Block iterative algorithms for the solution of parabolic optimal control problems

Christian E. Schaerer[1], Tarek Mathew[1], and Marcus Sarkis[1]

Instituto de Matemática Pura e Aplicada- IMPA,
Estrada Dona Castorina 110, CEP: 22460-320, Rio de Janeiro, Brazil.
cschaer@fluid.impa.br, tmathew@fluid.impa.br, msarkis@impa.br

Abstract. We consider block iterative methods for the solution of *large scale* linear-quadratic optimal control problems arising from the control of parabolic partial differential equations over a finite control horizon. This paradigm models new production strategies in oil and gas fields. To simulate the behavior in a reservoir under different scenarios, an optimal control problem can be formulated based on the constituent equations. After spatial discretization by finite element or finite difference methods, such problems typically require the optimal control of n coupled ordinary differential equations, where n can be quite large. Its solution by conventional methods can be prohibitively expensive in terms of computational cost and memory requirements.

We describe two iterative algorithms. The first algorithm employs a CG method to solve a symmetric positive definite reduced linear system for the unknown control variable. A preconditioner is described, which we prove yields a rate of convergence independent of the space and time discretization parameters, however, double iteration is required. A second algorithm is designed to avoid double iteration by introducing an auxiliary variable. It yields a symmetric indefinite system, and for this system a positive definite block preconditioner is described. We prove a rate of convergence independent of the space and time discretization parameters when MINRES acceleration is used. Numerical results are presented for test problems.

1 Introduction

Systems governed by parabolic partial differential equations arise in the modeling of various processes in the oil industry. An instance is the processes whose main objective is the displacement of a resident fluid (oil) by the injection of another fluid (gas) [14]. The associated equation for the pressure is parabolic. In this context, recent works have demonstrated that control strategies based on Optimal Control Theory (OCT) can potentially increase the production in oil and gas fields [14]. In addition, the efficiency of the OCT model makes it suitable for application to real reservoirs simulated using large scale models, in contrast to many existing techniques [12]. The main bottleneck in this approach, however, is the need to find a fast simulator to test all the necessary scenarios to decide an adequate strategy for each reservoir.

Our purpose in this paper, is to study iterative algorithms for the solution of finite time *linear-quadratic optimal control* problems governed by a parabolic partial differential equation. Such problems are computationally intensive and require the minimization of some quadratic objective functional $J(\cdot)$ (representing some cost to be minimized over time), subject to linear constraints given by a *stiff* system of n ordinary differential equations, where n is typically quite large. An application of the Pontryagin maximum principle to determine the optimal solution, see [9], results in a Hamiltonian system of ordinary differential equations, with initial and final conditions. This system is traditionally solved by reduction to a matrix Riccati equation for an unknown matrix function $P(t)$ of size n, on an interval $[0,T]$, see [9, 7, 11]. Solving the Riccati equation, and storing matrix $P(t)$ of size n on a time interval $[0,T]$ can become prohibitively expensive for large n. Instead, motivated by the *parareal* algorithm (of Lions, Maday and Turinici [6]) and iterative shooting methods in the control context [4, 13], we propose iterative algorithms for such control problems.

We formulate iterative algorithms for the parabolic optimal control problem based on a *saddle point* formulation [11]. We consider finite difference (or finite element) discretizations of the parabolic equation in space, and the θ-scheme in time. The cost functional is discretized in time using a trapezoidal or midpoint rule for the state variable and piecewise constant in time and space for the control variable. Lagrange multipliers (adjoint) variables are introduced to enforce the constraints, and the saddle point linear system is formulated for the optimal solution. Inspired by the reduction approach employed in [11] for elliptic control problems, we develop two algorithms whose rate of convergence does not deteriorate as the mesh parameters become small. The first algorithm uses a CG method to solve a symmetric positive definite reduced linear system for determining the unknown *control variable.* We show under specific assumptions that the resulting system has a condition number independent of the mesh parameters. For the second algorithm, we expand the reduced system consistently by introducing an auxiliary variable. We describe a block preconditioned algorithm using a MINRES method on the auxiliary and control variables. We analyze the convergence rates of these two proposed iterative algorithms.

Our discussion is organized as follows. In Section 2, we introduce the optimal control problem for the parabolic problem. In Section 3, we introduce the finite dimensional linear-quadratic optimal control problem. We also introduce the saddle point system obtained by a stable discretization of the parabolic control problem. In Section 4, we describe the preconditioners and theoretical results that justify the efficiency of the proposed methods. Finally, in Section 5, numerical results are presented which show that the rate of convergence of both proposed algorithms is independent of the space and time discretization.

2 The optimal control problem

Let $\mathcal{A}$ denote an operator from a space $L^2(t_o, t_f; Y)$ to $L^2(t_o, t_f; Y')$, where Y is a Hilbert space (in our case $Y = H_0^1(\Omega)$). The norm is the H-norm where H is a pivot Hilbert space with $Y \subset H \subset Y'$ and $H = L^2(\Omega)$. We consider the following state equation with $z(t) \in Y$:

$$\begin{cases} \partial_t z + \mathcal{A} z = \mathcal{B} v, \ \text{ for } t_0 < t < t_f \\ \qquad z(0) = z_o, \ \text{ on } \ \Omega, \end{cases} \tag{1}$$

where $z(\cdot) \in Y$ is known as the state variable and the operator $\mathcal{A}$ is coercive. The distributed control $v(\cdot)$ belongs to an admissible space $\mathcal{U} = L^2(t_o, t_f; \Omega)$ and $\mathcal{B}$ is an operator in $\mathcal{L}(\mathcal{U}, L^2(t_o, t_f; Y'))$. We assume that for each $v(\cdot)$, this problem is well posed; therefore we emphasize the dependence of z on $v \in \mathcal{U}$ using the notation $z(v)$. We associate the following cost function with the state equation (1):

