# Efficient and Parallel Solution of High-order Continuous Time Galerkin for Dissipative and Wave Propagation Problems[*]

Zhiming Chen[†]      Yong Liu[‡]

**Abstract.** We propose efficient and parallel algorithms for the implementation of the high-order continuous time Galerkin method for dissipative and wave propagation problems. By using Legendre polynomials as shape functions, we obtain a special structure of the stiffness matrix which allows us to extend the diagonal Padé approximation to solve ordinary differential equations with source terms. The unconditional stability, $hp$ error estimates, and $hp$ superconvergence at the nodes of the continuous time Galerkin method are proved. Numerical examples confirm our theoretical results.

**Key words.** Implicit time discretization; Padé approximation; Parallel implementation

**AMS classification**. 65M60

## 1   Introduction

In this paper, we study the following system of ordinary differential equations (ODEs)

$$\mathbf{Y}'(t) = \mathbb{D}\mathbf{Y}(t) + \mathbf{R}(t) \text{ in } (0,T), \quad \mathbf{Y}(0) = \mathbf{Y}_0, \tag{1.1}$$

which is obtained from the method-of-lines approach for linear partial differential equations (PDEs) after space discretization. Here $T > 0$ is the length of the time interval, $\mathbf{Y}, \mathbf{R} \in \mathbb{R}^M$, and $\mathbb{D}$ is an $M \times M$ real constant matrix, where $M$ is the number of degrees of freedom of the spatial discretization. Without loss of generality, we assume

$$\mathbb{D} + \mathbb{D}^T \le 0, \tag{1.2}$$

that is, $\mathbb{D} + \mathbb{D}^T$ is a semi-negative definite matrix. This condition is satisfied by a large class of linear PDEs including the dissipative problems such as the parabolic equations and the wave propagation problems such as the wave equation and Maxwell equations.

Let $0 = t_0 < t_1 < \cdots < t_N = T$ be a partition of $(0,T)$. If the source $\mathbf{R} = \mathbf{0}$ in (1.1), the exact solution in each time interval $(t_n, t_{n+1})$ is $\mathbf{Y}(t) = e^{\mathbb{D}(t-t_n)}\mathbf{Y}(t_n)$ for which Padé approximation to the exponential function can be used to construct and analyze

---

[†]LSEC, Institute of Computational Mathematics, Academy of Mathematics and System Sciences and School of Mathematical Science, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China. E-mail: zmchen@lsec.cc.ac.cn

[‡]LSEC, Institute of Computational Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P.R. China. E-mail: yongliu@lsec.cc.ac.cn

numerical schemes to solve (1.1). In [12], by using the partial fraction formula for the Padé approximation, the $[r/r]$, $r \geq 1$, Padé approximation leads to the following method

$$\mathbf{Y}(t_{n+1}) \approx \frac{P_r(\tau_n \mathbb{D})}{P_r(-\tau_n \mathbb{D})} \mathbf{Y}(t_n) = \left[ (-1)^r \mathbb{I} + \sum_{j=1}^{r} -\frac{P_r(-\zeta_j)}{P_r'(\zeta_j)} (\zeta_j \mathbb{I} + \tau_n \mathbb{D})^{-1} \right] \mathbf{Y}(t_n), \qquad (1.3)$$

where $\tau_n = t_{n+1} - t_n$, $\mathbb{I} \in \mathbb{R}^{M \times M}$ is the identity matrix, $P_r(z)$ is the numerator of the $[r/r]$ Padé approximation to the exponential function $e^z$, and $\{\zeta_1, \cdots, \zeta_r\}$ are zeros of $P_r(z)$ which are known to be simple and lie in the left-half plane. (1.3) indicates that one can compute the approximation of the solution $\mathbf{Y}(t_{n+1})$ in each time step by solving $k$ complex matrix problems and $r - 2k$ real matrix problems of the form $\zeta \mathbb{I} + \tau_n \mathbb{D}$ in parallel, where $k$, $0 \leq k \leq r/2$, is the number of complex zeros of $P_r(z)$ (see Remark 3.1 below). The purpose of this paper is to construct algorithms sharing this very desirable property for solving (1.1) with general nonzero sources $\mathbf{R}(t)$.

There exists a large literature on implicit single-step time-stepping methods for solving (1.1) (see, e.g.,[16] and the references therein). The following continuous time Galerkin method proposed in [17] is probably the simplest

$$\mathbf{Y}_r' = \mathbb{D}\mathcal{P}_{r-1}\mathbf{Y}_r + \mathcal{P}_{r-1}\mathbf{R} \quad \text{in } (t_n, t_{n+1}), \quad 0 \leq n \leq N - 1, \qquad (1.4)$$

where $\mathbf{Y}_r$ is a piecewise polynomial of degree $r \geq 1$ in each interval $(t_n, t_{n+1})$, continuous at the nodes $t = t_n$, and $\mathcal{P}_{r-1}$ is the local $L^2$ projection to the space of polynomials of degree $(r-1)$ in each interval. It is shown in [17] that (1.4) is equivalent to the $r$-stage Gauss collocation method at the nodes when $\mathbf{R} = \mathbf{0}$ and has the highest classical order $2r$ among all $r$-stage Runge-Kutta methods [16, Table 5.12]. The continuous time Galerkin method, together with finite element discretization in space, is used in [2] for the heat equation and in [11], [15] for the wave equation. We refer to [1] for a unified framework and the comparison of the most popular implicit single-step time-stepping methods including also the discontinuous time Galerkin method and various Runge-Kutta methods.

The difficulty in using the high-order continuous time Galerkin method or any implicit time Runge-Kutta methods is that a straightforward implementation requires to solve a system of linear equations of the size $rM \times rM$, which is not feasible in most time for PDE problems. In a recent work [23], efficient iterative algorithms are developed based on optimal preconditioning of the stage matrix for finding the stage vectors of the implicit Runge-Kutta methods for solving (1.1). For an $r$-stage implicit Runge-Kutta method, the stage matrix is an $r \times r$ block matrix with each block being a $M \times M$ matrix. One can find further references in [23] for developing efficient algorithms implementing the high order implicit time discretization methods in the literature. We also refer to [22], [19] for the implementation of the discontinuous time Galerkin method based on the block diagnalization of the stiffness matrix.

In this paper we propose an efficient realization of the method (1.4) which uses Legendre polynomials as shape functions to obtain a new stiffness matrix which is different from the stage matrix in [23] applying to the Gauss collocation method. By exploiting the special structure of the stiffness matrix, we construct an algorithm which computes the solution $\mathbf{Y}_r(t_n)$ at each node by solving $k$ complex matrix problems and $r - 2k$ real matrix problems in parallel, where $k$, $0 \leq k \leq r/2$, is the number of complex zeros of the $[r/r]$ Padé

2

numerator $P_r(z)$. Moreover, a parallel-in-time algorithm is proposed to compute the other coefficients of the solution $\mathbf{Y}_r$ in each time interval $(t_n, t_{n+1})$ which solves in parallel $kr$ complex matrix problems and $(r - 2k)r$ real matrix problems. For the dissipative system, in which $\mathbb{D} + \mathbb{D}^T$ is negative definite, in the parallel-in-time algorithm, only $k$ complex matrix problems and $(r - 2k)$ real matrix problems need to be solved. We remark that our parallel-in-time algorithm is different from the other parallel-in-time algorithms based on domain decomposition or space-time multigrid techniques in the literature (see, e.g., [13]).

As a by-product of our analysis, we obtain the following formula (Theorem 3.3) to compute the nodal values $\mathbf{Y}_r(t_{n+1})$, $0 \leq n \leq N - 1$, of the solution of (1.4)

$$\mathbf{Y}_r(t_{n+1}) = \frac{P_r(\tau_n \mathbb{D})}{P_r(-\tau_n \mathbb{D})} \mathbf{Y}_r(t_n) + \sum_{k=1}^{r} (-1)^{k+1} \frac{\phi_{k1}(\tau_n \mathbb{D})}{P_r(-\tau_n \mathbb{D})} \mathbf{b}_{k-1} + \tau_n \mathbf{R}_0, \tag{1.5}$$

where for $k = 1, \cdots, r$, $\phi_{k1}(\lambda)$ is a polynomial of degree $r$ satisfying some recurrence relations, and $\mathbf{b}_k$, $\mathbf{R}_0$ are vectors depending on the source $\mathbf{R}$. (1.5) can be viewed as a generalization of the $[r/r]$ Padé approximation (1.3) for solving the ODE system without sources.

The layout of the paper is as follows. In section 2 we introduce the continuous time Galerkin method for (1.1) and prove the strong stability and derive a $hp$ error estimate. In section 3 we propose our parallel algorithms to implement the continuous time Galerkin method. In section 4 we consider an alternative implementation for the dissipative system. In section 5 we prove the optimal stability and error estimates in terms of $r$ when $\mathbb{D}$ is a symmetric or skew-symmetric matrix. In section 6 we consider the application of the algorithms in this paper to solve the linear convection-diffusion equation by using the local discontinuous Galerkin method and the wave equation with discontinuous coefficients by using the unfitted finite element spatial discretization.

## 2 Implicit time discretization

In this section, we introduce the continuous time Galerkin method for solving (1.1). Let $0 = t_0 < t_1 < \ldots < t_N = T$ be a partition of the time interval $(0, T)$ with time steps $\tau_n = t_{n+1} - t_n$, $0 \leq n \leq N - 1$. We set $I_n = (t_n, t_{n+1})$ and $\tau = \max_{0 \leq n \leq N-1} \{\tau_n\}$. For any integer $m \geq 1$, we define the finite element space

$$\mathbf{V}_\tau^m := \{\mathbf{v} \in [C(0, T)]^M : \mathbf{v}|_{I_n} \in [P^m]^M, 0 \leq n \leq N - 1\},$$

where $P^m$ is the set of polynomials whose degree is at most $m$. Define the local projection $\mathcal{P}_m$, $m \geq 0$, such that in each time interval $I_n$, $\mathcal{P}_m : [L^2(I_n)]^M \to [P^m]^M$ satisfies

$$\int_{I_n} (\mathcal{P}_m \mathbf{v}, \mathbf{w}) \, dt = \int_{I_n} (\mathbf{v}, \mathbf{w}) \, dt \quad \forall \mathbf{w} \in [P^m]^M,$$

where we denote $(\cdot, \cdot)$ the inner product of $\mathbb{R}^M$. It is well-known (see, e.g., Schwab [21]) that for any $s \geq 0$, $m \geq 0$,

$$\|\mathbf{v} - \mathcal{P}_m \mathbf{v}\|_{L^2(I_n)} \leq C \frac{\tau_n^{\min(m+1, s)}}{(m+1)^s} \|\mathbf{v}\|_{H^s(I_n)} \quad \forall \mathbf{v} \in [H^s(I_n)]^M, \tag{2.1}$$

where the constant $C$ is independent of $m, \tau_n$ but may depend on $s$. In this paper, for any integer $d \geq 1$ and Banach space $X$, we denote $\| \cdot \|_X$ both the norm of $X$ and $[X]^d$.

For any integer $r \geq 1$, the continuous time Galerkin method for solving (1.1) is to find the function $\mathbf{Y}_r \in \mathbf{V}_\tau^r$ such that $\mathbf{Y}_r(0) = \mathbf{Y}_0$ and

$$\mathbf{Y}_r' = \mathbb{D}\mathcal{P}_{r-1}\mathbf{Y}_r + \mathcal{P}_{r-1}\mathbf{R} \ \text{ in } I_n, \ \ 0 \leq n \leq N-1. \tag{2.2}$$

The following stability lemma extends an idea in Griesmaier and Monk [15] where the continuous time Galerkin discretization in time and hybridizable discontinuous Galerkin method in space for the wave equation are considered.

**Lemma 2.1.** *The problem* (2.2) *has a unique solution* $\mathbf{Y}_r \in \mathbf{V}_\tau^r$ *which satisfies*

$$\max_{1 \leq n \leq N} \|\mathbf{Y}_r(t_n)\|_{\mathbb{R}^M} \leq \|\mathbf{Y}_0\|_{\mathbb{R}^M} + CT^{1/2}\|\mathbf{R}\|_{L^2(0,T)}, \tag{2.3}$$

$$\max_{0 \leq t \leq T} \|\mathbf{Y}_r\|_{\mathbb{R}^M} \leq Cr^2(\|\mathbf{Y}_0\|_{\mathbb{R}^M} + T^{1/2}\|\mathbf{R}\|_{L^2(0,T)}). \tag{2.4}$$

*where the constant $C$ is independent of $r, \tau, \mathbb{D}$ and $\mathbf{R}$.*

*Proof.* At each time step, (2.2) is equivalent to a linear system of equations whose existence and uniqueness of the solution follow from the stability estimate (2.4). To prove the stability estimates (2.3)-(2.4), we denote by $\{L_j\}_{j=0}^\infty$ the Legendre polynomials on $(-1, 1)$ and define $\widetilde{L}_j = L_j \circ \psi^{-1}$, where $\psi : (-1, 1) \to (t_n, t_{n+1})$ is the mapping $\psi(\xi) = \frac{t_n + t_{n+1}}{2} + \frac{t_{n+1} - t_n}{2}\xi$ for $\xi \in (-1, 1)$. Then $\{\widetilde{L}_j\}_{j=0}^\infty$ are orthogonal in $L^2(I_n)$, $\widetilde{L}_r(t_n) = (-1)^r, \widetilde{L}_r(t_{n+1}) = 1$, and

$$\int_{I_n} |\widetilde{L}_r|^2 dt = \frac{\tau_n}{2r+1}, \quad \int_{I_n} |\widetilde{L}_r'|^2 dt = \frac{2r(r+1)}{\tau_n}. \tag{2.5}$$

For $n = 0, \cdots, N-1$, let $\mathbf{Y}_r^n = \mathbf{Y}_r(t_n)$ and $\hat{\mathbf{Y}}_r \in [P^r]^M$ satisfy

$$\hat{\mathbf{Y}}_r' = \mathbb{D}\mathcal{P}_{r-1}\hat{\mathbf{Y}}_r \ \text{ in } I_n, \ \ \hat{\mathbf{Y}}_r(t_n) = \mathbf{Y}_r^n. \tag{2.6}$$

By multiplying (2.6) by $\hat{\mathbf{Y}}_r$ and integrating over $I_n$, we obtain easily by (1.2) that

$$\frac{1}{2}\|\hat{\mathbf{Y}}_r(t_{n+1})\|_{\mathbb{R}^M}^2 - \frac{1}{2}\|\mathbf{Y}_r^n\|_{\mathbb{R}^M}^2 = \int_{I_n} (\mathbb{D}\mathcal{P}_{r-1}\hat{\mathbf{Y}}_r, \mathcal{P}_{r-1}\hat{\mathbf{Y}}_r)dt \leq 0.$$

This implies

$$\|\hat{\mathbf{Y}}_r(t_{n+1})\|_{\mathbb{R}^M} \leq \|\mathbf{Y}_r^n\|_{\mathbb{R}^M}. \tag{2.7}$$

Since $\hat{\mathbf{Y}}_r \in [P^r]^M$ in $I_n$, we have the following decomposition introduced in [15]

$$\hat{\mathbf{Y}}_r = (-1)^r \mathbf{Y}_r^n \widetilde{L}_r + (t - t_n)\widetilde{\mathbf{Y}}_r, \ \ \widetilde{\mathbf{Y}}_r \in [P^{r-1}]^M. \tag{2.8}$$

