \end{pmatrix}, \quad \bar{f}_i = \begin{pmatrix} y_i^f \\ y_i^f(T) \end{pmatrix}, \quad i = N_t - 1. \tag{11}$$

For $i = 0, \ldots, N_t - 2$, the solution y_i of (5) is given by

$$y_i(t) = (\bar{E}_i \bar{y}_i)(t) + (\bar{F}_i \bar{u}_i)(t) + \bar{f}_i(t), \quad t \in (T_i, T_{i+1}) \tag{12}$$

and for $i = N_t - 1$, $y_i(t)$ is the first component of $\bar{E}_i \bar{y}_i + \bar{F}_i \bar{u}_i + \bar{f}_i$ evaluated at t. It is clear that $\bar{A}_i$, $\bar{B}_i$, $\bar{b}_i$ are closely related to $\bar{E}_i$, $\bar{F}_i$, $\bar{f}_i$. For example $\bar{A}_i \bar{y}_i = (\bar{E}_i \bar{y}_i)(T_{i+1})$.

We also need the operators

$$\bar{M}_i^z \in \mathscr{L}(\mathscr{Y}_i, \mathscr{Y}_i^*), \quad i = 0, \ldots, N_t - 2, \quad \bar{M}_i^z \in \mathscr{L}(\mathscr{Y}_i \times H, \mathscr{Y}_i^* \times H), \quad i = N_t - 1,$$

defined by

$$\langle \bar{M}_i^z y_i, w_i \rangle_{\mathscr{Y}_i^* \times \mathscr{Y}_i} = \int_{T_i}^{T_{i+1}} \alpha_1 \langle y_i(t), C(t)^* C(t) w_i(t) \rangle_{V^* \times V} \, \mathrm{d}t \quad \forall y_i, w_i \in \mathscr{Y}_i, \tag{13}$$

$i = 0, \ldots, N_t - 2$, and

$$\begin{aligned} &\left\langle \bar{M}_i^z \begin{pmatrix} y_i \\ \bar{y}_i \end{pmatrix}, \begin{pmatrix} w_i \\ \bar{w}_i \end{pmatrix} \right\rangle_{(\mathscr{Y}_i^* \times H) \times (\mathscr{Y}_i \times H)} \\ &\quad = \int_{T_i}^{T_{i+1}} \alpha_1 \langle y_i(t), C(t)^* C(t) w_i(t) \rangle_{V^* \times V} \, \mathrm{d}t \\ &\qquad + \alpha_2 \langle \bar{y}_i, C_T^* C_T \bar{w}_i \rangle_H \, \mathrm{d}t \quad \forall y_i, w_i \in \mathscr{Y}_i, \ \ \bar{y}_i, \bar{w}_i \in H, \end{aligned} \tag{14}$$

$i = N_t - 1$, and the vectors $\bar{z}_i = \alpha_1 C(\cdot)^* z_1 \in \mathscr{Y}_i$, $i = 0, \ldots, N_t - 2$, and

$$\bar{z}_i = \begin{pmatrix} \alpha_1 C(\cdot)^* z_1 \\ \alpha_2 C_T^* z_2 \end{pmatrix} \in \mathscr{Y}_i \times H.$$

We can now express the objective function (7) in terms of $\bar{y}_i, \bar{u}_i$:

$$\begin{aligned} &\sum_{i=0}^{N_t-1} \int_{T_i}^{T_{i+1}} \frac{1}{2} \|u_i(t)\|_U^2 + \frac{\alpha_1}{2} \|C(t) y_i(t) - z_1(t)\|_Z^2 \, \mathrm{d}t + \frac{\alpha_2}{2} \|C_T y_{N_t-1}(T_{N_t}) - z_2\|_{Z_T}^2 \\ &= \sum_{i=0}^{N_t-1} \left\{ \frac{1}{2} \langle \bar{u}_i, (I + \bar{F}_i^* \bar{M}_i^z \bar{F}_i) \bar{u}_i \rangle_{\mathscr{U}_i} + \langle \bar{y}_i, \bar{E}_i^* \bar{M}_i^z \bar{F}_i \bar{u}_i \rangle_H + \frac{1}{2} \langle \bar{y}_i, \bar{E}_i^* \bar{M}_i^z \bar{E}_i \bar{y}_i \rangle_H \right. \\ &\qquad \left. + \langle \bar{u}_i, \bar{F}_i^* (\bar{M}_i^z \bar{f}_i - \bar{z}_i) \rangle_{\mathscr{U}_i} + \langle \bar{y}_i, \bar{E}_i^* (\bar{M}_i^z \bar{f}_i - \bar{z}_i) \rangle_H \right\} + \text{const} \\ &\overset{\text{def}}{=} \sum_{i=0}^{N_t-1} \frac{1}{2} \langle \bar{y}_i, \bar{Q}_i \bar{y}_i \rangle_H + \langle \bar{c}_i, \bar{y}_i \rangle_H + \langle \bar{y}_i, \bar{R}_i \bar{u}_i \rangle_{\mathscr{U}_i} + \frac{1}{2} \langle \bar{u}_i, \bar{S}_i \bar{u}_i \rangle_{\mathscr{U}_i} + \langle \bar{d}_i, \bar{u}_i \rangle_{\mathscr{U}_i} \\ &\qquad + \frac{1}{2} \langle \bar{y}_{N_t}, \bar{Q}_{N_t} \bar{y}_{N_t} \rangle_H + \langle \bar{c}_{N_t}, \bar{y}_{N_t} \rangle_H + \text{const}, \end{aligned} \tag{15}$$

where $\bar{Q}_{N_t} = 0$, $\bar{c}_{N_t} = 0$, and 'const' represents all terms which are independent of the $\bar{y}_i$'s, $\bar{u}_i$'s, such as $\frac{1}{2}\alpha_2 \|z_2\|_{Z_T}^2$.

Remark 2. In the definition of operators and vectors we had to distinguish between the cases $i = 0, \ldots, N_t - 2$ and $i = N_t - 1$. This was necessary because of our reformulation (7) of the objective function. See Remark 1. If one uses the continuity condition $\bar{y}_{N_t} = y_{N_t-1}(T_{N_t})$ and express the

terminal observation $\|C_T y(T) - z_2\|^2_{Z_T}$ in (7) as $\|C_T \bar{y}_{N_t} - z_2\|^2_{Z_T}$, then the distinction between these two cases is not necessary. One would obtain problem (15) with $\bar{Q}_{N_t} = \alpha_2 C_T^* C_T$, $\bar{c}_{N_t} = -\alpha_2 C_T^* z_2$. Since we formally include $\bar{Q}_{N_t} = 0$ and $\bar{c}_{N_t} = 0$ in (15), the following discussions remain valid if this alternative formulation of the objective function is chosen.

We also note that our formalism applied in (7) can be easily translated to the discretized case.

From (9) and (7), (15) we see that the linear quadratic optimal control problem (1), (2) is equivalent to the problem

$$\min \frac{1}{2}\langle \bar{y}_{N_t}, \bar{Q}_{N_t}\bar{y}_{N_t}\rangle_H + \langle \bar{c}_{N_t}, \bar{y}_{N_t}\rangle_H + \sum_{i=0}^{N_t-1} \frac{1}{2}\langle \bar{y}_i, \bar{Q}_i \bar{y}_i\rangle_H + \langle \bar{c}_i, \bar{y}_i\rangle_H + \langle \bar{y}_i, \bar{R}_i \bar{u}_i\rangle_{\mathcal{U}_i} + \frac{1}{2}\langle \bar{u}_i, \bar{S}_i \bar{u}_i\rangle_{\mathcal{U}_i} + \langle \bar{d}_i, \bar{u}_i\rangle_{\mathcal{U}_i} \tag{16a}$$

$$\text{s.t. } \bar{y}_{i+1} = \bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i, \quad i = 0, \ldots, N_t,$$

$$\bar{y}_0 = y_0. \tag{16b}$$

Problem (16) is a discrete-time optimal-control problem in Hilbert space.

From the definition of $\bar{S}_i$ and $\bar{Q}_i$, $0 = 1, \ldots, N_t - 1$, $\bar{Q}_{N_t}$ in (15) we immediately obtain the following result.

Theorem 3. *The operators $\bar{S}_i$, $i = 0, \ldots, N_t - 1$, are strictly positive,*

$$\langle \bar{u}_i, \bar{S}_i \bar{u}_i\rangle_{\mathcal{U}_i} \geqslant \|\bar{u}_i\|^2_{\mathcal{U}_i} \quad \forall \bar{u}_i \in \mathcal{U}_i$$

and the operators $\bar{Q}_i$, $i = 0, \ldots, N_t$, are positive.

The augmented Lagrange function for (16) is given by

$$\begin{aligned} L_\rho(\vec{\bar{y}}, \vec{\bar{u}}, \vec{\bar{p}}) = {} & \tfrac{1}{2}\langle \bar{y}_{N_t}, \bar{Q}_{N_t}\bar{y}_{N_t}\rangle_H + \langle \bar{c}_{N_t}, \bar{y}_{N_t}\rangle_H \\ & + \sum_{i=0}^{N_t-1} \frac{1}{2}\langle \bar{y}_i, \bar{Q}_i \bar{y}_i\rangle_H + \langle \bar{c}_i, \bar{y}_i\rangle_H + \langle \bar{y}_i, \bar{R}_i \bar{u}_i\rangle_{\mathcal{U}_i} + \frac{1}{2}\langle \bar{u}_i, \bar{S}_i \bar{u}_i\rangle_{\mathcal{U}_i} + \langle \bar{d}_i, \bar{u}_i\rangle_{\mathcal{U}_i} \\ & + \sum_{i=0}^{N_t-1} \langle \bar{p}_{i+1}, -\bar{y}_{i+1} + \bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i\rangle_H \\ & + \frac{\rho}{2}\sum_{i=0}^{N_t-1} \|\bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i - \bar{y}_{i+1}\|^2_H, \end{aligned}$$

where $\vec{\bar{y}} = (\bar{y}_1, \ldots, \bar{y}_{N_t})$, $\vec{\bar{u}} = (u_0, \ldots, u_{N_t})$, $\vec{\bar{p}} = (\bar{p}_1, \ldots, \bar{p}_{N_t})$ and $\rho \geqslant 0$ is the augmentation parameter.

Theorem 4. *If $\vec{\bar{y}} \in H^{N_t-1}$ $\vec{\bar{u}} \in \mathcal{U}_0 \times \cdots \times \mathcal{U}_{N_t-1}$ solves (16), then there exists $\vec{\bar{p}} \in H^{N_t-1}$ such that the following equations are satisfied.*

State equation:

$$\bar{y}_{i+1} = \bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i, \quad i = 0, \ldots, N_t - 1,$$
$$\bar{y}_0 = y_0. \tag{17a}$$

Adjoint equation:

$$\bar{p}_{N_t} = \bar{Q}_{N_t} \bar{y}_{N_t} + c_{N_t} - \rho(\bar{A}_{N_t-1} \bar{y}_{N_t-1} + \bar{B}_{N_t-1} \bar{u}_{N_t-1} + \bar{b}_{N_t-1} - \bar{y}_{N_t}),$$

$$\bar{p}_i = \bar{A}_i^* \bar{p}_{i+1} + \bar{Q}_i \bar{y}_i + \bar{R}_i \bar{u}_i + \bar{c}_i + \rho \bar{A}_i^* (\bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i - \bar{y}_{i+1})$$
$$- \rho(\bar{A}_{i-1} \bar{y}_{i-1} + \bar{B}_{i-1} \bar{u}_{i-1} + \bar{b}_{i-1} - \bar{y}_i), \quad i = N_t - 1, \ldots, 1. \tag{17a}$$

Gradient equation:

$$\bar{S}_i \bar{u}_i + \bar{R}_i^* \bar{y}_i + \bar{B}_i^* \bar{p}_{i+1} + \bar{d}_i + \rho \bar{B}_i^* (\bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i - \bar{y}_{i+1}) = 0, \quad i = 0, \ldots, N_t - 1. \tag{17b}$$

On the other hand, if $\vec{\bar{y}} \in H^{N_t-1}$ $\vec{\bar{u}} \in \mathcal{U}_0 \times \cdots \times \mathcal{U}_{N_t-1}$ *and* $\vec{\bar{p}} \in H^{N_t-1}$ *satisfy* (17), *then* $\vec{\bar{y}} \in H^{N_t-1}$ $\vec{\bar{u}} \in \mathcal{U}_0 \times \cdots \times \mathcal{U}_{N_t-1}$ *solve* (16).