$$\begin{aligned} J(z(u), u) := {} & \frac{q}{2} \| z(v) - z_* \|^2_{L^2(t_o, t_f; L^2(\Omega))} + \frac{r}{2} \| v \|^2_{L^2(t_o, t_f; \Omega)} \\ & + \frac{s}{2} \| z(v)(t_f, x) - z_*(t_f, x) \|^2_{L^2(\Omega)}, \end{aligned} \tag{2}$$

where z_* is a given target. The optimal control problem for equation (1) consists of finding a controller $u \in \mathcal{U}$ which minimizes the cost function (2):

$$u = \text{argmin }_{v \in \mathcal{U}} J(z(v), v). \tag{3}$$

Since the terms $\frac{r}{2} \| v \|_{L^2(t_o, t_f; \Omega)} > 0$ and $\frac{q}{2} \| z(v) - z_\star \|_{L^2(t_o, t_f; L^2(\Omega))} \geq 0$ in the cost function (2), for $r > 0$ and $q > 0$, following [7], the optimal control (3) is well posed. To discretize state equation (1) we apply the finite element method to its weak formulation for every fixed $t \in (t_o, t_f)$. Hence, $z \in L^2(t_o, t_f; Y))$ is a weak solution of (1) provided its weak derivative $\dot{z} \in L^2(t_o, t_f; Y'))$ and

$$(\dot{z}(t), \eta) + (\mathcal{A} z(t), \eta) = (\mathcal{B} u(t), \eta) \quad \text{for all } \eta \in Y \text{ and } t \in (t_o, t_f). \tag{4}$$

In what follows, the form $(\mathcal{A} z, \eta)$ is assumed to be continuous on $Y \times Y$ and Y-elliptic. So let $Y_h(\Omega) \subset Y = H_o^1(\Omega)$ and let $z_{ho} \in Y_h$ be a good approximation for $z(t_o)$, the $L^2(\Omega)$-projection for instance. The bilinear form $(\mathcal{B} u, \eta)$ is assumed to be continuous on $U \times Y$. So let $U_h(\Omega) \subset U$ be a subspace for approximating u. Then the semi-discretization is given by

$$(\dot{z}_h(t), \eta_h) + (\mathcal{A} z_h(t), \eta_h) = (\mathcal{B} u_h(t), \eta_h) \quad \text{for all } \eta_h \in Y_h \text{ and } t \in (t_o, t_f), \tag{5}$$

$$z_h(t_o) = z_{ho}. \tag{6}$$

Let $\{\phi_1(x), ..., \phi_n(x)\}$ a basis of Y_h and $\{\varphi_1(x), ..., \varphi_m(x)\}$ a basis of U_h. Consequently, $z_h(t) = \sum_{j=1}^n \phi_j(x) \xi_j(t)$ and $u_h(t) = \sum_{j=1}^m \varphi_j(x) \mu_j(t)$. Then, for any $t \in (t_o, t_f)$, the discrete variational equality (5) is equivalent to

$$\sum_{j=1}^n (\phi_j, \phi_i) \, \dot{\xi}_j(t) + \sum_{j=1}^n (\mathcal{A} \phi_j, \phi_i) \, \xi_j(t) = \sum_{j=1}^m (\mathcal{B} \varphi_j, \phi_i) \, \mu_j(t) \quad \text{for all } i \in \{1, .., n\}.$$

Denoting by $\hat{A}_h := (\mathcal{A}\phi_j, \phi_i)_{i,j}$, $\hat{M}_h := (\phi_j, \phi_i)_{i,j}$, $\hat{B}_h = (\mathcal{B}\varphi_j, \phi_i)_{i,j}$, $\xi = (\xi_j(t))_j$, $\mu = (\mu_j(t))_j$ and $\xi_o = \xi(t_0)$. We obtain the following system of ordinary differential equations:

$$\hat{M}_h\dot{\xi} + \hat{A}_h\xi = \hat{B}_h\mu, \ \ t \in (t_o, t_f) \ \text{ and } \xi(t_o) = \xi_o. \tag{7}$$

By analogy with the spatial discretization of the state equation (1), the spatial discretization of the functional (2) is:

$$\begin{aligned} J_h(\xi, u) = \frac{q}{2}\int_{t_o}^{t_f} & (\xi - \xi_*)^T(t)\hat{M}_h(\xi - \xi_*)(t) \\ +\frac{r}{2}\int_{t_o}^{t_f} & u^T(t)R_h u(t) \\ & +\frac{s}{2}(\xi - \xi_*)^T(t_f)\hat{M}_h(\xi - \xi_*)(t_f), \end{aligned} \tag{8}$$

where both R_h and $\hat{M}_h$ are mass matrices. We assume that $z(u)$ is discretized using a piecewise linear function while u is discretized using discontinuous piecewise constant functions. Since the matrix $\hat{M}_h$ is symmetric positive definite, we factorize as $\hat{M}_h = U_h^T U_h$ and introduce new variables $y = U_h\xi$ and $u = \mu$, then the functional (8) takes the form:

$$\begin{aligned} J_h(y, u) = \frac{q}{2}\int_{t_o}^{t_f} (y - y*)^T(y - y_*) + \frac{r}{2}\int_{t_o}^{t_f} u^T R_h u \\ +\frac{s}{2}(y - y_*)^T(y - y_*)(t_f), \end{aligned} \tag{9}$$

and the state equation (7) is reduced to:

$$\begin{cases} \quad \dot{y} = A\,y + B\,u, \ \ t \in (0, t_f) \\ y(t_o) = y_0, \end{cases} \tag{10}$$

where $A := U_h^{-T}\hat{A}_h U_h^{-1}$ and $B := U_h^{-T}\hat{B}_h$.