Notice that $\mathcal{P}_{r-1}\hat{\mathbf{Y}}_r = \mathcal{P}_{r-1}[(t - t_n)\widetilde{\mathbf{Y}}_r]$, substituting this decomposition to (2.6), we have

$$(-1)^r \mathbf{Y}_r^n \widetilde{L}_r' + \widetilde{\mathbf{Y}}_r + (t - t_n)\widetilde{\mathbf{Y}}_r' = \mathbb{D}\mathcal{P}_{r-1}[(t - t_n)\widetilde{\mathbf{Y}}_r] \ \text{ in } I_n.$$

Multiply the equation by $\widetilde{\mathbf{Y}}_r \in [P^{r-1}]^M$ and integrate over $I_n$, we have by (2.5) that

$$\frac{1}{2}\|\widetilde{\mathbf{Y}}_r\|_{L^2(I_n)}^2 + \frac{1}{2}\tau_n\|\widetilde{\mathbf{Y}}_r(t_{n+1})\|_{\mathbb{R}^M}^2 \leq C\tau_n^{-1/2}r\|\mathbf{Y}_r^n\|_{\mathbb{R}^M}\|\widetilde{\mathbf{Y}}_r\|_{L^2(I_n)},$$

4

where we have used the fact that by (1.2)

$$\int_{I_n}(\mathbb{D}\mathcal{P}_{r-1}[(t-t_n)\widetilde{\mathbf{Y}}_r],\widetilde{\mathbf{Y}}_r)dt = \int_{I_n}(t-t_n)(\mathbb{D}\widetilde{\mathbf{Y}}_r,\widetilde{\mathbf{Y}}_r)dt \le 0. \qquad (2.9)$$

This yields $\|\widetilde{\mathbf{Y}}_r\|_{L^2(I_n)} \le C\tau_n^{-1/2}r\|\mathbf{Y}_r^n\|_{\mathbb{R}^M}$ and thus by using (2.5)

$$\|\hat{\mathbf{Y}}_r\|_{L^2(I_n)} \le C\tau_n^{1/2}r\|\mathbf{Y}_r^n\|_{\mathbb{R}^M}. \qquad (2.10)$$

On the other hand, it follows from (2.2) and (2.6) that

$$(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)' = \mathbb{D}\mathcal{P}_{r-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r) + \mathcal{P}_{r-1}\mathbf{R} \quad \text{in } I_n, \quad (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)(t_n) = 0. \qquad (2.11)$$

Then $\mathbf{Y}_r - \hat{\mathbf{Y}}_r = (t - t_n)\mathbf{W}_r$ for some $\mathbf{W}_r \in [P^{r-1}]^M$. By substituting this relation into the equation (2.11) we have

$$\mathbf{W}_r + (t-t_n)\mathbf{W}_r' = \mathbb{D}\mathcal{P}_{r-1}[(t-t_n)\mathbf{W}_r] + \mathcal{P}_{r-1}\mathbf{R} \quad \text{in } I_n.$$

By multiplying the equation by $\mathbf{W}_r$ and integrating over $I_n$, we obtain by a similar bound as in (2.9) that

$$\frac{1}{2}\|\mathbf{W}_r\|_{L^2(I_n)}^2 + \frac{1}{2}\tau_n\|\mathbf{W}_r(t_{n+1})\|_{\mathbb{R}^M}^2 \le \|\mathbf{R}\|_{L^2(I_n)}\|\mathbf{W}_r\|_{L^2(I_n)}.$$

This yields $\|\mathbf{W}_r\|_{L^2(I_n)} \le 2\|\mathbf{R}\|_{L^2(I_n)}$ and thus

$$\|\mathbf{Y}_r - \hat{\mathbf{Y}}_r\|_{L^2(I_n)} \le 2\tau_n\|\mathbf{R}\|_{L^2(I_n)}. \qquad (2.12)$$

Now by multiplying (2.11) by $\mathbf{Y}_r - \hat{\mathbf{Y}}_r$ and integrating over $I_n$ we obtain by (1.2) and (2.12) that

$$\frac{1}{2}\|(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)(t_{n+1})\|_{\mathbb{R}^M}^2 \le \int_{I_n}(\mathcal{P}_{r-1}\mathbf{R},\mathbf{Y}_r - \hat{\mathbf{Y}}_r)dt \le 2\tau_n\|\mathbf{R}\|_{L^2(I_n)}^2,$$

which implies by the triangle inequality and (2.7) that

$$\|\mathbf{Y}_r(t_{n+1})\|_{\mathbb{R}^M} \le \|\mathbf{Y}_r^n\|_{\mathbb{R}^M} + 2\tau_n^{1/2}\|\mathbf{R}\|_{L^2(I_n)}.$$

This yields (2.3). Next by using the triangle inequality, (2.10), and (2.12), we have

$$\|\mathbf{Y}_r\|_{L^2(I_n)} \le C\tau_n^{1/2}r\|\mathbf{Y}_r^n\|_{\mathbb{R}^M} + 2\tau_n\|\mathbf{R}\|_{L^2(I_n)}, \qquad (2.13)$$

which implies by the $hp$ inverse estimate that

$$\max_{t_n \le t \le t_{n+1}}\|\mathbf{Y}_r\|_{\mathbb{R}^M} \le C\tau_n^{-1/2}r\|\mathbf{Y}_r\|_{L^2(I_n)} \le Cr^2\|\mathbf{Y}_r^n\|_{\mathbb{R}^M} + C\tau_n^{1/2}r\|\mathbf{R}\|_{L^2(I_n)}.$$

This shows (2.4) and completes the proof of the lemma. $\qquad \Box$

To derive an $hp$ a priori error estimate for the continuous time Galerkin method (2.2), we first recall an interpolation operator in the literature (see, e.g., [21, Theorem 3.17]).

5

**Lemma 2.2.** *There exists an interpolation operator $\Pi_r : [H^1(0,T)]^M \to \mathbf{V}_\tau^r$ such that for any $\mathbf{v} \in [W^{1+s,\infty}(0,T)]^M$, $s \geq 1$, and $n = 0, 1, \cdots, N-1$,*

$$(\Pi_r \mathbf{v})(t_n) = \mathbf{v}(t_n), \quad (\Pi_r \mathbf{v})(t_{n+1}) = \mathbf{v}(t_{n+1}), \quad (\Pi_r \mathbf{v})' = \mathcal{P}_{r-1} \mathbf{v}' \quad \text{in } I_n, \quad (2.14)$$

$$\|\mathbf{v} - \Pi_r \mathbf{v}\|_{L^2(I_n)} \leq C \frac{\tau^{\min(r+1,s)}}{r^s} \|\mathbf{v}\|_{H^s(I_n)}, \quad (2.15)$$

$$\max_{t_n \leq t \leq t_{n+1}} \|\mathbf{v} - \Pi_r \mathbf{v}\|_{\mathbb{R}^M} \leq C \frac{\tau^{\min(r,s)+1}}{r^s} \|\mathbf{v}'\|_{W^{s,\infty}(I_n)}, \quad (2.16)$$

*where the constant $C$ is independent of $\tau, r$ but may depend on $s$.*

*Proof.* The interpolation operator is defined as

$$\Pi_r \mathbf{v} = \mathbf{v}(t_n) + \int_{t_n}^t (\mathcal{P}_{r-1} \mathbf{v}') dt \quad \forall t \in I_n.$$

(2.14) follows easily from this definition. Next by using (2.1), we have

$$\max_{t_n \leq t \leq t_{n+1}} \|\mathbf{v} - \Pi_r \mathbf{v}\|_{\mathbb{R}^M} \leq \tau_n^{1/2} \|\mathbf{v}' - \mathcal{P}_{r-1} \mathbf{v}'\|_{L^2(I_n)} \leq C \frac{\tau_n^{\min(r,s)+1}}{r^s} \|\mathbf{v}'\|_{W^{s,\infty}(I_n)}.$$

This shows (2.16).

The estimate (2.15) is proved for $s \geq 2$ in [21]. Here we use the duality argument to show (2.15) also from $s \geq 1$. Let $\mathbf{w} \in H_0^1(I_n)$ be the solution of the problem

$$-\mathbf{w}'' = \mathbf{v} - \Pi_r \mathbf{v} \quad \text{in } I_n.$$

It is easy to see that $\|\mathbf{w}\|_{H^2(I_n)} \leq C \|\mathbf{v} - \Pi_r \mathbf{v}\|_{L^2(I_n)}$. Since $(\mathbf{v} - \Pi_r \mathbf{v})(t_n) = \mathbf{0}, (\mathbf{v} - \Pi_r \mathbf{v})(t_{n+1}) = \mathbf{0}$, we multiply the equation by $\mathbf{v} - \Pi_r \mathbf{v}$, integrate over $I_n$, and use (2.14) to obtain

$$
\begin{aligned}
\|\mathbf{v} - \Pi_r \mathbf{v}\|_{L^2(I_n)}^2 = \int_{I_n} (\mathbf{w}', \mathbf{v}' - (\Pi_r \mathbf{v})') dt &= \int_{I_n} (\mathbf{w}' - \mathcal{P}_{r-1} \mathbf{w}', \mathbf{v}' - \mathcal{P}_{r-1} \mathbf{v}') dt \\
&\leq C \frac{\tau^{\min(r+1,s)}}{r^s} \|\mathbf{w}\|_{H^2(I_n)} \|\mathbf{v}\|_{H^s(I_n)}.
\end{aligned}
$$

This completes the proof by using $\|\mathbf{w}\|_{H^2(I_n)} \leq C \|\mathbf{v} - \Pi_r \mathbf{v}\|_{L^2(I_n)}$. $\qquad \square$

The following theorem on the *hp* error estimate is the main result of this section.

**Theorem 2.1.** *Let $s \geq 1$. Assume that $\mathbf{R} \in [H^s(0,T)]^M$, $\mathbf{Y} \in [W^{1+s,\infty}(0,T)]^M$ and $\mathbf{Y}_r \in \mathbf{V}_\tau^r$ is the solution of the problem (2.2), we have*

$$\max_{1 \leq n \leq N} \|(\mathbf{Y} - \mathbf{Y}_r)(t_n)\|_{\mathbb{R}^M} \leq C T^{1/2} \frac{\tau^{\min(r+1,s)}}{r^s} \|\mathbb{D}\mathbf{Y}\|_{H^s(0,T)},$$

$$\max_{0 \leq t \leq T} \|\mathbf{Y} - \mathbf{Y}_r\|_{\mathbb{R}^M} \leq C(1 + T^{1/2}) \frac{\tau^{\min(r+1,s)}}{r^{s-2}} (T^{1/2} \|\mathbf{Y}\|_{W^{s+1,\infty}(0,T)} + \|\mathbf{R}\|_{H^s(0,T)}),$$

*where the constant $C$ is independent of $\tau, r, \mathbb{D}$ but may depend on $s$.*

6

*Proof.* Let $\Pi_r \mathbf{Y} \in \mathbf{V}_\tau^r$ be the interpolation of $\mathbf{Y}$ defined in Lemma 2.2. Since $(\Pi_r \mathbf{Y})' = \mathcal{P}_{r-1} \mathbf{Y}'$ in $I_n$, we have

$$(\Pi_r \mathbf{Y})' = \mathcal{P}_{r-1}(\mathbb{D}\mathbf{Y} + \mathbf{R}) = \mathbb{D}\mathcal{P}_{r-1}(\mathbf{Y} - \Pi_r \mathbf{Y}) + \mathbb{D}\mathcal{P}_{r-1}(\Pi_r \mathbf{Y}) + \mathcal{P}_{r-1}\mathbf{R} \ \ \text{in } I_n.$$

Thus by (2.2) we have

$$\mathbf{Y}_r' - (\Pi_r \mathbf{Y})' = \mathbb{D}\mathcal{P}_{r-1}(\mathbf{Y}_r - \Pi_r \mathbf{Y}) - \mathbb{D}\mathcal{P}_{r-1}(\mathbf{Y} - \Pi_r \mathbf{Y}). \tag{2.17}$$

As $(\mathbf{Y}_r - \Pi_r \mathbf{Y})(0) = \mathbf{0}$, we use (2.3) and (2.15) to obtain

$$\begin{aligned}
\max_{1 \le n \le N} \|(\mathbf{Y}_r - \Pi_r \mathbf{Y})(t_n)\|_{\mathbb{R}^M} &\le CT^{1/2} \|\mathbb{D}\mathcal{P}_{r-1}(\mathbf{Y} - \Pi_r \mathbf{Y})\|_{L^2(0,T)} \\
&\le CT^{1/2} \frac{\tau^{\min(r+1,s)}}{r^s} \|\mathbb{D}\mathbf{Y}\|_{H^s(0,T)}.
\end{aligned}$$

This shows the first estimate as $\mathbf{Y}(t_n) = \Pi_r \mathbf{Y}(t_n)$. The second estimate can be proved similarly by using (2.4), (2.16), and $\mathbf{Y}' = \mathbb{D}\mathbf{Y} + \mathbf{R}$. $\qquad\square$

We remark that the first estimate in Theorem 2.1 is optimal in $\tau$ and $r$ and the second estimate in the theorem is optimal in $\tau$ but suboptimal in $r$ which is due to the stability estimate (2.4) in Lemma 2.1. In section 5 we will show that the stability in the $L^2$ norm can be improved to remove the dependence on $r$ in (2.13) when $\mathbb{D}$ is symmetric or skew-symmetric by using the explicit formulas of $\mathbf{Y}_r(t)$ in section 3. We remark that many spatial discretization matrices of the wave-like equations satisfy the property that $\mathbb{D}$ is skew-symmetric, such as the energy conserving mixed finite element methods for solving the Hodge wave equation in Wu and Bai [25] and the unfitted finite element method of the acoustic wave equation in Chen et. al. [7].

The classical order of Runge-Kutta methods is the convergence order at the nodes $t = t_n$. For the continuous time Galerkin method, it is proved to be $2r$ when $r \ge 2$ in Hulme [17] for nonlinear ODEs and in Aziz and Monk [2] for parabolic equations. The following theorem shows the $hp$ superconvergence of the continuous time Galerkin method at the nodes by using the idea of quasi-projection in [2, §4].

**Theorem 2.2.** *Let $s \ge 1$. Assume that $\mathbb{D}^r \mathbf{Y} \in [H^s(0,T)]^M$ and $\mathbf{Y}_r \in \mathbf{V}_\tau^r$ is the solution of the problem* (2.2), *we have*

$$\max_{1 \le n \le N} \|(\mathbf{Y} - \mathbf{Y}_r)(t_n)\|_{\mathbb{R}^M} \le CT^{1/2} \frac{\tau^{\min(2r,s+r-1)}}{r^s} \|\mathbb{D}^r \mathbf{Y}\|_{H^s(0,T)},$$

*where the constant $C$ is independent of $\tau, r$ but may depend on $s$.*

*Proof.* If $r = 1$, the theorem follows from the first estimate of Theorem 2.1. Now we assume $r \ge 2$. Let $\Pi_r \mathbf{Y} \in \mathbf{V}_\tau^r$ be the interpolation of $\mathbf{Y}$ defined in Lemma 2.2. Denote $\boldsymbol{\omega}_0 = \mathbf{Y} - \Pi_r \mathbf{Y}$. For $1 \le i \le r - 1$, we define correction functions $\boldsymbol{\omega}_i$ such that

$$\boldsymbol{\omega}_i(t_n) = \mathbf{0}, \ \boldsymbol{\omega}_i' = \mathbb{D}\mathcal{P}_{r-1}\boldsymbol{\omega}_{i-1} \ \ \text{in } I_n, \ \ n = 0, 1, \cdots, N - 1. \tag{2.18}$$

We claim that $\boldsymbol{\omega}_i(t_{n+1}) = \mathbf{0}$ so that $\boldsymbol{\omega}_i \in \mathbf{V}_\tau^r$. In fact, by (2.14), we have $(\boldsymbol{\omega}_0', \mathbf{v})_{I_n} = 0$ for any $\mathbf{v} \in [P^{r-1}]^M$, where $(\cdot, \cdot)_{I_n}$ is the inner product of $[L^2(I_n)]^M$. Since $\boldsymbol{\omega}_0(t_n) = \boldsymbol{\omega}_0(t_{n+1}) = \mathbf{0}$,

7

we obtain by integration by parts that $(\boldsymbol{\omega}_1', \mathbf{v}')_{I_n} = (\mathbb{D}\boldsymbol{\omega}_0, \mathbf{v}')_{I_n} = 0$. Therefore, $(\boldsymbol{\omega}_1', \mathbf{v})_{I_n} = 0$ for any $\mathbf{v} \in [P^{r-2}]^M$ and consequently, $\boldsymbol{\omega}_1(t_{n+1}) = \int_{I_n} \boldsymbol{\omega}_1' dt = \mathbf{0}$. By mathematical induction, we know easily by the same argument that $(\boldsymbol{\omega}_i', \mathbf{v})_{I_n} = 0$ for any $\mathbf{v} \in [P^{r-i-1}]^M$ and $\boldsymbol{\omega}_i(t_{n+1}) = \mathbf{0}$. This shows the claim.