Proof. The assertion follows from a straightforward application of the Lagrange multiplier theorem in Hilbert space [16, Theorem 26.1]; [25]. The lower bidiagonal (in the $\bar{y}_i$'s) structure of the constraints (16b) immediately implies their surjectivity. The optimality conditions (17) are obtained by setting the partial gradients of L_ρ with respect to $\bar{p}_{i+1}$, $\bar{y}_i$, and $\bar{u}_i$ to zero. Since we deal with convex problems the optimality conditions (17) are necessary and sufficient. □

We close this section with an interpretation of the optimality conditions (17). The operators $\bar{A}_i, \ldots, \bar{S}_i$ and the vectors $\bar{b}_i, \ldots, \bar{d}_i$, are introduced for theoretical purposes, but they do not have to be formed in actual computations. This will be indicated by the following remark and it will be discussed more in future sections. First, we need the adjoints of the operators $\bar{A}_i$, $\bar{B}_i$, $\bar{E}_i$, $\bar{F}_i$, $i = 0, \ldots, N_t - 1$.

Consider the differential equation

$$-\frac{\partial}{\partial t} p_i(t) + A(t)^* p_i(t) = g(t), \quad t \in [T_i, T_{i+1}], \tag{18a}$$
$$p(T_{i+1}) = \bar{p}_{i+1}, \tag{18b}$$

where $g \in L^2(T_i, T_{i+1}; V^*)$ and $\bar{p}_{i+1} \in H$.

Lemma 5. (i) *The adjoints* $\bar{A}_i^*$, $\bar{B}_i^*$, $i = 0, \ldots, N_t - 1$, *of the operators defined in* (8) *are given by*

$$\bar{A}_i^* \bar{p}_{i+1} = p_i(T_i), \quad \bar{B}_i^* \bar{p}_{i+1} = B(t)^* p_i(t),$$

where p_i *is the solution of* (18) *with* $g = 0$.

(ii) *The adjoints* $\bar{E}_i^*$, $\bar{F}_i^*$, $i = 0, \ldots, N_t - 2$, *of the operators defined in* (10) *are given by*

$$\bar{E}_i^* g = p_i(T_i), \quad \bar{F}_i^* g = B(t)^* p_i(t),$$

where p_i *is the solution of* (18) *with* $\bar{p}_{i+1} = 0$.

(iii) *The adjoints* $\bar{E}_i^*$, $\bar{F}_i^*$, $i = N_t - 1$, *of the operators defined in* (11) *are given by*

$$\bar{E}_i^* \begin{pmatrix} g \\ \bar{p}_{i+1} \end{pmatrix} = p_i(T_i), \quad \bar{F}_i^* \begin{pmatrix} g \\ \bar{p}_{i+1} \end{pmatrix} = B(t)^* p_i(t),$$

where p_i *is the solution of* (18).

Proof. We only prove (iii). All other statements can be shown similarly.

Let $i=N_t-1$, i.e., $T_{i+1}=T$. Furthermore, let y_i and p_i be the solutions of (5) and (18), respectively. Then

$$\int_{T_i}^{T_{i+1}} \left\langle \frac{\partial}{\partial t} y_i(t), p_i(t) \right\rangle_H + \langle A(t) y_i(t), p_i(t) \rangle_{V^* \times V} - \langle B(t)\bar{u}_i(t) + f(t), p_i(t) \rangle_{V^* \times V} \, \mathrm{d}t = 0,$$

$$\int_{T_i}^{T_{i+1}} \left\langle -\frac{\partial}{\partial t} p_i(t), y_i(t) \right\rangle_H + \langle A(t)^* p_i(t), y_i(t) \rangle_{V^* \times V} - \langle g(t), y_i(t) \rangle_{V^* \times V} \, \mathrm{d}t = 0.$$

Subtracting both equations and using (5b), (18b) gives

$$\begin{aligned} &\langle y_i(T_{i+1}), \bar{p}_{i+1} \rangle_H - \langle \bar{y}_i, p_i(T_i) \rangle_H \\ &\quad = \int_{T_i}^{T_{i+1}} \langle B(t)\bar{u}_i(t) + f(t), p_i(t) \rangle_{V^* \times V} - \langle g(t), y_i(t) \rangle_{V^* \times V} \, \mathrm{d}t. \end{aligned} \tag{19}$$

Now, let $\bar{y}_i, \bar{p}_{i+1} \in H$, $g \in L^2(T_i, T_{i+1}; V^*)$ be arbitrary, $\bar{u}_i = 0$, $f = 0$, and let y_i, p_i solve (5) and (18), respectively. Definition (11) of $\bar{E}_{N_t-1}$ and (19) imply

$$\begin{aligned} \left\langle \begin{pmatrix} g \\ \bar{p}_{i+1} \end{pmatrix}, \bar{E}_i \bar{y}_i \right\rangle_{(\mathscr{Y}_i^* \times H) \times (\mathscr{Y}_i \times H)} &= \int_{T_i}^{T_{i+1}} \langle g(t), y_i(t) \rangle_{V^* \times V} \, \mathrm{d}t + \langle y_i(T_{i+1}), \bar{p}_{i+1} \rangle_H \\ &= \langle \bar{y}_i, p_i(T_i) \rangle_H = \left\langle \bar{E}_i^* \begin{pmatrix} g \\ \bar{p}_{i+1} \end{pmatrix}, \bar{y}_i \right\rangle_H. \end{aligned}$$

This proves the first part of (iii).

To prove the second part of (iii), we let $\bar{u}_i \in \mathscr{U}_i$, $\bar{p}_{i+1} \in H$, $g \in L^2(T_i, T_{i+1}; V^*)$ be arbitrary, $\bar{y}_i = 0$, $f = 0$, and let y_i, p_i solve (5) and (18), respectively. Definition (11) of $\bar{F}_{N_t-1}$ and (19) imply

$$\begin{aligned} \left\langle \begin{pmatrix} g \\ \bar{p}_{i+1} \end{pmatrix}, \bar{F}_i \bar{u}_i \right\rangle_{(\mathscr{Y}_i^* \times H) \times (\mathscr{Y}_i \times H)} &= \int_{T_i}^{T_{i+1}} \langle g(t), y_i(t) \rangle_{V^* \times V} \, \mathrm{d}t + \langle y_i(T_{i+1}), \bar{p} \rangle_H \\ &= \int_{T_i}^{T_{i+1}} \langle B(t)^* p_i(t), \bar{u}_i(t) \rangle_{U^* \times U} \, \mathrm{d}t \\ &= \left\langle \bar{F}_i^* \begin{pmatrix} g \\ \bar{p}_{i+1} \end{pmatrix}, \bar{u}_i \right\rangle_{\mathscr{U}_i^* \times \mathscr{Y}_i}. \end{aligned}$$

□

Remark 6. Let $\bar{A}_i$, $\bar{B}_i$, $\bar{b}_i$, $i=1,\ldots,N_t-1$, be defined by (8) and let $\bar{Q}_i$, $\bar{R}_i$, $\bar{S}_i$, $\bar{c}_i$, $\bar{d}_i$, $i=1,\ldots,N_t-1$, $\bar{Q}_{N_t}$, $\bar{c}_{N_t}$ be defined by (15).

(i) Computation of $\bar{y}_{i+1}=\bar{A}_i\bar{y}_i+\bar{B}_i\bar{u}_i+\bar{b}_i$, $i=0,\ldots,N_t-1$: From (12), (8) we see that $\bar{y}_{i+1}=y_i(T_{i+1})$, where y_i is the solution of (5). Individual quantities $\bar{A}_i\bar{y}_i$, $\bar{B}_i\bar{u}_i$, or $\bar{b}_i$ can be computed by setting the appropriate parts of the input variables $\bar{y}_i$, $\bar{u}_i$, and f to zero.

(ii) Computation of $\bar{p}_i=\bar{A}_i^*\bar{p}_{i+1}+\bar{Q}_i\bar{y}_i+\bar{R}_i\bar{u}_i+\bar{c}_i$, $i=1,\ldots,N_t-1$: From (15) we see that

$$\begin{aligned}\bar{A}_i^*\bar{p}_{i+1}+\bar{Q}_i\bar{y}_i+\bar{R}_i\bar{u}_i+\bar{c}_i &= \bar{A}_i^*\bar{p}_{i+1}+\bar{E}_i^*(\bar{M}_i^z(\bar{E}_i\bar{y}_i+\bar{F}_i\bar{u}_i+\bar{f}_i)-\bar{z}_i),\\ &= \bar{A}_i^*\bar{p}_{i+1}+\bar{E}_i^*(\bar{M}_i^z w_i-\bar{z}_i),\end{aligned}$$

where w_i is the solution of (5).

Definitions (13), (14) of M_i^z and Lemma 5 imply that $\bar{p}_i=p_i(T_i)$, where p_i is the solution of

$$-\frac{\partial}{\partial t}p_i(t)+A(t)^*p_i(t)=\alpha_1 C(t)^*(C(t)w_i(t)-z_1),\quad t\in[T_i,T_{i+1}] \tag{20a}$$

with final condition

$$p_i(T_{i+1})=\begin{cases}\bar{p}_{i+1}, & i=0,\ldots,N_t-2,\\ \bar{p}_{i+1}+\alpha_2 C_T^*(C_T w_i(T)-z_2), & i=N_t-1.\end{cases} \tag{20b}$$

(iii) For $i=N_t$, $\bar{p}_{N_t}=\bar{Q}_{N_t}\bar{y}_{N_t}+\bar{c}_{N_t}=0$ since $\bar{Q}_{N_t}=0$, $\bar{c}_{N_t}=0$.

(iv) Computation of $\bar{v}_i=\bar{S}_i\bar{u}_i+\bar{R}_i^*\bar{y}_i+\bar{B}_i^*\bar{p}_{i+1}+\bar{d}_i$, $i=0,\ldots,N_t-1$: From (15) we see that

$$\begin{aligned}\bar{S}_i\bar{u}_i+\bar{R}_i^*\bar{y}_i+\bar{B}_i^*\bar{p}_{i+1}+\bar{d}_i &= \bar{B}_i^*\bar{p}_{i+1}+\bar{u}_i+\bar{F}_i^*(\bar{M}_i^z(\bar{E}_i\bar{y}_i+\bar{F}_i\bar{u}_i+\bar{f}_i)-\bar{z}_i)\\ &= \bar{B}_i^*\bar{p}_{i+1}+\bar{u}_i+\bar{F}_i^*(\bar{M}_i^z w_i-\bar{z}_i),\end{aligned}$$

where w_i is the solution of (5).

Definitions (13), (14) of M_i^z and Lemma 5 imply that

$$\bar{v}_i(t)=B(t)^*p_i(t)+\bar{u}_i(t),$$

where p_i solves (18).

