A FAST AND STABLE PRECONDITIONED ITERATIVE METHOD FOR OPTIMAL CONTROL PROBLEM OF WAVE EQUATIONS *

BUYANG LI†, JUN LIU‡, AND MINGQING XIAO§

Abstract. In this paper, we develop a new central finite difference scheme in terms of both time and space for solving the first-order necessary optimality systems that characterize the optimal control of wave equations. The obtained new scheme is proved to be unconditionally convergent with a second-order accuracy, without the requirement of Courant-Friedrichs-Lewy (CFL) condition on the corresponding grid ratio. An efficient preconditioned iterative method is further developed for solving the discretized sparse linear system based on the relationship between the resultant matrix structure and the coupled PDE optimality system. Numerical examples are presented to verify the theoretical analysis and to demonstrate the high efficiency of the proposed preconditioned iterative solver.

Key words. Wave equation; optimal control; implicit scheme; preconditioner; GMRES.

AMS subject classifications. 49K20, 65N06, 65N12, 65N22, 65F08, 65F10.

1. Introduction. Optimal control problems governed by time-dependent partial differential equations (PDEs) [3, 4, 16, 27, 39] have been very attractive to the scientific computing community in the last decade due to that solving this type of problems requires highly extensive computing. Such problems appear in a wide range of applications such as flow control design [10], aerodynamic shape optimization [17], and photoacoustic tomography [2]. In order to achieve various desirable control purposes, both efficiency and accuracy of numerical computing become not only critical but also essential. It requires a fully understanding of the subtle interplay among numerical optimization, numerical PDEs, and numerical linear algebra in order to develop approachable computing. In this paper, we develop a new implicit scheme in time, associated with a well-structured preconditioned iterative fast solver, to solve the optimality systems arising from optimal control problems governed by wave equations.

Let $\Omega = (0, 1)^d$, $d \geq 1$, be the spatial domain with boundary $\Gamma := \partial \Omega$. Given a finite period of time $T > 0$, define $Q = \Omega \times (0, T)$ and $\Sigma = \Gamma \times (0, T)$. We consider the following standard optimal control problem [27] of minimizing a tracking-type quadratic cost functional

$$
J(y, u) = \frac{1}{2} \| y - g \|_{L^2(Q)}^2 + \frac{\gamma}{2} \| u \|_{L^2(Q)}^2
$$

subject to the linear wave equation:

$$
\begin{align*}
\frac{\partial^2 y}{\partial t^2} - \Delta y &= f + u \quad \text{in } Q, \\
y &= 0 \quad \text{on } \Sigma, \\
y(\cdot, 0) &= y_0 \quad \text{in } \Omega, \\
y_t(\cdot, 0) &= y_1 \quad \text{in } \Omega,
\end{align*}
$$

where $u \in U := L^2(Q)$ is the distributed control function, $g \in L^2(Q)$ is the desired tracking trajectory, $\gamma > 0$ represents either the weight of the cost of control or the Tikhonov regularization parameter, $f \in L^2(Q)$, and the initial conditions $y_0 \in H^1_0(\Omega)$ and $y_1 \in L^2(\Omega)$. The existence, uniqueness and regularity of the solution for the optimal control problem (1.1)-(1.2) are well

---

*Project was supported in part by NSF 1021203, 1419028 of the United States and in part by NSFC 11301262 of China. The research stay of B. Li at Universität Tübingen was supported by the Alexander von Humboldt Foundation. J. Liu was also supported in part by the Dissertation Research Assistantship of Southern Illinois University.

†Department of Mathematics, Nanjing University, 210093 Nanjing, Jiangsu, P.R. China. buyangli@nju.edu.cn

‡Department of Mathematics and Statistical Sciences, Jackson State University, Jackson, MS 39217, USA. jun.liu@jsums.edu

§Department of Mathematics, Southern Illinois University, Carbondale, IL 62901, USA. mxiao@siu.edu
established [27]. By defining an appropriate Lagrange functional and making use of the strict convexity, the optimal solution pair \((y, u)\) to (1.1)-(1.2) is shown to be completely characterized by the unique solution triplet \((y, p, u)\) to the following optimality system

\[
\begin{align*}
\begin{cases}
y_{tt} - \Delta y - u &= f \quad \text{in } Q, \\
y(\cdot, 0) &= y_0 \quad \text{in } \Omega, \\
y_t(\cdot, 0) &= y_1 \quad \text{in } \Omega,
\end{cases}
\quad \begin{cases}
p_{tt} - \Delta p + y &= g \quad \text{in } Q, \\
p(\cdot, T) &= 0 \quad \text{in } \Omega, \\
p_t(\cdot, T) &= 0 \quad \text{in } \Omega,
\end{cases}
\gamma u - p &= 0 \quad \text{in } Q,
\end{align*}
\]

(1.3)

where the state \(y\) evolves forward in time and the adjoint state \(p\) marches backward in time. The control \(u = p/\gamma\) can be eliminated from the above optimality system as following

\[
\begin{align*}
\begin{cases}
y_{tt} - \Delta y - p/\gamma &= f \quad \text{in } Q, \\
y(\cdot, 0) &= y_0 \quad \text{in } \Omega, \\
y_t(\cdot, 0) &= y_1 \quad \text{in } \Omega,
\end{cases}
\quad \begin{cases}
p_{tt} - \Delta p + y &= g \quad \text{in } Q, \\
p(\cdot, T) &= 0 \quad \text{in } \Omega, \\
p_t(\cdot, T) &= 0 \quad \text{in } \Omega,
\end{cases}
\end{align*}
\]

(1.4)

It is well-known that the main challenge for solving (1.4) results from the fact that the state \(y\) and the adjoint state \(p\) are marching in opposite orientations. Its numerical discretizations will create an enormously huge algebraic system of equations as we have to resolve all time steps simultaneously [12].

Different from elliptic and parabolic cases, there are few available results on fast computing of optimal control of wave equations. There are some developments of numerical algorithms for the optimal control of hyperbolic or wave equations [5, 7, 8, 9, 19, 20, 21, 22, 23, 24, 25, 28, 40]. Some comprehensive and interesting results are given, for example, in [23]. In their work, the authors analyzed the superlinear convergence of the semismooth Newton method that is employed to treat the inequality control constraints in optimal control problems governed by the wave equation. The discretization is through finite element approach for distributed control, Neumann boundary control, and Dirichlet boundary control, respectively. The original second-order wave equation is formulated as a first-order system for their discretizations, in which the time variable is discretized by the Crank-Nicolson scheme based on the trapezoidal rule. However, there is no discussion for the implementation of the proposed algorithms for solving the resultant discretized systems. Furthermore, the reformulated first-order system introduces two extra dependent variables, which increases the computational burdens. Another notable work can be found in [28]. In their approach, the authors applied the finite volume element method to the distributed control problems governed by second-order hyperbolic equations, where the optimal error estimates in certain norm were proved for the spatially semi-discrete optimality system, but the convergence of the full-discrete scheme is not seen. In the given numerical experiments, the spatial and temporal step sizes are chosen to satisfy the Courant-Friedrichs-Lewy (CFL) condition that may not be desirable for an efficient algorithm. For solving the discretized system, a nice fixed-point iterative algorithm is provided in [33]. However, when the regularization (or penalization) parameter \(\gamma\) in the cost functional becomes small, the approach for our underlying problem may suffer from slow convergence or even divergence. More recently, numerical methods were developed for the optimal control of nonlinear hyperbolic system with possible discontinuous solutions [6, 13]. However, the implementation of fast computing for solving the discretized optimality system has not been discussed. Although these second-order schemes are available in literature, to the best of our knowledge, the study of an efficient numerical implementation (fast solver) for the optimal control problem of wave equations has not been seen yet.

Generally speaking, an efficient numerical implementation includes two steps: the first step is to seek a numerical discretization that is not only convergent but also can provide a well-structured
discretization, and the second step is to develop an efficient iteration for the obtained large algebraic systems. These two steps are inevitably correlated closely. If a high-order numerical scheme is developed with a poor structure, then the design of an efficient numerical implementation will become very difficult, if it is not impossible. Therefore, in order to have an efficient implementation, it is essential to develop the numerical schemes that can be suitably adapted to the later construction of iterative linear solvers [14, 7, 29, 30, 32, 34] so that it can handle large-scale degrees of freedom and high dimensions.

In this paper we develop a new modified central difference scheme for both time and spatial variables. The proposed numerical scheme for solving (1.4) is not only shown to be unconditionally stable but also to have a nice discretized structure. It is not required to satisfy the CFL condition that usually is necessary in classical theory for solving hyperbolic equations by standard explicit scheme[26, 38]. Based on our setting, we construct an effective preconditioned iterative solver for solving the resultant discretized linear system.

The paper is organized as follows. In next section we give the standard explicit scheme in time with a central finite difference scheme in space for discretizing the optimality system (1.4). As a development, we present a new implicit scheme in time and provide the error estimate of the resulting full-discrete scheme in Section 3. Section 3.3 discusses the construction of an effective block upper triangular preconditioner by the well-known GMRES method that is suitable for solving the corresponding discretized system. Numerical experiments are performed in Section 4 to validate our theoretical outcome and to demonstrate the effectiveness of the proposed preconditioner. Finally, the paper ends with concluding remarks in Section 5. To simplify the notations, we only present the analysis for the case $d = 2$, but the results presented in this paper hold for arbitrary $d \geq 1$.

**2. A standard central difference scheme.** We partition the time interval $[0, T]$ uniformly into $0 = t_0 < t_1 < \cdots < t_N = T$ with $t_k - t_{k-1} = \tau = T/N$, and discretize the space domain $\Omega$ uniformly into $0 = \xi_0 < \xi_1 < \cdots < \xi_{M_1} = 1$ and $0 = \zeta_0 < \zeta_1 < \cdots < \zeta_{M_2} = 1$, with $h_1 = \xi_i - \xi_{i-1}$, $h_2 = \zeta_j - \zeta_{j-1}$. Let $h = \max(h_1,h_2)$. We define the discrete inner product $(\varphi^n, \phi^n) = \sum_{i,j=1}^{M_1-1,M_2-1} \varphi^n_i \phi^n_j h_1 h_2$ and the corresponding discrete $L^2(\Omega)$ norm $\|\phi^n\| = \sqrt{(\phi^n, \phi^n)}$.

We also define the discrete gradient

$$\nabla_h \varphi^n = \left( \frac{\varphi^n_{i+1,j} - \varphi^n_{i-1,j}}{h_1}, \frac{\varphi^n_{i,j+1} - \varphi^n_{i,j-1}}{h_2} \right)_{i=1,j=1}^{M_1,M_2},$$

and the discrete Laplacian (in 2D)

$$(\Delta_h Y^n)_{ij} = \frac{Y^n_{i-1,j} - 2Y^n_{i,j} + Y^n_{i+1,j}}{h_1^2} + \frac{Y^n_{i,j-1} - 2Y^n_{i,j} + Y^n_{i,j+1}}{h_2^2}.$$

We discretize the equations (1.4) by the leap-frog scheme in time with a standard five-point second order central difference discretization in space

\begin{align}
Y_{n+1} - 2Y_n + Y_{n-1} &\quad \frac{\tau^2}{2}, \quad n = 0, 1, 2, \cdots, N - 1, \tag{2.1} \\
P_{n+1} - 2P_n + P_{n-1} &\quad \frac{\tau^2}{2}, \quad n = 1, 2, \cdots, N - 1, N \quad \tag{2.2}
\end{align}

where $Y^n = (Y^n_{ij})_{i=1,j=1}^{M_1-1,M_2-1}$ and $P^n = (P^n_{ij})_{i=1,j=1}^{M_1-1,M_2-1}$ with $Y^n_{ij}$ and $P^n_{ij}$ being the discrete approximation of $y(t_i, t_j, t_n)$ and $p(t_i, t_j, t_n)$, respectively. Similarly notations are used for $f^n$ and $g^n$. The initial conditions are derived by Taylor expansions and using (1.4) to represent $y_{tt}$ and $p_{tt}$, that is,

\begin{align}
Y^0_{i,j} &= y_0(\xi_i, \zeta_j), \\
Y^1_{i,j} &= y_0(\xi_i, \zeta_j) + y_1(\xi_i, \zeta_j) \tau + \frac{\tau^2}{2} (\Delta y_0(\xi_i, \zeta_j) + f_0^{i,j} + \frac{1}{\gamma} P^0_{i,j}), \tag{2.3}
\end{align}
(2.4) \[ P_{i,j}^N = 0, \quad P_{i,j}^{N-1} = \frac{\tau^2}{2}(-Y_{i,j}^N + y_{i,j}^N), \]

where we have used the final time conditions \( p(\cdot, T) = 0 \) and \( p_t(\cdot, T) = 0 \).