Let $\boldsymbol{\omega} = \sum_{i=1}^{r-1} \boldsymbol{\omega}_i \in \mathbf{V}_\tau^r$. By (2.17) and (2.18), we have

$$\mathbf{Y}_r' - (\Pi_r \mathbf{Y})' + \boldsymbol{\omega}' = \mathbb{D}\mathcal{P}_{r-1}(\mathbf{Y}_r - \Pi_r \mathbf{Y} + \boldsymbol{\omega}) - \mathbb{D}\mathcal{P}_{r-1}(\boldsymbol{\omega}_{r-1}).$$

As $(\mathbf{Y}_r - \Pi_r \mathbf{Y} + \boldsymbol{\omega})(0) = \mathbf{0}$, we use (2.3) to obtain

$$\|(\mathbf{Y}_r - \Pi_r \mathbf{Y} + \boldsymbol{\omega})(t_n)\|_{\mathbb{R}^M} \leq CT^{1/2}\|\mathbb{D}\mathcal{P}_{r-1}(\boldsymbol{\omega}_{r-1})\|_{L^2(0,T)} \leq CT^{1/2}\|\mathbb{D}\boldsymbol{\omega}_{r-1}\|_{L^2(0,T)}.$$

Now it follows from (2.18) that

$$\|\mathbb{D}\boldsymbol{\omega}_i\|_{L^2(I_n)} \leq \tau\|\mathbb{D}^2\boldsymbol{\omega}_{i-1}\|_{L^2(I_n)}, \quad 1 \leq i \leq r-1.$$

By using (2.15) we have then

$$\|\mathbb{D}\boldsymbol{\omega}_{r-1}\|_{L^2(I_n)} \leq \tau^{r-1}\|\mathbb{D}^r\boldsymbol{\omega}_0\|_{L^2(I_n)} \leq C\frac{\tau^{\min(2r,s+r-1)}}{r^s}\|\mathbb{D}^r\mathbf{Y}\|_{H^s(I_n)}.$$

This completes the proof since by (2.14) and (2.18), $(\mathbf{Y}_r - \Pi_r \mathbf{Y} + \boldsymbol{\omega})(t_n) = (\mathbf{Y}_r - \mathbf{Y})(t_n)$, $1 \leq n \leq N$. $\square$

The correction function $\boldsymbol{\omega} = \sum_{i=1}^{r-1} \boldsymbol{\omega}_i$ is introduced in [2] which is related to the idea of quasi-projection in Douglas Jr. et al [10]. Our new observation is that $\boldsymbol{\omega}_i = \mathbf{0}$ at the nodes for $1 \leq i \leq r-1$, which simplifies the proof.

To conclude this section, we recall some facts about Padé approximation to the exponential function which can be found in Saff and Varga [20] and the references therein. For any integers $m, n \geq 0$, the $[m/n]$ Padé approximation to $e^z$ is defined as the polynomials $P_m(z) \in P^m$, $Q_n(z) \in P^n$, $Q_n(0) = 1$, for which

$$e^z - \frac{P_m(z)}{Q_n(z)} = O(|z|^{m+n+1}) \quad \text{as } |z| \to 0.$$

It is known that

$$P_m(z) = \sum_{j=0}^{m} \frac{(m+n-j)!m!z^j}{(m+n)!j!(m-j)!}, \quad Q_n(z) = \sum_{j=0}^{n} \frac{(m+n-j)!n!(-z)^j}{(m+n)!j!(n-j)!}. \tag{2.19}$$

Obviously, $Q_n(z) = P_n(-z)$. When $m = n$, $P_m(z), Q_m(z)$ are called diagonal Padé numerator and denominator of type $[m/m]$ for $e^z$. The following lemma follows easily from (2.19)

**Lemma 2.3.** *The diagonal Padé numerator of type $[m/m]$ for $e^z$ satisfies $P_0(z) = 1$, $P_1(z) = 1 + \frac{1}{2}z$, $P_2(z) = 1 + \frac{1}{2}z + \frac{1}{12}z^2$, and*

$$P_m(z) = P_{m-1}(z) + \frac{z^2}{4(2m-1)(2m-3)}P_{m-2}(z), \quad m \geq 2.$$

8

The following lemma is proved in Hairer and Wanner [16, Theorem 4.12], [20, Theorem 2.4]. It is essential in proving the A-stability of numerical methods for ODEs based on the Padé approximation of the exponential function.

**Lemma 2.4.** *All zeros of the diagonal Padé numerator of type $[m/m]$, $m \geq 1$, for $e^z$ are simple and lie in the half-plane $\{z \in \mathbb{C} : \mathrm{Re}(z) \leq -2\}$.*

For $m \geq 1$, denote $\zeta_1, \cdots, \zeta_m \in \mathbb{C}$ the zeros of $P_m(z)$, the diagonal Padé numerator of type $[m/m]$ for $e^z$, then by (2.19)

$$P_m(z) = \frac{m!}{(2m)!}(z - \zeta_1) \cdots (z - \zeta_m), \quad Q_m(z) = (-1)^m \frac{m!}{(2m)!}(z + \zeta_1) \cdots (z + \zeta_m).$$

Recall that any polynomial $F \in P^{m-1}$ can be expanded as the Lagrange interpolation function at $m$ distinct zeros of $P_m(-z)$. This yields the following partial fraction formula (see, e.g., Szegö [24, Theorem 3.3.5])

$$\frac{F(z)}{P_m(-z)} = \sum_{j=1}^{m} -\frac{F(-\zeta_j)}{P_m'(\zeta_j)} \frac{1}{z + \zeta_j}. \tag{2.20}$$

Since $P_m(z) - (-1)^m P_m(-z) \in P^{m-1}$, we obtain the partial fraction formula for the $[m/m]$ Padé approximation of $e^z$ (see Gallopoulos and Saad [12])

$$R_{m,m}(z) = \frac{P_m(z)}{P_m(-z)} = (-1)^m + \sum_{j=1}^{m} -\frac{P_m(-\zeta_j)}{P_m'(\zeta_j)} \frac{1}{z + \zeta_j}. \tag{2.21}$$

Recall that if $p(z) = \sum_{i=0}^{m} a_i z^i$ is a polynomial of degree $m$, then $p(\mathbb{X}) := \sum_{i=0}^{m} a_i \mathbb{X}^i$ for any matrix $\mathbb{X} \in \mathbb{R}^{d \times d}$, $d \geq 1$. Obviously, if $p(z) = p_1(z) + p_2(z)$ or $q(z) = p_1(z) \cdot p_2(z)$, where $p_1, p_2$ are polynomials, then $p(\mathbb{X}) = p_1(\mathbb{X}) + p_2(\mathbb{X})$, $q(\mathbb{X}) = p_1(\mathbb{X}) \cdot p_2(\mathbb{X})$. It follows now from (2.20)-(2.21) that for any $F \in P^{m-1}$ and any matrix $\mathbb{X} \in \mathbb{R}^{d \times d}$ such that $P_m(-\mathbb{X})$ is invertible,

$$\frac{F(\mathbb{X})}{P_m(-\mathbb{X})} = \sum_{j=1}^{m} -\frac{F(-\zeta_j)}{P_m'(\zeta_j)}(\zeta_j \mathbb{I} + \mathbb{X})^{-1}, \tag{2.22}$$

$$\frac{P_m(\mathbb{X})}{P_m(-\mathbb{X})} = (-1)^m \mathbb{I} + \sum_{j=1}^{m} -\frac{P_m(-\zeta_j)}{P_m'(\zeta_j)}(\zeta_j \mathbb{I} + \mathbb{X})^{-1}. \tag{2.23}$$

The identity (2.23) is the basis of the method (1.3) in the introduction.

# 3 Parallel implementation

In this section, we propose parallel algorithms to implement the problem (2.2) based on finding analytic formulas of the determinant and all factors of the stiffness matrix of the continuous time Galerkin method at each time step. To form the stiffness matrix, we use the Legendre polynomials $\{\widetilde{L}_j\}_{j=0}^{r}$ in $I_n$ as the basis functions. We assume

$$\mathbf{Y}_r(t) = \sum_{j=0}^{r} \mathbf{a}_j \widetilde{L}_j(t), \quad \mathcal{P}_{r-1}\mathbf{R} = \sum_{j=0}^{r-1} \mathbf{R}_j \widetilde{L}_j(t),$$

9

where $\mathbf{a}_j, \mathbf{R}_j \in \mathbb{R}^M$. Since $\mathcal{P}_{r-1}\widetilde{L}_r = 0$ in $I_n$, from (2.2), we have

$$\sum_{j=0}^{r} \mathbf{a}_j \widetilde{L}'_j(t) = \mathbb{D} \sum_{j=0}^{r-1} \mathbf{a}_j \widetilde{L}_j(t) + \sum_{j=0}^{r-1} \mathbf{R}_j \widetilde{L}_j(t).$$

For any $k \geq 1$, multiplying the equation by $(t - t_n)(t_{n+1} - t)\widetilde{L}'_k(t)$ and integrating over $I_n$, we obtain

$$\mathbf{a}_k \frac{\tau_n}{2} \frac{k(k+1)}{k+\frac{1}{2}} = \mathbb{D} \sum_{j=0}^{r-1} \int_{t_n}^{t_{n+1}} \mathbf{a}_j \widetilde{L}_j(t)(t - t_n)(t_{n+1} - t)\widetilde{L}'_k(t)\, dt$$

$$+ \sum_{j=0}^{r-1} \int_{t_n}^{t_{n+1}} \mathbf{R}_j \widetilde{L}_j(t)(t - t_n)(t_{n+1} - t)\widetilde{L}'_k(t)\, dt, \qquad (3.1)$$

where we have used the fact that

$$\int_{t_n}^{t_{n+1}} \widetilde{L}'_j \widetilde{L}'_k (t - t_n)(t_{n+1} - t)\, dt = \frac{\tau_n}{2} \int_{-1}^{1} L'_j L'_k (1 - t^2)\, dt = \frac{\tau_n}{2} \frac{k(k+1)}{k+\frac{1}{2}} \delta_{j,k}.$$

Here $\delta_{j,k}$ is the Kronecker delta function. By the recursion relation $(2k + 1)\widetilde{L}_k(t) = \frac{\tau_n}{2}(\widetilde{L}'_{k+1}(t) - \widetilde{L}'_{k-1}(t))$,

$$\int_{t_n}^{t_{n+1}} \widetilde{L}_j(t)(t - t_n)(t_{n+1} - t)\widetilde{L}'_k(t)\, dt$$

$$= \frac{1}{2j+1} \int_{t_n}^{t_{n+1}} \frac{\tau_n}{2}(\widetilde{L}'_{j+1}(t) - \widetilde{L}'_{j-1}(t))\widetilde{L}'_k(t)(t - t_n)(t_{n+1} - t)\, dt$$

$$= \frac{\tau_n^2}{4} \frac{k(k+1)}{k+\frac{1}{2}} \left( \frac{1}{2k-1}\delta_{j+1,k} - \frac{1}{2k+3}\delta_{j-1,k} \right).$$

Substituting the identity into (3.1), we have

$$\mathbf{a}_k = \frac{\tau_n}{2}\mathbb{D}\left( \frac{\mathbf{a}_{k-1}}{2k-1} - \frac{\mathbf{a}_{k+1}}{2k+3} \right) + \frac{\tau_n}{2}\left( \frac{\mathbf{R}_{k-1}}{2k-1} - \frac{\mathbf{R}_{k+1}}{2k+3} \right), \quad 1 \leq k \leq r-2, \qquad (3.2)$$

$$\mathbf{a}_k = \frac{\tau_n}{2}\mathbb{D}\frac{\mathbf{a}_{k-1}}{2k-1} + \frac{\tau_n}{2}\frac{\mathbf{R}_{k-1}}{2k-1}, \quad k = r-1, r. \qquad (3.3)$$

By the condition $\mathbf{Y}_r(t_n) = \mathbf{Y}_r^n$, we also have

$$\sum_{j=0}^{r}(-1)^j \mathbf{a}_j = \mathbf{Y}_r^n. \qquad (3.4)$$

(3.2)-(3.4) can be written as a system of linear equations

$$\mathbb{A}\mathbf{X} = \mathbf{B}, \qquad (3.5)$$

where $\mathbf{X} = (\mathbf{a}_0^T, \mathbf{a}_1^T, \cdots, \mathbf{a}_r^T)^T$, $\mathbf{B} = (\mathbf{b}_0^T, \mathbf{b}_1^T, \cdots, \mathbf{b}_r^T)^T$ with

$$\mathbf{b}_{k-1} = \begin{cases} -\frac{\tau_n}{2}\left(\frac{\mathbf{R}_{k-1}}{2k-1} - \frac{\mathbf{R}_{k+1}}{2k+3}\right) & \text{if } 1 \leq k \leq r-2, \\ -\frac{\tau_n}{2}\frac{\mathbf{R}_{k-1}}{2k-1} & \text{if } k = r-1, r, \\ \mathbf{Y}_r^n & \text{if } k = r+1, \end{cases}$$

and

$$\mathbb{A} = \begin{pmatrix} \frac{\tau_n}{2}\mathbb{D} & -\mathbb{I} & -\frac{\tau_n}{2}\mathbb{D}\frac{1}{5} & \cdots & \cdots & \cdots & 0 \\ 0 & \frac{\tau_n}{2}\mathbb{D}\frac{1}{3} & -\mathbb{I} & -\frac{\tau_n}{2}\mathbb{D}\frac{1}{7} & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & \frac{\tau_n}{2}\mathbb{D}\frac{1}{2r-5} & -\mathbb{I} & -\frac{\tau_n}{2}\mathbb{D}\frac{1}{2r-1} & 0 \\ 0 & \cdots & \cdots & \cdots & \frac{\tau_n}{2}\mathbb{D}\frac{1}{2r-3} & -\mathbb{I} & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \frac{\tau_n}{2}\mathbb{D}\frac{1}{2r-1} & -\mathbb{I} \\ \mathbb{I} & -\mathbb{I} & \mathbb{I} & -\mathbb{I} & \cdots & \cdots & (-1)^{r+2}\mathbb{I} \end{pmatrix}.$$

Here $\mathbb{I} \in \mathbb{R}^{M \times M}$ is the identity matrix. By Lemma 2.1, (3.5) has a unique solution. Since $\mathbb{A} \in \mathbb{R}^{M(r+1) \times M(r+1)}$, it is expensive to solve (3.5) directly when $M \gg 1$. Here we propose efficient and parallel algorithms to solve (3.5).