4. Iterative solution of the optimality system

4.1. Optimality system

We group Eqs. (17) in the following way:

$$\begin{aligned}&\bar{S}_0\bar{u}_0+\bar{R}_0^*\bar{y}_0+\bar{B}_0^*\bar{p}_1+\bar{d}_0+\rho\bar{B}_0^*(\bar{A}_0\bar{y}_0+\bar{B}_0\bar{u}_0+\bar{b}_0-\bar{y}_1)=0,\\ &\bar{A}_0\bar{y}_0+\bar{B}_0\bar{u}_0+\bar{b}_0-\bar{y}_1=0,\\ &-\bar{p}_i+\bar{A}_i^*\bar{p}_{i+1}+\bar{Q}_i\bar{y}_i+\bar{R}_i\bar{u}_i+\bar{c}_i+\rho\bar{A}_i^*(\bar{A}_i\bar{y}_i+\bar{B}_i\bar{u}_i+\bar{b}_i-\bar{y}_{i+1})\\ &\quad-\delta(\bar{A}_{i-1}\bar{y}_{i-1}+\bar{B}_{i-1}\bar{u}_{i-1}+\bar{b}_{i-1}-\bar{y}_i)=0,\end{aligned} \tag{21a}$$

$$\bar{S}_i\bar{u}_i + \bar{R}_i^*\bar{y}_i + \bar{B}_i^*\bar{p}_{i+1} + \bar{d}_i + \rho\bar{B}_i^*(\bar{A}_i\bar{y}_i + \bar{B}_i\bar{u}_i + \bar{b}_i - \bar{y}_{i+1}) = 0,$$
$$\bar{A}_i\bar{y}_i + \bar{B}_i\bar{u}_i + \bar{b}_i - \bar{y}_{i+1} = 0,$$
$$i = 1,\ldots,N_t - 1, \tag{21b}$$

$$-\bar{p}_{N_t} + \bar{Q}_{N_t}\bar{y}_{N_t} + c_{N_t} - \delta(\bar{A}_{N_t-1}\bar{y}_{N_t-1} + \bar{B}_{N_t-1}\bar{u}_{N_t-1} + \bar{b}_{N_t-1} - \bar{y}_{N_t}) = 0. \tag{21c}$$

Systems (17) and (21) are identical if $\rho = \delta \geqslant 0$. Here we introduce the second parameter δ to better distinguish the terms in (17) corresponding to ρ, δ, respectively. In particular, ρ is associated with terms $\bar{A}_i\bar{y}_i + \bar{B}_i\bar{u}_i + \bar{b}_i - \bar{y}_{i+1}$ linking state information from time subdomain $[T_i, T_{i+1}]$ to the initial data $\bar{y}_{i+1}$ for the state in $[T_{i+1}, T_{i+2}]$, while δ is associated with terms $\bar{A}_{i-1}\bar{y}_{i-1} + \bar{B}_{i-1}\bar{u}_{i-1} + \bar{b}_{i-1} - \bar{y}_i$ linking state information from time subdomain $[T_{i-1}, T_i]$ to the initial data $\bar{y}_i$ for the state in $[T_i, T_{i+1}]$. We will see later (Theorem 8) that these terms may influence the convergence of iterative schemes for the solution of (21) differently. This has motivated the introduction of the second parameter δ at this point. We assume that $\rho, \delta \geqslant 0$.

Now we arrange Eqs. (21) into a block system

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{22}$$

where the variables $\mathbf{x}$ and the right-hand side $\mathbf{b}$ are given by

$$\mathbf{x} = \begin{pmatrix} \bar{y}_1 \\ \bar{u}_0 \\ \hline \bar{y}_2 \\ \bar{u}_1 \\ \bar{p}_1 \\ \hline \bar{y}_3 \\ \bar{u}_2 \\ \bar{p}_2 \\ \hline \vdots \\ \vdots \\ \hline \bar{y}_{N_t} \\ \bar{u}_{N_t-1} \\ \bar{p}_{N_t-1} \\ \hline \bar{p}_{N_t} \end{pmatrix}, \quad \mathbf{b} = - \begin{pmatrix} \bar{d}_0 + (\bar{R}_0^* + \rho\bar{B}_0^*\bar{A}_0)y_0 + \rho\bar{B}_0^*\bar{b}_0 \\ \bar{b}_0 + \bar{A}_0 y_0 \\ \hline \bar{c}_1 + \rho\bar{A}_1^*\bar{b}_1 - \delta\bar{b}_0 \\ \bar{d}_1 + \rho\bar{B}_1^*\bar{b}_1 \\ \bar{b}_1 \\ \hline \bar{c}_2 + \rho\bar{A}_2^*\bar{b}_2 - \delta\bar{b}_1 \\ \bar{d}_2 + \rho\bar{B}_2^*\bar{b}_2 \\ \bar{b}_2 \\ \hline \vdots \\ \vdots \\ \hline \bar{c}_{N_t-1} + \rho\bar{A}_{N_t-1}^*\bar{b}_{N_t-1} - \delta\bar{b}_{N_t-2} \\ \bar{d}_{N_t-1} + \rho\bar{B}_{N_t-1}^*\bar{b}_{N_t-1} \\ \bar{b}_{N_t-1} \\ \hline \bar{c}_{N_t} \end{pmatrix}$$

$$
\left(\begin{array}{cc|ccc|ccc|ccc|ccc|c}
-\rho\bar{B}_0^* & \bar{S}_0^\rho & 0 & 0 & \bar{B}_0^* & & & & & & & & & & \\
-I & \bar{B}_0 & 0 & 0 & 0 & & & & & & & & & & \\
\hline
\bar{Q}_1^{\rho,\delta} & 0 & -\rho\bar{A}_1^* & \bar{R}_1^\rho & -I & 0 & 0 & \bar{A}_1^* & & & & & & & \\
(\bar{R}_1^\rho)^* & 0 & -\rho\bar{B}_1^* & \bar{S}_1^\rho & 0 & 0 & 0 & \bar{B}_1^* & & & & & & & \\
\bar{A}_1 & 0 & -I & \bar{B}_1 & 0 & 0 & 0 & 0 & & & & & & & \\
\hline
-\delta\bar{A}_1 & 0 & \bar{Q}_2^{\rho,\delta} & -\delta\bar{B}_1 & 0 & -\rho\bar{A}_2^* & \bar{R}_2^\rho & -I & 0 & 0 & \bar{A}_2^* & & & & \\
0 & 0 & (\bar{R}_2^\rho)^* & 0 & 0 & -\rho\bar{B}_2^* & \bar{S}_2^\rho & 0 & 0 & 0 & \bar{B}_2^* & & & & \\
0 & 0 & \bar{A}_2 & 0 & 0 & -I & \bar{B}_2 & 0 & 0 & 0 & 0 & & & & \\
\hline
 & & & & & & & & \ddots & & & & & & \\
 & & & & & & & & & \ddots & & & & & \\
 & & & & & & & & & & \ddots & & & & \\
\hline
 & & & & & -\delta\bar{A}_{N_t-2} & 0 & 0 & \bar{Q}_{N_t-1}^{\rho,\delta} & -\delta\bar{B}_{N_t-2} & 0 & -\rho\bar{A}_{N_t-1}^* & \bar{R}_{N_t-1}^\rho & -I & \bar{A}_{N_t-1}^* \\
 & & & & & 0 & 0 & 0 & (\bar{R}_{N_t-1}^\rho)^* & 0 & 0 & -\rho\bar{A}_{N_t-1}^* & \bar{S}_{N_t-1}^\rho & 0 & \bar{B}_{N_t-1}^* \\
 & & & & & 0 & 0 & 0 & \bar{A}_{N_t-1} & 0 & 0 & -I & \bar{B}_{N_t-1} & 0 & 0 \\
\hline
 & & & & & & & & -\delta\bar{A}_{N_t-1} & 0 & 0 & \bar{Q}_{N_t}+\delta I & -\delta\bar{B}_{N_t-1} & 0 & -I
\end{array}\right)
$$

Fig. 1. The operator $\mathbf{A}$ ($\bar{Q}_i^{\rho,\delta} = \bar{Q}_i + \rho\bar{A}_i^*\bar{A}_i + \delta I$, $\bar{R}_i^\rho = \bar{R}_i + \rho\bar{A}_i^*\bar{B}_i$, $\bar{S}_i^\rho = \bar{S}_i + \rho\bar{B}_i^*\bar{B}_i$).

and where $\mathbf{A}$ is an operator with one super-blockdiagonal and two sub-blockdiagonals shown in Fig. 1.

Since the operators $\bar{S}_i$, $i = 0,\ldots,N_t - 1$, are strictly positive (cf. Theorem 3) the diagonal blocks of $\mathbf{A}$ in Fig. 1 are continuously invertible. With this ordering of the equations and unknowns the operator $\mathbf{A}$ in Fig. 1 is nonselfadjoint. As we will discuss in Section 4.2.1, the diagonal blocks are related to a linear quadratic optimal control problem. This linear quadratic optimal control problem is essentially given by the restriction of (1), (2) onto the time subdomain $[T_i, T_{i+1}]$, but with initial conditions and augmentation of the objective function that tie the states and adjoints of this time subdomain optimal control problem to linear quadratic optimal control problems on adjacent time subdomains.

4.2. GS iterations

4.2.1. Solution of the block diagonal systems

We use Remark 6 to examine the diagonal blocks of $\mathbf{A}$ in Fig. 1. In particular, we will show that the diagonal blocks of $\mathbf{A}$ are invertible and that the application of the inverse of a diagonal block corresponds to the solution of a linear quadratic optimal control problem restricted to the time subdomain $[T_i, T_{i+1}]$.

For $i \in \{0, \ldots, N_t - 1\}$ let y_i be the solution of

$$\frac{\partial}{\partial t} y_i(t) + A(t) y_i(t) = B(t)\bar{u}_i(t) + f(t), \quad t \in (T_i, T_{i+1}), \tag{23a}$$

$$y_i(T_i) = \bar{y}_i. \tag{23b}$$

Remark 6 shows that

$$\bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i = y_i(T_{i+1}).$$

For $i \in \{0, \ldots, N_t - 1\}$ let p_i be the solution of

$$-\frac{\partial}{\partial t} p_i(t) + A(t)^* p_i(t) = \alpha_1 C(t)^* (C(t) y_i(t) - z_1), \quad t \in (T_i, T_{i+1}), \tag{23c}$$

$$p_i(T_{i+1}) = \begin{cases} \bar{p}_{i+1} + \rho(y_i(T_{i+1}) - \bar{y}_{i+1}), & i = 0, \ldots, N_t - 2, \\ \bar{p}_{i+1} + \rho(y_i(T_{i+1}) - \bar{y}_{i+1}) + \alpha_2 C_T^* (C_T y_i(T) - z_2), & i = N_t - 1, \end{cases} \tag{23d}$$

where y_i solves (23a), (23b). Remark 6 shows that

$$\bar{A}_i^* \bar{p}_{i+1} + \bar{Q}_i \bar{y}_i + \bar{R}_i \bar{u}_i + \bar{c}_i + \rho \bar{A}_i^* (\bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i - \bar{y}_{i+1}) = p_i(T_i).$$

Finally, for $i \in \{0, \ldots, N_t - 1\}$ the equation

$$0 = B(t)^* p_i(t) + \bar{u}_i(t), \quad t \in (T_i, T_{i+1}), \tag{23e}$$

where p_i solves (23c), (23d), is just the equation

$$\bar{S}_i \bar{u}_i + \bar{R}_i^* \bar{y}_i + \bar{B}_i^* \bar{p}_{i+1} + \bar{d}_i + \rho \bar{B}_i^* (\bar{A}_i \bar{y}_i + \bar{B}_i \bar{u}_i + \bar{b}_i - \bar{y}_{i+1}) = 0$$

(see Remark 6).