In summary, spatial discretization transforms the constraints (1) into a system of n linear ordinary differential equations (10), where $y(\cdot) \in \mathbb{R}^n$ denotes state space variables having initial value y_0, while $u(\cdot) \in \mathbb{R}^m$ denotes control variables. Although, A, B are $n \times n$ and $n \times m$ matrix functions, respectively, we consider them as time-invariants and given by a symmetric and *negative* definite matrix A of size n (with n large). In the case $\mathcal{A} = -\Delta$, matrix A will correspond to a discrete Laplacian, and its eigenvalues will lie in an interval $[-c, -d]$ where $c = O(h^{-2})$ and $d = O(1)$ (for grid size h in the spatial discretization of the parabolic equation).

The discrete optimal control problem seeks $y(\cdot) \in \mathbb{R}^n$ and $u(\cdot) \in \mathbb{R}^m$ satisfying (10) and *minimizing* a non-negative quadratic cost functional $J(.,.)$, possibly more general than (9), given by:

$$\begin{cases} J(y, u) \equiv \int_{t_o}^{t_f} l(y, u)\,dt + \psi(y(t_f)), \quad \text{where} \\ l(y, u) \equiv \frac{1}{2}\left(e(t)^T Q(t)e(t) + u(t)^T R(t)u(t)\right), \\ \psi(y(t_f)) \equiv \frac{1}{2}\left(y(t_f) - y_*(t_f)\right)^T C\left(y(t_f) - y_*(t_f)\right), \end{cases} \tag{11}$$

where $e(t) := y(t) - y_*(t)$, Q is an $n \times n$ symmetric positive semi-definite matrix function, $y_*(\cdot) \in \mathbb{R}^n$ is a given *tracking* function, C is an $n \times n$ symmetric positive semidefinite matrix, and R is an $m \times m$ symmetric positive definite matrix function. The linear-quadratic optimal control problem, thus, seeks the minimum of $J(\cdot)$ in (11) subject to the constraints (10). Given the tracking function $y_*(\cdot)$, the optimal control $u(\cdot)$ must ideally yield $y(\cdot)$ "close" to $y_*(\cdot)$.

3 The basic saddle point system

In this section, we consider stable time-discretization of the optimal control problem given by:

$$\min\, J(y,u), \tag{12}$$

subject to

$$\begin{cases} \dot{y} = A\,y + B\,u, & \text{for } \; t_o < t < t_f \\ y(t_o) = y_0, \end{cases} \tag{13}$$

where $J(y,u)$ is defined in (11) with matrices Q, R and S being time invariant. We discretize the time domain $t \in [t_o,\, t_f]$ using $(l-1)$ interior grid points, so that the time step is $\tau = (t_f - t_o)/(l)$ with $t_i = i\tau$. The state variable y at the time t_i is denoted by $y_i := y(t_i)$. We assume that the controller u is constant on each interval $(t_i,\, t_{i+1}]$ with the value $u_{i+1/2} = u(t_{i+1/2})$. Hence, a stable discretization of equation (13) using the θ-scheme can be written as:

$$F_1 y_{i+1} = F_0 y_i + \tau B u_{i+1/2}, \;\; y_0 = y(t_o), \;\; i = 0, 1, ..., l-1, \tag{14}$$

where $F_1, F_0 \in \Re^{n\times n}$ are (constant) matrices given by $F_0 := I + \tau(1-\theta)A$ and $F_1 := I - \tau\theta A$. Using a full discretization in time, equation (13) takes the matrix form:

$$E\mathbf{y} + N\mathbf{u} = \mathbf{f}, \tag{15}$$

where the discrete state vector $\mathbf{y} \in \Re^{nl}$ and control vector $\mathbf{u} \in \Re^{ml}$ are:

$$\mathbf{y} := [y_1, \ldots, y_l]^T \quad \text{and} \quad \mathbf{u} := [u_{1/2}, \ldots, u_{l-1/2}]^T, \tag{16}$$

respectively. The input vector $\mathbf{f} \in \Re^{nl}$ is given by $\mathbf{f} := [-F_0 y_o, 0, ..., 0]^T$, and the matrices $E \in \Re^{(nl)\times(nl)}$ and $N \in \Re^{(ml)\times(ml)}$ have the following block structure:

$$E := \begin{bmatrix} -F_1 & & & \\ F_0 & -F_1 & & \\ & \ddots & \ddots & \\ & & F_0 & -F_1 \end{bmatrix} \quad \text{and} \quad N := \tau \begin{bmatrix} B & & \\ & \ddots & \\ & & B \end{bmatrix}. \tag{17}$$

The discretization of the performance functional $J(y,u)$ takes the form:

$$J_h(y,u) \equiv \frac{1}{2}(\mathbf{u}^T G \mathbf{u}^T + \mathbf{e}^T Z \mathbf{e} + e^T C e(t_f)), \tag{18}$$

where vector $\mathbf{e} \in \Re^{nl}$ is defined in terms of the discrete error at time t_i defined as $e_i := y(i\tau) - y_*(i\tau)$ for $i = 1, ..., l$. Hence, the discrete error vector is defined as $\mathbf{e} := [e_1^T, \ldots, e_l^T]^T$. In the numerical experiments, we consider matrix G to be diagonal since we approximate the controller using piecewise constant functions in time and also in space. The error $\mathbf{e}$ is approximated using piecewise linear functions in both time and space, hence matrix Z is block tri-diagonal where each block is matrix $\hat{M}_h$. The discrete Lagrangian $\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p})$ has the matrix form:

$$\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p}) = \frac{1}{2}(\mathbf{u}^T G \mathbf{u}^T + \mathbf{e}^T K \mathbf{e}) + \mathbf{p}^T (E\mathbf{y} + N\mathbf{u} - \mathbf{f}), \tag{19}$$

where K is defined as $K := Z + \varGamma$ and $\varGamma = diag(0, 0, ..., 0, C)$. To obtain a discrete saddle formulation of (19), optimality conditions analogous to Section 2 are used. Hence, the discrete saddle point system has the matrix form:

$$\begin{bmatrix} K & 0 & E^T \\ 0 & G & N^T \\ E & N & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} K\mathbf{g} \\ 0 \\ \mathbf{f} \end{bmatrix} \tag{20}$$

where $\mathbf{g} := [g_i]$, $g_i = y_*(i\tau)$. In the following, we study the condition number of the evolution matrix E.