To illustrate the structure of the discretized system, we formulate the above explicit scheme into a two-by-two block structured symmetric indefinite linear system

\[ (2.5) \quad S_h \begin{bmatrix} y_h \\ p_h \end{bmatrix} := \begin{bmatrix} I_h & F_h^T \\ F_h & -I_h/\gamma \end{bmatrix} \begin{bmatrix} y_h \\ p_h \end{bmatrix} = \begin{bmatrix} g_h \\ f_h \end{bmatrix}, \]

where

\[ (2.6) \quad F_h = \frac{1}{\tau^2} \begin{bmatrix} I & 0 & \cdots & 0 \\ -2I - \tau^2 \Delta_h & I & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & I & -2I - \tau^2 \Delta_h \end{bmatrix}, \]

\[ \hat{I}_h = \begin{bmatrix} I/2 & 0 & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & I & 0 \end{bmatrix}, \quad \hat{I}_h = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & I & 0 \end{bmatrix}, \]

\[ (2.7) \quad f_h = \begin{bmatrix} f^0/2 + y_1/\tau + (I/\tau^2 + \Delta_h/2)y_0 \\ \vdots \\ f^{N-2} \\ f^{N-1} \end{bmatrix}, \]

(2.8) \[ y_h = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad G = \begin{bmatrix} g^0 \\ g^1 \\ \vdots \\ g^N \end{bmatrix}, \quad \text{and} \quad p_h = \begin{bmatrix} p^0 \\ p^1 \\ \vdots \\ p^N \end{bmatrix}. \]

Here \( I \) is an identity matrix of appropriate size and the vectors \( y_0, y_1, f^n, g^n, y^n, \) and \( p^n \) are the lexicographic ordering (vectorization) of the corresponding function approximations over spatial grid points. Notice that the submatrix \( F_h \) and its transpose \( F_h^T \) become very ill-conditioned when the CFL condition \( (\tau \leq h) \) does not hold (detailed argument is provided in Appendix A), which is expected since this corresponds to the instability of the explicit scheme applied to a single wave equation. Although, due to the coupling effects, the whole system seems to be well-conditioned regardless of the CFL condition, it is difficult to design efficient iterative solvers by existing approaches, such as by preconditioned GMRES [34], with the ill-conditioned \( F_h \). We shall modify the above scheme in next section to remove the requirement of CFL condition and to avoid solving the ill-conditioned matrix \( F_h \) resulted from standard discretization.
3. A new scheme with a fast preconditioned solver. One critical observation is that the above explicit scheme in time prevents us from solving the coupled system in an efficient way by various time-marching algorithms as usually seen in handling a single wave equation. This motivates us to change the explicit scheme that relies on CFL condition to an implicit scheme so that restrictions on mesh size ratios can be eliminated. It is expected that implicit schemes are more suitable for developing robust iterative solvers. In this section, we introduce an implicit central difference scheme for optimality system, which not only allows us to show the convergence but also to construct an effective preconditioner of the corresponding discretized system.

3.1. The modified scheme. We propose the following scheme (averaging the Laplacian term)

\[
Y^{n+1} - 2Y^n + Y^{n-1} = \frac{\Delta h Y^{n+1} + \Delta h Y^{n-1}}{2} - \frac{P^n}{\gamma} = f^n, \quad n = 1, 2, \ldots, N - 1
\]

\[
P^{n+1} - 2P^n + P^{n-1} = \frac{\Delta h P^{n+1} + \Delta h P^{n-1}}{2} + Y^n = g^n, \quad n = 1, 2, \ldots, N - 1
\]

where \( Y^n = (Y^n)_{i,j=1}^{M_1-1,M_2-1} \) and \( P^n = (P^n)_{i,j=1}^{M_1-1,M_2-1} \) with \( Y^n_{i,j} \) and \( P^n_{i,j} \) are the discrete approximations of \( y(\xi_i, \zeta_j, t_n) \) and \( p(\xi_i, \zeta_j, t_n) \), respectively. Compared to the standard central difference scheme (2.1)-(2.4), we artificially introduce a second-order approximation over three consecutive time steps, i.e.,

\[
\Delta y^n = \frac{\Delta y^{n+1} + \Delta y^{n-1}}{2} + O(\tau^2),
\]

\[
\Delta p^n = \frac{\Delta p^{n+1} + \Delta p^{n-1}}{2} + O(\tau^2),
\]

which maintains the same second order of accuracy as the standard central difference scheme does. The initial conditions are derived based on Taylor expansions up to the order \( O(\tau^3) \), by using (3.4) to represent \( y_{tt} \) and \( p_{tt} \), i.e.,

\[
Y^{0}_{i,j} = y_0(\xi_i, \zeta_j), \quad \left(1 - \frac{\tau^2}{2}\Delta h\right) Y^1_{i,j} = y_0(\xi_i, \zeta_j) + y_1(\xi_i, \zeta_j)\tau + \frac{\tau^2}{2}(f^{0}_{i,j} + \frac{1}{\gamma}P^{0}_{i,j}),
\]

\[
P^0_{i,j} = 0, \quad \left(1 - \frac{\tau^2}{2}\Delta h\right) P^{N-1}_{i,j} = \frac{\tau^2}{2}(-Y^N_{i,j} + g^N_{i,j}).
\]

where we have used implicit schemes in approximating \( Y^1_{i,j} \) and \( P^{N-1}_{i,j} \). In Section 3.3 we will see that the implicit schemes used in (3.5)-(3.6) leads to the successful construction of an effective preconditioner.

By denoting \( D_h = I - \frac{\tau^2}{2}\Delta h \), the scheme above can be formulated as a symmetric indefinite linear system

\[
M_h \begin{bmatrix}
  y_h \\
  p_h
\end{bmatrix} = \begin{bmatrix}
  \mathbf{I}_h & L_h^T \\
  L_h & -\mathbf{I}_h/\gamma
\end{bmatrix} \begin{bmatrix}
  y_h \\
  p_h
\end{bmatrix} = \begin{bmatrix}
  g_h \\
  f_h
\end{bmatrix},
\]

where

\[
D_h = \begin{bmatrix}
  D_h & 0 & 0 & \cdots & 0 \\
  -2I & D_h & 0 & \cdots & 0 \\
  D_h & -2I & D_h & \cdots & 0 \\
  0 & \cdots & \cdots & \cdots & 0 \\
  0 & \cdots & D_h & -2I & D_h \\
  0 & 0 & \cdots & D_h & -2I
\end{bmatrix},
\]

\[
L_h = \frac{1}{\tau^2} \begin{bmatrix}
  0 & 0 & 0 & \cdots & 0 \\
  0 & D_h & 0 & \cdots & 0 \\
  0 & 0 & D_h & \cdots & 0 \\
  0 & 0 & 0 & \cdots & D_h \\
  0 & 0 & \cdots & 0 & 0
\end{bmatrix},
\]
where the first two components of $f_h$ are different from previous standard scheme. The matrix $L_h$ has a much smaller condition number than $F_h$ in (2.5), which will be crucial to the development of efficient iterative solvers. In the following Table 1, we report the numerically estimated condition numbers of the standard and modified central difference scheme for the given example 1 presented in Section 4 with $T = 2$ using MATLAB’s build-in function condest. Here we set $\tau = 2h$, which clearly violates the CFL condition ($\tau \leq h$). From this example one can see that $F_h$ is highly ill-conditioned compared with $L_h$, which will inevitably incapacitate any numerical methods that rely on computing or approximating $F_h^{-1}v$ for solving the algebraic system (2.5). Our modified scheme does not suffer from this drawback since the corresponding $L_h$ is even much more well-conditioned than the whole system $M_h$. For our approach, the computation of $L_h^{-1}v$ plays an important role in the implementation of a desirable preconditioner. In Appendix, we provide a formal argument in which it shows that the condition number of $F_h$ grows exponentially when the CFL condition is not met.

### Table 1

The condition numbers of the explicit and implicit scheme for Ex. 1 (with $T = 2$ and $\gamma = 10^{-2}$).

<table>
<thead>
<tr>
<th></th>
<th>Explicit Scheme</th>
<th>Implicit Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cond($F_h$)</td>
<td>cond($S_h$)</td>
</tr>
<tr>
<td>(8,8)</td>
<td>1.59e+09</td>
<td>1.53e+03</td>
</tr>
<tr>
<td>(16,16)</td>
<td>2.47e+18</td>
<td>6.57e+03</td>
</tr>
<tr>
<td>(32,32)</td>
<td>5.00e+36</td>
<td>2.85e+04</td>
</tr>
<tr>
<td>(64,64)</td>
<td>2.01e+73</td>
<td>1.26e+05</td>
</tr>
<tr>
<td>(128,128)</td>
<td>3.26e+146</td>
<td>5.56e+05</td>
</tr>
<tr>
<td>(256,256)</td>
<td>8.55e+292</td>
<td>2.44e+06</td>
</tr>
</tbody>
</table>

**3.2. Error estimates of the numerical solution.** In this section, we present error estimates for the numerical solution given by the scheme (3.1)-(3.6). The discrete version of Poincare inequality (e.g., see [18]) will be used, i.e. there exists a positive constant $C_0$, independent of $h$, such that if $y = (y_{ij})$ satisfies the boundary condition $y_{0,j} = y_{M_1,j} = y_{i,0} = y_{i,M_2} = 0$ for $i = 1, \cdots, M_1 - 1$ and $j = 1, \cdots, M_2 - 1$, then

\[
\|y\| \leq C_0 \|\nabla_h y\|.
\]

Throughout our approach, the following discrete version of integration by parts will be used:

\[
(-\Delta_h z, w) = (\nabla_h z, \nabla_h w)
\]

where the functions $z, w$ are defined on the mesh points and vanish on the boundary $\partial \Omega$. We first introduce the following two lemmas that will be used in the proof of Theorem 3.3.

**Lemma 3.1 (Lemma 5.1 of [15], Discrete Gronwall’s inequality).** Let $\tau = T/N$. If $E^n \geq 0$ for $n = 0, 1, \cdots, N - 1$ and $E^k \leq \alpha + \beta \sum_{n=0}^{k-1} \tau E^n$ for $0 \leq k \leq N - 1$, then

\[
\max_{0 \leq n \leq N-1} E^n \leq C_{\beta,T} \alpha,
\]
where the constant $C_{\beta,T}$ only depends on $\beta$ and $T$.