Notice that $\mathbb{A}$ is a $(r+1) \times (r+1)$ block matrix. For $\lambda \in \mathbb{R}$, we define $\mathbb{E}_{r+1}(\lambda) \in \mathbb{R}^{(r+1) \times (r+1)}$ by

$$\mathbb{E}_{r+1}(\lambda) := \begin{pmatrix} a_1 & -1 & b_1 & \cdots & \cdots & \cdots & 0 \\ 0 & a_2 & -1 & b_2 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & a_{r-2} & -1 & b_{r-2} & 0 \\ 0 & \cdots & \cdots & \cdots & a_{r-1} & -1 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & a_r & -1 \\ c_1 & c_2 & \cdots & \cdots & \cdots & c_r & c_{r+1} \end{pmatrix} \tag{3.6}$$

with

$$a_k = \frac{\lambda}{2}\frac{1}{2k-1}, \quad b_k = -\frac{\lambda}{2}\frac{1}{2k+3}, \quad c_k = (-1)^{k+1}, \quad k = 1, \cdots, r+1. \tag{3.7}$$

Then $\mathbb{A} = \mathbb{E}_{r+1}(\tau_n \mathbb{D})$ by replacing each element $e_{ij}(\lambda)$ of $\mathbb{E}_{r+1}(\lambda)$ by the $M \times M$ matrix $e_{ij}(\tau_n \mathbb{D})\mathbb{I}$, $i, j = 1, \cdots, r+1$.

We are going to solve (3.5) by extending the Cramer rule for the block matrices. We first introduce some notation. For any matrix $\mathbb{X} = (X_{ij})_{i,j=1}^d$, $d \geq 1$, we denote $\mathbb{X}_{i,j}$ the matrix obtained by removing the $i$-th row and $j$-th column, $i, j = 1, \ldots, d$. We also denote $\mathbb{X}_{(i_1, \cdots, i_k),(j_1, \cdots, j_l)}$ the matrix obtained by removing the $i_1, \cdots, i_k$-th rows and the $j_1, \cdots, j_l$-th columns, where $1 \leq i_1 < \cdots < i_k \leq d, 1 \leq j_1 < \cdots < j_l \leq d$. The following property about the adjugate matrix is well known

$$(\det \mathbb{X})\delta_{i,j} = \sum_{k=1}^d (-1)^{i+k}(\det \mathbb{X}_{k,i})X_{kj}. \tag{3.8}$$

Denote by $\mathbb{H} = \mathbb{E}_{r+1}(\lambda)_{(r-1,r,r+1),(r-1,r,r+1)} \in \mathbb{R}^{(r-2)\times(r-2)}$. Then the matrix $\mathbb{E}_{r+1}(\lambda)$ can be partitioned as

$$\mathbb{E}_{r+1}(\lambda) = \left( \begin{array}{c|ccc} & b_{r-3} & 0 & 0 \\ \mathbb{H} & -1 & b_{r-2} & 0 \\ & a_{r-1} & -1 & 0 \\ & 0 & a_r & -1 \\ \hline c_1 \quad \cdots \quad c_{r-2} & c_{r-1} & c_r & c_{r+1} \end{array} \right). \tag{3.9}$$

Similarly, we have

$$\mathbb{E}_r(\lambda) = \left( \begin{array}{c|cc} & b_{r-3} & 0 \\ \mathbb{H} & -1 & 0 \\ & a_{r-1} & -1 \\ \hline c_1 \quad \cdots \quad c_{r-2} & c_{r-1} & c_r \end{array} \right), \quad \mathbb{E}_{r-1}(\lambda) = \left( \begin{array}{c|c} & 0 \\ \mathbb{H} & -1 \\ \hline c_1 \quad \cdots \quad c_{r-2} & c_{r-1} \end{array} \right). \tag{3.10}$$

The following simple identities will play an important role in our analysis

$$\mathbb{E}_{r+1}(\lambda)_{r,(r,r+1)} = \mathbb{E}_r(\lambda)_{*,r}, \tag{3.11}$$

where for any $\mathbb{X} \in \mathbb{R}^{d \times d}$, we denote $\mathbb{X}_{*,j} \in \mathbb{R}^{d \times (d-1)}$ the matrix by removing the $j-$th column of $\mathbb{X}$. Similarly, we denote $\mathbb{X}_{i,*} \in \mathbb{R}^{(d-1) \times d}$ the matrix by removing the $i$-th row of $\mathbb{X}$.

The following elementary lemma is useful in our analysis.

**Lemma 3.1.** *For any $r \geq 3$ and $1 \leq j \leq r-2$, we have*

$$\det[\mathbb{E}_{r+1}(\lambda)_{(r-2,r,r+1),(j,r,r+1)}] = -a_{r-1} \det[\mathbb{E}_{r-1}(\lambda)_{r-1,j}]$$
$$= a_{r-1} c_{r-1} \det[\mathbb{E}_{r-1}(\lambda)_{r-2,j}], \tag{3.12}$$
$$\det[\mathbb{E}_{r+1}(\lambda)_{(r-1,r,r+1),(j,r,r+1)}] = -\det[\mathbb{E}_r(\lambda)_{r,j}]$$
$$= c_r \det[\mathbb{E}_r(\lambda)_{r-1,j}]. \tag{3.13}$$

*Proof.* We only prove (3.12). The identity (3.13) can be proved similarly. By (3.11),

$$\det[\mathbb{E}_{r+1}(\lambda)_{(r-2,r,r+1),(j,r,r+1)}] = \det[\mathbb{E}_r(\lambda)_{(r-2,r),(j,r)}]$$
$$= \det \left( \begin{array}{c|c} & \\ \mathbb{H}_{r-2,j} & b_{r-3} \\ & a_{r-1} \end{array} \right)$$
$$= a_{r-1} \det[\mathbb{H}_{r-2,j}].$$

On the other hand, it is easy to see from (3.11) that

$$\det[\mathbb{E}_{r-1}(\lambda)_{r-1,j}] = -\det \mathbb{H}_{r-2,j}, \quad \det[\mathbb{E}_{r-1}(\lambda)_{r-2,j}] = c_{r-1} \det[\mathbb{H}_{r-2,j}].$$

This completes the proof. $\qquad\square$

To proceed, we note that

$$\mathbb{E}_2(\lambda) = \begin{pmatrix} a_1 & -1 \\ 1 & -1 \end{pmatrix}, \quad \mathbb{E}_3(\lambda) = \begin{pmatrix} a_1 & -1 & 0 \\ 0 & a_2 & -1 \\ 1 & -1 & 1 \end{pmatrix}. \tag{3.14}$$

**Lemma 3.2.** *Let $\varphi_r(\lambda) = \det \mathbb{E}_{r+1}(\lambda)$ be the determinant of $\mathbb{E}_{r+1}(\lambda)$. Then $\varphi_1(\lambda) = 1 - a_1$, $\varphi_2(\lambda) = 1 - a_1 + a_1 a_2$, and*

$$\varphi_r(\lambda) = \varphi_{r-1}(\lambda) + a_r a_{r-1} \varphi_{r-2}(\lambda), \quad r \geq 3. \tag{3.15}$$

*Moreover, $\varphi_r(\lambda) = P_r(-\lambda)$, where $P_r(\lambda)$ is the numerator of $[r/r]$ Padé approximation of $e^\lambda$.*

*Proof.* The determinants of $\mathbb{E}_{r+1}(\lambda)$ for $r = 1, 2$ follow easily from (3.14). Since by (3.7), $a_r = -b_{r-2}$, we obtain by adding the $r$-th row to the $(r-2)$-th row of $\mathbb{E}_{r+1}(\lambda)$ in (3.9) and then expanding the determinant by the $r$-th row that

$$
\begin{aligned}
\varphi_r(\lambda) &= \det \begin{pmatrix} \mathbb{H} & \begin{matrix} b_{r-3} & 0 & 0 \\ -1 & 0 & -1 \\ a_{r-1} & -1 & 0 \\ 0 & a_r & -1 \end{matrix} \\ \begin{matrix} c_1 & \cdots & c_{r-2} \end{matrix} & \begin{matrix} c_{r-1} & c_r & c_{r+1} \end{matrix} \end{pmatrix} \\
&= \det \mathbb{E}_r(\lambda) + a_r \det \begin{pmatrix} \mathbb{H} & \begin{matrix} b_{r-3} & 0 \\ -1 & -1 \\ a_{r-1} & 0 \end{matrix} \\ \begin{matrix} c_1 & \cdots & c_{r-2} \end{matrix} & \begin{matrix} c_{r-1} & c_r \end{matrix} \end{pmatrix} \\
&= \det \mathbb{E}_r(\lambda) + a_r a_{r-1} \det \mathbb{E}_{r-1}(\lambda),
\end{aligned}
$$

where in the last equality we have expanded the determinant by the $(r-1)$-th row. Since $a_1 = \lambda/2$, $a_2 = \lambda/6$, we use Lemma 2.3 to conclude $\varphi_r(\lambda) = P_r(-\lambda)$, where $P_r(\lambda)$ is the numerator of $[r/r]$ Padé approximation of $e^\lambda$. $\square$

**Theorem 3.1.** *The matrix $\varphi_r(\tau_n \mathbb{D})$ is invertible. The solution of (3.5) $\mathbf{X} = (\mathbf{a}_0^T, \cdots, \mathbf{a}_r^T)^T$ satisfies that for $i = 1, \cdots, r+1$,*

$$\mathbf{a}_{i-1} = (-1)^r \delta_{i,r+1} \mathbf{b}_r + \sum_{j=1}^{r} (\zeta_j \mathbb{I} + \tau_n \mathbb{D})^{-1} \left( \sum_{k=1}^{r+1} (-1)^{i+k+1} \frac{\phi_{ki}(-\zeta_j)}{P'_r(\zeta_j)} \mathbf{b}_{k-1} \right),$$

*where $\phi_{ki}(\lambda) = \det[\mathbb{E}_{r+1}(\lambda)_{k,i}]$, $k, i = 1, \cdots, r+1$, are the minors of $\mathbb{E}_{r+1}(\lambda)$.*

*Proof.* By Lemma 3.2, $\varphi_r(\lambda) = P_r(-\lambda) = (-1)^r \frac{r!}{(2r)!}(\lambda + \zeta_1) \cdots (\lambda + \zeta_r)$, where $\zeta_1, \cdots, \zeta_r \in \mathbb{C}$ are zeros of the diagonal Padé numerator of type $[r/r]$ for $e^\lambda$. By Lemma 2.4, $\mathrm{Re}(\zeta_k) \leq -2$, $k = 1, \cdots, r$. On the other hand, (1.2) implies that the eigenvalues of $\mathbb{D}$ lie in the left half-plane. Thus the eigenvalues of $\zeta_k \mathbb{I} + \tau_n \mathbb{D}$, $1 \leq k \leq r$, lie in the half-plane $\{z \in \mathbb{C} : \mathrm{Re}(z) \leq -2\}$. This shows $\varphi_r(\tau_n \mathbb{D})$ is invertible.

13

Now by (3.8) we have

$$[\det \mathbb{E}_{r+1}(\lambda)]\delta_{i,j} = \sum_{k=1}^{r+1} (-1)^{i+k} [\det \mathbb{E}_{r+1}(\lambda)_{k,i}] e_{kj}(\lambda),$$

where $e_{kj}(\lambda)$ is the $(k,j)$ element of $\mathbb{E}_{r+1}(\lambda)$. By replacing $\lambda$ by $\tau_n\mathbb{D}$ in above equality, we have

$$\varphi_r(\tau_n\mathbb{D})\delta_{i,j} = \sum_{k=1}^{r+1} (-1)^{i+k}\phi_{ki}(\tau_n\mathbb{D})e_{kj}(\tau_n\mathbb{D}). \qquad (3.16)$$

From (3.5) we have

$$\sum_{j=1}^{r+1} e_{ij}(\tau_n\mathbb{D})\mathbf{a}_{j-1} = \mathbf{b}_{i-1}, \quad i = 1, \cdots, r+1.$$

Thus multiplying (3.16) by $\mathbf{a}_{j-1}$ and summing over $j$ from 1 to $r+1$, we obtain

$$\begin{aligned}
\varphi_r(\tau_n\mathbb{D})\mathbf{a}_{i-1} &= \sum_{k=1}^{r+1}(-1)^{i+k}\phi_{ki}(\tau_n\mathbb{D}) \cdot \sum_{j=1}^{r+1} e_{kj}(\tau_n\mathbb{D})\mathbf{a}_{j-1} \\
&= \sum_{k=1}^{r+1}(-1)^{i+k}\phi_{ki}(\tau_n\mathbb{D})\mathbf{b}_{k-1}. \qquad (3.17)
\end{aligned}$$

Note that for $(k,i) \neq (r+1,r+1)$, $\phi_{ki}(\lambda) \in P^m$, $m \leq r-1$, by (2.22) we have

$$\frac{\phi_{ki}(\tau_n\mathbb{D})}{\varphi_r(\tau_n\mathbb{D})} = \sum_{j=1}^{r} -\frac{\phi_{ki}(-\zeta_j)}{P'_r(\zeta_j)}(\zeta_j\mathbb{I} + \tau_n\mathbb{D})^{-1}.$$

For $(k,i) = (r+1,r+1)$, we have from (3.6) that $\det[E_{r+1}(\lambda)_{r+1,r+1}] = \prod_{j=1}^{r} a_j = \frac{r!}{(2r)!}\lambda^r$. Thus $\phi_{r+1,r+1}(\lambda) - (-1)^r\varphi_r(\lambda) \in P^{r-1}$, by using (2.22) again we obtain

$$\frac{\phi_{r+1,r+1}(\tau_n\mathbb{D})}{\varphi_r(\tau_n\mathbb{D})} = (-1)^r\mathbb{I} + \sum_{j=1}^{r} -\frac{\phi_{r+1,r+1}(-\zeta_j)}{P'_r(\zeta_j)}(\zeta_j\mathbb{I} + \tau_n\mathbb{D})^{-1}.$$

This completes the proof of the theorem. $\qquad\qquad\square$

We remark that (3.17) can also be proved by using an abstract result in Brown [4, Theorem 2.19 and Corollary 2.21] where linear algebra when matrix elements are defined over a space of commuting matrices are studied.

From this theorem we know that the discrete problem (2.2) can be solved by solving $r(r+1)$ linear systems of equations of order $M \times M$ in parallel once all minors of $\mathbb{E}_{r+1}(\lambda)$ at $\lambda = -\zeta_j, j = 1, \cdots, r$, are known. In the following, we will find recursive formulas to computing these minors.

Let $\mathbb{G}_r(\lambda) = \mathbb{E}_{r+1}(\lambda)_{r,r+1}$, $r \geq 1$. The determinants of $\mathbb{G}_r(\lambda)$ for $r = 1, 2$ can be calculated by (3.14). We have the following lemma for $\det \mathbb{G}_r(\lambda)$ for $r \geq 3$.