Theorem 1. *Let matrix A be a $n \times n$ symmetric negative definite with eigenvalues $\lambda_i(A)$ for $1 \leq i \leq n$ and let the evolution matrix E be as defined in (17) with matrices F_0 and F_1 given by:*

$$F_0 := I + \tau(1-\theta)A \quad \textit{and} \quad F_1 := I - \tau\theta A \tag{21}$$

for $0 \leq \theta \leq 1$. Then, there are no stability restrictions on τ when $\theta \geq \frac{1}{2}$, while if $\theta < \frac{1}{2}$, then (14) will be stable only if $\tau \leq 2/\left((1-2\theta)\rho_{max}\right)$. In addition matrix EE^T has the condition number:

$$\operatorname{cond}(EE^T) \leq \frac{4\left(1 + \tau\theta\rho_{\max}\right)^2}{(\tau\rho_{\min})^2} \tag{22}$$

where $\rho_{\max} := \max \mid \lambda_i \mid$ and $\rho_{\min} := \min \mid \lambda_i \mid$.

Proof. Part 1. Consider the marching scheme for equation (1) given by:

$$y_{k+1} = \varPhi y_k + F_1^{-1} \tau B u \tag{23}$$

where $\varPhi$ is the marching matrix given by

$$\varPhi := (I - \tau\theta A)^{-1}(I + \tau(1-\theta)A). \tag{24}$$

The stability condition for (23) is given by

$$\mid (1 - \tau\theta\lambda_i)^{-1}(1 + \tau(1-\theta)\lambda_i) \mid \leq 1 \tag{25}$$

or equivalently,

$$1 + \tau(1-\theta)\lambda_i \leq 1 - \tau\theta\lambda_i \tag{26}$$

$$-1 - \tau(1-\theta)\lambda_i \leq 1 - \tau\theta\lambda_i. \tag{27}$$

From (26), we obtain $\tau\lambda_i \leq 0$ and $\tau \mid \lambda_i \mid (1-2\theta) \leq 2$ since $\lambda_i < 0$. In the case $\theta \geq 1/2$, there is no restriction on τ, consequently the marching scheme is unconditionally stable. On other hand, if $\theta < 1/2$ then $0 < (1-2\theta)$ and in order for the scheme to be stable it is necessary that $\tau \leq 2/\left((1-2\theta)\rho_{\max}\right)$. In this case, the marching scheme is conditionally stable.

Part 2. To analyze the condition number of matrix EE^T, we consider the structure of the block matrix given by:

$$EE^T = \begin{bmatrix} F_1F_1^T & -F_1F_0^T & & & \\ -F_0F_1^T & F_0F_0^T + F_1F_1^T & -F_1F_0^T & & \\ & -F_0F_1^T & F_0F_0^T + F_1F_1^T & -F_1F_0^T & \\ & & \ddots & \ddots & \ddots \\ & & & -F_0F_1^T & F_0F_0^T + F_1F_1^T \end{bmatrix}. \tag{28}$$

We define $Q^TAQ = \Lambda = \operatorname{diag}(\lambda_i)$ as the eigendecomposition of A where matrix $Q = [q_1, \ldots, q_n]$ is orthogonal. The diagonalization of F_0 and F_1 using Q is denoted $\Lambda_0 = Q^TF_0Q = Q^T(I - \tau\theta A)Q$ and $\Lambda_1 = Q^TF_1Q = Q^T(I + \tau(1-\theta)A)Q$, respectively. Then, the block sub-matrices of EE^T are diagonalized and EE^T can be expressed in the form:

$$EE^T \to \begin{bmatrix} \Lambda_1^2 & -\Lambda_0\Lambda_1 & & & \\ -\Lambda_0\Lambda_1 & \Lambda_0^2 + \Lambda_1^2 & -\Lambda_1\Lambda_0 & & \\ & -\Lambda_0\Lambda_1 & \Lambda_0^2 + \Lambda_1^2 & -\Lambda_1\Lambda_0 & \\ & & \ddots & \ddots & \ddots \\ & & & -\Lambda_0\Lambda_1 & \Lambda_0^2 + \Lambda_1^2 \end{bmatrix}. \tag{29}$$

We next permute matrix (29) by a matrix P by ordering the eigenvalues to obtain blocks with structure:

$$\Theta_i := (PEE^TP^T)_i = \begin{bmatrix} a_i^2 & -a_ib_i & & & \\ -a_ib_i & a_i^2 + b_i^2 & -a_ib_i & & \\ & -a_ib_i & a_i^2 + b_i^2 & -a_ib_i & \\ & & \ddots & \ddots & \ddots \\ & & & -a_ib_i & a_i^2 + b_i^2 \end{bmatrix}, \tag{30}$$

where $b_i := (1 + \tau(1-\theta)\lambda_i)$ and $a_i := (1 - \tau\theta\lambda_i)$. Gershgorin Theorem [2] yields:

$$\mid \mu(\Theta_i) - a_i^2 \mid \leq \mid a_ib_i \mid \quad \text{or} \quad \mid \mu(\Theta_i) - a_i^2 - b_i^2 \mid \leq 2 \mid a_ib_i \mid \tag{31}$$