Now we are able to discuss the solution of the bock diagonal systems. Let $i = 0$. If $y_0, \bar{u}_0, p_0$ solves (23), then the solution $(\bar{y}_1, \bar{u}_0)$ of (21a) is given by

$$\bar{y}_1 = y_0(T_1).$$

Let $i \in \{1, \ldots, N_t - 1\}$. If $y_i, \bar{u}_i, p_i$ solves (23) and if y_{i-1} solves (23a) with i replaced by $i - 1$, (23b), then the solution $(\bar{y}_{i+1}, \bar{u}_i, \bar{p}_i)$ of (21b) is given by

$$\bar{y}_{i+1} = y_i(T_{i+1}), \quad \bar{p}_i = p_i(T_i) + \delta(\bar{y}_i - y_{i-1}(T_i)).$$

Finally, since $\bar{Q}_{N_t} = 0$, $\bar{c}_{N_t} = 0$ the solution of (21c) is

$$\bar{p}_{N_t} = \delta(\bar{y}_{N_t} - y_{N_t - 1}(T_{N_t})),$$

where $y_{N_t - 1}$ solves (23a), (23b) with $i = N_t - 1$.

Notice that system (23) for $i = 1, \ldots, N_t - 1$ is the optimality system for the quadratic optimization problem

$$\min \frac{1}{2} \int_{T_i}^{T_{i+1}} \|\bar{u}_i(t)\|_U^2 \, \mathrm{d}t + \frac{\alpha_1}{2} \int_{T_i}^{T_{i+1}} \|C(t) y_i(t) - z_1(t)\|_Z^2 \, \mathrm{d}t$$
$$+ \langle y_i(T_{i+1}), \bar{p}_{i+1} + \rho(y_i(T_{i+1}) - \bar{y}_{i+1}) \rangle_H$$

$$\text{s.t. } \frac{\partial}{\partial t} y_i(t) + A(t) y_i(t) = B(t)\bar{u}_i(t) + f(t), \quad t \in (T_i, T_{i+1}),$$
$$y_i(T_i) = \bar{y}_i, \tag{24}$$

where for $i = N_t - 1$, the term $\langle y_i(T_{i+1}), \bar{p}_{i+1}\rangle_H$ in the objective function has to be replaced by $\langle y_i(T_{i+1}), \bar{p}_{i+1} + \rho(y_i(T_{i+1}) - \bar{y}_{i+1}) + \alpha_2 C_T^*(C_T y_i(T) - z_2)\rangle_H$.

4.2.2. GS iterations

Let $\mathbf{D}$, $-\mathbf{L}$, $-\mathbf{U}$ be the block diagonal part, the strictly lower block triangular part, and the strictly upper block triangular part of $\mathbf{A}$, respectively. Thus,

$$\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}.$$

In the previous section, we have established the invertibility of $\mathbf{D}$. Hence, we can apply block GS iterations. We will look at three GS iterations (see, e.g., [2,19,35,38]). One iteration of the forward GS method is given by $\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{L})^{-1}(\mathbf{b} + \mathbf{U}\mathbf{x}_k)$, one iteration of the backward GS method is given by $\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{U})^{-1}(\mathbf{b} + \mathbf{L}\mathbf{x}_k)$, and one iteration of the forward–backward GS method consists of a forward GS iteration followed by a backward GS iteration:

$$\mathbf{x}_{k+1/2} = (\mathbf{D} - \mathbf{L})^{-1}(\mathbf{b} + \mathbf{U}\mathbf{x}_k),$$

$$\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{U})^{-1}(\mathbf{b} + \mathbf{L}\mathbf{x}_{k+1/2}).$$

More generally, we can consider the block SOR method. However, since the block SOR iterates can be computed from the block GS iterations, we restrict ourselves to the above cases. The GS method depends on the ordering of the variables and equations. Therefore other orderings might be useful. We discuss some of those later. First, we study the implementation of the block GS methods.

4.2.3. Interpretation of the GS iterations

Forward GS: One sweep of the forward GS method is given as follows:

Computation of $\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{L})^{-1}(\mathbf{b} + \mathbf{U}\mathbf{x}_k)$:

(a) Solve (21a) for $\bar{y}_1, \bar{u}_0$.

(b) For $i = 1, \ldots, N_t - 1$:

Solve (21b) for $\bar{y}_{i+1}, \bar{u}_i, \bar{p}_i$.

(c) Compute $\bar{p}_{N_t}$ from (21c). (25)

In (25) we overwrite the components of $\mathbf{x}_k$ by those of $\mathbf{x}_{k+1}$ as soon as they become available. Therefore, we omit the index k in steps (a)–(c) of (25).

Theorem 7. (i) *The operator-vector product* $(\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}\mathbf{x}$ *is independent of* ρ *and* δ.

(ii) *The null-space of* $(\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}$ *satisfies*

$$\mathcal{N}((\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}) \supset H \times L^2(T_0, T_1, U) \times H \times L^2(T_1, T_2, U) \times \{0\}$$
$$\times \cdots \times H \times L^2(T_{N_t-1}, T_{N_t}, U) \times \{0\} \times \{0\}.$$

Proof. Because of the third equation in (21b), all terms in (21b) involving ρ will be zero and, since we perform a forward sweep, all terms in (21b), (21c) involving δ will be zero. Hence, the forward GS method is independent of ρ and δ.

The second assertion follows immediately from the fact that all column blocks in the operator $\mathbf{U}$ corresponding to $\bar{y}_{i+1}, \bar{u}_i$, $i = 0, \ldots, N_t - 1$, are equal to zero (cf. Fig. 1). □

Because of Theorem 7 it is sufficient to consider $\rho = \delta = 0$.

Using our discussions in Section 4.2.1 we can formulate the forward GS method as follows:

Computation of $\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{L})^{-1}(\mathbf{b} + \mathbf{U}\mathbf{x}_k)$:

For $i = 0, \ldots, N_t - 1$:

Solve (24) (or (23)).

Set $\bar{y}_{i+1} = y_i(T_{i+1})$, $\bar{p}_i = p_i(T_i)$.

(If $i = 0$, only $\bar{y}_1, \bar{u}_0$ are computed.)

Set $\bar{p}_{N_t} = 0$. (26)

In the forward GS method, the states computed as the solutions of (23a), (23b) are continuous in time in the sense

$$y_i(T_{i+1}) = \bar{y}_{i+1} = y_{i+1}(T_i), \quad i = 0, \ldots, N_t - 1.$$

The adjoints and the controls, however, are in general not continuous at a given GS iteration. Only as the GS iterations converges (assuming it does) will the jumps in adjoints and controls at the time domain interfaces T_i vanish. If we perform one iteration of the forward GS method with starting value $\bar{p}_i = 0$, $i = 1, \ldots, N_t$, then the states y_i (dashed), controls $\bar{u}_i$ (solid), and the adjoints p_i (dotted) at the end of the forward GS iteration are sketched in Fig. 2. The $\bar{y}_i$, $\bar{p}_i$ components of the new iterate are indicated by • and ○, respectively.

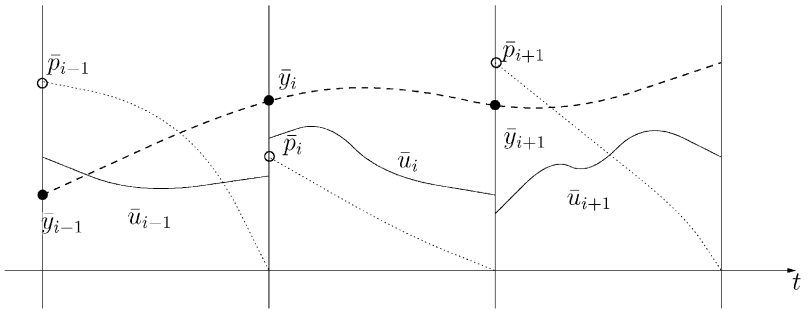

Fig. 2. Sketch of the states (dashed), the controls (solid) and the adjoints (dotted) after iteration k of the forward GS iteration (26).

Backward GS: One sweep of the backward GS method is given as follows:

Computation of $\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{U})^{-1}(\mathbf{b} + \mathbf{L}\mathbf{x}_k)$:

(a) Compute $\bar{p}_{N_t}$ from (21c).

(b) For $i = N_t - 1, \ldots, 1$:

Solve (21b) for $\bar{y}_{i+1}, \bar{u}_i, \bar{p}_i$.

(c) Solve (21a) for $\bar{y}_1, \bar{u}_0$. (27)

Again we overwrite the components of $\mathbf{x}_k$ by those of $\mathbf{x}_{k+1}$ as soon as they become available. Therefore, we omit the index k in steps a.–c. of (27).

Theorem 8. (i) *The operator-vector product* $(\mathbf{D} - \mathbf{U})^{-1}\mathbf{L}\mathbf{x}$ *is independent of* ρ.
(ii) *The null-space of* $(\mathbf{D} - \mathbf{U})^{-1}\mathbf{L}$ *satisfies*

$$\mathcal{N}((\mathbf{D} - \mathbf{U})^{-1}\mathbf{L}) \supset \{0\} \times L^2(T_0, T_1, U) \times \{0\} \times \{0\} \times H \times \cdots \times \{0\} \times \{0\} \times H \times H.$$

If $\delta = 0$,

$$\mathcal{N}((\mathbf{D} - \mathbf{U})^{-1}\mathbf{L}) \supset \{0\} \times L^2(T_0, T_1, U) \times \{0\} \times L^2(T_1, T_2, U) \times H \times \cdots \times \{0\} \times L^2(T_{N_t-1}, T_{N_t}, U) \times H \times H.$$

Proof. The third equation (21b) implies that all terms in (21b) involving ρ are zero. This proves (i).

The first assertion in (ii) follows immediately from the fact that all column blocks in the operator $\mathbf{L}$ corresponding to $\bar{u}_0$ $\bar{p}_i$, $i = 1, \ldots, N_t$, are equal to zero (cf. Fig. 1). If $\delta = 0$, then all column blocks in the operator $\mathbf{L}$ corresponding to $\bar{u}_i, \bar{p}_{i+1}$, $i = 0, \ldots, N_t - 1$, are equal to zero (cf. Fig. 1). □

Our discussions in Section 4.2.1 allow us to formulate the backward GS method as follows:

Computation of $\mathbf{x}_{k+1} = (\mathbf{D} - \mathbf{U})^{-1}(\mathbf{b} + \mathbf{L}\mathbf{x}_k)$:

If $\delta > 0$ solve (23a), (23b) for $i = N_t - 1$.

Compute $\bar{p}_{N_t} = \delta(\bar{y}_{N_t} - y_{N_t-1}(T_{N_t}))$.

For $i = N_t - 1, \ldots, 1$:

If $\delta > 0$ solve (23a), (23b) with i replaced by $i - 1$

Solve (24) (or (23)).

Set $\bar{y}_{i+1} = y_i(T_{i+1})$, $\bar{p}_i = p_i(T_i) + \delta(\bar{y}_i - y_{i-1}(T_i))$.

(If $i = 0$, only $\bar{y}_1, \bar{u}_0$ are computed.) (28)

In the backward GS method with $\delta = 0$, the adjoints are continuous in time in the sense

$$p_i(T_{i+1}) = \bar{p}_{i+1} = p_{i+1}(T_{i+1}), \quad i = 0, \ldots, N_t - 1.$$

The states, however, are in general discontinuous at a given iteration of the backward GS method.

4.2.4. Connection between the GS iteration and instantaneous control

Our presentation of the forward GS method (26) now allows for a different interpretation of instantaneous control methods in their simplest form. If one step of the forward GS method (26) with starting value $\bar{y}_i = 0$, $\bar{u}_i = 0$, $\bar{p}_i = 0$, $i = 1, \ldots, N_t$, is applied, then problem (24) to be solved in the ith substep of the forward GS method is identical to the original optimal control problem (1), (2) restricted to (T_i, T_{i+1}) with initial conditions $y(T_i) = \bar{y}_i$. Thus, instantaneous control methods are equivalent to one step of the forward GS method (26) with starting value $\bar{y}_i = 0$, $\bar{u}_i = 0$, $\bar{p}_i = 0$, $i = 1, \ldots, N_t$. If more than one GS iteration is performed, then the objective function term $\langle y_i(T_{i+1}), \bar{p}_{i+1} + \rho(y_i(T_{i+1}) - \bar{y}_{i+1})\rangle_H$ in (26) allows to propagate information from the time interval $[T_{i+1}, T]$ backward to the time interval $[T_i, T_{i+1}]$ via the adjoint $\bar{p}_{i+1}$ and the state $\bar{y}_{i+1}$, which are both associated with $[T_{i+1}, T_{i+2}]$.