14

**Lemma 3.3.** *For $r \geq 3$, we have*

$$\det \mathbb{G}_r(\lambda) = \det \mathbb{G}_{r-1}(\lambda) - a_{r-1}b_{r-2} \det \mathbb{G}_{r-2}(\lambda) + c_r \prod_{k=1}^{r-1} a_k.$$

*Proof.* By definition and the partition in (3.9), we know that

$$\mathbb{G}_r(\lambda) = \left( \begin{array}{c|ccc} & & b_{r-3} & 0 \\ & \mathbb{H} & -1 & b_{r-2} \\ & & a_{r-1} & -1 \\ \hline c_1 & \cdots \quad c_{r-2} & c_{r-1} & c_r \end{array} \right).$$

By expanding the determinant by the $(r-1)$-th row and use (3.11), we obtain

$$
\begin{aligned}
\det \mathbb{G}_r(\lambda) &= \det[\mathbb{E}_{r+1}(\lambda)_{(r-1,r),(r,r+1)}] + a_{r-1} \det \left( \begin{array}{c|c} & \\ \mathbb{H} & b_{r-2} \\ \hline c_1 \quad \cdots \quad c_{r-2} & c_r \end{array} \right) \\
&= \det[\mathbb{E}_r(\lambda)_{r-1,r}] + a_{r-1}(c_r \det \mathbb{H} - b_{r-2} \det[\mathbb{E}_{r+1}(\lambda)_{(r-2,r-1.r),(r-1,r,r+1)}]) \\
&= \det \mathbb{G}_{r-1}(\lambda) + c_r \prod_{k=1}^{r-1} a_k - a_{r-1}b_{r-2} \det[\mathbb{E}_r(\lambda)_{(r-2,r-1),(r-1,r)}] \\
&= \det \mathbb{G}_{r-1}(\lambda) + c_r \prod_{k=1}^{r-1} a_k - a_{r-1}b_{r-2} \det[\mathbb{E}_{r-1}(\lambda)_{r-2,r-1}] \\
&= \det \mathbb{G}_{r-1}(\lambda) - a_{r-1}b_{r-2} \det \mathbb{G}_{r-2}(\lambda) + c_r \prod_{k=1}^{r-1} a_k,
\end{aligned}
$$

where we have used the fact that $\det \mathbb{H} = \Pi_{k=1}^{r-2} a_k$ and expanded the determinant by the last column in the second equality. This completes the proof. $\square$

The minors of $\mathbb{G}_r(\lambda)$ for $r = 1, 2$ can be computed directly by (3.14). The following lemma gives the recursive formulas for some of the minors of $\mathbb{G}_r(\lambda)$ which will be used to compute the minors of $\mathbb{E}_{r+1}(\lambda)$.

**Lemma 3.4.** *For $r \geq 3$, we have*

$$
\begin{aligned}
\det[\mathbb{G}_r(\lambda)_{i,r}] &= (-1)^{r-i-1} \frac{a_1 \cdots a_{r-1}}{a_1 \cdots a_i} \det \mathbb{G}_i(\lambda), \quad 1 \leq i \leq r-1, \\
\det[\mathbb{G}_r(\lambda)_{r-1,j}] &= \det[\mathbb{E}_r(\lambda)_{r-1,j}] - b_{r-2} \det[\mathbb{G}_{r-1}(\lambda)_{r-2,j}], \quad 1 \leq j \leq r-1,
\end{aligned}
$$

*Proof.* For $i = 1, \cdots, r - 2$, by definition, we have

$$
\det[\mathbb{G}_r(\lambda)_{i,r}] \;=\; \det \begin{pmatrix}
a_1 & -1 & b_1 & & & & & & \\
& \ddots & \ddots & \ddots & & & & & \\
& & a_{i-1} & -1 & b_{i-1} & & & & \\
& & & 0 & a_{i+1} & -1 & b_{i+1} & & \\
& & & & \ddots & \ddots & \ddots & \ddots & \\
& & & & & 0 & a_{r-3} & -1 & b_{r-3} \\
& & & & & & 0 & a_{r-2} & -1 \\
& & & & & & & 0 & a_{r-1} \\
c_1 & \cdots & \cdots & \cdots & \cdots & \cdots & c_{r-3} & c_{r-2} & c_{r-1}
\end{pmatrix}
$$

$$
= \;(-a_{r-1})\cdots(-a_{i+1})\det \begin{pmatrix}
a_1 & -1 & b_1 & & \\
& \ddots & \ddots & \ddots & \\
& & a_{i-2} & -1 & b_{i-2} \\
& & & a_{i-1} & -1 \\
c_1 & \cdots & \cdots & c_{i-1} & c_i
\end{pmatrix}
$$

$$
= \;(-1)^{r-i-1}\Big( \prod_{k=i+1}^{r-1} a_k \Big) \det \mathbb{G}_i(\lambda).
$$

Finally, for $\mathbb{G}(\lambda)_{i,r}$, $i = r - 1$, we have by the definition and using the first identity in (3.11)

$$
\mathbb{G}_r(\lambda)_{r-1,r} = \mathbb{E}_{r+1}(\lambda)_{(r-1,r),(r,r+1)} = \mathbb{E}_r(\lambda)_{r-1,r} = \mathbb{G}_{r-1}(\lambda).
$$

This shows the first equality of the lemma. To show the second equality, for any $1 \leq j \leq r - 2$, we have by using the partition (3.9) that

$$
\mathbb{G}_r(\lambda)_{r-1,j} = \left( \begin{array}{c|cc}
& b_{r-3} & 0 \\
\mathbb{H}_{*,j} & -1 & b_{r-2} \\
\hline
c_1 \;\cdots\; c_{j-1} \; c_{j+1} \; \cdots \; c_{r-2} \; c_{r-1} & c_r &
\end{array} \right).
$$

Thus by expanding the determinant by the last column, we have by using (3.11), (3.13) and the definition of $\mathbb{G}_{r-1}(\lambda)$ that

$$
\begin{aligned}
&\det[\mathbb{G}_r(\lambda)_{r-1,j}] \\
=\;& c_r \det[\mathbb{E}_{r+1}(\lambda)_{(r-1,r,r+1),(j,r,r+1)}] - b_{r-2}\det[\mathbb{E}_{r+1}(\lambda)_{(r-2,r-1,r),(j,r,r+1)}] \\
=\;& \det[\mathbb{E}_r(\lambda)_{r-1,j}] - b_{r-2}\det[\mathbb{E}_r(\lambda)_{(r-2,r-1),(j,r)}] \\
=\;& \det[\mathbb{E}_r(\lambda)_{r-1,j}] - b_{r-2}\det[\mathbb{G}_{r-1}(\lambda)_{r-2,j}].
\end{aligned}
$$

This completes the proof. $\qquad\square$

The minors of $\mathbb{E}_{r+1}(\lambda)$ for $r = 1, 2$ can be easily computed from (3.14). The following theorem gives the recursive formulas for computing all minors of $\mathbb{E}_{r+1}(\lambda)$ for $r \geq 3$.

16

**Theorem 3.2.** *Let $r \geq 3$, we have*
*1° For $i = 1, \cdots, r-2$,*

$$\det[\mathbb{E}_{r+1}(\lambda)_{i,j}] = \det[\mathbb{E}_r(\lambda)_{i,j}] + a_r a_{r-1} \det[\mathbb{E}_{r-1}(\lambda)_{i,j}], \quad 1 \leq j \leq r-2,$$

$$\det[\mathbb{E}_{r+1}(\lambda)_{i,j}] = (-1)^{j-i-1} \frac{a_1 \cdots a_{j-1}}{a_1 \cdots a_i} \det \mathbb{G}_i(\lambda), \quad j = r-1, r, r+1.$$

*2° For $i = r-1$, we have*

$$\det[\mathbb{E}_{r+1}(\lambda)_{r-1,j}] = -c_{r+1} a_r \det[\mathbb{E}_r(\lambda)_{r,j}] + \det[\mathbb{G}_r(\lambda)_{r-1,j}], \quad 1 \leq j \leq r-1,$$

$$\det[\mathbb{E}_{r+1}(\lambda)_{r-1,j}] = (-1)^{j-i-1} \frac{a_1 \cdots a_{j-1}}{a_1 \cdots a_i} \det \mathbb{G}_i(\lambda), \quad j = r, r+1.$$

*3° For $i = r$, we have*

$$\det[\mathbb{E}_{r+1}(\lambda)_{r,j}] = \det[\mathbb{E}_r(\lambda)_{r-1,j}] - a_{r-1} b_{r-2} \det[\mathbb{E}_{r-1}(\lambda)_{r-2,j}], \quad 1 \leq j \leq r-2,$$

$$\det[\mathbb{E}_{r+1}(\lambda)_{r,j}] = c_{r+1}(-1)^{i-j} \prod_{k=1}^{j-1} a_k, \quad j = r-1, r,$$

$$\det[\mathbb{E}_{r+1}(\lambda)_{r,r+1}] = \det \mathbb{G}_{r-1}(\lambda) - a_{r-1} b_{r-2} \det \mathbb{G}_{r-2}(\lambda) + c_r \prod_{k=1}^{r-1} a_k.$$

*4° For $i = r+1$, we have*

$$\det[\mathbb{E}_{r+1}(\lambda)_{r+1,j}] = -\det[\mathbb{E}_r(\lambda)_{r,j}] - a_{r-1} b_{r-2} \det[\mathbb{E}_{r-1}(\lambda)_{r-1,j}], \quad 1 \leq j \leq r-2,$$

$$\det[\mathbb{E}_{r+1}(\lambda)_{r+1,j}] = (-1)^{i-j} \prod_{k=1}^{j-1} a_k, \quad j = r-1, r, r+1.$$

*Proof.* The proof is divided into 4 steps.

STEP 1. The first equality in 1° can be proved by the same argument as that in Lemma 3.2. Here we omit the details. We only prove the second equality in 1° when $j = r+1$. The other cases can be proved similarly. By the partition in (3.9), we know that for $1 \leq i \leq r-2$,

$$\mathbb{E}_{r+1}(\lambda)_{i,r+1} = \left( \begin{array}{cc|cc} & \mathbb{H}_{i,*} & & \mathbb{F}_{i,*} \\ \hline & & a_{r-1} & -1 \\ & & 0 & a_r \\ c_1 & \cdots \quad c_{r-2} & c_{r-1} & c_r \end{array} \right), \quad \mathbb{F} = \left( \begin{array}{ccc} 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ b_{r-3} & 0 & 0 \\ -1 & b_{r-2} & 0 \end{array} \right),$$

where $\mathbb{F} \in \mathbb{R}^{(r-2) \times 3}$. By expanding the determinant first by the $r$-th and then by the $(r-1)$-th row, we know by (3.11) that

$$
\begin{aligned}
\det[\mathbb{E}_{r+1}(\lambda)_{i,r+1}] &= a_r a_{r-1} \det[\mathbb{E}_{r+1}(\lambda)_{(i,r-1,r),(r-1,r,r+1)}] \\
&= a_r a_{r-1} \det[\mathbb{E}_r(\lambda)_{(i,r-1),(r-1,r)}] \\
&= a_r a_{r-1} \det[\mathbb{G}_{r-1}(\lambda)_{i,r-1}].
\end{aligned}
$$

17

This shows the second equality in $1°$ by the first identity in Lemma 3.4.

STEP 2. We only prove the first equality in $2°$ when $1 \le j \le r - 2$. The other cases can be proved similarly. By the partition in (3.9), we have for $1 \le j \le r - 2$,

$$\mathbb{E}_{r+1}(\lambda)_{r-1,j} = \left(\begin{array}{c|ccc} & & b_{r-3} & 0 & 0 \\ \multicolumn{1}{c}{} & \mathbb{H}_{*,j} & -1 & b_{r-2} & 0 \\ \hline & & 0 & a_r & -1 \\ c_1 & \cdots \quad c_{r-2} & c_{r-1} & c_r & c_{r+1} \end{array}\right)$$

Expanding the determinant by the last column, we obtain

$$\begin{aligned} \det[\mathbb{E}_{r+1}(\lambda)_{r-1,j}] &= c_{r+1}a_r \det[\mathbb{E}_{r+1}(\lambda)_{(r-1,r,r+1),(j,r,r+1)}] + \det[\mathbb{E}_{r+1}(\lambda)_{(r-1,r),(j,r+1)}] \\ &= -c_{r+1}a_r \det[\mathbb{E}_r(\lambda)_{r,j}] + \det[\mathbb{G}_r(\lambda)_{r-1,j}]. \end{aligned}$$

This shows the first equality in $2°$ for $1 \le j \le r - 2$.

STEP 3. The second equality in $3°$ can be easily proved. The last equality is shown in Lemma 3.3 since $\mathbb{E}_{r+1}(\lambda)_{r,r+1} = \mathbb{G}_r(\lambda)$ by definition. To show the first equality in $3°$, we again use the partition in (3.9) to obtain for $1 \le j \le r - 2$,

$$\mathbb{E}_{r+1}(\lambda)_{r,j} = \left(\begin{array}{c|ccc} & & b_{r-3} & 0 & 0 \\ \multicolumn{1}{c}{} & \mathbb{H}_{*,j} & -1 & b_{r-2} & 0 \\ & & a_{r-1} & -1 & 0 \\ \hline c_1 \; \cdots \; c_{j-1} \; c_{j+1} \; \cdots \; c_{r-2} & c_{r-1} & c_r & c_{r+1} \end{array}\right).$$

By expanding the determinant successively by the last columns, we have

$$\begin{aligned} &\det[\mathbb{E}_{r+1}(\lambda)_{r,j}] \\ = \; &-c_{r+1} \det[\mathbb{E}_{r+1}(\lambda)_{(r-1,r,r+1),(j,r,r+1)}] - c_{r+1}b_{r-2} \det[\mathbb{E}_{r+1}(\lambda)_{(r-2,r,r+1),(j,r,r+1)}]. \end{aligned}$$

This shows the first equality in $3°$ by Lemma 3.1.

STEP 4. The first equality in $4°$ can be proved by the same argument as that Step 3. The second equality can be easily proved. Here we omit the details. $\square$

This theorem indicates that all minors of $\mathbb{E}_{r+1}(\lambda)$ can be computed once one knows all minors of $\mathbb{E}_m(\lambda)$, $1 \le m \le r$, and the minors $\det[\mathbb{G}_r(\lambda)_{r-1,j}]$, $1 \le j \le r - 1$, which can be computed by Lemma 3.4 recursively based on the information of the minors $E_m(\lambda)$, $1 \le m \le r$.

The following lemma indicates that the nodal values of the solution to (2.2) depends only on the coefficient $\mathbf{a}_0$.

**Lemma 3.5.** *Let* $\mathbf{Y}_r(t)$, $r \ge 1$, *be the solution of the problem* (2.2). *Then*

$$\mathbf{Y}_r(t_{n+1}) = \mathbf{Y}_r(t_n) + \tau_n \mathbb{D}\mathbf{a}_0 + \tau_n \mathbf{R}_0, \quad n = 1, \cdots, N - 1.$$

*Proof.* We integrate (2.2) over $I_n$ and use the orthogonality of Legendre polynomials to obtain

$$\mathbf{Y}_r(t_{n+1}) = \mathbf{Y}_r(t_n) + \int_{I_n} \mathbb{D} \left( \sum_{j=0}^{r-1} \mathbf{a}_j \widetilde{L}_j(t) \right) dt + \int_{I_n} \sum_{j=0}^{r-1} \mathbf{R}_j \widetilde{L}_j(t) \, dt,$$

$$= \mathbf{Y}_r(t_n) + \tau_n \mathbb{D} \mathbf{a}_0 + \tau_n \mathbf{R}_0.$$

This completes proof. $\qquad\square$

By Theorem 3.1, we have then

$$\mathbf{Y}_r(t_{n+1}) = \mathbf{Y}_r(t_n) + \sum_{j=1}^{r} (\zeta_j \mathbb{I} + \tau_n \mathbb{D})^{-1} \left( \sum_{k=1}^{r+1} (-1)^k \frac{\phi_{k1}(-\zeta_j)}{P_r'(\zeta_j)} (\tau_n \mathbb{D}) \, \mathbf{b}_{k-1} \right) + \tau_n \mathbf{R}_0.$$

This leads to the following parallel algorithm to compute the nodal values of the solution $\mathbf{Y}_r$ to the problem (2.2).