**Lemma 3.2.** For any function $w$ defined on the mesh points of $\Omega$ vanishing on the boundary $\partial \Omega$, we have

\begin{align}
(3.11) \quad & \| D_h^{-1}w \| \leq \| w \|, \\
(3.12) \quad & \| D_h^{-1/2}w \| \leq \| D_h^{-1/2}w \|, \\
(3.13) \quad & \| \nabla_h D_h^{-1}w \| \leq \| \nabla_h w \|, \quad \text{and} \\
(3.14) \quad & \tau^2 \| \nabla_h D_h^{-1}w \| \leq \frac{1}{2} \| w \|^2.
\end{align}

**Proof.** Since $-\Delta_h$ is symmetric and positive definite, we denote by $\xi_j$, $j = 1, \cdots, (M_1 - 1)(M_2 - 1)$, the orthonormal eigenfunctions of $-\Delta_h$ corresponding to the positive eigenvalues $\lambda_j$, $j = 1, \cdots, (M_1 - 1)(M_2 - 1)$, respectively. For any function $u = \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} \alpha_j \xi_j$ we have

\[
\| \nabla_h D_h^{-1}u \|^2 = (\nabla_h (1 - \tau^2 \Delta_h/2)^{-1}u, \nabla_h (1 - \tau^2 \Delta_h/2)^{-1}u) \\
= (-\Delta_h (1 - \tau^2 \Delta_h/2)^{-1}u, (1 - \tau^2 \Delta_h/2)^{-1}u) \\
= \left( \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} \frac{\alpha_j \lambda_j}{1 + \tau^2 \lambda_j/2} \xi_j, \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} \frac{\alpha_j}{1 + \tau^2 \lambda_j/2} \xi_j \right) \\
= \left( \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} \frac{|\alpha_j|^2 \lambda_j}{(1 + \tau^2 \lambda_j/2)^2} \right) \leq \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} |\alpha_j|^2 \lambda_j = (\Delta_h u, u) = \| \nabla_h u \|^2.
\]

Similarly, we have

\[
\| D_h^{-1}u \|^2 = \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} \frac{|\alpha_j|^2}{(1 + \tau^2 \lambda_j/2)^2} \leq \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} |\alpha_j|^2 = \| u \|^2,
\]

\[
\| D_h^{-1/2}u \|^2 = \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} \frac{|\alpha_j|^2}{1 + \tau^2 \lambda_j/2} \leq \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} |\alpha_j|^2 = \| u \|^2,
\]

and

\[
\tau^2 \| \nabla_h D_h^{-1}u \|^2 = \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} \frac{|\alpha_j|^2 \tau^2 \lambda_j}{(1 + \tau^2 \lambda_j/2)^2} \leq \sum_{j=1}^{(M_1 - 1)(M_2 - 1)} |\alpha_j|^2 = \frac{1}{2} \| u \|^2.
\]

Therefore, $\| D_h^{-1}w \| = \| D_h^{-1/2}D_h^{-1/2}w \| \leq \| D_h^{-1/2}w \|$, and the proof is thus completed. $\blacksquare$

**Theorem 3.3.** Assume the solution $y, p \in C^{4,4}(\overline{Q})$ and the time step size $\tau \leq \gamma^4$. Then there exists a positive constant $C_* := C_*(\gamma, T)$, independent of $h$ and $\tau$, such that

\[
(3.15) \quad \max_{0 \leq n \leq N} \{ \gamma \| Y^n - y^n \| + \| P^n - p^n \| \} \leq C_*(\tau^2 + h^2).
\]

**Proof.** To simplify the notations, we denote by $C_1, C_2, \cdots$, positive constants which do not depend on $\tau, h, n$ or $k$ in the following arguments.
Notice that the exact solution $y^n_{i,j} = y(\xi_i, \zeta_j, t_n)$ and $p^n_{i,j} = p(\xi_i, \zeta_j, t_n)$ satisfy the equations

\begin{align}
3.16 \quad & \frac{y^{n+1} - 2y^n + y^{n-1}}{\tau^2} - \frac{\Delta_h y^{n+1} + \Delta_h y^{n-1}}{2} - p^n / \gamma = f^n + F^n, \quad n = 1, 2, \cdots, N - 1 \\
3.17 \quad & \frac{p^{n+1} - 2p^n + p^{n-1}}{\tau^2} - \frac{\Delta_h p^{n+1} + \Delta_h p^{n-1}}{2} + y^n = g^n + G^n, \quad n = 1, 2, \cdots, N - 1
\end{align}

and

\begin{align}
3.18 \quad & y^0_{i,j} = y_0(\xi_i, \zeta_j), \quad p^0_{i,j} = 0 \\
3.19 \quad & \left(1 - \frac{\tau^2}{2} \Delta_h\right) y^1_{i,j} = y_0(\xi_i, \zeta_j) + y_1(\xi_i, \zeta_j) + \tau^2 \left(f^{0}_{i,j} + \frac{1}{\gamma} p^{0}_{i,j}\right) + F^{0}_{i,j}, \\
3.20 \quad & \left(1 - \frac{\tau^2}{2} \Delta_h\right) p^{N-1}_{i,j} = -\frac{1}{2} y^{N-1 \gamma}_{i,j} + \frac{1}{2} \varphi^{N \gamma}_{i,j} + G^{N}_{i,j},
\end{align}

where (3.18) and (3.19) are derived via Taylor expansions by using (1.4) to represent $y_{tt}$ and $p_{tt}$. Also, $F^n$ and $G^n$ denote the truncation errors, which satisfy

\[ \|F^n\| + \|G^n\| \leq C_1(\tau^2 + h^2) \quad \text{for} \quad n = 1, 2, \cdots, N - 1, \]

\[ \|F^0\| + \|G^N\| + \|\nabla_h F^0\| + \|\nabla_h G^N\| \leq C_1(\tau^3 + \tau h^3), \]

for some positive constant $C_1$ when the solution $y, p \in C^{4,4}(\Omega)$.

Let $e^n = Y^n - y^n$ and $\eta^n = P^n - p^n$. Then the difference between (3.1)-(3.6) and (3.16)-(3.19) gives

\begin{align}
3.21 \quad & e^{n+1} - 2e^n + e^{n-1} - \frac{\Delta_h e^{n+1} + \Delta_h e^{n-1}}{2} - \eta^n / \gamma = -F^n, \\
3.22 \quad & \eta^{n+1} - 2\eta^n + \eta^{n-1} - \frac{\Delta_h \eta^{n+1} + \Delta_h \eta^{n-1}}{2} + e^n = -G^n,
\end{align}

for $n = 1, 2, \cdots, N - 1$ and

\begin{align}
3.23 \quad & e^0 = 0, \quad \left(1 - \frac{\tau^2}{2} \Delta_h\right) e^1 = \frac{1}{2 \gamma} \varphi^0 \tau^2 - F^0, \\
3.24 \quad & \eta^N = 0, \quad \left(1 - \frac{\tau^2}{2} \Delta_h\right) \eta^{N-1} = -\frac{1}{2} e^N \tau^2 - G^N.
\end{align}

The discrete inner product of (3.20) and $e^{n+1} - e^{n-1}$ yields

\begin{align}
3.25 \quad & \frac{\|e^{n+1} - e^n\|^2 - \|e^n - e^{n-1}\|^2}{\tau^2} + \|\nabla_h e^{n+1}\|^2 - \|\nabla_h e^{n-1}\|^2 - \frac{1}{2} e^{n+1} - e^{n-1}, \eta^n / \gamma \\
& = -(F^n, e^{n+1} - e^{n-1}),
\end{align}