Using condition (25), we guarantee stability when $\mid b_i \mid \leq \mid a_i \mid$ obtaining

$$\mu(\Theta_i) \leq \max\left(\mid a_i \mid (\mid a_i \mid + \mid b_i \mid), (\mid a_i \mid + \mid b_i \mid)^2\right) \leq \max 4 \mid a_i \mid^2 \tag{32}$$

and

$$\mu(\Theta_i) \geq \min\left((\mid a_i \mid^2 - \mid a_i \mid\mid b_i \mid), (\mid a_i \mid - \mid b_i \mid)^2\right) \geq \min(\mid a_i \mid - \mid b_i \mid)^2 \tag{33}$$

To obtain an upper bound for $\mu(\Theta_i)$ from (32), we define $\rho_{\max} := \max \mid \lambda_i \mid$, therefore we have $\mu(\Theta_i) \leq 4\,(1+\tau\theta\rho_{\max})^2$. To obtain a lower bound for $\mu(\Theta_i)$, from (33) we define $\rho_{\min} := \min \mid \lambda_i \mid$ obtaining $\mu(\Theta_i) \geq (\tau\rho_{\min})^2$. Therefore, the condition number for the matrix EE^T in terms of the upper and lower bound is given by:

$$\operatorname{cond}(EE^T) \leq 4\left(\frac{1+\tau\theta\rho_{\max}}{\tau\rho_{\min}}\right)^2. \tag{34}$$

This completes the proof.

Remark. Notice that for both finite difference or finite element discretizations on a domain of size $O(1)$, the eigenvalues of matrix A satisfies the bounds $\alpha_1 \leq |\lambda_i(A)| \leq \alpha_2 h^{-2}$ Then using (34) we obtain:

$$\operatorname{cond}(EE^T) \approx \left(\frac{1+\tau\theta\alpha_2 h^{-2}}{\tau\alpha_1}\right)^2. \tag{35}$$

Therefore, matrix E is ill-conditioned $O(h^{-4})$ when τ and h are refined. To solve system (20) using Uzawa's method, it is necessary to solve at each iteration $-(EK^{-1}E^T + NG^{-1}N^T)\mathbf{p} = \mathbf{f} - E\mathbf{g}$. Matrix $S := (EK^{-1}E^T + NG^{-1}N^T)$ is the Schur complement of system (20) with respect to the Lagrange multiplier $\mathbf{p}$.

Next, we analyze the condition number of S. Notice that due to the positive semi-definiteness of matrix C in (18), we obtain in the sense of quadratic forms that $K^{-1} = (Z + \Gamma)^{-1} \leq Z^{-1}$ and apply it in the following estimate for the condition number of the Schur complement S. Henceforth, we normalize $q = 1$.

Lemma 1. *Let the upper and lower bound for the singular values of EE^T be given by $4(1+\tau\theta\rho_{max})^2$ and $(\tau\rho_{min})^2$, respectively. Let us assume, using (9) and (10), that the mass matrices Z, G, N, and Γ satisfy*

$$\begin{aligned}
c_1\tau\mathbf{y}^T\mathbf{y} \leq \quad & \mathbf{y}^T Z\mathbf{y} \quad \leq c_2\tau\mathbf{y}^T\mathbf{y} && (36)\\
c_3 r\tau h^d\mathbf{u}^T\mathbf{u} \leq \quad & \mathbf{u}^T G\mathbf{u} \quad \leq c_4 r\tau h^d\mathbf{u}^T\mathbf{u}, && (37)\\
c_5\tau^2 h^d\mathbf{p}^T\mathbf{p} \leq\ & \mathbf{p}^T NN^T\mathbf{p} \leq c_6\tau^2 h^d\mathbf{p}^T\mathbf{p} \quad \textit{and} && (38)\\
0 \leq \quad & \mathbf{y}^T\Gamma\mathbf{y} \quad \leq c_7 s\mathbf{y}^T\mathbf{y}. && (39)
\end{aligned}$$

Then, the condition number of matrix S is given by:

$$\operatorname{cond}(S) = \frac{c_4 r(c_5\tau + c_7 s)}{c_1\tau c_3 r}\,\frac{4c_3 r(1+\rho_{\max}\,\tau\theta)^2 + c_6\tau^2 c_1}{c_4 r(\tau\rho_{\min})^2 + c_5\tau(c_2\tau + c_7 s)} \tag{40}$$

where $S := EK^{-1}E^T + NG^{-1}N^T$ is the Schur complement.

Proof. Using the upper and lower bounds for K, EE^T, NN^T and G we obtain:
Upper bound:

$$\mathbf{p}^T S\mathbf{p} = \mathbf{p}^T EK^{-1}E^T\mathbf{p} + \mathbf{p}^T NG^{-1}N^T\mathbf{p} \tag{41}$$

$$\leq \mathbf{p}^T EZ^{-1}E^T\mathbf{p} + \mathbf{p}^T NG^{-1}N^T\mathbf{p} \tag{42}$$

$$\leq \frac{1}{c_1\tau}\mathbf{p}^T EE^T\mathbf{p} + \frac{1}{c_3 r\tau h^d}\mathbf{p}^T NN^T\mathbf{p} \tag{43}$$

$$\leq \left(\frac{4}{c_1\tau}(1+\tau\theta\rho_{\max})^2 + \frac{c_6\tau^2 h^d}{c_3 r\tau h^d}\right)\mathbf{p}^T\mathbf{p} \tag{44}$$

$$= \left(\frac{4}{c_1\tau}(1+\tau\theta\rho_{\max})^2 + \frac{c_6\tau}{c_3 r}\right)\mathbf{p}^T\mathbf{p}. \tag{45}$$