**Algorithm 3.1.** *Given* $\mathbf{Y}_r(t_0) = \mathbf{Y}_0$. *For* $n = 1, \cdots, N - 1$, *do the following.*
$1°$ *Compute* $\mathbf{v}_j \in \mathbb{R}^M$, $j = 1, \cdots, r$, *in parallel, where*

$$\mathbf{v}_j = \sum_{k=1}^{r+1} (-1)^k \frac{\phi_{k1}(-\zeta_j)}{P_r'(\zeta_j)} (\tau_n \mathbb{D}) \, \mathbf{b}_{k-1}.$$

$2°$ *Solve* $(\tau_n \mathbb{D} + \zeta_j \mathbb{I}) \mathbf{w}_j = \mathbf{v}_j$, $j = 1, \cdots, r$, *in parallel.*
$3°$ *Compute*

$$\mathbf{Y}_r(t_{n+1}) = \mathbf{Y}_r(t_n) + \sum_{j=0}^{r} \mathbf{w}_j + \tau_n \mathbf{R}_0.$$

The following parallel-in-time algorithm computes the solution of the problem (2.2) inside each time interval.

**Algorithm 3.2.** *Given* $\mathbf{Y}_r(t_0) = \mathbf{Y}_0$.
$1°$ *Call Algorithm 3.1 to obtain* $\mathbf{Y}_r(t_n)$, $n = 1, \ldots, N$.
$2°$ *Compute the coefficients* $\mathbf{a}_1, \cdots, \mathbf{a}_r$ *of* $\mathbf{Y}_r$ *in each time interval* $I_n$, $n = 1, \cdots, N - 1$, *in parallel as follows.*
(i) *Compute* $\mathbf{v}_{ij} \in \mathbb{R}^M$, $i = 2, \cdots, r + 1$, $j = 1, \cdots r$, *in parallel, where*

$$\mathbf{v}_{ij} = \sum_{k=1}^{r+1} (-1)^{i+k+1} \frac{\phi_{ki}(-\zeta_j)}{P_r'(\zeta_j)} \, \mathbf{b}_{k-1}.$$

(ii) *Solve* $(\tau_n \mathbb{D} + \zeta_j \mathbb{I}) \mathbf{w}_{ij} = \mathbf{v}_{ij}$, $i = 2, \cdots, r + 1, j = 1, \cdots, r$, *in parallel.*
(iii) *Compute* $\mathbf{a}_{i-1} = (-1)^r \delta_{i,r+1} \mathbf{b}_r + \sum_{j=1}^{r} \mathbf{w}_{ij}$, $i = 2, \cdots, r + 1$, *in parallel.*

**Remark 3.1.** *In [12], it is observed that the zeros of $P_r(z)$ come in complex conjugate pairs if they are complex. If $\zeta_{j'} = \bar{\zeta}_j$, $j, j' = 1, \cdots, r$, then $\mathbf{v}_j = \bar{\mathbf{v}}_{j'}$, and*

$$\mathbf{w}_j + \mathbf{w}_{j'} = \frac{\mathbf{v}_j}{\tau_n \mathbb{D} + \zeta_j \mathbb{I}} + \frac{\bar{\mathbf{v}}_j}{\tau_n \mathbb{D} + \bar{\zeta}_j \mathbb{I}} = 2 \operatorname{Re} \left[ \frac{\mathbf{v}_j}{\tau_n \mathbb{D} + \zeta_j \mathbb{I}} \right].$$

*Thus one need only to solve $k$ complex matrix problems instead of $2k$ in Algorithm 3.1 $(2°)$ and $kr$ complex matrix problems instead of $2kr$ in Algorithm 3.2 $(2°)$, where $2k$, $0 \leq k \leq r/2$, are the number of complex zeros of $P_r(z)$.*

**Remark 3.2.** *If $\zeta = a + \mathbf{i}b$ is a complex zero of $P_r(z)$, then $a \leq -2$ by Lemma 2.3. By Remark 3.1, without loss of generality, we can choose one of the zeros such that $b < 0$. Let $\mathbf{w} = \mathbf{w}_1 + \mathbf{i}\mathbf{w}_2$, $\mathbf{v} = \mathbf{v}_1 + \mathbf{i}\mathbf{v}_2$, where $\mathbf{v}_i, \mathbf{w}_i \in \mathbb{R}^M$, $i = 1, 2$, satisfy $(\tau_n \mathbb{D} + \zeta \mathbb{I})\mathbf{w} = \mathbf{v}$. Then*

$$\widetilde{\mathbb{D}} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} := \begin{pmatrix} \tau_n \mathbb{D} + a\mathbb{I} & -b\,\mathbb{I} \\ -b\,\mathbb{I} & -(\tau_n \mathbb{D} + a\mathbb{I}) \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ -\mathbf{v}_2 \end{pmatrix}.$$

*Let $\mathbb{F} = \operatorname{diag}(\tau_n \mathbb{D} + (a + b)\,\mathbb{I}, \tau_n \mathbb{D} + (a + b)\,\mathbb{I}) \in \mathbb{R}^{2M \times 2M}$ be the diagonal matrix. It is shown in Chen et al [5, Lemma 4.1] that the condition number $\kappa(\mathbb{F}^{-1}\widetilde{\mathbb{D}}) \leq \sqrt{2}$. Therefore, the complex system $(\tau_n \mathbb{D} + \zeta \mathbb{I})\mathbf{w} = \mathbf{v}$ can be efficiently solved if one has the efficient solver for the real matrix $\tau_n \mathbb{D} + (a + b)\,\mathbb{I}$, where $a + b \leq -2$. Notice that the eigenvalues of $\mathbb{D}$ lie in the left half-plane due to the assumption $\mathbb{D} + \mathbb{D}^T \leq 0$.*

**Remark 3.3.** *For non-standard ODE systems of the form,*

$$\mathbb{M}\mathbf{Y}' = \mathbb{D}\mathbf{Y} + \mathbf{R} \quad in \ (0, T), \quad \mathbf{Y}(0) = \mathbf{Y}_0, \tag{3.18}$$

*one can use the transformations $\widetilde{\mathbf{Y}} = \mathbb{M}^{\frac{1}{2}}\mathbf{Y}$, $\widetilde{\mathbb{D}} = \mathbb{M}^{-\frac{1}{2}}\mathbb{D}\mathbb{M}^{-\frac{1}{2}}$, $\widetilde{\mathbf{R}} = \mathbb{M}^{-\frac{1}{2}}\mathbf{R}$ to transform the problem (3.18) to (1.1) and use above algorithms to solve the transformed problem. This leads to the following algorithm which is similar to Algorithm 3.1 to find the nodal values of the solution of the continuous time Galerkin method for solving (3.18). A similar algorithm to Algorithm 3.2 can also be formulated.*

**Algorithm 3.3.** *Given $\mathbf{Y}_r(t_0) = \mathbf{Y}_0$. For $n = 1, \cdots, N - 1$, do the following.*
$1°$ *Compute $\mathbf{v}_j \in \mathbb{R}^M$, $j = 1, \cdots, r$, in parallel, where*

$$\mathbf{v}_j = \sum_{k=1}^{r+1} (-1)^k \frac{\phi_{k1}(-\zeta_j)}{P_r'(\zeta_j)} (\tau_n \mathbb{D}) \mathbb{M}^{-1} \mathbf{b}_{k-1}.$$

$2°$ *Solve $(\tau_n \mathbb{D} + \zeta_j \mathbb{M})\mathbf{w}_j = \mathbf{v}_j$, $j = 1, \cdots, r$, in parallel.*
$3°$ *Compute*

$$\mathbf{Y}_r(t_{n+1}) = \mathbf{Y}_r(t_n) + \sum_{j=1}^{r} \mathbf{w}_j + \tau_n \mathbb{M}^{-1} \mathbf{R}_0.$$

To conclude this section, we prove the following theorem for finding the nodal values of the solution (2.2) which extends (1.3) for solving the ODE system (1.1) when $\mathbf{R} = \mathbf{0}$.

**Theorem 3.3.** *Let* $\mathbf{Y}_r \in \mathbf{V}_\tau^r$, $r \geq 1$, *be the solution of the problem* (2.2). *Then for* $n = 1, \cdots, N - 1$,

$$\mathbf{Y}_r(t_{n+1}) = \frac{P_r(\tau_n\mathbb{D})}{P_r(-\tau_n\mathbb{D})}\mathbf{Y}_r(t_n) + \sum_{k=1}^{r}(-1)^{k+1}\frac{\phi_{k1}(\tau_n\mathbb{D})}{P_r(-\tau_n\mathbb{D})}\mathbf{b}_{k-1} + \tau_n\mathbf{R}_0,$$

*where* $\phi_{k1}(\lambda) = \det[\mathbb{E}_{r+1}(\lambda)_{k,1}]$.

*Proof.* By Lemma 3.2 and (3.17) we have

$$P_r(-\tau_n\mathbb{D})\mathbf{a}_0 = \sum_{k=1}^{r+1}(-1)^{k+1}\phi_{k1}(\tau_n\mathbb{D})\mathbf{b}_{k-1}.$$

Since $\mathbf{b}_{r+1} = \mathbf{Y}_r(t_n)$, by Lemma 3.5,

$$\begin{aligned}
\mathbf{Y}_r(t_{n+1}) &= \left[\mathbb{I} + (-1)^r\frac{\tau_n\mathbb{D}\phi_{r+1,1}(\tau_n\mathbb{D})}{P_r(-\tau_n\mathbb{D})}\right]\mathbf{Y}_r(t_n) \\
&\quad + \sum_{k=1}^{r}(-1)^{k+1}\phi_{k1}(\tau_n\mathbb{D})\mathbf{b}_{k-1} + \tau_n\mathbf{R}_0.
\end{aligned}$$

Denote by $\psi_r(\lambda) = (-1)^r\lambda\phi_{r+1,1}(\lambda) = (-1)^r\lambda\det[\mathbb{E}_{r+1}(\lambda)_{r+1,1}]$. By Theorem 3.2, 4°, we know that $\psi_r(\lambda)$ satsifies

$$\psi_r(\lambda) = \psi_{r-1}(\lambda) + \frac{\lambda^2}{4}\frac{1}{(2r-1)(2r-3)}\psi_{r-2}(\lambda), \quad r \geq 3.$$

On the other hand, by (3.14), we have $\psi_1(\lambda) = \lambda, \psi_2(\lambda) = \lambda$. This implies by Lemma 2.3 that $\psi_r(\lambda) = P_r(\lambda) - P_r(-\lambda)$. Thus

$$\mathbb{I} + (-1)^r\frac{\tau_n\mathbb{D}\phi_{r+1,1}(\tau_n\mathbb{D})}{P_r(-\tau_n\mathbb{D})} = \mathbb{I} + \frac{P_r(\tau_n\mathbb{D}) - P_r(-\tau_n\mathbb{D})}{P_r(-\tau_n\mathbb{D})} = P_r(\tau_n\mathbb{D}).$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

## 4 The dissipative system

In this section, we propose an alternative way to compute the coefficients $\mathbf{a}_1, \cdots, \mathbf{a}_r$ of the solution $\mathbf{Y}_r$ of the problem (2.2) when the ODE system (1.1) is dissipative $\mathbb{D} + \mathbb{D}^T < 0$. The algorithm is based on the block tridiagonal structure of the matrix and is less expensive than the step 2° in Algorithm 3.2.

Let $\widetilde{\mathbf{X}} = (\mathbf{a}_1^T, \cdots, \mathbf{a}_r^T)^T \in \mathbb{R}^{rM}$ and $\widetilde{\mathbf{B}} = (\widetilde{\mathbf{b}}_1^T, \mathbf{b}_2^T, \cdots, \mathbf{b}_r^T)^T$ with $\widetilde{\mathbf{b}}_1 = \mathbf{b}_1 - \frac{1}{2}(\tau_n\mathbb{D})\mathbf{a}_0$. It follows from (3.5) that $\widetilde{\mathbf{X}} \in \mathbb{R}^{rM}$ satisfies

$$\widetilde{\mathbb{A}}\widetilde{\mathbf{X}} = \widetilde{\mathbf{B}}, \qquad\qquad\qquad\qquad\qquad\qquad (4.1)$$

where

$$\widetilde{\mathbb{A}} = \begin{pmatrix} -\mathbb{I} & -\frac{\tau_n}{2}\mathbb{D}\frac{1}{5} & \cdots & \cdots & \cdots & 0 \\ \frac{\tau_n}{2}\mathbb{D}\frac{1}{3} & -\mathbb{I} & -\frac{\tau_n}{2}\mathbb{D}\frac{1}{7} & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \cdots & \cdots & \frac{\tau_n}{2}\mathbb{D}\frac{1}{2r-5} & -\mathbb{I} & -\frac{\tau_n}{2}\mathbb{D}\frac{1}{2r-1} & 0 \\ \cdots & \cdots & \cdots & \frac{\tau_n}{2}\mathbb{D}\frac{1}{2r-3} & -\mathbb{I} & 0 \\ \cdots & \cdots & \cdots & \cdots & \frac{\tau_n}{2}\mathbb{D}\frac{1}{2r-1} & -\mathbb{I} \end{pmatrix}.$$

It is easy to see that $\widetilde{\mathbb{A}} = \mathbb{E}_{r+1}(\tau_n\mathbb{D})_{r+1,1}$, let $\phi_r(\lambda) = \det[\mathbb{E}_{r+1}(\lambda)_{r+1,1}]$. The goal of this section is to show that $\phi_r(\tau_n\mathbb{D}) \in \mathbb{R}^{M\times M}$ is invertible so that the standard chasing algorithm for block tridiagonal matrices (cf., e.g., Golub and Van Load [14, §4.5]) can be used to solve (4.1).

We start with the following lemma.

**Lemma 4.1.** *Let $u_n \in P^n$, $n \geq 0$, such that* (i) $u_0 = 0, u_1 = 1$ *or* (ii) $u_0 = 1, u_1 = 1 + A_1 t$, $A_1 \in \mathbb{R}$, *and for $n \geq 2$,*

$$u_n(t) = (1 + A_n t)u_{n-1}(t) - C_n t^2 u_{n-2}, \tag{4.2}$$

*where $A_n \in \mathbb{R}, C_n > 0$ for $n \geq 2$. Then we have, for $n \geq 2$,*

$$u_n(u_{n-1} + tu'_{n-1}) - tu_{n-1}u'_n > 0 \quad in\ \mathbb{R}\backslash\{0\}.$$

*Proof.* We set, for $n \geq 1$,

$$G_n(t, t') = \frac{t'u_n(t)u_{n-1}(t') - tu_{n-1}(t)u_n(t')}{t' - t}.$$

It is easy to see by (4.2) that $G_1(t, t') = u_0$, and for $n \geq 2$,

$$G_n(t, t') = u_{n-1}(t)u_{n-1}(t') + C_n tt' G_{n-1}(t, t').$$

This implies easily

$$G_n(t, t') = u_{n-1}(t)u_{n-1}(t') + \sum_{k=0}^{n-2} C_n \cdots C_{k+2}(tt')^{n-1-k}u_k(t)u_k(t'),$$

where we have used $G_1(t, t') = u_0(t)u_0(t')$ in both cases (i) and (ii). By letting $t' \to t$, we obtain for $n \geq 2$,

$$\lim_{t' \to t} G_n(t, t') = u_n(t)[tu_{n-1}(t)]' - tu_{n-1}(t)u'_n(t) > 0 \ \ in\ \mathbb{R}\backslash\{0\},$$

where we have used the condition $u_1 = 1$ in the case (i) and $u_0 = 1$ in the case (ii). This completes the proof. $\square$

The following theorem is the main result of this section.