and by summing up the equations for $n = 1, \cdots, N - 1$, one can get

\begin{align}
3.26 \quad & \sum_{n=1}^{N-1} (e^{n+1} - e^{n-1}, \eta^n) / \gamma - \sum_{n=1}^{N-1} (F^n, e^{n+1} - e^{n-1})
\end{align}
which in together with (3.22) implies that
\[
\frac{\|e^N - e^{N-1}\|^2}{\tau^2} + \frac{1}{2} \|\nabla_h e^N\|^2 + \frac{1}{2} \|\nabla_h e^{N-1}\|^2 = \sum_{n=1}^{N-1} (e^{n+1} - e^{n-1}, \eta^n) / \gamma - \sum_{n=1}^{N-1} (F^n, e^{n+1} - e^{n-1})
\]
(3.26)
\[
+ \left| D_h^{-1} \left( \frac{1}{2\gamma} \eta^0 \tau - F^0 / \tau \right) \right|^2 + \frac{\tau^2}{2} \left\| \nabla_h D_h^{-1} \left( \frac{1}{2\gamma} \eta^0 \tau - F^0 / \tau \right) \right|^2.
\]
Similarly, the discrete inner product of (3.21) and \(\eta^{n+1} - \eta^{n-1}\) gives
\[
\frac{\|\eta^{n+1} - \eta^n\|^2}{\tau^2} + \frac{1}{2} \|\nabla_h \eta^0\|^2 + \frac{1}{2} \|\nabla_h \eta^{n-1}\|^2 = \sum_{n=1}^{N-1} (\eta^{n+1} - \eta^{n-1}, e^n) + \sum_{n=1}^{N-1} (G^n, \eta^{n+1} - \eta^{n-1})
\]
(3.27)
\[
= -(G^n, \eta^{n+1} - \eta^{n-1}),
\]
and by summing up the equations for \(n = 1, \ldots, N - 1\), we have
\[
\frac{\|\eta^1 - \eta^0\|^2}{\tau^2} + \frac{1}{2} \|\nabla_h \eta^0\|^2 + \frac{1}{2} \|\nabla_h \eta^1\|^2 = \sum_{n=1}^{N-1} (\eta^{n+1} - \eta^{n-1}, e^n) + \sum_{n=1}^{N-1} (G^n, \eta^{n+1} - \eta^{n-1})
\]
(3.28)
\[
+ \left| D_h^{-1} \left( \frac{1}{2} e^N \tau + G^N / \tau \right) \right|^2 + \frac{\tau^2}{2} \left\| \nabla_h D_h^{-1} \left( \frac{1}{2} e^N \tau + G^N / \tau \right) \right|^2.
\]
By using Lemma 3.2, the sum of (3.28) and \(\gamma \times (3.26)\) implies
\[
\frac{\gamma \|e^N - e^{N-1}\|^2 + \|\eta^1 - \eta^0\|^2}{\tau^2} + \frac{\gamma}{2} \|\nabla_h e^N\|^2 + \frac{\gamma}{2} \|\nabla_h e^{N-1}\|^2 + \frac{1}{2} \|\nabla_h \eta^0\|^2 + \frac{1}{2} \|\nabla_h \eta^1\|^2
\]
\[
= \sum_{n=1}^{N-1} (e^{n+1} - e^{n-1}, \eta^n) + \sum_{n=1}^{N-1} (\eta^{n+1} - \eta^{n-1}, e^n)
\]
\[
- \sum_{n=1}^{N-1} (G^n, e^{n+1} - e^{n-1}) + \sum_{n=1}^{N-1} (G^n, \eta^{n+1} - \eta^{n-1})
\]
\[
+ \frac{\gamma}{2} \left| D_h^{-1} \left( \frac{\tau}{2\gamma} \eta^0 \tau - F^0 / \tau \right) \right|^2 + \frac{\gamma}{2} \left| D_h^{-1} \left( \frac{\tau}{2\gamma} e^N \tau + G^N / \tau \right) \right|^2
\]
\[
+ \frac{\gamma}{2} \left| \nabla_h D_h^{-1} \left( \frac{\tau^2}{2\gamma} \eta^0 \tau - F^0 \right) \right|^2 + \frac{\gamma}{2} \left| \nabla_h D_h^{-1} \left( \frac{\tau^2}{2\gamma} e^N \tau + G^N \right) \right|^2
\]
\[
= (e^N, \eta^{N-1}) - (e^1, \eta^0) - \sum_{n=1}^{N-1} (G^n, e^{n+1} - e^{n-1}) + \sum_{n=1}^{N-1} (G^n, \eta^{n+1} - \eta^{n-1})
\]
\[
+ \frac{\gamma}{2} \left| D_h^{-1} \left( \frac{\tau}{2\gamma} \eta^0 \tau - F^0 / \tau \right) \right|^2 + \frac{\gamma}{2} \left| D_h^{-1} \left( \frac{\tau}{2\gamma} e^N \tau + G^N / \tau \right) \right|^2
\]
\[
+ \frac{\gamma}{2} \left| \nabla_h D_h^{-1} \left( \frac{\tau^2}{2\gamma} \eta^0 \tau - F^0 \right) \right|^2 + \frac{\gamma}{2} \left| \nabla_h D_h^{-1} \left( \frac{\tau^2}{2\gamma} e^N \tau + G^N \right) \right|^2
\]
\[
\leq (e^N, -\frac{1}{2} D_h^{-1} e^N \tau^2 - D_h^{-1} G^N) - \left( \frac{1}{2\gamma} D_h^{-1} \eta^0 \tau^2 - D_h^{-1} F^0, \eta^0 \right)
\]
\[
+ \left( \sum_{n=1}^{N-1} \tau \|G^n\|^2 \right)^{\frac{1}{2}} \left( \sum_{n=1}^{N-1} \tau \frac{1}{\tau^2} (e^{n+1} - e^n)^2 + \frac{1}{\tau^2} (e^n - e^{n-1})^2 \right)^{\frac{1}{2}}
\]
\[
+ \left( \sum_{n=1}^{N-1} \tau \|G^n\|^2 \right)^{\frac{1}{2}} \left( \sum_{n=1}^{N-1} \tau \left\| \eta^{n+1} - \eta^n \right\|^2 \right)^{\frac{1}{2}} + \frac{\tau^4}{4\gamma} \| \nabla h \eta^0 \|^2 + \frac{\tau^4}{4} \| \nabla h e_N \|^2 + \| \nabla h G^N \|^2 \\
+ 2\gamma \left\| \frac{1}{\gamma} D_h^{-1} \eta^0 \right\|^2 + 2\gamma \left\| D_h^{-1} F^0 \right\|^2 + 2 \left( \frac{\tau^2}{2} \right) D_h^{-1} e_N \left\| \right. + 2 \left. \left\| D_h^{-1} G^N \right\| \tau \right]^2 \\
+ \frac{\tau^4}{4\gamma} \| \nabla h \eta^0 \|^2 + \gamma \| \nabla h F^0 \|^2 + \frac{\tau^4}{4} \| \nabla h e_N \|^2 + \| \nabla h G^N \|^2 \\
\leq \| e_N \| \| G^N \| - \frac{\tau^2}{2} \left\| D_h^{-1/2} e_N \right\|^2 + \| F^0 \| \| \eta^0 \| - \frac{\tau^2}{2\gamma} \left\| D_h^{-1/2} \eta^0 \right\|^2 \\
+ \left( \sum_{n=1}^{N-1} \tau \| G^n \|^2 \right)^{\frac{1}{2}} \left( \sum_{n=1}^{N-1} \tau \left\| \eta^{n+1} - \eta^n \right\|^2 \right)^{\frac{1}{2}} + \frac{\tau^2}{2\gamma} \left\| D_h^{-1} \eta^0 \right\|^2 + \frac{\tau^2}{2} \left\| D_h^{-1} e_N \right\|^2 + 2 \left\| G^N \right\| \tau \right]^2 \\
+ \frac{\tau^4}{4\gamma} \| \nabla h \eta^0 \|^2 + \gamma \| \nabla h F^0 \|^2 + \frac{\tau^4}{4} \| \nabla h e_N \|^2 + \| \nabla h G^N \|^2 \\
\leq C_1 (\tau^3 + \tau^2 h) (\| e_N \| + \| \eta^0 \|) \\
+ 2(\sqrt{\gamma} + 1) C_1 \sqrt{T} (\tau^2 + h^2) \left( \sum_{n=0}^{N-1} \tau \frac{\gamma}{\tau} \left\| e^{n+1} - e^n \right\|^2 + \left\| \eta^{n+1} - \eta^n \right\|^2 \right) \left\| \nabla h e_N \right\|^2 + \frac{1}{2} \left\| \nabla h \eta^0 \right\|^2 \\
+ \frac{\tau^4}{4\gamma} \| \nabla h \eta^0 \|^2 + \gamma \| \nabla h F^0 \|^2 + \frac{\tau^4}{4} \| \nabla h e_N \|^2 \\
\leq 2C_1 \max(C_0/\gamma, C_0) (\tau^3 + \tau^2 h) \left( \frac{\gamma}{2} \left\| \nabla h e_N \right\|^2 + \frac{1}{2} \left\| \nabla h \eta^0 \right\|^2 \right) \\
+ 2(\sqrt{\gamma} + 1) C_1 \sqrt{T} (\tau^2 + h^2) \left( \sum_{n=0}^{N-1} \tau \frac{\gamma}{\tau} \left\| e^{n+1} - e^n \right\|^2 + \left\| \eta^{n+1} - \eta^n \right\|^2 \right) \left\| \nabla h e_N \right\|^2 + \frac{1}{2} \left\| \nabla h \eta^0 \right\|^2 \\
+ \frac{\tau^4}{4\gamma} \| \nabla h \eta^0 \|^2 + \gamma \| \nabla h F^0 \|^2 + \frac{\tau^4}{4} \| \nabla h e_N \|^2 \\
\leq 4C_1 \max(C_0/\gamma, C_0) \tau + 2(\sqrt{\gamma} + 1) C_1 \sqrt{T} \left( \tau^2 + h^2 \right) \left\{ \frac{\gamma}{2} \left\| \nabla h e_N \right\|^2 + \frac{1}{2} \left\| \nabla h \eta^0 \right\|^2 \\
+ \left( \sum_{n=0}^{N-1} \tau \frac{\gamma}{\tau} \left\| e^{n+1} - e^n \right\|^2 + \left\| \eta^{n+1} - \eta^n \right\|^2 \right) \right\} \\
+ (\gamma + 1) C_1 \left( \tau^3 + \tau^2 h \right) \left( \frac{\gamma}{2} \left\| \nabla h e_N \right\|^2 + \frac{1}{2} \left\| \nabla h \eta^0 \right\|^2 \right) \\
\leq C_2 (\tau^2 + h^2) E_{\tau, h} + C_2 (\tau^4 + h^4) \left( \frac{\tau^4}{4\gamma} \| \nabla h \eta^0 \|^2 + \frac{\tau^4}{4} \| \nabla h e_N \|^2 \right),
\]

where $C_2$ is some positive constant and $E_{\tau, h} := \max_{0 \leq n \leq N-1} E_n^{\tau, h}$ with

\[
(E_n^{\tau, h})^2 = \frac{\gamma}{\tau} \left\| e^{n+1} - e^n \right\|^2 + \left\| \eta^{n+1} - \eta^n \right\|^2 + \frac{\gamma}{2} \left\| \nabla h e^{n+1} \right\|^2 + \frac{\gamma}{2} \left\| \nabla h e_n \right\|^2 + \frac{1}{2} \left\| \nabla h \eta^{n+1} \right\|^2 + \frac{1}{2} \left\| \nabla h \eta^n \right\|^2.
\]

When $\tau \leq \frac{1}{4}$, the two terms $\frac{\tau^4}{4\gamma} \| \nabla h \eta^0 \|^2$ and $\frac{\tau^4}{4} \| \nabla h e_N \|^2$ are eliminated by the left-handed side,
and the last inequality thus reduces to

\[
\gamma \|e^N - e^{N-1}\|^2 + \|\eta^1 - \eta^0\|^2 + \frac{\gamma}{4} \|\nabla_h e^N\|^2 + \frac{\gamma}{4} \|\nabla_h e^{N-1}\|^2 + \frac{1}{4} \|\nabla_h \eta^0\|^2 + \frac{1}{4} \|\nabla_h \eta^1\|^2
\]

(3.31) \quad \leq C_3 (\tau^2 + h^2) E_{r,h} + C_3 (\tau^4 + h^4).

By using the last inequality, one can convert the analysis of the forward and backward boundary value problem to the analysis of an initial-value problem, where the initial errors \( e^0, e^1, \eta^0 \) and \( \eta^1 \) are well controlled.

Let \( 1 \leq k \leq N - 1 \) be fixed and sum up (3.27) for \( n = 1, \ldots, k \). Then we obtain that

\[
\frac{\|\eta^{k+1} - \eta^k\|^2}{\tau^2} + \frac{\|\nabla_h \eta^{k+1}\|^2}{2} + \frac{\|\nabla_h \eta^k\|^2}{2} - \frac{\|\nabla_h \eta^0\|^2}{2} = - \sum_{n=1}^{k} (\eta^{n+1} - \eta^{n-1}, e^n) - \sum_{n=1}^{k} (G^n, \eta^{n+1} - \eta^{n-1}).
\]

(3.32)

By summing up (3.24) for \( n = 1, \ldots, k \), we derive

\[
\frac{\|e^{k+1} - e^k\|^2 - \|e^1 - e^0\|^2}{\tau^2} + \frac{\|\nabla_h e^{k+1}\|^2}{2} + \frac{\|\nabla_h e^k\|^2}{2} - \frac{\|\nabla_h e^0\|^2}{2} = \sum_{n=1}^{k} (e^{n+1} - e^{n-1}, \eta^n) / \gamma - \sum_{n=1}^{k} (F^n, e^{n+1} - e^{n-1}).
\]

(3.33)

Next adding (3.32) with \( \gamma \times (3.33) \), and by making use of (3.31), we have (note \( e^0 = 0 \) and \( \tau^4 \leq \gamma \))

\[
\gamma \|e^{k+1} - e^k\|^2 + \|\eta^{k+1} - \eta^k\|^2 + \frac{\gamma}{2} \|\nabla_h e^{k+1}\|^2 + \frac{\gamma}{2} \|\nabla_h e^k\|^2 + \frac{1}{2} \|\nabla_h \eta^{k+1}\|^2 + \frac{1}{2} \|\nabla_h \eta^k\|^2
\]

\[
= \frac{\|e^{k+1} - e^k\|^2 + \|\eta^{k+1} - \eta^k\|^2}{\tau^2} + \gamma \|\nabla_h e^{k+1}\|^2 + \frac{\gamma}{2} \|\nabla_h e^k\|^2 + \frac{1}{2} \|\nabla_h \eta^{k+1}\|^2 + \frac{1}{2} \|\nabla_h \eta^k\|^2
\]

\[
+ \sum_{n=1}^{k} (e^{n+1} - e^{n-1}, \eta^n) - \sum_{n=1}^{k} (\eta F^n, e^{n+1} - e^{n-1}) - \sum_{n=1}^{k} (\eta^{n+1} - \eta^{n-1}, e^n) - \sum_{n=1}^{k} (G^n, \eta^{n+1} - \eta^{n-1})
\]

\[
= \frac{\|\eta^{k+1} - \eta^k\|^2}{\tau^2} + \frac{1}{2} \|\nabla_h \eta^{k+1}\|^2 + \frac{1}{2} \|\nabla_h \eta^k\|^2 + \sum_{n=1}^{k} (e^{n+1} - e^{n-1}, \eta^n) - \sum_{n=1}^{k} (\eta F^n, e^{n+1} - e^{n-1})
\]

\[- \sum_{n=1}^{k} (\eta^{n+1} - \eta^{n-1}, e^n) - \sum_{n=1}^{k} (G^n, \eta^{n+1} - \eta^{n-1})
\]

\[
\leq \frac{\tau^2}{4 \gamma} C_0^2 \|\nabla_h \eta^0\|^2 + 2 \gamma C_1^2 (\tau^2 + \tau h)^2 + \frac{\tau^4}{8 \gamma} \|\nabla_h \eta^0\|^2 + \frac{C^2_{\gamma} (\tau^3 + \tau h)^2}{2} + \frac{\|\eta^1 - \eta^0\|^2}{\tau^2}
\]

\[
+ \frac{1}{2} \|\nabla_h \eta^1\|^2 + \frac{1}{2} \|\nabla_h \eta^0\|^2 + \sum_{n=1}^{k} (e^{n+1} - e^{n-1}, \eta^n) - \sum_{n=1}^{k} (\eta F^n, e^{n+1} - e^{n-1})
\]