Lower bound:

$$\mathbf{p}^T S\mathbf{p} \geq \frac{1}{(c_2\tau + c_7)}\mathbf{p}^T EE^T\mathbf{p} + \frac{1}{c_4 r\tau h^d}\mathbf{p}^T NN^T\mathbf{p} \tag{46}$$

$$\geq \left(\frac{(\tau\rho_{\min})^2}{(c_2\tau + c_7 s)} + \frac{c_5\tau^2 h^d}{c_4 r\tau h^d}\right)\mathbf{p}^T\mathbf{p} \tag{47}$$

$$= \left(\frac{(\tau\rho_{\min})^2}{(c_2\tau + c_7 s)} + \frac{c_5\tau}{c_4 r}\right)\mathbf{p}^T\mathbf{p}. \tag{48}$$

Therefore, the condition number of matrix S can be estimated by:

$$\operatorname{cond}(S) = \frac{c_4 r(c_5\tau + c_7 s)}{c_1\tau c_3 r}\,\frac{4c_3 r(1+\rho_{\max}\ \tau\theta)^2 + c_6\tau^2 c_1}{c_4 r(\tau\rho_{\min})^2 + c_5\tau(c_2\tau + c_7 s)}. \tag{49}$$

Remark The estimation given in (49) shows that matrix S is ill-conditioned. Indeed, let all the constants $c_i = 1$. Then the expression (49) reduces to:

$$\operatorname{cond}(S) \approx \frac{\tau + s}{\tau}\,\frac{r(1+h^{-2}\tau\theta)^2 + \tau^2}{r\tau^2 + \tau^2 + \tau s}. \tag{50}$$

Taking $\theta = 1/2$ and $h \approx \tau$, and with the reasonable assumption that $0 < O(h^4) \leq r \leq O(s/\tau)$, we obtain $\operatorname{cond}(S) \approx O(rh^{-4})$.

4 The reduced system for u

We next consider an algorithm based on the solution of a reduced Schur complement for the control variable $\mathbf{u}$. Assuming that $G \neq 0$ and solving the first and third block row in (20) will yield $\mathbf{p} = -E^{-T}K\mathbf{y} + E^{-T}K\mathbf{g}$ and $\mathbf{y} = -E^{-1}N\mathbf{u} + E^{-1}\mathbf{f}$, respectively. Then, system (20) can be reduced to the following Schur complement system for $\mathbf{u}$:

$$(G + N^T E^{-T}KE^{-1}N)\mathbf{u} = N^T E^{-T}KE^{-1}\mathbf{f} - N^T E^{-T}K\mathbf{g}. \tag{51}$$

Matrix $(G + N^T E^{-T}KE^{-1}N)$ is symmetric and positive definite. In the next Lemma, we show that $(G + N^T E^{-T}KE^{-1}N)$ is spectrally equivalent to G.

Lemma 2. *Let the bounds for G, E, K, N and Γ be as presented in Lemma 1. Then, there exist constants μ_{mim} and μ_{max}, independent of h and u, such that*

$$\mu_{min}\mathbf{u}^T G\mathbf{u} \leq \mathbf{u}^T(N^T E^{-T} K E^{-1} N)\mathbf{u} \leq \mu_{max}\mathbf{u}^T G\mathbf{u} \tag{52}$$

Proof. Using the upper and lower bounds for K, EE^T, NN^T and G we obtain:
Upper bound:

$$\begin{aligned}
\mathbf{u}^T N^T E^{-T} K E^{-1} N\mathbf{u} &\leq (c_2\tau + c_7 s)\mathbf{u}^T N^T E^{-T} E^{-1} N\mathbf{u} && (53)\\
&\leq \frac{(c_2\tau + c_7 s)}{(\tau\rho_{\text{min}})^2}\mathbf{u}^T N^T N\mathbf{u} && (54)\\
&\leq \frac{(c_2\tau + c_7 s)c_6\tau^2 h^d}{(\tau\rho_{\text{min}})^2}\mathbf{u}^T\mathbf{u} && (55)\\
&= \frac{(c_2\tau + c_7 s)c_6 h^d}{(\rho_{\text{min}})^2}\mathbf{u}^T\mathbf{u} && (56)\\
&\leq \frac{(c_2\tau + c_7 s)c_6}{(\rho_{\text{min}})^2 c_3 r\tau}\mathbf{u}^T G\mathbf{u} && (57)\\
&= \mu_{\text{max}}\ \mathbf{u}^T G\mathbf{u}. && (58)
\end{aligned}$$

Lower bound:

$$\begin{aligned}
\mathbf{u}^T N^T E^{-T} K E^{-1} N\mathbf{u} &\geq (c_1\tau)\mathbf{u}^T N^T E^{-T} E^{-1} N\mathbf{u} && (59)\\
&\geq \frac{c_1\tau}{4(1+\tau\rho_{\text{max}}\theta)^2}\mathbf{u}^T N^T N\mathbf{u} && (60)\\
&\geq \frac{c_1 c_5\tau^3 h^d}{4(1+\tau\rho_{\text{max}}\theta)^2}\mathbf{u}^T\mathbf{u} && (61)\\
&\geq \frac{c_1 c_5\tau^2 h^d}{4(1+\tau\rho_{\text{max}}\theta)^2 c_4 r}\mathbf{u}^T G\mathbf{u} && (62)\\
&= \mu_{\text{min}}\ \mathbf{u}^T G\mathbf{u}. && (63)
\end{aligned}$$

This completes the proof.