**Theorem 4.1.** *Let $\mathbb{D} \in \mathbb{R}^{M\times M}$ satisfy $\mathbb{D} + \mathbb{D}^T < 0$, that is, $\mathbb{D} + \mathbb{D}^T$ is negative definite. Then the matrix $\phi_r(\tau_n\mathbb{D}) \in \mathbb{R}^{M\times M}$, $r \geq 1$, $1 \leq n \leq N - 1$, is invertible.*

22

*Proof.* We are going to show that all zeros of $\phi_r(\lambda)$ locate at the imaginary axis $\{z \in \mathbb{C} : \mathrm{Re}(z) = 0\}$. This implies easily $\phi_r(\tau_n \mathbb{D})$ is invertible since the eigenvalues of $\mathbb{D}$ lie in the left-half plane due to the dissipative property $\mathbb{D} + \mathbb{D}^T < 0$.

To study the property of the zeros of $\phi_r(\lambda)$, we denote $\psi_r(\lambda) = (-1)^r \phi_r(\lambda)$. By (3.14) and Theorem 3.2 (4°) we know that $\psi_1(\lambda) = 1, \psi_2(\lambda) = 1$, and

$$\psi_r(\lambda) = \psi_{r-1}(\lambda) + d_r \lambda^2 \psi_{r-2}(\lambda), \quad d_r = \frac{1}{4(2r-1)(2r-3)}, \quad \text{for } r \geq 3. \qquad (4.3)$$

We note that (4.3) is also valid for $r = 2$ if we define $\psi_0(\lambda) = 0$. Set $t = \lambda^2$ and for $m \geq 0$, define $f_m(t) = \psi_{2m}(\lambda), g_m = \psi_{2m+1}(\lambda)$. Then

$$f_m(t) = g_{m-1}(t) + d_{2m} t f_{m-1}(t), \quad f_0(t) = 0, f_1(t) = 1,$$
$$g_m(t) = f_m(t) + d_{2m+1} t g_{m-1}(t), \quad g_0(t) = 1, g_1(t) = 1 + d_3 t.$$

This implies that, for $m \geq 2$,

$$f_m(t) = (1 + A_m t) f_{m-1}(t) - C_m t^2 f_{m-2}(t), \qquad (4.4)$$
$$g_m(t) = (1 + \tilde{A}_m t) g_{m-1}(t) - \tilde{C}_m t^2 f_{m-2}(t), \qquad (4.5)$$

where $A_m = d_{2m} + d_{2m-1}, C_m = d_{2m-1} d_{2m-2}, \tilde{A}_m = d_{2m+1} + d_{2m}, \tilde{C}_m = d_{2m} d_{2m-1}$. We observe that $f_m, g_m$ satisfy the same recurrence relation but with different coefficients and initial values. In the following we will only prove $f_m, m \geq 2$, has $m$ real zeros in $(-\infty, 0)$ which then implies that $\phi_{2m}(\lambda) = f_m(\lambda^{1/2})$ has all zeros on the imaginary axis. The proof for $\phi_{2m+1}(\lambda), m \geq 1$, is similar and we omit the details.

We extend the argument in [24, §3.3 (4)] for orthogonal polynomials to show that $f_m, m \geq 2$, has $m$ zeros in $(-\infty, 0)$ by using Sturm theorem (cf., e.g., Perron [18, pp.7-9]) based on the recurrence formula (4.4). We first note that if $f_m(t) = \sum_{k=0}^{m} \theta_k t^k$, then $\theta_m > 0$ since by (4.3) the leading coefficients of the polynomial $\psi_r(\lambda)$ are positive, and $\theta_0 = f_m(0) = 1$ by (4.4). Now we claim that

$$f_m(t), f_{m-1}(t), \cdots, f_1(t) \qquad (4.6)$$

form a Sturmian sequence in $[-M, -\delta]$ for sufficiently large $M > 0$ and sufficiently small $\delta > 0$ in the following sense. (i) $f_1(t) = 1$ has no zeros in $[-M, -\delta]$. (ii) $f_m(-M) f_m(-\delta) \neq 0$ for $M \gg 1$ and $\delta \ll 1$ since $\theta_m > 0$ and $\theta_0 = 1$. (iii) If $c \in [-M, -\delta]$ is a zero of $f_k(t)$, $1 \leq k \leq m-1$, then $f_{k+1}(c) f_{k-1}(c) < 0$. In fact, By Lemma 4.1, $f_{k-1}(c) \neq 0$. By (4.4), $f_{k+1}(c) = -C_m c^2 f_{k-1}(c)$. (iv) If $c \in [-M, -\delta]$ such that $f_m(c) = 0$, then $f_m'(c) f_{m-1}(c) > 0$, which is a direct consequence of Lemma 4.1. Now the number of variations of sign in (4.6) at $t = -M$ is $m$ for sufficiently large $M$ since $\theta_m > 0$; it is zero at $t = -\delta$ for sufficiently small $\delta > 0$ since $\theta_0 = 1$. Thus by Sturm theorem we conclude that $f_m(t)$ has exactly $m$ zeros in $[-M, -\delta]$. This completes the proof. □

Based on this theorem, we can use the following parallel-in-time algorithm to compute the solution $\mathbf{Y}_r$ to the problem (2.2).

**Algorithm 4.1.** *Given* $\mathbf{Y}_r(t_0) = \mathbf{Y}_0$.
1° *Call Algorithm 3.1 to obtain* $\mathbf{Y}_r(t_n)$ *for* $n = 1, \ldots, N$.
2° *Compute* $\mathbf{Y}_r$ *in each time interval* $I_n$, $n = 1, \cdots, N - 1$, *in parallel by solving* (4.1) *to obtain* $\mathbf{a}_1, \cdots, \mathbf{a}_r$ *using the LU decomposition for block tridiagonal matrices.*

We remark that the algorithm of *LU* decomposition for block tridiagonal matrices for solving (4.1) requires to solve $r$ systems of linear equations of size $M$ in sequential instead of to solve $r^2$ systems of linear equations of size $M$ in parallel in Step 2° of Algorithm 3.2.

## 5   Optimal stability and error esstimates

In this section, we show optimal stability and error estimates of the continuous time Galerkin method (2.2) in terms of $r$ when $\mathbb{D}$ is a symmetric or skew-symmetric matrix. This will be achieved by using the explicit formulas in Theorem 3.1. We start by studying further properties of the minors of the stiffness matrix of the continuous time Galerkin method $\mathbb{A} = \mathbb{E}_{r+1}(\tau_n \mathbb{D})$.

Let $\chi_{r+1,j}(\lambda) = (-1)^{r+1} \det[\mathbb{E}_{r+1}(\lambda)_{r+1,j}]$, then by (3.14) we have

$$\chi_{2,1} = -1, \chi_{2,2} = a_1, \chi_{3,1} = -1, \chi_{3,2} = a_1, \chi_{3,3} = -a_1 a_2. \tag{5.1}$$

For $r \geq 3$, by 4° in Theorem 3.2, we have the following recursive formulas

$$\chi_{r+1,j}(\lambda) = \chi_{r,j}(\lambda) + a_r a_{r-1} \chi_{r-1,j}(\lambda), \quad 1 \leq j \leq r - 2, \tag{5.2}$$

$$\chi_{r+1,j}(\lambda) = (-1)^j \prod_{k=1}^{j-1} a_k, \quad j = r - 1, r, r + 1. \tag{5.3}$$

Let $\varphi_0 = 1$ and $\varphi_r(\lambda) = \det \mathbb{E}_{r+1}(\lambda)$, $r \geq 1$. Then by Lemma 3.2, $\varphi_1 = 1 - a_1$,

$$\varphi_{r+1} = \varphi_r + a_r a_{r+1} \varphi_{r-1}, \quad r \geq 1. \tag{5.4}$$

**Lemma 5.1.** *For any* $r \geq 1$ *and* $\lambda \in \mathbb{R}$, *we have*

$$\sum_{j=1}^{r-1} [\chi_{r,j}(-\lambda)\chi_{r+1,j}(\lambda) + \chi_{r,j}(\lambda)\chi_{r+1,j}(-\lambda)] \frac{1}{2j - 1}$$
$$= \varphi_{r-1}(-\lambda)\varphi_r(\lambda) + \varphi_{r-1}(\lambda)\varphi_r(-\lambda) + \frac{(-1)^r 2r}{2r - 1}(a_1 \cdots a_{r-1})^2, \tag{5.5}$$

$$\sum_{j=1}^{r+1} \chi_{r+1,j}(-\lambda)\chi_{r+1,j}(\lambda) \frac{1}{2j - 1} = \varphi_r(-\lambda)\varphi_r(\lambda) + \frac{(-1)^{r+1} 2r}{2r + 1}(a_1 \cdots a_r)^2. \tag{5.6}$$

*Proof.* We denote

$$A_r : = \sum_{j=1}^{r-1} [\chi_{r,j}(-\lambda)\chi_{r+1,j}(\lambda) + \chi_{r,j}(\lambda)\chi_{r+1,j}(-\lambda)] \frac{1}{2j - 1}$$
$$- [\varphi_{r-1}(-\lambda)\varphi_r(\lambda) + \varphi_{r-1}(\lambda)\varphi_r(-\lambda)],$$

$$B_{r+1} : = \sum_{j=1}^{r+1} \chi_{r+1,j}(-\lambda)\chi_{r+1,j}(\lambda)\frac{1}{2j-1} - \varphi_r(-\lambda)\varphi_r(\lambda).$$

We will argue by induction. First (5.5)-(5.6) are obvious for $r = 1, 2$ by (5.1). Now we assume (5.5)-(5.6) are valid for all $r \leq n$, $n \geq 2$. Since by (5.3), $\chi_{n+2,n}(\lambda) = \chi_{n+1,n}(\lambda)$, we have by (5.2) and (5.4) that

$$A_{n+1} = 2B_{n+1} - 2\chi_{n+1,n+1}(-\lambda)\chi_{n+1,n+1}(\lambda)\frac{1}{2n+1} + a_n a_{n+1} A_n.$$

Now by (5.3) and the induction assumption that (5.5)-(5.6) are valid for $r = n$, we obtain

$$A_{n+1} = \frac{(-1)^{n+1}2(n+1)}{2n+1}(a_1 \cdots a_n)^2.$$

This shows (5.5) for $r = n + 1$. Similarly, we can prove by (5.2)-(5.4) that

$$\begin{aligned} B_{n+2} & = B_{n+1} + (a_n a_{n+1})^2 B_n + a_n a_{n+1} A_n \\ & \quad + \chi_{n+2,n+2}(-\lambda)\chi_{n+2,n+2}(\lambda)\frac{1}{2n+3} - (a_n a_{n+1})^2 \chi_{n,n}(-\lambda)\chi_{n,n}(\lambda)\frac{1}{2n-1}. \end{aligned}$$

Now by the induction assumption (5.6) for $r = n, n+1$ and (5.5) for $r = n$, we obtain by using (5.3) that

$$B_{n+2} = (-1)^{n+2}\frac{2(n+1)}{2n+3}(a_1 \cdots a_{n+1})^2.$$

This completes the proof. $\qquad\square$

**Lemma 5.2.** *Let $r \geq 1$. For any $\lambda \leq 0$, we have*

$$\sum_{j=1}^{r-1} \chi_{r,j}(\lambda)\chi_{r+1,j}(\lambda)\frac{1}{2j-1} \leq \varphi_{r-1}(\lambda)\varphi_r(\lambda), \quad \sum_{j=1}^{r+1} \chi_{r+1,j}(\lambda)^2 \frac{1}{2j-1} \leq \varphi_r(\lambda)^2. \quad (5.7)$$

*Proof.* For any $r \geq 1$, we denote

$$C_r := \sum_{j=1}^{r-1} \chi_{r,j}\chi_{r+1,j}\frac{1}{2j-1} - \varphi_{r-1}\varphi_r, \quad D_{r+1} := \sum_{j=1}^{r+1} \chi_{r+1,j}^2 \frac{1}{2j-1} - \varphi_r^2.$$

We again argue by induction. First (5.7) is obvious for $r = 1, 2$ by (5.1) since $\lambda \leq 0$. Now we assume (5.7) is valid for all $r \leq n$, $n \geq 2$. By (5.2)-(5.4), it is easy to see that

$$C_{n+1} = -\chi_{n+1,n+1}^2 \frac{1}{2n+1} + D_{n+1} + a_n a_{n+1} C_n,$$

where we have used $\chi_{n+2,n}(\lambda) = \chi_{n+1,n}(\lambda)$. Thus if $C_n \leq 0, D_{n+1} \leq 0$, then $C_{n+1} \leq 0$.
On the other hand, by (5.2)-(5.4), we have

$$D_{n+2} = D_{n+1} + a_n a_{n+1} D_n + 2a_n a_{n+1} C_n + (a_1 \cdots a_{n+1})^2 \left(\frac{1}{2n+3} - \frac{1}{2n-1}\right).$$

Thus $D_{n+2} \leq 0$ if $D_n \leq 0, D_{n+1} \leq 0$, and $C_n \leq 0$. This completes the proof. $\qquad\square$

25

The following theorem is the main result of this section.

**Theorem 5.1.** *Let $\mathbb{D}$ be a symmetric or skew-symmetric matrix and $\mathbf{Y}_r \in \mathbf{V}_\tau^r$ is the solution of the problem (2.2). Then we have*

$$\|\mathbf{Y}_r\|_{L^2(0,T)} \leq T^{1/2}\|\mathbf{Y}_0\|_{\mathbb{R}^M} + CT\|\mathbf{R}\|_{L^2(0,T)}, \tag{5.8}$$

$$\max_{0 \leq t \leq T} \|\mathbf{Y}_r\|_{\mathbb{R}^M} \leq Cr(\|\mathbf{Y}_0\|_{\mathbb{R}^M} + T^{1/2}\|\mathbf{R}\|_{L^2(0,T)}), \tag{5.9}$$

*where the constant $C$ is independent of $\tau, r, \mathbb{D}$, and $\mathbf{R}$.*

*Proof.* Let $\hat{\mathbf{Y}}_r \in [P^r]^M$ be defined in (2.6) of Lemma 2.1, we claim that

$$\|\hat{\mathbf{Y}}_r\|_{L^2(I_n)} \leq \tau_n^{1/2}\|\mathbf{Y}_r^n\|_{\mathbb{R}^M}, \tag{5.10}$$

which improves the bound (2.10) in the proof of Lemma 2.1. To show (2.6), by Theorem 3.1, we have

$$\hat{\mathbf{Y}}_r(t) = \sum_{j=1}^{r+1} \mathbf{a}_{j-1}\tilde{L}_{j-1}(t) = \sum_{j=1}^{r+1}(-1)^j \frac{\chi_{r+1,j}(\tau_n\mathbb{D})}{\varphi_r(\tau_n\mathbb{D})}\mathbf{Y}_r^n\tilde{L}_{j-1}(t) \quad \forall t \in I_n.$$

Thus by (2.5),

$$\|\hat{\mathbf{Y}}_r\|_{L^2(I_n)}^2 = \sum_{j=1}^{r+1}\|\chi_{r+1,j}(\tau\mathbb{D})\varphi_r(\tau_n\mathbb{D})^{-1}\mathbf{Y}_r^n\|_{\mathbb{R}^M}^2 \frac{\tau_n}{2j-1}. \tag{5.11}$$

Denote $\mathbf{Z}_r^n = \varphi_r(\tau_n\mathbb{D})^{-1}\mathbf{Y}_r^n$. If $\mathbb{D}$ is skew-symmetric $\mathbb{D}^T = -\mathbb{D}^T$, by (2.5), (5.6), we have

$$\|\hat{\mathbf{Y}}_r\|_{L^2(I_n)}^2 = \sum_{j=1}^{r+1}\|\chi_{r+1,j}(\tau_n\mathbb{D})\mathbf{Z}_r^n\|_{\mathbb{R}^M}^2 \frac{\tau_n}{2j-1}$$

$$= \tau_n\|\varphi_r(\tau_n\mathbb{D})\mathbf{Z}_r^n\|_{\mathbb{R}^M}^2 - \tau_n\left\|\sqrt{\frac{2r}{2r+1}}\frac{r!}{(2r)!}(\tau_n\mathbb{D})^r\mathbf{Z}_r^n\right\|_{\mathbb{R}^M}^2$$

$$\leq \tau_n\|\mathbf{Y}_r^n\|_{\mathbb{R}^M}^2.$$

This shows the claim (5.10) when $\mathbb{D}$ is skew-symmetric.