\[- \sum_{n=1}^{k} (\eta^{n+1} - \eta^{n-1}, e^n) - \sum_{n=1}^{k} (G^n, \eta^{n+1} - \eta^{n-1})
\]

\[
\leq \frac{1}{4 \sqrt{\gamma}} C_0^2 \|\nabla_h \eta^0\|^2 + \gamma C_1^2 (\tau^2 + \tau h)^2 + \frac{1}{8} \|\nabla_h \eta^0\|^2 + \frac{C^{2 \gamma} (\tau^2 + \tau h)^2}{2}.
\]
\[ + \frac{\|\eta - \eta^{n}\|^2}{\tau^2} + \frac{1}{2}\|\nabla_h \eta_1\|^2 + \frac{1}{2}\|\nabla_h \eta_0\|^2 + \sum_{n=1}^{k} (e^{n+1} - e^n, \eta^n) - \sum_{n=1}^{k} (\gamma F^n, e^{n+1} - e^n) \]
\[- \sum_{n=1}^{k} (\eta^{n+1} - \eta^n, e^n) - \sum_{n=1}^{k} (G^n, \eta^{n+1} - \eta^n) \]
\[ \leq C_4(\tau^2 + h^2)E_{r,h} + C_4(\tau^4 + h^4) + \left( \sum_{n=1}^{k} \gamma\|F^n\|^2 \right)^{\frac{1}{2}} \left( \sum_{n=1}^{k} \gamma\|e^{n+1} - e^n\|^2 + \gamma\|e^n - e^{n-1}\|^2 \right)^{\frac{1}{2}} \]
\[ + \left( \sum_{n=1}^{k} \tau\|G^n\|^2 \right)^{\frac{1}{2}} \left( \sum_{n=1}^{k} \tau\|\eta^{n+1} - \eta^n\|^2 + \|\eta^n - \eta^{n-1}\|^2 \right)^{\frac{1}{2}} + \frac{1}{4T} \sum_{n=1}^{k} \sum_{n=0}^{\infty} \frac{\gamma\|e^{n+1} - e^n\|^2 + \|\eta^{n+1} - \eta^n\|^2}{\tau^2} \]
\[ \leq C_4(\tau^2 + h^2)E_{r,h} + C_4(\tau^4 + h^4) + \frac{1}{2} \frac{C_3^2 T}{\gamma} \sum_{n=1}^{k} \sum_{n=0}^{\infty} T\|e^{n+1} - e^n\|^2 + \|\eta^{n+1} - \eta^n\|^2 + \|\eta^n - \eta^{n-1}\|^2 \]
\[ \quad + \left( \frac{1}{2T} + \frac{C_3^2 T}{\gamma} \right) \sum_{n=0}^{k-1} \gamma\|e^{n+1} - e^n\|^2 + \|\eta^{n+1} - \eta^n\|^2 + \frac{1}{2} \frac{C_3^2 T}{\tau^2} \sum_{n=1}^{k} \sum_{n=0}^{\infty} \|\nabla_h \eta^n\|^2 \right) . \]
\( (3.34) \)

By moving the term \( \frac{1}{2} \frac{\gamma\|e^{k+1} - e^k\|^2 + \|\eta^{k+1} - \eta^k\|^2}{\tau^2} \) to the left-handed side and using the definition of \( E_{r,h}^n \), we reduce (3.34) to

\[ |E_{r,h}^n|^2 \leq 4C_4(\tau^2 + h^2)E_{r,h} + 2C_4(\tau^4 + h^4) + \left( \frac{1}{T} + \frac{4C_3^2 T}{\gamma} \right) \sum_{n=0}^{k-1} |E_{r,h}^n|^2 , \]
\( (3.35) \)

which holds for any \( 1 \leq k \leq N - 1 \). Next we apply discrete Gronwall’s inequality to (3.35) that leads to

\[ |E_{r,h}|^2 \leq C_5(\tau^2 + h^2)E_{r,h} + C_5(\tau^4 + h^4) \leq \frac{1}{2} |E_{r,h}|^2 + \frac{1}{2} \frac{C_3^2 (\tau^2 + h^2)^2}{\gamma} + C_5(\tau^4 + h^4) , \]
\( (3.36) \)

and we have

\[ |E_{r,h}|^2 \leq C_5^2 (\tau^2 + h^2)^2 + 2C_5(\tau^4 + h^4) . \]
\( (3.37) \)

Finally, by applying the discrete Poincare inequality, the conclusion follows, and the proof of Theorem 3.3 is thus completed. \( \blacksquare \)

### 3.3. A fast preconditioned iterative solver

Next we are ready to utilize the preconditioned Krylov subspace methods, such as GMRES method [34], to solve the symmetric indefinite sparse linear system (3.7), i.e.,

\[ M_h \begin{bmatrix} \eta_h \\ p_h \end{bmatrix} := \begin{bmatrix} I_h & L_h^T \\ L_h & -I_h/\gamma \end{bmatrix} \begin{bmatrix} \eta_h \\ p_h \end{bmatrix} = \begin{bmatrix} g_h \\ f_h \end{bmatrix} . \]
\( (3.38) \)
For this type of two-by-two block sparse linear system with a saddle point structure, various numerical methods have been summarized in [1]. Our main goal here is to find an effective and efficient preconditioner which can speed up the convergence under GMRES approach by altering the spectrum distribution of the original system in a desirable way. Inspired by the framework presented in [36], we construct the following symmetric indefinite constrained preconditioner

$$ P_h = \begin{bmatrix} 0 & L_h^T \\ L_h & -I_h/\gamma \end{bmatrix}, $$

where $L_h$ has a block upper triangular structure with the same diagonal block $D_h$. Here $L^{-1}v$ can be quickly computed by applying a block forward substitution as well as the well-known FFT algorithm to solve each diagonal block. In particular, the preconditioning step $P^{-1}v$ can be done with $NM \log(M)$ operations for 1D case and with $NM_1M_2 \log(M_1M_2)$ operations for 2D case. From the structural entries of $L_h$ in $P_h$ we know that $P_h$ is nonsingular. Moreover, the right preconditioned system is given by

$$ M_h P_h^{-1} = \begin{bmatrix} I_h + \gamma^{-1} \hat{I}_h L_h^{-1} \hat{I}_h L_h^{-T} - \hat{I}_h L_h^{-1} & I_h \\ 0 & I_h \end{bmatrix}. $$

Clearly, half of the eigenvalues of $M_h P_h^{-1}$ are ones, while the remaining half are determined by

$$ R_h := (I_h + \gamma^{-1} \hat{I}_h L_h^{-1} \hat{I}_h L_h^{-T}). $$

By exploring the connection between the matrices $L_h^{-1}$ and $L_h^{-T}$ and the underlying discretized linear system (3.1)-(3.2), we are able to show the following theorem, which implies that all eigenvalues of the preconditioned coefficient matrix $M_h P_h^{-1}$ are real numbers and they are uniformly greater than one and less than an upper bound which depends only on the parameters $\gamma$ and $T$.

**Theorem 3.4.** Let $\lambda(R_h)$ be any eigenvalue of $R_h$, then $\lambda(R_h) \in \mathbb{R}$ and

$$ 1 < \lambda(R_h) < 1 + \kappa/\gamma, $$

where $\kappa$ is a positive constant that is independent of $\tau$ and $h$.

**Proof.** Using the fact that $\lambda(AB) = \lambda(BA)$, we have

$$ \lambda(\hat{I}_h L_h^{-1} \hat{I}_h L_h^{-T}) = \lambda(I_h^{1/2} L_h^{-1} \hat{I}_h^{1/2} L_h^{-T} \hat{I}_h^{1/2}) = \lambda((I_h^{1/2} L_h^{-1} \hat{I}_h^{1/2})(\hat{I}_h^{1/2} L_h^{-1} \hat{I}_h^{1/2})^T), $$

which indicates that $\lambda(\hat{I}_h L_h^{-1} \hat{I}_h L_h^{-T})$ is a real number and so is $\lambda(R_h)$. To show the lower bound, we need to invoke the fact that $\lambda(A A^T) > 0$ for any nonsingular matrix $A$. Obviously, the matrix $(\hat{I}_h^{1/2} L_h^{-1} \hat{I}_h^{1/2})$ is nonsingular and thus it follows

$$ \lambda(\hat{I}_h L_h^{-1} \hat{I}_h L_h^{-T}) = \lambda((I_h^{1/2} L_h^{-1} \hat{I}_h^{1/2})(\hat{I}_h^{1/2} L_h^{-1} \hat{I}_h^{1/2})^T) > 0. $$

To prove the upper bound of the eigenvalues, we first prove the boundedness of the matrix norm $\|L_h^{-T}\|_{\infty,2}$ induced by the following vector norm

$$ \|\tilde{q}\|_{\infty,2} := \max_{1 \leq k \leq N} \|q^n\|, $$

where $\tilde{q} = (q^1, q^2, \cdots, q^N)^T$. Let $\tilde{\psi} := (\psi^{N-2}, \psi^{N-1})^T$ be a solution of the linear system

$$ L_h \tilde{\psi} = \tilde{q}. $$
Recall that for the corresponding finite difference discretizations (3.2) of \( L^T_h \), the solution \( \tilde{\psi} \) solves

\[
\psi^{n+1} - 2\psi^n + \psi^{n-1} = -\frac{\Delta t}{2} (\psi_{n+1} + \psi_{n-1}) = q^n, \quad n = 1, 2, \ldots, N - 1
\]

with the initial conditions \( \psi^{N-1} = \tau^2 D^{-1}_h q^N / 2 \) and \( \psi^N = 0 \).

We consider the discrete inner product of (3.42) with \( \psi^{n+1} - \psi^{n-1} \), which gives

\[
\frac{\| \psi^{n+1} - \psi^n \|^2}{\tau^2} - \| \psi^n - \psi^{n-1} \|^2 + \frac{\| \nabla_h \psi^{n+1} \|^2 - \| \nabla_h \psi^{n-1} \|^2}{2} = (q^n, \psi^{n+1} - \psi^{n-1}),
\]

and by summing up the equations for \( n = k, \ldots, N - 1 \), we obtain

\[
\frac{\| \psi^k - \psi^{k-1} \|^2}{\tau^2} + \frac{\| \nabla_h \psi^k \|^2 + \| \nabla_h \psi^{k-1} \|^2}{2} = \sum_{n=k}^{N-1} \frac{\| q^n \|^2}{\tau^2} + \frac{\| \psi^n - \psi^{n-1} \|^2}{\tau^2} + \| \nabla_h \psi^n \|^2 + \| \nabla_h \psi^{n-1} \|^2 \leq \sum_{n=k}^{N-1} \frac{\| q^n \|^2}{\tau^2} + \frac{\| \psi^n - \psi^{n-1} \|^2}{\tau^2} + \| \nabla_h \psi^n \|^2 + \| \nabla_h \psi^{n-1} \|^2
\]

\[
+ \frac{\tau^2 \| D^{-1}_h q^N \|^2}{4} + \frac{\tau^2 \| \nabla_h D^{-1}_h q^N \|^2}{4}
\]

\[
(3.44)
\]

where we have used Lemma 3.2 in the last step. By applying Gronwall's inequality, we derive

\[
(3.45) \quad \max_{1 \leq k \leq N} \left( \frac{\| \psi^k - \psi^{k-1} \|^2}{\tau^2} + \frac{\| \nabla_h \psi^k \|^2 + \| \nabla_h \psi^{k-1} \|^2}{2} \right) \leq C_6 \sum_{n=k}^{N} \tau \| q^n \|^2 \leq C_6 \max_{1 \leq k \leq N} \| q^n \|^2,
\]

where \( C_6 \) is some positive constant which is independent of \( \tau \) and \( h \) (but may depend on \( T \)). The last inequality above, together with (3.10), yields

\[
\| \tilde{\psi} \|_{\infty, 2} = \| L^{-T} \tilde{q} \|_{\infty, 2} \leq C_7 \| \tilde{q} \|_{\infty, 2},
\]

that is \( \| L^{-T} \|_{\infty, 2} \leq C_7 \) for some positive constant \( C_7 \). Similarly, one can show that \( \| L^{-1} \|_{\infty, 2} \leq C_8 \) also holds for some positive constant \( C_8 \). Moreover, it is obvious that \( \| \tilde{I} \|_{\infty, 2} = \| \tilde{I} \|_{\infty, 2} \leq 1 \).