First Algorithm. The Schur complement system (51) can be solved using a CG algorithm (conjugate gradient) using the matrix G as a preconditioner. We note that

$$\mathbf{u}^T G\mathbf{u} \leq \mathbf{u}^T(G + N^T E^{-T} K E^{-1} N)\mathbf{u} \leq (1+\mu_{\text{max}})\mathbf{u}^T G\mathbf{u}. \tag{64}$$

Since the ρ_{min} is $O(1)$ and ρ_{max} is $O(h^{-4})$, it is easy to see that $\mu_{\text{min}} = O(\frac{h^4}{r})$ and $\mu_{\text{max}} = O(\frac{1+s/\tau}{r})$. Hence, the rate of convergence of this algorithm is independent of h and with a condition number estimate bounded by $O(1+\frac{1+\frac{s}{\tau}}{r})$. This algorithm is simple to implement however but has two drawbacks. It is a double iteration algorithm (require two applications of E^{-1}) and it is not directly parallelizable.

Second Algorithm. To overcome the mentioned drawbacks, we define $\hat{\mathbf{b}} := -N^T E^{-T} K E^{-1} \mathbf{f} + N^T E^{-T} K \mathbf{g}$ and the auxiliary variable $\mathbf{w} := -E^{-T} K E^{-1} N \mathbf{u}$. Hence system (51) can be written in the form:

$$\begin{bmatrix} EK^{-1}E^T & N \\ N^T & -G \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{b}} \end{bmatrix}. \tag{65}$$

In this case, the action of E^{-1} is required only in a pre-computed step to assemble the right hand side input vector $\hat{\mathbf{b}}$. System (65) is symmetric and indefinite. Thus, it can be solved iteratively using MINRES with a positive definite block diagonal preconditioner diag$(E_o K_o^{-1} E_o^T, G_o)$, where K_o is any matrix spectrally equivalent to the mass matrix K, matrix E_o is any matrix spectrally equivalent (or a preconditioner) to the evolution matrix E [6, 13, 3], and matrix G_o is a preconditioner for matrix G. The following Theorem estimates the condition number of the preconditioned system.

Theorem 2. *Let the bounds for matrices G, E, K, N and Γ be as presented in lemma 1 and denote $\mathcal{P} := blockdiag(EK^{-1}E^T, G)$ the block diagonal preconditioner and $\mathcal{H}$ the coefficient matrix of system (65). Then, the condition number of the preconditioned system satisfies the bound:*

$$\kappa(\mathcal{P}^{-1}\mathcal{H}) \leq O\left(\left(1 + \frac{1 + s/\tau}{r}\right)^{1/2}\right). \tag{66}$$

Proof. Since the preconditioner $\mathcal{P}$ is positive definite, we consider the generalizes eigenvalue problem given by:

$$\begin{bmatrix} EK^{-1}E^T & N \\ N^T & -G \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{u} \end{bmatrix} = \lambda \begin{bmatrix} EK^{-1}E^T & \\ & G \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{u} \end{bmatrix}, \tag{67}$$

We obtain the equations

$$(\lambda - 1) EK^{-1}E^T \mathbf{w} = N\mathbf{u} \;\text{ and }\; (\lambda + 1) G\mathbf{u} = N^T \mathbf{w}. \tag{68}$$

From these equations we obtain $N^T E^{-T} K E^{-1} N u = (\lambda^2 - 1) G u$ where $(\lambda^2 - 1)$ is the generalized eigenvalue of $N^T E^{-T} K E^{-1} N$ with respect to G. Using Lemma 2, we obtain bounds for λ as follows:

$$max|\lambda| \leq (1 + \mu_{max})^{1/2} = O\left(\left(1 + \frac{1 + s/\tau}{r}\right)^{1/2}\right) \tag{69}$$

$$min|\lambda| \geq (1 + \mu_{min})^{1/2} = O(1). \tag{70}$$

The theorem follows, since:

$$\kappa(\mathcal{P}^{-1}\mathcal{H}) \leq \frac{max|\lambda|}{min|\lambda|}. \tag{71}$$

This completes the proof.

Remark. Generalization of this theorem for matrices G_o and $E_o K_o^{-1} E_o^T$ (spectrally equivalent to G and $EK^{-1}E^T$ respectively) follows directly from ([5]).

Remark. Applying matrix E is very unstable, but applying matrix E^{-1} is stable. The algorithms presented here do not require application of E or E^T since

$$\mathcal{PH} = \begin{bmatrix} I & E^{-T}KE^{-1}N \\ G^{-1}N^T & -I \end{bmatrix}. \tag{72}$$

5 Numerical Experiments

In this section, we consider the numerical solution of the optimal control of the 1D-heat equation. In this case, the constraints are given by:

$$\partial_t z - \partial_{xx} z = v, \quad 0 < x < 1, \quad t > 0$$

with boundary conditions $z(t,0) = 0$ and $z(t,1) = 0$ for $t \geq 0$, with initial data $z(0,x) = 0$ for $x \in [0,1]$, and with the performance function $z_* = sin(\pi x)$ for all $t \in [0,1]$. Following [8], we take $q = 1$ and $r = 0.0001$. The trapezoidal rule discretization is considered in the numerical experiments. As a stopping criteria for the iterative solvers, we take $\|\mathbf{r_k}\|/\|\mathbf{r_0}\| \leq 10^{-3}$ where $\mathbf{r_k}$ is the residual at each iteration k.

Table 1. Number of CG iterations for Algorithm 1. The parameters $s = 0$ ($s = 1$).