As we have mentioned in the introduction, the instantaneous control methods applied in the literature [7,13,14,22–24] are somewhat different in that they may use a receeding time horizon, they use a different time-subinterval objective function, they use inexact solutions of subproblems (24), and they are applied to nonlinear problems. Moreover, 'convergence' of the instantaneous control approaches in [7,13,14,22–24] means usually that an objective function evaluated at time T is sufficiently small. Inexact subproblem (24) solves can be integrated into our formulation and our approach can be extended to nonlinear problems (although in more than one way, see Section 6). However, our final time T is fixed. Thus, our interpretation of instantaneous control as one iteration of the GS method can only be applied to explain the behavior of instantaneous control after the final time T is determined. If the GS method converges, which is often not the case (see next section) and for which no sufficient conditions exist for our applications, then convergence means that the states $\bar{y}_i$, controls $\bar{u}_i$, and adjoints $\bar{p}_i$, $i = 1, \ldots, N_t$, converge to the optimal states $y^*(T_i)$, the optimal controls $u^*|_{(T_i, T_{i+1})}$, and the optimal adjoints $y^*(T_i)$, $i = 1, \ldots, N_t$, respectively. This is different from the notion of convergence used in instantaneous control. It also needs to be re-emphasized that since the time horizon T in instantaneous control is not fixed a-priori, instantaneous control techniques are not intended for use in an open-loop control context, which is the basis for our time-domain decomposition techniques. Therefore, a precise comparison between existing instantaneous control techniques and our approach is impossible.

4.3. GS preconditioners

Even if the GS method converges, the convergence is rather slow. Therefore, we propose to use the GS method as a preconditioner in a Krylov subspace method. Let

$$\mathbf{A} = \mathbf{M} - \mathbf{N},$$

where $\mathbf{M} = \mathbf{D} - \mathbf{L}$ in the forward GS method, $\mathbf{M} = \mathbf{D} - \mathbf{U}$ in the backward GS method, and $\mathbf{M} = (\mathbf{D} - \mathbf{L})\mathbf{D}^{-1}(\mathbf{D} - \mathbf{U})$ in the forward–backward GS method (see, e.g., [19,35]). Left preconditioning

with the preconditioner $\mathbf{M}$ means that we apply a Krylov subspace method to the system

$$(\mathbf{I}-\mathbf{M}^{-1}\mathbf{N})\mathbf{x}=\mathbf{M}^{-1}\mathbf{A}\mathbf{x}=\mathbf{M}^{-1}\mathbf{b}. \tag{29}$$

The application of a Krylov subspace method to the preconditioned system requires the computation of

$$(\mathbf{I}-\mathbf{M}^{-1}\mathbf{N})\mathbf{x}-\mathbf{M}^{-1}\mathbf{b}=\mathbf{x}-\mathbf{M}^{-1}(\mathbf{N}\mathbf{x}+\mathbf{b})$$

with given $\mathbf{x}$ and, except in the initial iteration, $\mathbf{b}=\mathbf{0}$. The computation of $\mathbf{M}^{-1}(\mathbf{N}\mathbf{x}+\mathbf{b})$ for the forward, backward and forward–backward GS method can be performed using (26), (28). If $\mathbf{b}=\mathbf{0}$, then (26), (28) have to be executed with $\bar{y}_0=0, \tilde{f}_j=0, z_{1,j}=0, z_2=0$ in (23) and (24).

4.4. Parallelism and hierarchical GS

Let $\delta=0$. In the forward and backward GS sweep the solution $\bar{y}_{i+1}, \bar{u}_i, \bar{p}_i$ of the system (21b) depends only on $\bar{y}_{j+1}, \bar{u}_j, \bar{p}_j$ with $j=i\pm1$. Thus, we can solve in parallel the diagonal block systems (22) corresponding to even indices i and then we can solve in parallel the diagonal block systems (22) corresponding to odd indices i. This corresponds to a symmetric block permutation of system (22), that groups the blocks with even indices and the ones with odd indices. This approach is completely analogous to the red–black-ordering of linear systems arising from the discretization of PDEs. Of course, since the GS method depends on the ordering of the system, the red–black-ordering of (22) will influence the convergence behavior of the GS method.

The optimal control problems (24) corresponding to the block diagonal systems in (22) are of the same type as the original problem (1), (2). Hence, the solution approach discussed in this section can also be applied for the solution of the block diagonal systems in (22).

5. Numerical experiments

5.1. Neumann control of the one-dimensional heat equation

We consider the minimization problem

$$\min \frac{1}{2}\int_0^T u^2(t)\,\mathrm{d}t+\frac{\alpha_1}{2}\int_0^T\int_a^b (y(t,x)-z_1(t,x))^2\,\mathrm{d}x\,\mathrm{d}t+\frac{\alpha_2}{2}\int_0^1 (y(T,x)-z_2(x))^2\,\mathrm{d}x$$

governed by the one-dimensional linear heat equation

$$\begin{aligned}
&\frac{\partial}{\partial t}y(t,x)-\frac{\partial^2}{\partial x^2}y(t,x)=f(t,x), \quad t\in(0,T),\ x\in(0,1),\\
&\frac{\partial}{\partial x}y(t,0)=u(t), \quad t\in(0,T),\\
&\frac{\partial}{\partial x}y(t,1)=r(t), \quad t\in(0,T),\\
&y(0,x)=y_0(x), \quad x\in(0,1).
\end{aligned} \tag{30}$$

We set $H=L^2(0,1)$, $V=H^1(0,1)$, $U=L^2(0,T)$, and $Z=Z_T=L^2(a,b)$.

Table 1
Problem specifications

Case	α_1	α_2	$[a,b]$
1	10^3	10^3	$[0,1]$
2	10^3	0	$[0,1]$
3	10^3	0	$[0.7,1]$
4	0	10^3	$[0,1]$

This is a small example, that allows us to explore the spectra of the GS iteration matrices numerically.

For the spatial discretization we use piecewise linear finite elements on a uniform grid $x_j = (j-1)/(n_x-1)$, $j = 1,\ldots,n_x$. The observations $z_1(t,\cdot)$, $t \in [0,T]$, and z_2 are replaced by their interpolations. For the discretization in time we use the backward Euler method with step size $1/n_t$. We use uniform time subintervals $[T_i, T_{i+1}]$ of length $1/N_t = k/n_t$. In all cases $n_x = 11$, $n_t = 30$. Other problem parameters are specified in Table 1. The right-hand side, the given boundary data, and the initial conditions are chosen to be

$$f(t,x) = (4\pi^2(1-\mathrm{e}^{-t}) + \mathrm{e}^{-t})\sin(2\pi x), \quad r(t) = 2\pi(1-\mathrm{e}^{-t}), \quad y_0(x) = 0.$$

These data are chosen so that if $u = r$, then $y(t,x) = \sin(2\pi x)(1-\mathrm{e}^{-t})$ solves the state equation (30). The desired data are $z_1 = 1$ and $z_2 = 1$.

Cases 3 and 4 are expected to be more difficult for the GS (or the instantaneous control) method, since the observation region is on the right-hand side of the spatial interval, whereas control is applied at $x = 0$ (case 3), or because only final time observations are present, which need to be 'transmitted' to the control over the entire time horizon (case 4). Table 2 shows the spectral radii of the forward GS iteration matrices $\mathbf{M}^{-1}\mathbf{N} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}$ in Fig. 1 for varying time subdomains. We use $\rho = \delta = 0$. This table shows that the GS method alone rarely converges and that even if it converges the convergence tends to be slow, since the spectral radius of the iteration matrix tends to be close to one.

As Table 2 shows, the GS method fails to converge most of the time. However, a look at the spectrum of the GS iteration matrices $\mathbf{M}^{-1}\mathbf{N}$ for this example reveals that, while the largest absolute eigenvalue is usually greater than one, the absolute eigenvalues tend to go to zero rather fast. Fig. 3 shows the absolute eigenvalues for several GS iteration matrices for case 4 (again with $\delta = \rho = 0$). The corresponding plots for cases 1–3 were very similar and are not shown.

Table 2
Spectral radii

N_t	Case 1	Case 2	Case 3	Case 4
3	4.27e − 1	3.97e − 1	2.83e + 0	1.94e + 0
10	6.50e − 1	5.02e − 1	2.20e + 0	3.88e + 0
15	7.27e − 1	6.18e − 1	1.78e + 0	2.78e + 0
30	1.22e + 0	7.83e − 1	1.35e + 0	5.09e + 0

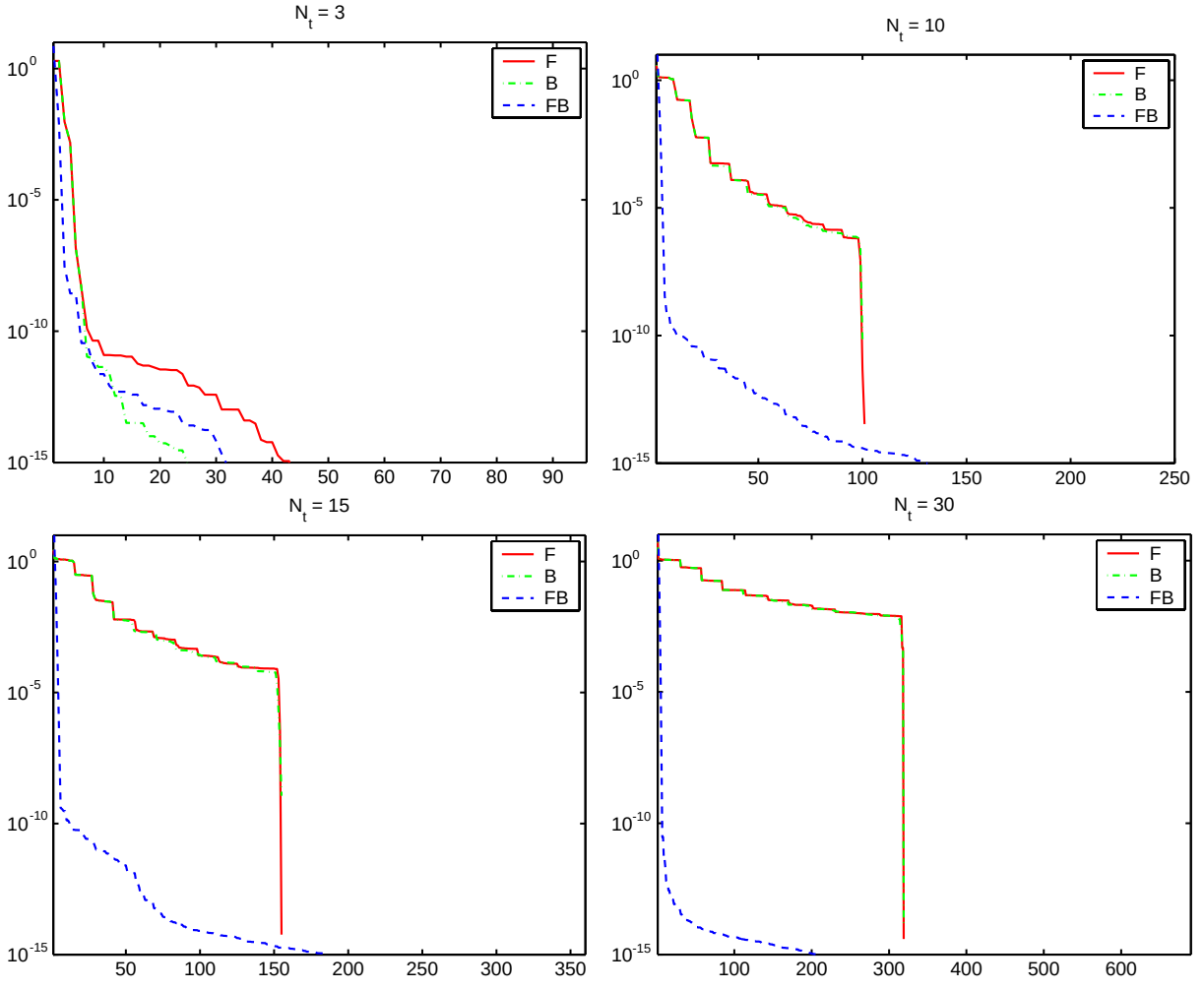

Fig. 3. Absolute eigenvalues of the forward (F), backward (B), and forward–backward (FB) GS Iteration matrices for Case 4.