Therefore,

\[
\lambda(\tilde{I}_h L^{-1}_h \tilde{I}_L^{-1} h L^{-T}_h) \leq \| \tilde{I}_h L^{-1}_h \tilde{I}_L^{-1} h L^{-T}_h \|_{\infty, 2} \leq C_7 C_8 =: \kappa,
\]

where \( \kappa \) is a positive constant that is independent of \( \tau \) and \( h \). The proof is thus completed. 

To illustrate the effect of our theoretical estimates, in the following Fig. 1 and Fig. 2, we plot the numerically computed eigenvalues of \( M_h \) and \( M_h P^{-1}_h \) for Ex. 1 with \( \gamma = 10^{-2} \) using \( M = N = 16 \) and \( M = N = 32 \), respectively. As anticipated, the eigenvalues of preconditioned systems are highly accumulated around one within a uniformly bounded interval, which reasonably envisions a fast convergence with the preconditioned GMRES method. Notice that the eigenvalue distributions of \( M_h P^{-1}_h \) fall nicely to our estimated bounds. Such desirable clustered spectrum distributions after preconditioning are achievable with our proposed implicit scheme, which is hardly possible for the standard explicit scheme. According to our estimates, the preconditioned GMRES method may show a slower convergence rate as the regularization parameter \( \gamma \) decreases to zero, but this is an inherent problem due to the very weak convexity of the underlying cost function. How to come up with an effective and regularization parameter robust preconditioner in such a case is another active and widely open research topic, with many recent contributions [31, 35, 37] as well as references therein.
4. Numerical examples. In this section, we will provide several numerical examples to validate the obtained theoretical results and to demonstrate the high efficiency of our proposed approach. All simulations are implemented using MATLAB R2014a on a laptop PC with Intel(R) Core(TM) i3-3120M CPU@2.50GHz and 12GB RAM. The CPU time (in seconds) is estimated by timing functions tic/toc.

For simplicity, we will denote the discrete $L^2$ norm on $Q$ in short by $\| \cdot \|$, that is $\| \cdot \| := \| \cdot \|_{L^2(Q)}$. Based on the error estimates, we also defined the discrete $L^\infty(L^2)$ norm $\| \cdot \|_{L^\infty(L^2)}$. We first compute the discrete $L^\infty(L^2)$ norms of state and adjoint state approximation errors $e^h_y = \| y_h - y \|_{L^\infty(L^2)}$ and $e^h_p = \| p_h - p \|_{L^\infty(L^2)}$ and then estimate the experimental order of accuracy by calculating the logarithmic ratios of the approximation errors between two successive refined meshes, i.e.,

$$\text{Order} = \log_2 \left( \frac{e^{2h}}{e^h} \right),$$

which should be close to two for a second-order accuracy. For initialization of the iterative methods, the state $y$ and the adjoint state $p$ are set to be zero, and the stopping criterion is chosen to be

$$\text{Rel. Res.} := \sqrt{\| r^{(k)}_y \|^2 + \| r^{(k)}_p \|^2} \leq \text{tol},$$

where $r^{(k)}_y$ and $r^{(k)}_p$ denote the residuals after $k$-th iteration. In our numerical simulations, according to the level of discretization errors as well as the regularization parameter $\gamma$, we set $\text{tol} = \gamma \times 10^{-6}$ and $\text{tol} = \gamma \times 10^{-4}$ for 1D and 2D examples, respectively.

Example 1. Let $\Omega = (0,1)$ and $T = 2$. Choose $y_0(x) = \sin(\pi x)$, $y_1(x) = 0$,

$$f = -\pi^2 \sin(\pi x) \cos(\pi t) + \pi^2 \sin(\pi x) \cos(\pi t) - \sin(\pi x) (t - T)^2 / \gamma,$$
and
\[ g(x, t) = 2 \sin(\pi x) + \pi^2 \sin(\pi x)(t - T)^2 + \sin(\pi x) \cos(\pi t), \]
such that the exact solution is
\[ y(x, t) = \sin(\pi x) \cos(\pi t) \]
and
\[ p(x, t) = \sin(\pi x)(t - T)^2. \]

The numerical results of our implicit scheme solving by the preconditioned GMRES method with regularization parameter \( \gamma = 10^{-2} \) and \( \gamma = 10^{-4} \) are reported in Table 2 and 3, respectively. The implicit scheme delivers a clear second-order accuracy, which validates our proved error estimates for the implicit scheme. The required number of iterations for achieving convergence criterion is independent of mesh size and the computational CPU time grows roughly as a linearithmic function \( O(m \log(m)) \) with respect to the degrees of freedom \( m \), which shows the excellent effectiveness of our proposed preconditioner. However, comparing the column ‘Iter’ in Table 2 and 3, it takes more iterations for a smaller \( \gamma \), which is reasonable by our Theorem 3.4.

For comparison, we also give the corresponding results of the same implicit scheme solved by MATLAB’s backslash sparse direct solver in Tables 4 and 5. Notice that MATLAB’s sparse direct solver is highly optimized by making the best use of the sparsity. Hence it seems reasonable to use it as a benchmark. Nevertheless, our proposed implicit scheme with preconditioned GMRES is computationally more efficient than the sparse direct solver as the size of the discretized system is increased. For such 1D problems, it seems that sparse direct solver still has certain marginal advantage in CPU time when the size of the discretized system is small, but this is not the case for handling more complex systems such as the 2D problems appeared in next following example.

### Table 2

**Results for Ex. 1 with \( \gamma = 10^{-2} \) (Implicit scheme with preconditioned GMRES).**

<table>
<thead>
<tr>
<th>((M, N))</th>
<th>(e_y^h) Order</th>
<th>(e_p^h) Order</th>
<th>Iter</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(128,128)</td>
<td>1.69e-03</td>
<td>1.00e-04</td>
<td>8</td>
<td>0.332</td>
</tr>
<tr>
<td>(256,256)</td>
<td>4.23e-04</td>
<td>2.50e-05</td>
<td>8</td>
<td>0.851</td>
</tr>
<tr>
<td>(512,512)</td>
<td>1.06e-04</td>
<td>6.26e-06</td>
<td>8</td>
<td>2.661</td>
</tr>
<tr>
<td>(1024,1024)</td>
<td>2.64e-05</td>
<td>1.56e-06</td>
<td>8</td>
<td>7.655</td>
</tr>
<tr>
<td>(2048,2048)</td>
<td>6.61e-06</td>
<td>3.91e-07</td>
<td>8</td>
<td>31.419</td>
</tr>
</tbody>
</table>

### Table 3

**Results for Ex. 1 with \( \gamma = 10^{-4} \) (Implicit scheme with preconditioned GMRES).**

<table>
<thead>
<tr>
<th>((M, N))</th>
<th>(e_y^h) Order</th>
<th>(e_p^h) Order</th>
<th>Iter</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(128,128)</td>
<td>1.73e-03</td>
<td>5.53e-06</td>
<td>21</td>
<td>0.836</td>
</tr>
<tr>
<td>(256,256)</td>
<td>4.33e-04</td>
<td>1.38e-06</td>
<td>21</td>
<td>2.581</td>
</tr>
<tr>
<td>(512,512)</td>
<td>1.08e-04</td>
<td>3.40e-07</td>
<td>21</td>
<td>7.385</td>
</tr>
<tr>
<td>(1024,1024)</td>
<td>2.71e-05</td>
<td>7.98e-08</td>
<td>21</td>
<td>24.218</td>
</tr>
<tr>
<td>(2048,2048)</td>
<td>6.77e-06</td>
<td>1.80e-08</td>
<td>22</td>
<td>116.744</td>
</tr>
</tbody>
</table>

### Table 4

**Results for Ex. 1 with \( \gamma = 10^{-2} \) (Implicit scheme with sparse direct solver).**

<table>
<thead>
<tr>
<th>((M, N))</th>
<th>(e_y^h) Order</th>
<th>(e_p^h) Order</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(128,128)</td>
<td>1.69e-03</td>
<td>1.00e-04</td>
<td>0.439</td>
</tr>
<tr>
<td>(256,256)</td>
<td>4.23e-04</td>
<td>2.50e-05</td>
<td>3.820</td>
</tr>
<tr>
<td>(512,512)</td>
<td>1.06e-04</td>
<td>6.26e-06</td>
<td>32.508</td>
</tr>
<tr>
<td>(1024,1024)</td>
<td>2.64e-05</td>
<td>1.57e-06</td>
<td>322.786</td>
</tr>
</tbody>
</table>
Table 5
Results for Ex. 1 with $\gamma = 10^{-4}$ (Implicit scheme with sparse direct solver).

<table>
<thead>
<tr>
<th>$(M, N)$</th>
<th>$e_y^h$ Order</th>
<th>$e_p^h$ Order</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(128,128)</td>
<td>1.73e-03 -</td>
<td>5.54e-06 -</td>
<td>0.224</td>
</tr>
<tr>
<td>(256,256)</td>
<td>4.33e-04 2.00</td>
<td>1.39e-06 2.00</td>
<td>1.445</td>
</tr>
<tr>
<td>(512,512)</td>
<td>1.08e-04 2.00</td>
<td>3.47e-07 2.00</td>
<td>7.399</td>
</tr>
<tr>
<td>(1024,1024)</td>
<td>2.70e-05 2.00</td>
<td>8.68e-08 2.00</td>
<td>54.660</td>
</tr>
</tbody>
</table>

Table 6
Results for Ex. 2 with $\gamma = 10^{-2}$ (Implicit scheme with preconditioned GMRES).

<table>
<thead>
<tr>
<th>$(M_1, M_2, N)$</th>
<th>$e_y^h$ Order</th>
<th>$e_p^h$ Order</th>
<th>Iter</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(32,32,32)</td>
<td>1.98e-02 -</td>
<td>2.02e-03 -</td>
<td>7</td>
<td>0.355</td>
</tr>
<tr>
<td>(64,64,64)</td>
<td>4.99e-03 1.99</td>
<td>5.04e-04 2.00</td>
<td>7</td>
<td>1.969</td>
</tr>
<tr>
<td>(128,128,128)</td>
<td>1.25e-03 2.00</td>
<td>1.26e-04 2.00</td>
<td>7</td>
<td>14.539</td>
</tr>
<tr>
<td>(256,256,256)</td>
<td>3.13e-04 2.00</td>
<td>3.15e-05 2.00</td>
<td>7</td>
<td>124.004</td>
</tr>
</tbody>
</table>

Table 7
Results for Ex. 2 with $\gamma = 10^{-4}$ (Implicit scheme with preconditioned GMRES).

<table>
<thead>
<tr>
<th>$(M_1, M_2, N)$</th>
<th>$e_y^h$ Order</th>
<th>$e_p^h$ Order</th>
<th>Iter</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(32,32,32)</td>
<td>3.87e-02 -</td>
<td>7.47e-05 -</td>
<td>16</td>
<td>0.775</td>
</tr>
<tr>
<td>(64,64,64)</td>
<td>9.69e-03 2.00</td>
<td>1.94e-05 1.94</td>
<td>16</td>
<td>5.222</td>
</tr>
<tr>
<td>(128,128,128)</td>
<td>2.42e-03 2.00</td>
<td>4.71e-06 2.04</td>
<td>16</td>
<td>35.181</td>
</tr>
<tr>
<td>(256,256,256)</td>
<td>6.05e-04 2.00</td>
<td>1.11e-06 2.09</td>
<td>17</td>
<td>374.137</td>
</tr>
</tbody>
</table>

Table 8
Results for Ex. 2 with $\gamma = 10^{-2}$ (Implicit scheme with sparse direct solver).