Nx \ Nt	32	64	128	256	512
32	10 (9)	11 (9)	11 (10)	11 (11)	11 (11)
64	10 (9)	11 (9)	11 (10)	11 (11)	11 (11)
128	10 (9)	11 (10)	11 (11)	11 (11)	11 (11)
256	10 (9)	11 (10)	11 (11)	11 (11)	11 (11)
512	10 (9)	11 (10)	11 (11)	11 (11)	11 (11)

Algorithm 1: Reduction to u. In the first case we consider $s = 0$ in the cost function. We use matrix G as a preconditioner and CG for solving the preconditioned resulting system. Table 1 presents the number of iterations when both time and space grid are refined. Notice that, the number of iterations is independent of both the time discretization τ and space discretization h. Hence, the algorithm is scalable both in time and space grid parameters, as predicted by the analysis developed in Section 4. Table 1 also presents the number of iterations, within parenthesis, when parameter $s = 1$, and the number of iterations also remains constant.

Algorithm 2. Table 2 presents the number of iterations when both time and space grid are refined. Notice that, as predicted by the analysis, when the time grid is refined more than $\tau = 1/128$, the number of iterations remains bounded when both time discretization τ and space discretization h are refined.

In Table 3, we make a comparison of both algorithms for different values of r and s. We note that both algorithms are robust for a large range of values r and

Table 2. Number of MINRES iterations for algorithm 2. Parameter $s = 0$ ($s = 1$).

Nx \ Nt	32	64	128	256	512
32	20 (18)	20 (18)	22 (18)	22 (18)	22 (20)
64	20 (18)	20 (18)	22 (18)	22 (18)	22 (20)
128	20 (18)	20 (18)	22 (18)	22 (18)	22 (20)
256	20 (18)	20 (18)	22 (18)	22 (18)	22 (20)
512	20 (18)	20 (18)	22 (18)	22 (18)	22 (20)

s, therefore the analysis developed on Section 4 is not sharp with repect to the dependence on the coefficient r and s. Notice that when s and r are increased, the number of iterations decrease however the solution is not acceptable since it is not close enough to the target function z_*. In addition, if the term $\frac{s}{2}\|z(v)(t_f,x) - z_*(t_f,x)\|^2_{L^2(\Omega)}$ is considered, depending on the scaling of the matrices R, B and K the solution may exhibit *boundary layer* character. This analysis is beyond the scope of this article.

Table 3. Comparison of the number of iterations for different values of r and s. The constant $s = 10$. Acronym: Algoritm 1: Alg. 1, Algoritm 2: Alg. 2 and *: Non acceptable solution.

Acronym \ r	10^{-2}	10^{-3}	10^{-4}	10^{-5}
$s = 0$ Alg.1	7	11	14	14
Alg.2	12	20	26	26
$s = 1$ Alg.1	6^*	9	14	14
Alg.2	12^*	18	20	20
$s = 10$ Alg.1	5^*	5	5	5
Alg.2	10^*	10	10	10

6 Concluding Remarks

In this paper we have described two approaches for iteratively solving the linear quadratic parabolic optimal control problem. The first method is based on the CG solution of a Schur complement. This is obtained by reducing the saddle point system to the system associated with the control variable. This method is simple to implement but requires double iteration. The second method avoids double iteration introducing an auxiliary variable. The resulting system is symmetric and indefinite, so that MINRES can be used. The structure of this method also allows parallel block preconditioners. The preconditioners described yield a rate of convergence independent of the time and space parameters.

Acknowledgments. C.E.S. acknowledges Etereldes Gonçalves Junior for a fruitful discussion on topics of this work.

References

1. G. Biros and O. Gattas, *Parallel Lagrange-Newton-Krylov-Schur Methods for PDE-Constrained Optimization. Part I: The Krylov-Schur Solver*, SIAM Journal on Scientific Computing, 27(2), pp. 687–713, 2005.
2. J. W. Demmel, *Applied Numerical Linear ALgebra*, SIAM, Philadelphia, 1997.
3. M. J. Gander and S. Vandewalle, *On the super linear and linear convergence of the parareal algorithm*, Proceedings of the 16th International Conference on Domain Decomposition Methods, 2005.
4. M. Heinkenschloss, *A time-domain decomposition iterative method for the solution of distributed linear quadratic optimal control problems*, Journal of Computational and Applied Mathematics, Volume 173, Issue 1, No 1, Pages 169-198, 2005.
5. A. Klawon, *Preconditioners for Indefinite Problems.* Reports CS-NYU, TR1996-716, March, 1996.
6. J. L. Lions and Y. Maday and G. Turinici, *Résolution d'EDP par un schéma en Temps Pararéel*, C.R. Acad. Sci. Paris, t. 332, Série I, pp. 661–668, 2001.
7. J.L. Lions *Optimal Control of Systems Governed by Partial Differential Equations problems*, Springer, Berlin, 1971.
8. A. Locatelli, *Otimal Control: An Introduction*, Birkhäuser, Berlin, 2001.
9. D. Luenberger, *Introduction to Dynamic Systems: Theory, Models and Applications*, Wiley, New York, 1979.
10. Y. Maday and G. Turinici, *A parareal in time procedure for the control of partial differential equations*, C.R.Acad. Sci. Paris, t. 335, Ser. I, pp. 387–392, 2002.
11. T. Mathew and M. Sarkis and C.E. Schaerer , *Block matrix preconditioners foir elliptic optimal control problems.*, submitted to Numerical Linear Algebra with Applications, 2005.
12. P. Sarma, K. Aziz L.J. Durlofsky, *Implementation of adjoint solution for optimal control of smart well*, SPE reservoir Simulation Symposium, Texas-USA, January 31- February 2, 2005.
13. C.E. Schaerer and E. Kaszkurewicz, *The shooting method for the numerical solution of ordinary differential equations: a control theoretical perspective*, International Journal of Systems Science, Vol. 32, No. 8, pp. 1047-1053, 2001.
14. B. Sudaryanto and Y. C. Yortsos, *Optimization of fluid fronts dynamics in porous media using rate control. I. Equal mobility fluids* Physics of Fluids, Vol. 12, No 7, pp. 1656–1670, 2000.