For our discretization, the sizes of $\bar{y}_i$, $\bar{p}_i$ are n_x, the size of $\bar{u}_i$ is n_t/N_t. Theorems 7, 8 guarantee that at least $N_t n_x + n_t$ eigenvalues of the forward GS and backward GS iteration matrices are equal to zero. Hence, at most $N_t n_x$ eigenvalues of the forward GS and backward GS iteration matrices are nonzero. This is reflected in the plots of Fig. 3. There are few qualitative differences in the performance of forward GS and backward GS. It is remarkable, however, that the eigenvalues of the forward–backward (FB) GS iteration matrices decay fast, even if the time subinterval length is decreased, i.e., N_t is increased. The intuitive explanation for this observation is that in the forward sweep problem information is propagated from $t = 0$ to T via the state, whereas in the backward sweep problem information is propagated from $t = T$ to 0 via the adjoint. Since the FB-GS method combines both sweeps, information exchange is faster. A theoretical justification for the fast decay of eigenvalues is still missing. Because it is not clearly visible, we remark that the spectral

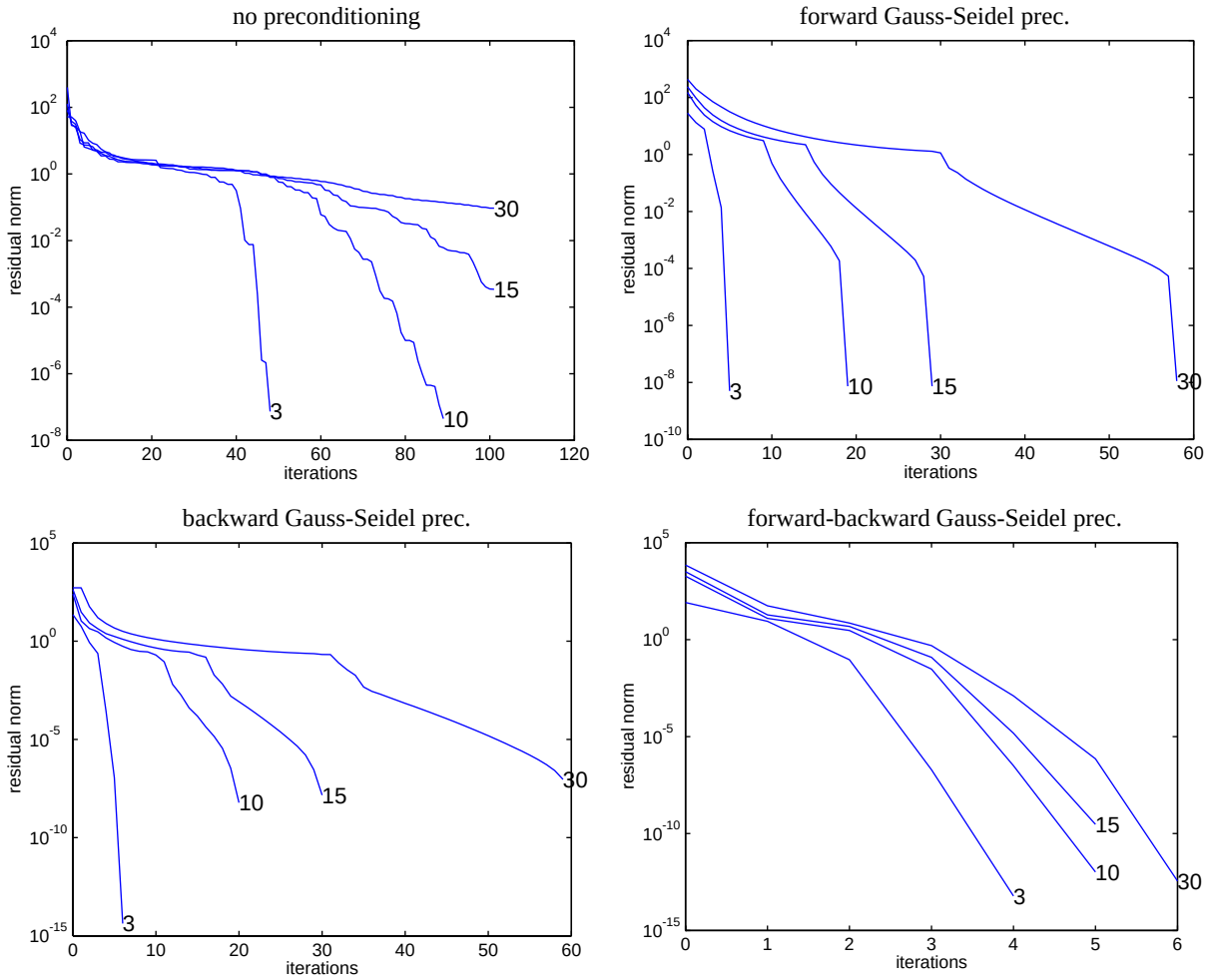

Fig. 4. Convergence history of GMRES (Example 1, case 4, $N_t = 3, 10, 15, 30$).

radii of all GS iteration matrices whose absolute eigenvalues are portrayed in Fig. 3 are larger than one.

If the eigenvalues of the GS iteration matrix $\mathbf{M}^{-1}\mathbf{N}$ cluster at zero, this means that the eigenvalues of $\mathbf{A}$ preconditioned with that GS method cluster at one (see [16]). Since roughly speaking the convergence of Kryov subspace methods tends to be the better the more the eigenvalues of the preconditioned system matrix cluster (in the case of GMRES see, e.g., [10,39]), it seems attractive to use the GS methods as preconditioners. We use forward (F), backward (B) and forward –backward (FB) GS as a preconditioner in GMRES. The GMRES iteration is truncated if the residual norm is less than 10^{-7} or if 100 iterations are performed. The results for case 4 are documented in Fig. 4. The numerical results for the other cases were similar and are therefore not displayed. We observe that GMRES applied to (22) fails to converge within the allowed number of iterations. GMRES with F-GS or B-GS preconditioner converges. However, we note that the number

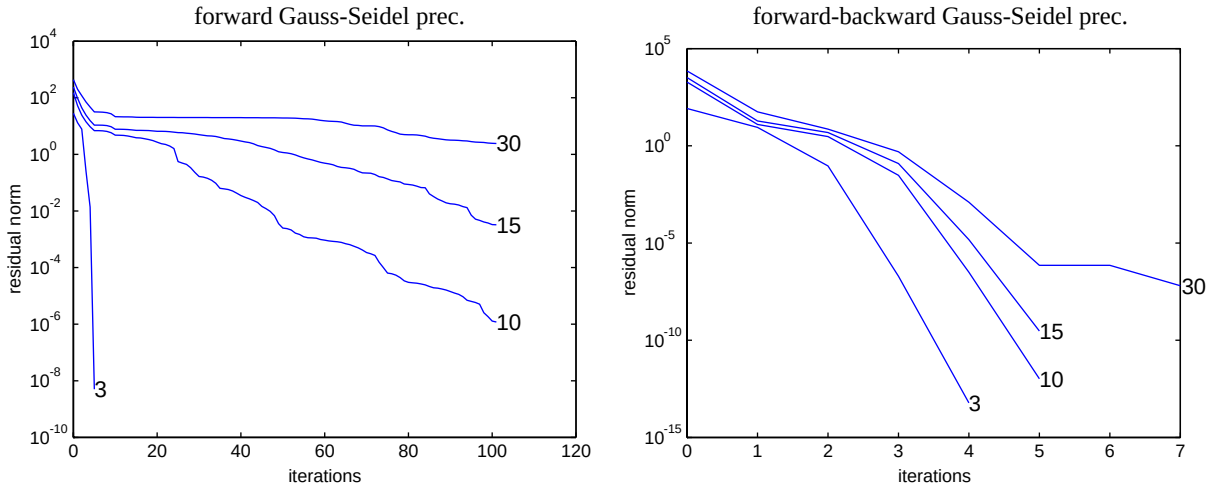

Fig. 5. Convergence history of GMRES(5) (Example 1, case 4, $N_t = 3, 10, 15, 30$).

of iterations increases as the number of time-subintervals increases. This is expected from Fig. 3. Fig. 3 shows that the absolute eigenvalues of the F-GS or B-GS iteration matrix decrease the slower the more time-subintervals we have. For the F-GS or B-GS preconditioner the number of preconditioned GMRES iterations seems to increase roughly linear with the number N_t of time-subintervals used in the preconditioner. If the FB-GS preconditioner is used, however, the number of preconditioned GMRES iterations is much lower than in all other cases and the number increases very little when the number N_t of time-subintervals increases. Thus, the FB-GS preconditioner is preferable, even though one preconditioning step is twice as expensive as the F-GS or the B-GS preconditioning step.

The amount of storage required in GMRES grows linearly with the number of iterations. Since for the target applications $\mathbf{x}$ is very large and storage is a severe bottleneck, we also use restarted GMRES(5). Without preconditioning GMRES(5) essentially stagnated; we do not show the results. The performance of GMRES(5) with left backward GS preconditioning is very similar to that of GMRES(5) with left forward GS preconditioning. Therefore, Fig. 5 contains only results for the latter. Fig. 5 shows that unless the number of subintervals are very small, GMRES(5) with the F-GS preconditioner fails to converge. In case of the FB-GS preconditioner the iterations for $N_t = 3, 10, 15$ convergence in five iterations and are therefore not effected by the restart. For $N_t = 30$, the restart slows down the rate of decrease of the residual norms, but GMRES(5) converges. For comparison, we also add the results obtained with BiCGStab, instead of GMRES. The storage requirements and the amount of work per iteration is constant in BiCGStab, but BiCGStab requires two matrix-vector products and two applications of the preconditioner per iteration. Fig. 6 shows that GMRES(5) combined with the FB-GS preconditioner is superior to BiCGStab combined with the FB-GS preconditioner. If only F-GS or B-GS preconditioning is used, BiCGStab converges, but the number of iterations required increases linearly with the number N_T of time intervals.