<table>
<thead>
<tr>
<th>$(M_1, M_2, N)$</th>
<th>$e_y^h$ Order</th>
<th>$e_p^h$ Order</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(8,8,8)</td>
<td>2.81e-01 -</td>
<td>3.21e-02 -</td>
<td>0.012</td>
</tr>
<tr>
<td>(16,16,16)</td>
<td>7.68e-02 1.87</td>
<td>8.10e-03 1.99</td>
<td>0.538</td>
</tr>
<tr>
<td>(32,32,32)</td>
<td>1.98e-02 1.95</td>
<td>2.02e-03 2.01</td>
<td>43.670</td>
</tr>
</tbody>
</table>

Table 9
Results for Ex. 2 with $\gamma = 10^{-4}$ (Implicit scheme with sparse direct solver).

<table>
<thead>
<tr>
<th>$(M_1, M_2, N)$</th>
<th>$e_y^h$ Order</th>
<th>$e_p^h$ Order</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(8,8,8)</td>
<td>6.03e-01 -</td>
<td>1.01e-03 -</td>
<td>0.015</td>
</tr>
<tr>
<td>(16,16,16)</td>
<td>1.53e-01 1.98</td>
<td>2.67e-04 1.92</td>
<td>0.661</td>
</tr>
<tr>
<td>(32,32,32)</td>
<td>3.87e-02 1.98</td>
<td>7.47e-05 1.83</td>
<td>40.224</td>
</tr>
</tbody>
</table>

Example 2. Let $\Omega = (0,1)^2$ and $T = 2$. Choose

$$y_0(x_1, x_2) = \sin(\pi x_1) \sin(\pi x_2), \quad y_1(x_1, x_2) = \sin(\pi x_1) \sin(\pi x_2),$$

$$f = (1 + 2\pi^2)e^t \sin(\pi x_1) \sin(\pi x_2) - (t - T)^2 \sin(\pi x_1) \sin(\pi x_2)/\gamma,$$

and

$$g(x_1, x_2, t) = (e^t + 2 + 2\pi^2(t - T)^2) \sin(\pi x_1) \sin(\pi x_2),$$
such that the exact solution is

\[ y(x_1, x_2, t) = e^t \sin(\pi x_1) \sin(\pi x_2) \quad \text{and} \quad p(x_1, x_2, t) = (t - T)^2 \sin(\pi x_1) \sin(\pi x_2). \]

The numerical results for Ex. 2 are reported in Table 6, 7, 8, and 9. Similar conclusions can be drawn as in the previous example. With a laptop PC, the sparse direct solver can not handle a $64 \times 64 \times 64$ mesh (about 262,144 unknowns) due to the requirement of high memory costs, while our preconditioned GMRES method can solve a $256 \times 256 \times 256$ mesh (about 16,777,216 unknowns) in about two minutes. The key difference is that the preconditioned GMRES approach has a linearithmic time complexity while the sparse direct solver usually does not. This shows a marvelous advantage of iterative methods over (sparse) direct solvers in handling large-scale problems such as the one we are confronting here.

Next we demonstrate by numerical examples that our proposed approach is also valid for the cases that the tracking trajectory $g$ is discontinuous as well as the control constraint is appeared, though theoretical arguments are not available at this point. This indicates that it may be a worthy of a further study under our framework, in particular, for the control constraint cases.

**Table 10**

*Results for Ex. 3 with $\gamma = 10^{-2}$ (Implicit scheme with preconditioned GMRES).*

<table>
<thead>
<tr>
<th>$(M_1, M_2, N)$</th>
<th>$|y_h - g|$</th>
<th>Rel. Res.</th>
<th>Iter</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(32,32,32)</td>
<td>5.12272</td>
<td>8.56e-07</td>
<td>8</td>
<td>0.515</td>
</tr>
<tr>
<td>(64,64,64)</td>
<td>5.23732</td>
<td>8.59e-07</td>
<td>8</td>
<td>3.044</td>
</tr>
<tr>
<td>(128,128,128)</td>
<td>5.29387</td>
<td>8.20e-07</td>
<td>8</td>
<td>17.821</td>
</tr>
<tr>
<td>(256,256,256)</td>
<td>5.32197</td>
<td>7.94e-07</td>
<td>8</td>
<td>150.479</td>
</tr>
</tbody>
</table>

**Table 11**

*Results for Ex. 3 with $M_1 = M_2 = N = 32$ (Implicit scheme with preconditioned GMRES).*

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$|y_h - g|$</th>
<th>Rel. Res.</th>
<th>Iter</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0e-01</td>
<td>5.54918</td>
<td>1.01e-07</td>
<td>5</td>
<td>0.342</td>
</tr>
<tr>
<td>1.0e-02</td>
<td>5.12272</td>
<td>8.56e-07</td>
<td>8</td>
<td>0.448</td>
</tr>
<tr>
<td>1.0e-03</td>
<td>4.10046</td>
<td>5.45e-07</td>
<td>16</td>
<td>0.797</td>
</tr>
<tr>
<td>1.0e-04</td>
<td>2.97933</td>
<td>6.77e-07</td>
<td>33</td>
<td>1.767</td>
</tr>
<tr>
<td>1.0e-05</td>
<td>2.10683</td>
<td>9.17e-07</td>
<td>61</td>
<td>4.072</td>
</tr>
<tr>
<td>1.0e-06</td>
<td>1.41148</td>
<td>9.69e-07</td>
<td>108</td>
<td>10.159</td>
</tr>
<tr>
<td>1.0e-07</td>
<td>1.02305</td>
<td>9.72e-07</td>
<td>160</td>
<td>16.465</td>
</tr>
<tr>
<td>1.0e-08</td>
<td>0.93657</td>
<td>9.91e-07</td>
<td>186</td>
<td>18.938</td>
</tr>
</tbody>
</table>

**Fig. 3.** Target state $g$ and computed optimal state $y_h$ with $\gamma = 10^{-4}$ at $t = T/2$ in Ex. 3 ($M_1 = M_2 = N = 32$)
Example 3 [23]. Let $\Omega = (0,1)^2$ and $T = 2$. Choose

$$y_0(x_1,x_2) = \sin(\pi x_1) \sin(\pi x_2), \quad y_1(x_1,x_2) = 0, \quad f \equiv 0,$$

and a non-attainable discontinuous target function $g$ given by

$$g(x_1,x_2,t) = \begin{cases} 
10x_2 & \text{if } x_1 < 0.5, \\
1 & \text{otherwise.}
\end{cases}$$

In this example, since the exact solutions are not known, we will report the relative residuals of computed approximations to verify the convergence of our preconditioned GMRES iterative solver. We further compute the difference norm $\|y - g\|$ to show how the computed optimal state $y$ approaches the desired target $g$, as the regularization parameter $\gamma$ is decreased. Here we use a fixed $\text{tol} = 10^{-6}$ so that the iteration numbers are comparable as $\gamma$ varies. With a fixed $\gamma = 10^{-2}$, Table 10 presents a quite similar mesh-independent convergence result as previous cases.

With varying $\gamma$, the computational results with a fixed mesh $M_1 = M_2 = N = 32$ are given in Table 11. As expected from our Theorem 3.4, the required iteration numbers in column “Iter” grows almost linearly as $\gamma$ is decreased. Moreover, Fig. 3 and 4 illustrate the corresponding desired state $y$ with $\gamma = 10^{-4}$, $\gamma = 10^{-6}$, and $\gamma = 10^{-8}$, respectively. Clearly, a smaller $\gamma$ leads to smaller tracking errors, but with more computational costs.

Example 4. In this example, we numerically investigate the applicability and efficiency of our proposed approach for the case with bilateral box constraints on control (see, e.g., [23] and references therein). More specifically, we enforce the constraint $u \in U_{ad} := \{u \in U | u_a \leq u \leq u_b\}$ with $u_a, u_b \in U$. We implement our proposed algorithm within the framework of semismooth Newton (SSN) method [23]. The generalization of our modified finite difference scheme is straightforward. In each Newton iteration, our preconditioned iterative solver is employed to approximate the corresponding linearized Jacobian system, which has a very similar structure as in our unconstrained case.

Let $\Omega = (0,1)$ and $T = 2$. Choose $y_0(x) = \sin(\pi x)$, $y_1(x) = 0$, $u_a = 5$, $u_b = 10$

$$f = -\pi^2 \sin(\pi x) \cos(\pi t) + \pi^2 \sin(\pi x) \cos(\pi t) - \max\{u_a, \min\{u_b, \sin(\pi x)(t - T)^2 / \gamma\}\},$$

and

$$g(x,t) = 2 \sin(\pi x) + \pi^2 \sin(\pi x)(t - T)^2 + \sin(\pi x) \cos(\pi t),$$

such that the exact solution is $y(x,t) = \sin(\pi x) \cos(\pi t)$ and $p(x,t) = \sin(\pi x)(t - T)^2$. The corresponding optimal control can be derived from the projection $u = \max\{u_a, \min\{u_b, p / \gamma\}\}$. 
Based on numerical experiments, we choose $\text{tol} = 10^{-8}$ and $\text{tol} = 10^{-4}$ for the relative residual based the stopping criteria of the outer SSN iterations as well as the inner GMRES iterations, respectively. To balance the outer and inner tolerances usually delivers a better overall performance. For simplicity, we set the initial guesses of the outer SSN iterations and the inner GMRES iterations to be zero. Such a choice seems to work very well for our example due to a reasonable convexity (i.e. $\gamma$ is not too small) even though the SSN method usually is locally convergent in general.

**Table 12**

<table>
<thead>
<tr>
<th>$(M, N)$</th>
<th>$e^h_{\text{up}}$</th>
<th>Order</th>
<th>$e^h_{\text{up}}$</th>
<th>Order</th>
<th>SSN Iter</th>
<th>GMRES Iters</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(128,128)</td>
<td>1.80e-03</td>
<td>–</td>
<td>2.62e-04</td>
<td>–</td>
<td>7</td>
<td>[2;2;3;2;2;2;2]</td>
<td>1.420</td>
</tr>
<tr>
<td>(256,256)</td>
<td>4.49e-04</td>
<td>2.00</td>
<td>6.56e-05</td>
<td>2.00</td>
<td>6</td>
<td>[2;2;3;2;2;2;2]</td>
<td>2.440</td>
</tr>
<tr>
<td>(512,512)</td>
<td>1.12e-04</td>
<td>2.00</td>
<td>1.64e-05</td>
<td>2.00</td>
<td>6</td>
<td>[2;2;3;2;2;2;2]</td>
<td>7.133</td>
</tr>
<tr>
<td>(1024,1024)</td>
<td>2.80e-05</td>
<td>2.01</td>
<td>3.89e-06</td>
<td>2.07</td>
<td>5</td>
<td>[2;2;3;2;2]</td>
<td>17.240</td>
</tr>
<tr>
<td>(2048,2048)</td>
<td>6.91e-06</td>
<td>2.02</td>
<td>8.37e-07</td>
<td>2.22</td>
<td>5</td>
<td>[2;2;3;2;2;2;2]</td>
<td>61.240</td>
</tr>
</tbody>
</table>