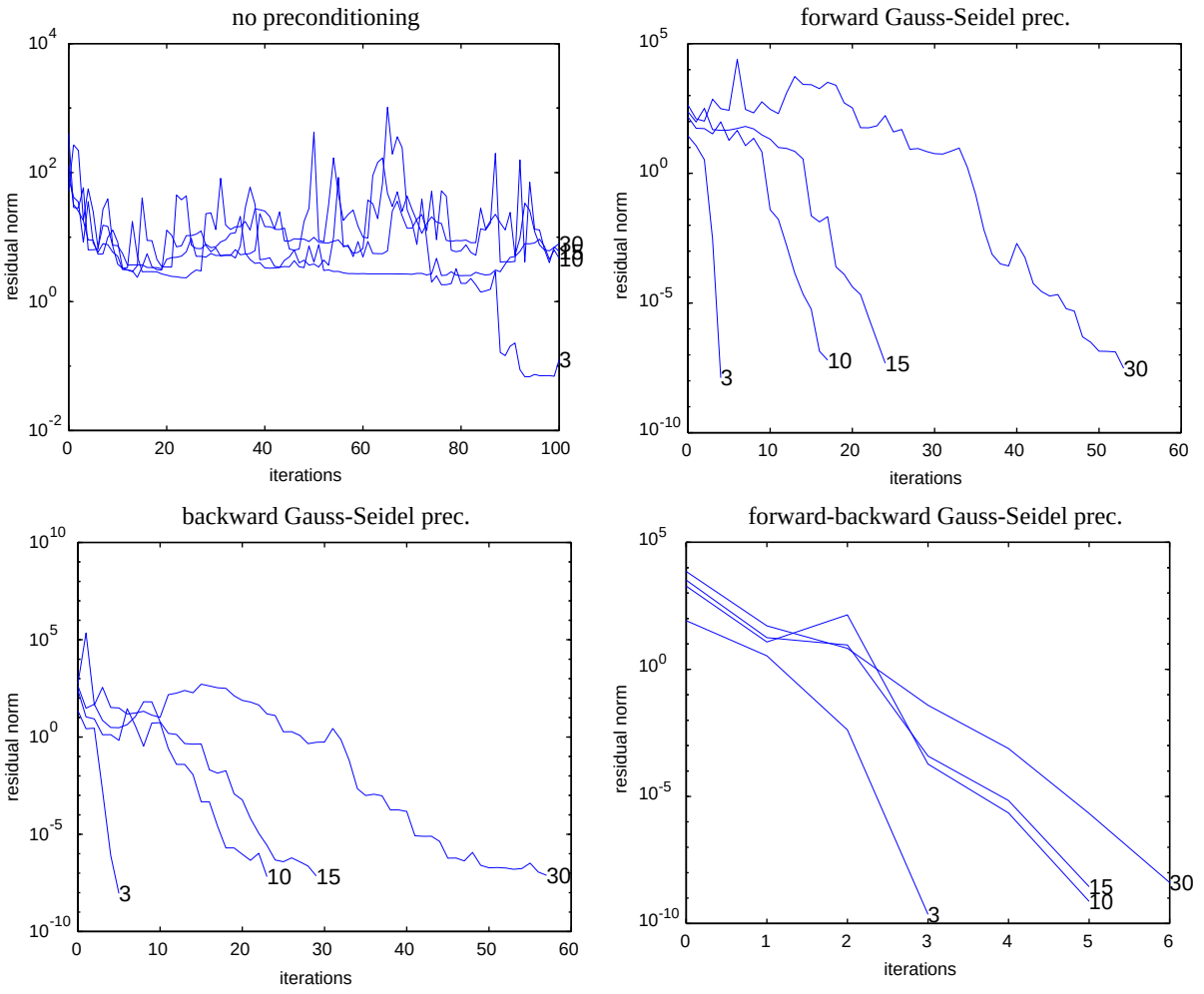

Fig. 6. Convergence history of BiCGStab (Example 1, case 4, $N_t = 3, 10, 15, 30$).

5.2. Dirichlet control of the two-dimensional heat equation

Our second example is

$$\min \frac{1}{2}\int_0^T \|u^2(t,\cdot)\|^2_{L^2(\Gamma)}\,\mathrm{d}t + \frac{\alpha_1}{2}\int_0^T \|(y(t,\cdot)-z_1(t,\cdot)\|^2_{L^2(\Omega)}\,\mathrm{d}t + \frac{\alpha_2}{2}\|y(T,\cdot)-z_2\|^2_{-1}$$

$$\text{s.t. } \frac{\partial}{\partial t}y(t,x) - \nu\Delta y(t,x) = f(t,x), \quad t\in(0,T),\ x\in\Omega,$$

$$y(t,x) = u(t,x), \quad t\in(0,T),\ x\in\Gamma_0,$$

$$y(t,x) = 0, \quad t\in(0,T),\ x\in\partial\Omega\setminus\Gamma_0,$$

$$y(0,x) = y_0(x), \quad x\in\Omega. \tag{31}$$

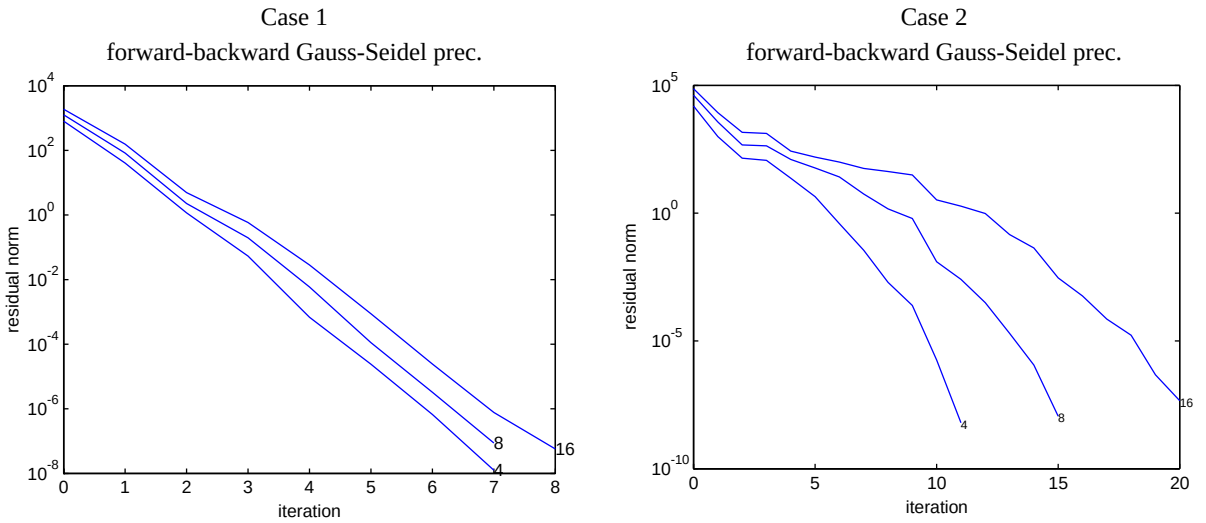

Fig. 7. Convergence history of GMRES (Example 2, $N_t = 4, 8, 16$).

Here $\Omega = (0,1)^2$ and $\|\cdot\|_{-1}$ denotes the norm in $H^{-1}(\Omega)$, which defined by $\|y\|_{-1} = \|\nabla\phi\|_{L^2}$, where $\phi \in H_0^1(\Omega)$ is the solution of $\int_\Omega \nabla\phi \cdot \nabla\theta = \langle y, \theta\rangle_{H^{-1}\times H_0^1}$ for all $\theta \in H_0^1(\Omega)$. The objective function with $\alpha_1 = 0$ is an approximate controllability problem studied in [11,18, Section 2]. For $u \in U = L^2(0,T;L^2(\partial\Omega))$ problem (31) has a unique solution y in $L^2(0,T;L^2(\Omega)) \cap C^0([0,T];(H^1(\Omega))^*)$ with y_t in $L^2(0,T;(H^2(\Omega))^*)$ (see [18, Section 2]). While this optimal control problem is well posed, it does not fit into the framework of Section 2. Therefore, we apply our domain decomposition based methods to the semi-discrete problem in which we first discretize optimal control problem in space.

Our spatial discretization of the problem follows [11,18, Section 2.6], who use linear finite elements on a uniform triangulation which is constructed by dividing Ω into n_x^2 squares of equal size and then cutting each square from the lower left to the top right into two triangles. In our computations $n_x = 16$. We use the backward Euler in time using $n_t = kN_t = 32$ time steps.

Our problem data are those of the second test problem in [18, p. 182]. In particular, $\Gamma_0 = (0,1) \times \{0\}$, $v = 1/(2\pi^2)$, $z_1(t,x_1,x_2) = \min\{x_1, x_2, 1-x_1, 1-x_2\}$, $z_2(x_1,x_2) = \min\{x_1, x_2, 1-x_1, 1-x_2\}$, $f = 0$, and $y_0 = 0$. We consider two cases. In case 1 we set $\alpha_1 = \alpha_2 = 10^5$ and in case 2 we set $\alpha_1 = 0, \alpha_2 = 10^5$.

Because of the problem size we did not compute the eigenvalues of the GS iteration matrices. Instead, we only solved (22) using GMRES with left preconditioning. The preconditioned GMRES iteration was truncated when the residual norm was less than 10^{-7}. Without preconditioning or even with forward GS or backward GS preconditioning the preconditioned GMRES did not converge within 100 iterations. The GMRES iterations with forward–backward GS preconditioning are documented in Fig. 7. For case 1 we observe at most a slight increase in GMRES iterations, when the number of time-subdomains is increased. For case 2, which is the problem in which only end-time observations are present, and which is significantly more ill-conditioned, the number of GMRES iterations seems to increase roughly linearly with the number of time-subdomains. The total number of preconditioned GMRES iterations, however, still remains reasonably small.

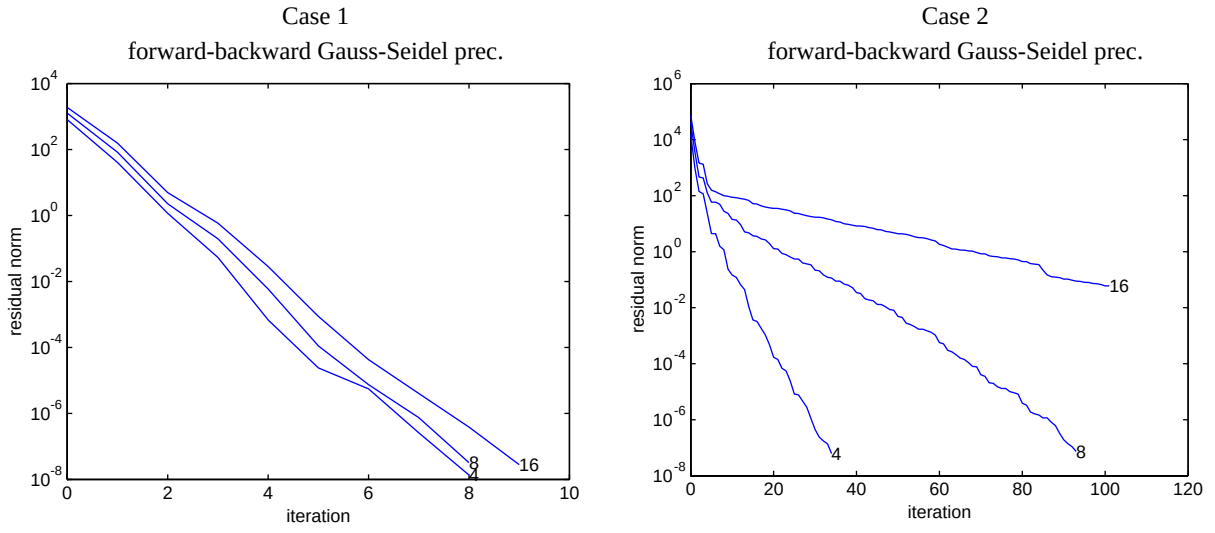

Fig. 8. Convergence history of GMRES(5) (Example 2, $N_t = 4, 8, 16$).