**Table 13**

<table>
<thead>
<tr>
<th>$(M, N)$</th>
<th>$e^h_{\text{up}}$</th>
<th>Order</th>
<th>$e^h_{\text{up}}$</th>
<th>Order</th>
<th>SSN Iter</th>
<th>GMRES Iters</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(128,128)</td>
<td>1.79e-03</td>
<td>–</td>
<td>2.57e-04</td>
<td>–</td>
<td>4</td>
<td>[2;2;2;2]</td>
<td>0.484</td>
</tr>
<tr>
<td>(256,256)</td>
<td>4.49e-04</td>
<td>2.00</td>
<td>6.44e-05</td>
<td>2.00</td>
<td>4</td>
<td>[2;2;2;2]</td>
<td>1.235</td>
</tr>
<tr>
<td>(512,512)</td>
<td>1.12e-04</td>
<td>2.00</td>
<td>1.61e-05</td>
<td>2.00</td>
<td>3</td>
<td>[2;2;2]</td>
<td>2.781</td>
</tr>
<tr>
<td>(1024,1024)</td>
<td>2.80e-05</td>
<td>2.00</td>
<td>4.02e-06</td>
<td>2.00</td>
<td>3</td>
<td>[2;2;2]</td>
<td>8.900</td>
</tr>
<tr>
<td>(2048,2048)</td>
<td>7.01e-06</td>
<td>2.00</td>
<td>1.01e-06</td>
<td>2.00</td>
<td>3</td>
<td>[2;2;2]</td>
<td>32.858</td>
</tr>
</tbody>
</table>

**Fig. 5.** Computed state $y_h$ and control $u_h = \max\{5, \min\{10, p_h/\gamma\}\}$ with $\gamma = 10^{-2}$ in Ex. 4 ($M = N = 64$)

The numerical results for Ex. 4 are reported in Tables 12 and 13, from which one can see a clear second-order accuracy of our modified scheme as well as a mesh-independent convergence of the embed SSN approach. The column ‘GMRES Iters’ lists the corresponding numbers of preconditioned GMRES iterations that work on solving the Jacobian linear system at each Newton iteration. It shows that only about two preconditioned GMRES iterations are sufficient to approximately solve the algebraic system up to the given tolerance $10^{-4}$. For this particular example, the preconditioned GMRES convergence rate also seems to be independent of the mesh size and insensitive to the regularization parameter $\gamma$. This nice convergent property indicates that our
preconditioned iterative solver works well in this case, though theoretical verification is not available yet. Fig. 5 depicts the computed optimal state and optimal control under the given control constraints.

5. Concluding Remarks. In this paper, we have shown the second-order accuracy of a new implicit scheme in time for a system of forward-and-backward coupled wave equations arising from wave optimal control problems. The proposed scheme is unconditionally stable and the well-developed discretized structure allows us to develop a fast solver with an effective preconditioner by using the preconditioned iterative method. Numerical experiments demonstrate a high degree of consistency with our theoretical analysis.

In view of the analysis given in [18], the hypothesis $y, p \in C^{4,4}(Q)$ may be further relaxed by using integral representations of the truncation errors instead of Taylor series expansions. If we denote $S_h$ as the standard finite element space of piece-wise polynomials defined on $\Omega$, and define the operator $\Delta_h : S_h \to S_h$ by

$$(\Delta_h u_h, v_h) = -(\nabla u_h, \nabla v_h), \quad \forall v_h \in S_h,$$

then the scheme (3.1)-(3.2) and (3.5)-(3.6) consists of finite difference time discretization and finite element space discretization. The proposed implicit time-stepping scheme as well as the obtained preconditioner are also expected to work seamlessly with finite element discretizations in space. However, it is not clear at this point whether our preconditioning strategy can be extended to finite element or discontinuous Galerkin discretizations in time, since the system discretization structure may not be preserved.

The preliminary numerical results in Example 4 suggest that our proposed method is also very promising when applied to the constrained case. A further discussion will be reported elsewhere.

Appendix A: the conditioning number of $F_h$.

In this appendix, we show that, under certain condition, the conditioning number of $F_h$ appeared in section 2 grows exponentially as $h \to 0$.

**Lemma A.1.** Let $\lambda_1$ be the maximal eigenvalue of the positive definite discrete operator $-\Delta_h$. If $\tau^2 \lambda_1 = \kappa > 16/3$, then there holds

$$\text{cond}(F_h) := \|F_h\|_2 \|F_h^{-1}\|_2 \geq \frac{1}{2K} e^{\sigma \kappa \tau \sqrt{\kappa}},$$

where $c_k = \sqrt{K - 4} - \sqrt{\kappa}/2$, $\sigma_K = \frac{\ln(1 + c_K \sqrt{\kappa})}{c_K \sqrt{\kappa}} > 0$, and $\| \cdot \|_2$ denotes the matrix 2-norm.

**Proof of Lemma A.1.** Consider the linear system of the 2D case

(A.1) $F_h \vec{y} = \vec{f},$

where $\vec{y} = (\vec{y}_0, \ldots, \vec{y}_{N-1})^T$ is a column vector such that each $\vec{y}_n$ is a row vector $\vec{y}_n = (\bar{y}_{n1}, \ldots, \bar{y}_{nM})$, with $M = (M_1 - 1)(M_2 - 1)$, and similarly, $\vec{f} = (\bar{f}_0, \ldots, \bar{f}_{N-1})^T$ is a column vector such that each $\bar{f}_n$ is a row vector $\bar{f}_n = (\bar{f}_{n1}, \ldots, \bar{f}_{nM})$. Note that the solution of (A.1) is given by

(A.2) $\vec{y}_0 = \tau^2 \bar{f}_0,$
(A.3) $\vec{y}_1 = (2 + \tau^2 \Delta_h)\vec{y}_0 + \tau^2 \bar{f}_1,$
(A.4) $\vec{y}_n = (2 + \tau^2 \Delta_h)\vec{y}_{n-1} - \vec{y}_{n-2} + \tau^2 \bar{f}_n, \quad n \geq 2,$
Choose an orthogonal system of eigenvectors of $\Delta_h$, say $\phi_j$, $j = 1, \cdots, M$, which corresponds to the eigenvalues $\lambda_j$, $j = 1, \cdots, M$. Expressing $\bar{y}_n$ and $\bar{f}_n$ as
$$
\bar{y}_n = \sum_{j=1}^{M} \alpha_{n,j} \phi_j \quad \text{and} \quad \bar{f}_n = \sum_{j=1}^{M} \gamma_{n,j} \phi_j,
$$
and substituting these identities into (A.2)-(A.4) by setting $\bar{f}_n = 0$ for $n \geq 2$, we obtain
\begin{align*}
\alpha_{0,j} &= \tau^2 \gamma_{0,j}, \quad n \geq 2, \\
\alpha_{1,j} &= (2 - \tau^2 \lambda_j) \tau^2 \gamma_{0,j} + \tau^2 \gamma_{1,j}, \\
\alpha_{n,j} &= (2 - \tau^2 \lambda_j) \alpha_{n-1,j} - \alpha_{n-2,j}, \quad n \geq 2,
\end{align*}
whose iterative solution is given by
$$
\alpha_{n,j} = \frac{\alpha_{1,j} - \alpha_{0,j} b_j}{a_j - b_j} \phi_j^n + \frac{\alpha_{0,j} a_j - \alpha_{1,j} b_j}{a_j - b_j} \phi_j^n,
$$
where
$$
a_j = 1 - \frac{1}{2} \tau^2 \lambda_j + \sqrt{-4 \tau^2 \lambda_j + \tau^4 \lambda_j^2}, \quad b_j = 1 - \frac{1}{2} \tau^2 \lambda_j - \sqrt{-4 \tau^2 \lambda_j + \tau^4 \lambda_j^2}.
$$
Suppose that $\lambda_1$ is the largest among $\lambda_j$, $j = 1, \cdots, M$. When $\tau^2 \lambda_1 = \kappa > 16/3$, we have the bounds
$$
1 + c_\kappa \tau \sqrt{\lambda_1} \leq a_1 \leq 1 + \kappa, \quad \text{with} \quad c_\kappa = \sqrt{\kappa - 4} - \sqrt{\kappa}/2 > 0.
$$
In this case, for the following chosen special data
$$
\gamma_{0,1} = 1, \quad \gamma_{1,1} = a_1 - 2 + \tau^2 \lambda_1 > 0, \quad \gamma_{n,1} = 0 \quad \text{for} \quad n \geq 2, \\
\gamma_{n,j} = 0 \quad \text{for all} \quad j \neq 1 \quad \text{and} \quad n \geq 0,
$$
we have (by straightforward verification)
\begin{align*}
\frac{\alpha_{1,1} - \alpha_{0,1} b_1}{a_1 - b_1} &= \tau^2 \quad \text{and} \quad \frac{\alpha_{0,1} a_1 - \alpha_{1,1} b_1}{a_1 - b_1} = 0, \\
\|\bar{f}\|_2 &= \sqrt{\sum_{n,j} |\gamma_{n,j}|^2} = 1 + a_1 - 2 + \tau^2 \lambda_1 \leq 2\kappa,
\end{align*}
and therefore (by the recursion formula and relations $N = T/\tau$, $\tau \sqrt{\lambda_1} = \sqrt{\kappa}$)
$$
\alpha_{N,1} = \tau^2 a_1^N \geq \tau^2 (1 + c_\kappa \tau \sqrt{\lambda_1})^N = \tau^2 e^{N \ln(1 + c_\kappa \tau \sqrt{\lambda_1})} = \tau^2 e^{\sigma_\kappa \epsilon \tau \sqrt{\lambda_1}} \geq \frac{\tau^2}{2\kappa} e^{\sigma_\kappa \epsilon \tau \sqrt{\lambda_1}} \|\bar{f}\|_2,
$$
where we have used (A.12) in the last inequality. Therefore, we obtain (for the above chosen $\bar{f}$)
$$
\|F_h^{-1}\bar{f}\|_2 = \|\bar{y}\|_2 = \sqrt{\sum_{n,j} |\alpha_{n,j}|^2} \geq |\alpha_{N,1}| \geq \frac{\tau^2}{2\kappa} e^{\sigma_\kappa \epsilon \tau \sqrt{\lambda_1}} \|\bar{f}\|_2,
$$
which, by the definition of matrix 2-norm, implies
$$
\|F_h^{-1}\|_2 \geq \frac{\tau^2}{2\kappa} e^{\sigma_\kappa \epsilon \tau \sqrt{\lambda_1}}.
$$
On the other hand, it is obvious that
\[ \| F_h \|_2 \geq 1/\tau^2. \]
The last two inequalities yield the conclusion and the proof of Lemma A.1 is completed.

**Remark A.1.** For the discrete operator \(-\Delta_h\), we always have \( \lambda_1 \geq c_*^2/h^2 \) for some \( c_* > 0 \). It follows that
\[ \text{cond}(F_h) \geq \frac{1}{2\kappa} e^{c_* \sigma c_* T/h}. \]
This implies \( \text{cond}(F_h) \) grows exponentially as \( h \) tends to zero, whenever the grid ratio \( r = \tau/h \) is such that \( \kappa > 16/3 \). For example, in 1D case with a central finite difference discretization, we in fact have \( |11| \lambda_1 = 4h^{-2} \cos^2(\pi h/2) \), which implies \( \kappa = \tau^2 \lambda_1 = 4\tau^2 \cos^2(\pi h/2) \). Hence, we thus obtain an exponential growth of \( \text{cond}(F_h) \) whenever the grid ratio becomes
\[ r > \frac{2}{\sqrt{3} \cos(\pi h/2)} > \frac{2}{\sqrt{3}} \approx 1.1547. \]
Similarly, for 2D case we get \( r > \frac{\sqrt{2}}{\sqrt{3}} \approx 0.8165 \) and for 3D case we get \( r > \frac{2}{3} \approx 0.6667 \). These conditions are slightly stronger than the corresponding CFL conditions, which seems to be reasonable since the constructed approach given by our proof may not necessarily produce the tightest bound.

**REFERENCES**