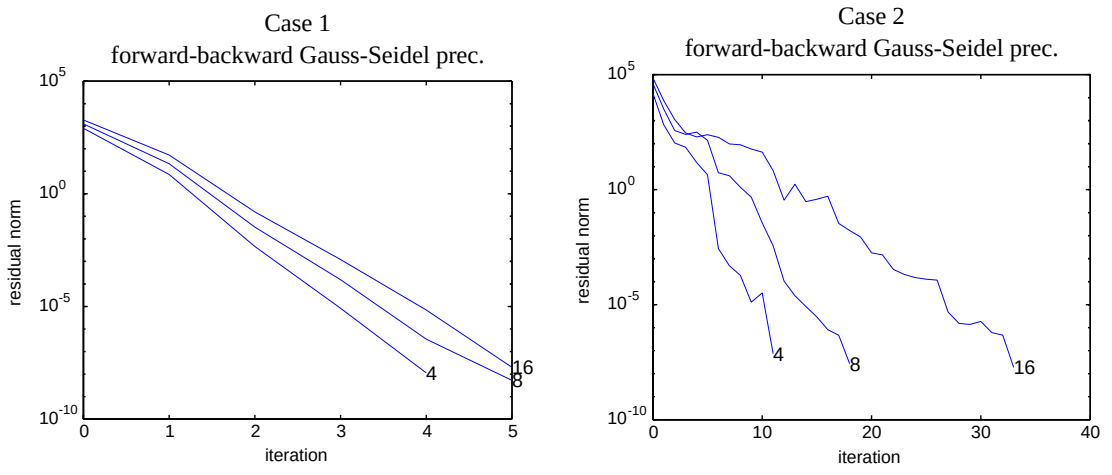

Fig. 9. Convergence history of BiCGStab (Example 2, $N_t = 4, 8, 16$).

Figs. 8 and 9 show the convergence plots for GMRES(5) and BiCGStab, respectively. For case 1, the number of iterations for GMRES(5) and BiCGStab with forward–backward GS preconditioning is almost independent of the number time subintervals. In terms of operator-vector products, all three methods, GMRES, GMRES(5) and BiCGStab are comparable. In case 2, the convergence properties of GMRES(5) deteriorate significantly when the number N_t of time subintervals is increased and it fails to converge within the maximum number of iterations, 100, for $N_t = 16$. In case 2, the number of BiCGStab, roughly grows linearly with N_t. In terms of operator-vector products GMRES outperforms BiCGStab by a factor of 2–3.5.

6. Conclusions and outlook

We have introduced a class of time-domain decomposition based methods for the solution of distributed linear quadratic optimal control problems (1), (2). These methods are derived from a multiple shooting-based reformulation of the distributed linear quadratic optimal control problem as a DTOC problem in Hilbert space. The optimality conditions for this DTOC problem lead to a linear system with block structure, which motivates the application of block GS methods for its solution. If the applications of the GS methods are stated in the framework of the original problem (1), (2), then they reveal an interesting connection between instantaneous control methods and the forward GS method. Instantaneous control methods can be interpreted as the application of one step of the forward GS method with starting values zero. However, our formulation also leads to different iterations, such as forward GS, backward GS, and forward–backward GS. Since convergence of the GS methods cannot be guaranteed for our class of problems, we propose to use them as preconditioners in a Krylov-subspace method. Numerical results show that the forward–backward GS preconditioner is vastly superior to forward GS and backward GS. Other orderings for the GS method, especially red–black orderings (Section 4.4) are interesting because they allow a parallel solution of the problem. We will explore this approach in the future.

Most of the material in Sections 3 and 4 can be generalized to problems with nonlinear state equations or nonquadratic objective functions. It can be shown that this leads to a nonlinear GS method and we can interpret instantaneous control methods as the application of one step of the forward nonlinear GS method with starting values zero. The nonlinear GS method has rather stringent convergence requirements (see [33]) and in our context we do not expect convergence in most cases. Moreover, for nonconvex problems we cannot simply solve the first-order necessary optimality conditions. Alternatively, we may apply Newton or sequential quadratic programming type methods. In each step of these methods a linear quadratic problem of type (1), (2) has to be solved and the techniques of this paper can be applied for this task.

Acknowledgements

The author's initial research on time-domain decomposition iterative methods for the solution of optimal control problems began jointly with Martin Berggren (Department of Scientific Computing, Uppsala University, and the Aeronautical Research Institute of Sweden). The paper [5] contains some ideas that are expanded on in the present paper. The author's interest in this topic was renewed by the work on instantaneous control of turbulent flows, which was brought to the author's attention by S. Scott Collis (Sandia National Laboratories). The author would like to thank Martin Berggren for the early discussions and Scott Collis for many stimulating interactions.

The careful reading of two anonymous referees is also greatly acknowledged.

References

[1] U.M. Ascher, R.M.M. Mattheij, R.D. Russel, Numerical Solution of Boundary Value Problems for Ordinary Differential Equations, in: Classics in Applied Mathematics, Vol. 13, SIAM, Philadelphia, 1995.

[2] O. Axelsson, Iterative Solution Methods, Cambridge University Press, Cambridge, London, New York, 1994.

[3] A. Bensoussan, G. Da Prato, M.C. Delfour, S.K. Mitter, Representation and Control of Infinite Dimensional Systems, Vol. I, Birkhäuser, Basel, Boston, Berlin, 1992.
[4] M. Berggren, Numerical solution of a flow-control problem: vorticity reduction by dynamic boundary action, SIAM J. Sci. Comput. 19 (1998) 829–860.
[5] M. Berggren, M. Heinkenschloss, Parallel solution of optimal-control problems by time-domain decomposition, in: M.-O. Bristeau, G. Etgen, W. Fitzgibbon, J.L. Lions, J. Periaux, M.F. Wheeler (Eds.), Computational Science for the 21st Century, Wiley, Chichester, 1997, pp. 102–112.
[6] J.T. Betts, Practical Methods for Optimal Control using Nonlinear Programming, Advances in Design and Control, SIAM, Philadelphia, 2001.
[7] T.R. Bewley, P. Moin, R. Temam, DNS-based predictive control of turbulence: an optimal benchmark for feedback algorithms, J. Fluid Mech. 447 (2001) 179–225.
[8] H.G. Bock, Randwertprobleme zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen, Preprint Nr. 442, Universität Heidelberg, Institut für Angewandte Mathematik, SFB 123, D-6900 Heidelberg, Germany, 1988.
[9] R. Bulirsch, Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblems und Aufgaben der optimalen Steuerung, Technical Report, Report of the Carl–Cranz Gesellschaft, 1971.
[10] S.L. Campbell, I.C.F. Ipsen, C.T. Kelley, C.D. Meyer, Z.Q. Xue, Convergence estimates for solution of integral equations with GMRES, J. Integral Equations Appl. 8 (1996) 19–34.
[11] C. Cartel, R. Glowinski, J.-L. Lions, On the exact and approximate boundary controllabilities for the heat equation: a numerical approach, J. Optim. Theory Appl. 82 (1994) 429–484.
[12] Y. Chang, Approximate models for optimal control of turbulent channel flow, Ph.D. Thesis, Department of Mechanical Engineering and Materials Science, Rice University, 2000.
[13] Y. Chang, S.S. Collis, Active control of turbulent channel flows by large eddy simulation, in: Proceedings of the 1999 ASME/JSME Joint Fluids Engineering Conference, San Francisco, CA, 1999.
[14] H. Choi, M. Hinze, K. Kunisch, Instantaneous control of backward-facing step flows, Appl. Numer. Math. IMACS J. 31 (2) (1999) 133–158.
[15] H. Choi, R. Temam, P. Moin, J. Kim, Feedback control for unsteady flow and its application to the stochastic Burgers equation, J. Fluid Mech. 253 (1993) 509–543.
[16] K. Deimling, Nonlinear Functional Analysis, Springer, Berlin, Heidelberg, New York, 1985.
[17] P. Deuflhard, Nonlinear equation solvers in boundary value problem codes, in: B. Childs (Ed.), Codes for the BVPs in ODEs, Springer Lecture Notes in Computer Science, Vol. 74, Springer, Berlin, 1979, pp. 40–66.
[18] R. Glowinski, J.-L. Lions, Exact and approximate controllability for distributed parameter systems, in: A. Iserles (Ed.), Acta Numerica 1995, Cambridge University Press, Cambridge, London, New York, 1995, pp. 159–333.
[19] A. Greenbaum, Iterative Methods for the Solution of Linear Systems, SIAM, Philadelphia, 1997.
[20] A. Griewank, Achieving logarithmic growth of temporal and spatial complexity reverse automatic differentiation, Optim. Methods Software 1 (1992) 35–54.
[21] M. Hinze, Optimal and instantaneous control of the instationary Navier–Stokes equations, Habilitation Thesis, Technical Report, Fachbereich Mathematik Technische Universität Berlin, Strasse des 17 Juni 136, D-10623 Berlin, Germany, 2000.
[22] M. Hinze, K. Kunisch, On suboptimal control strategies for the Navier–Stokes equations, in: ESIAM Proceedings, Control and Partial Differential Equations, Vol. 4, 1998, pp. 181–198.
[23] M. Hinze, S. Volkwein, Analysis of instantaneous control for the Burgers equation, Nonlinear Anal. Theory Methods Appl. Int. Multidisciplinary J. Ser. A Theory Methods 50 (1, Ser. A: Theory Methods) (2002) 1–26.
[24] L.S. Hou, Y. Yan, Dynamics and approximations of a velocity tracking problem for the Navier–Stokes flows with piecewise distributed controls, SIAM J. Control Optim. 35 (1997) 1847–1885.
[25] J. Jahn, Introduction to the Theory of Nonlinear Optimization, 2nd Edition, Springer, Berlin, Heidelberg, New York, 1996.
[26] J.E. Lagnese, G. Leugering, Time-domain decomposition of optimal control problems for the wave equation, System Control Lett. 48 (2003) 229–242.
[27] C. Lee, J. Kim, H. Choi, Suboptimal control of turbulent channel flow, J. Fluid Mech. 358 (1998) 245–258.
[28] D.B. Leineweber, I. Bauer, H.G. Bock, J.P. Schlöder, An efficient multiple shooting based reduced SQP strategy for large scale dynamic process optimization. Part I: theoretical aspects, Comput. Chem. Engrg. 27 (2003) 157–166.

[29] D.B. Leineweber, H. Schäfer, H.G. Bock, J.P. Schlöder, An efficient multiple shooting based reduced SQP strategy for large scale dynamic process optimization. Part II: software aspects and applications, Comput. Chem. Engrg. 27 (2003) 167–174.
[30] J.-L. Lions, Optimal Control of Systems Governed by Partial Differential Equations, Springer, Berlin, Heidelberg, New York, 1971.
[31] J.S. Logsdon, L.T. Biegler, Accurate determination of optimal reflux policies for the maximum distillate problem in batch distillation, Indust. Engrg. Chem. Res. 32 (1993) 692–700.
[32] H.J. Oberle, W. Grimm, BNDSCO—a program for the numerical solution of optimal control problems, Technical Report, Institute for Flight Systems Dynamics, DLR, Oberpfaffenhofen, Germany, 1989.
[33] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.
[34] J.M. Restrepo, G.K. Leaf, A. Griewank, Circumventing storage limitations in variational data assimilation studies, SIAM J. Sci. Comput. 19 (1998) 1586–1605.
[35] Y. Saad, Iterative Methods for Sparse Linear Systems, PWS Publishing Company, Boston, New York, Singapore, Toronto, 1996.
[36] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, 2nd Edition, Springer, New York, Berlin, Heidelberg, London, Paris, 1993.
[37] A. Unger, F. Tröltzsch, Fast solution of optimal control problems in the selective cooling of steel, Z. Angew. Math. Mech. (ZAMM) 81 (2001) 447–456.
[38] R.S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[39] H. Vorst, C. Vuik, The superlinear convergence behaviour of GMRES, J. Comput. Appl. Math. 48 (1993) 327–341.
[40] E. Zeidler, Nonlinear Functional Analysis and its Applications II/A: Linear Monotone Operators, Springer, Berlin, Heidelberg, New York, 1990.