If $\mathbb{D}$ is symmetric, the eigenvalues of $\mathbb{D}$ are non-positive since $\mathbb{D} + \mathbb{D}^T \leq 0$. By Lemma 5.2, it is easy to show that

$$\sum^{r+1}\|\chi_{r+1,j}(\tau_n\mathbb{D})\mathbf{Z}_r^n\|_{\mathbb{R}^M}^2 \frac{\tau_n}{2j-1} - \tau_n\|\varphi_r(\tau_n\mathbb{D})\mathbf{Z}_r^n\|_{\mathbb{R}^M}^2 \leq 0,$$

which yields

$$\sum_{j=1}^{r+1}\|\chi_{r+1,j}(\tau_n\mathbb{D})\varphi_r(\tau_n\mathbb{D})^{-1}\mathbf{Y}_r^n\|_{\mathbb{R}^M}^2 \frac{\tau_n}{2j-1} \leq \tau_n\|\mathbf{Y}_r^n\|_{\mathbb{R}^M}^2.$$

26

Now it follows from (5.11) that $\|\hat{\mathbf{Y}}_r\|_{L^2(I_n)} \leq \tau_n^{1/2}\|\mathbf{Y}_r^n\|_{\mathbb{R}^M}$. This shows the claim (5.10) when $\mathbb{D}$ is a symmetric matrix.

It follows from (5.10), (2.12) and (2.3) that

$$\begin{aligned}
\|\mathbf{Y}_r\|_{L^2(I_n)} &\leq \tau_n^{1/2}\|\mathbf{Y}_r^n\|_{\mathbb{R}^M} + 2\tau_n\|\mathbf{R}\|_{L^2(I_n)} \\
&\leq \tau_n^{1/2}\|\mathbf{Y}_0\|_{\mathbb{R}^M} + C\tau_n^{1/2}T^{1/2}\|\mathbf{R}\|_{L^2(0,T)}
\end{aligned}$$

This implies (5.8) easily. Now by the *hp* inverse estimate,

$$\max_{t_n \leq t \leq t_{n+1}} \|\mathbf{Y}_r\|_{\mathbb{R}^M} \leq C\tau_n^{-1/2}r\|\mathbf{Y}_r\|_{L^2(I_n)} \leq Cr(\|\mathbf{Y}_0\|_{\mathbb{R}^M} + T^{1/2}\|\mathbf{R}\|_{L^2(0,T)}).$$

This shows (5.9). This completes the proof. $\qquad\square$

The following theorem which improves the error estimates in Theorem 2.1 can be proved by the argument in Theorem 2.1 by using Theorem 5.1 instead of Lemma 2.1. Here we omit the details.

**Theorem 5.2.** *Let $\mathbb{D}$ be a symmetric or skew-symmetric matrix. Assume that $\mathbf{R} \in [H^s(0,T)]^M$, $\mathbf{Y} \in [W^{1+s,\infty}(0,T)]^M$, $s \geq 1$, and $\mathbf{Y}_r \in \mathbf{V}_\tau^r$ is the solution of the problem (2.2), we have*

$$\|\mathbf{Y} - \mathbf{Y}_r\|_{L^2(0,T)} \leq C(1+T)\frac{\tau^{\min(r+1,s)}}{r^s}(\|\mathbf{Y}\|_{H^s(0,T)} + \|\mathbb{D}\mathbf{Y}\|_{H^s(0,T)}),$$

$$\max_{0 \leq t \leq T} \|\mathbf{Y} - \mathbf{Y}_r\|_{\mathbb{R}^M} \leq C(1+T^{1/2})\frac{\tau^{\min(r+1,s)}}{r^{s-1}}(T^{1/2}\|\mathbf{Y}\|_{W^{s+1,\infty}(0,T)} + \|\mathbf{R}\|_{H^s(0,T)}),$$

*where the constant $C$ is independent of $\tau, r, \mathbb{D}$, and $\mathbf{R}$ but may depend on $s$.*

We remark that the first estimate in Theorem 5.2 is optimal both in $\tau$ and $r$.

# 6 Numerical examples

In this section, we provide some numerical examples to confirm the theoretical results in this paper.

**Example 1.** *(Dissipative problem) Let $\Omega = (0,1) \times (0,1)$ and $T = 1$. We consider the following constant coefficient convection-diffusion problem*

$$\begin{cases} u_t + \nabla \cdot (\boldsymbol{\beta}u - \epsilon\nabla u) = f & \text{in } \Omega \times (0,T), \\ u(\mathbf{x},0) = u_0(\mathbf{x}) & \text{in } \Omega. \end{cases} \tag{6.1}$$

*The boundary condition is set to be periodic. The source term $f$ is chosen such that the exact solution is $u(\mathbf{x},t) = \exp(-t)\sin(4\pi(x_1 - t))\cos(4\pi(x_2 - t))$.*

We choose $\boldsymbol{\beta} = (1,1)^T$ and $\epsilon = 1$ in (6.1). For spatial discretizations, we apply the local discontinuous Galerkin (LDG) method in Cockburn and Shu [9] by using purely upwind fluxes for convection terms and alternating fluxes for diffusion terms. For the sake of completeness, we recall the method for solving (6.1) here.

Let $\mathcal{M}$ denote a uniform Cartesian mesh of $\Omega$ with $h$ the length of the sides of the elements. $\mathcal{E} = \mathcal{E}^{\text{side}} \cup \mathcal{E}^{\text{bdy}}$, where $\mathcal{E}^{\text{side}} := \{e = \partial K \cap \partial K' : K, K' \in \mathcal{M}\}$, $\mathcal{E}^{\text{bdy}} := \{e = \partial K \cap \partial \Omega : K \in \mathcal{M}\}$. For any subset $\widehat{\mathcal{M}} \subset \mathcal{M}$ and $\widehat{\mathcal{E}} \subset \mathcal{E}$, we use the notation

$$(u, v)_{\widehat{\mathcal{M}}} = \sum_{K \in \widehat{\mathcal{M}}} (u, v)_K, \quad \langle u, v \rangle_{\widehat{\mathcal{E}}} = \sum_{e \in \widehat{\mathcal{E}}} \langle u, v \rangle_e,$$

where $(\cdot, \cdot)_K$ and $\langle \cdot, \cdot \rangle_e$ denote the inner product of $L^2(K)$ and $L^2(e)$, respectively.

For any $e \in \mathcal{E}$, we fix a unit normal vector $\mathbf{n}_e$ of $e$ with the convention that $\mathbf{n}_e$ is the unit outer normal to $\partial \Omega$ if $e \in \mathcal{E}^{\text{bdy}}$. For any $v \in H^1(\mathcal{M}) := \{v : v \in H^1(K), K \in \mathcal{M}\}$, we define the jump operator of $v$ across $e$:

$$\llbracket v \rrbracket_e := v^- - v^+ \quad \forall e \in \mathcal{E}^{\text{side}}, \quad \llbracket v \rrbracket_e := v^- \quad \forall e \in \mathcal{E}^{\text{bdy}},$$

where $v^{\pm}(\mathbf{x}) := \lim_{\varepsilon \to 0^+} v(\mathbf{x} \pm \varepsilon \mathbf{n}_e) \ \forall \mathbf{x} \in e$. For any integer $p \geq 0$, we define the finite element space

$$V_h^p := \{v \in L^2(\Omega) : v|_K \in Q^p(K), K \in \mathcal{M}\},$$

where $Q^p(K)$ denotes the space of polynomials of degree at most $p$ in each variable in $K$.

The semi-discrete problem is to find $(u_h, \mathbf{q}_h) \in [V_h^p]^3$ such that, for all test functions $(v_h, \mathbf{r}_h) \in [V_h^p]^3$,

$$\begin{aligned}
&(\partial_t u_h, v_h)_{\mathcal{M}} + \mathcal{G}(\boldsymbol{\beta} u_h, v_h) = \sqrt{\epsilon} \left[ -(\mathbf{q}_h, \nabla v_h)_{\mathcal{M}} + \langle \mathbf{q}_h^- \cdot \mathbf{n}, \llbracket v_h \rrbracket \rangle_{\mathcal{E}} \right] + (f, v_h)_{\mathcal{M}}, \\
&(\mathbf{q}_h, \mathbf{r}_h)_{\mathcal{M}} = \sqrt{\epsilon} \left[ -(u_h, \operatorname{div} \mathbf{r}_h)_{\mathcal{M}} + \langle u_h^+, \llbracket \mathbf{r}_h \rrbracket \cdot \mathbf{n} \rangle_{\mathcal{E}} \right], \\
&u_h(\mathbf{x}, 0) = (\mathcal{P}_h u_0)(\mathbf{x}) \quad \text{in } \Omega.
\end{aligned}$$

Here $\mathcal{P}_h : L^2(\Omega) \to V_h^p$ is the standard $L^2$ projection operator, and

$$\mathcal{G}(\boldsymbol{\beta} u_h, v_h) = -(\boldsymbol{\beta} u_h, \nabla v_h)_{\mathcal{M}} + \langle \check{u}_h \boldsymbol{\beta} \cdot \mathbf{n}, \llbracket v_h \rrbracket \rangle_{\mathcal{E}},$$

where $\check{u}_h$ is chosen as the upwind flux: $\check{u}_h = u_h^-$ if $\boldsymbol{\beta} \cdot \mathbf{n} > 0$, $\check{u} = u_h^+$ if $\boldsymbol{\beta} \cdot \mathbf{n} < 0$. For $e \in \mathcal{E}^{\text{bdy}}$, we use the periodic boundary condition to define $u_h^+$.

The optimal $L^2$-norm error estimate of order $p+1$ of the semi-discrete scheme for quasi-uniform Cartesian meshes can be found in Cheng et al [8, Theorem 2.4], where it is shown that $\max_{0 \leq t \leq T} \|u - u_h\|_{L^2(\Omega)} \leq C(1 + T)h^{p+1}$. Therefore, combined with the continuous time Galerkin scheme, we know that the fully discrete scheme has $O(h^{p+1} + \tau^{r+1})$ accuracy in the norm $\|\cdot\|_{L^\infty(0,T;L^2(\Omega))}$ and $O(h^{p+1} + \tau^{2r})$ in the $L^2$ norm at nodes $t = t_n, n = 1, \cdots, N$.

To test the accuracy at the nodes, we set $\tau = h^{\frac{p+1}{2r}}$ and thus $N = T/\tau = T\beta^{\frac{1}{r}}$, where $\beta = h^{-\frac{p+1}{2}}$. The numerator of the $[r/r]$ Padé approximation $P_r(z)$ has $2k$ complex zeros and 1 real root if $r = 2k + 1, k \geq 1$, and $2k$ complex zeros if $r = 2k, k \geq 1$. Denote by $C(2M)$ the costs of solving the matrix problem $\tau_n \mathbb{D} + \zeta_j \mathbb{I}$ with $\zeta_j$ being complex and $C(M)$ the costs of solving the matrix problem $\tau_n \mathbb{D} + \zeta_j \mathbb{I}$ with $\zeta_j$ real, where $\zeta_j, j = 1, \cdots, r$, are zeros of $P_r(z)$. Then the computational time in each time step of Algorithm 3.1 is proportion to $C(2M)$ for the parallel computation and proportion to $kC(2M) + (r - 2k)C(M)$ for the sequential computation. The wall time of using Algorithm 3.1 using parallel machines is then proportion to $N = T\beta^{\frac{1}{r}}$ which is decreasing in $r$. Thus high order time discretization is

28

preferred for parallel computations. On the other hand, for the sequential computation, the wall time of using Algorithm 3.1 is proportion to $rN = Tr\beta^{\frac{1}{r}}$ which minimizes at $r = \ln\beta$ for $r > 0$. This implies that the optimal choice of the order for the sequential computation is $r = \lfloor\ln\beta\rfloor + 1$, where $\lfloor a\rfloor$ is the maximum integer strictly less than $a > 0$. Table 1 shows the error $\|(u - u_h)(\cdot, T)\|_{L^2(\Omega)}$ at the terminal time when $r = \lfloor\ln\beta\rfloor + 1$. The optimal $(p+1)$-th order is observed which confirms our theoretical results. We observe that the errors of high order schemes are significant smaller than the low order schemes.

To test the accuracy in the $\|\cdot\|_{L^\infty(0,T;L^2(\Omega))}$ norm, we set $\tau = h^{\frac{p+1}{r+1}}$ and thus $N = T/\tau = T\gamma^{\frac{1}{r+1}}$, where $\gamma = h^{-(p+1)}$. The wall time of using Algorithm 3.1 and Algorithm 3.2 for the parallel computation is proportion to $N = T\gamma^{\frac{1}{r+1}}$ which decreases in $r$. On the other hand, for the sequential computation, the wall time of using Algorithm 3.1 and Algorithm 4.1 is $rN = Tr\gamma^{\frac{1}{r+1}}$ which is increasing in $r$ if $\ln\gamma \le 4$ and minimizes at $r^* = [-(2-\ln\gamma) + \sqrt{(2-\ln\gamma)^2 - 4}]/2$ if $\ln\gamma \ge 4$. Since $r^* \ge 1$ is equivalent to $\ln\gamma \ge 4$, the optimal choice of the order for minimizing the computation wall time is $r = \max(1, \lfloor r^*\rfloor + 1)$. We note that for the sequential computation, Algorithm 4.1 is cheaper than Algorithm 3.2. Table 2 shows the error

$$\max_{0\le n\le N-1, 1\le k\le 10} \|(u - u_h)(\cdot, t_n + 0.1k\tau_n)\|_{L^2(\Omega)}$$

as the approximation of $\|u - u_h\|_{L^\infty(0,T;L^2(\Omega))}$ when $r = \max(1, \lfloor r^*\rfloor + 1)$. We again observe the optimal $(p+1)$-th order convergence and that high order methods perform much better than low order methods.

Table 1: Example 1: numerical errors of $\|(u - u_h)(\cdot, T)\|_{L^2(\Omega)}$ and orders.

| $h$ | $p = 3$ | | $p = 4$ | | $p = 5$ | |
|---|---|---|---|---|---|---|
| | error | order | error | order | error | order |
| 1/4 | 8.06E-03 | – | 1.29E-03 | – | 1.71E-04 | – |
| 1/8 | 5.64E-04 | 3.84 | 4.40E-05 | 4.88 | 2.84E-06 | 5.91 |
| 1/16 | 3.55E-05 | 3.99 | 1.37E-06 | 5.01 | 4.33E-08 | 6.04 |
| 1/32 | 2.03E-06 | 4.13 | 4.26E-08 | 5.00 | 6.94E-10 | 5.96 |

Table 2: Example 1: numerical errors in $\|\cdot\|_{L^\infty(0,T;L^2(\Omega))}$ norm and orders.

| $h$ | $p = 3$ | | $p = 4$ | | $p = 5$ | |
|---|---|---|---|---|---|---|
| | error | order | error | order | error | order |
| 1/4 | 2.77E-02 | – | 3.99E-03 | – | 5.81E-04 | – |
| 1/8 | 2.00E-03 | 3.79 | 1.36E-04 | 4.87 | 1.02E-05 | 5.83 |
| 1/16 | 1.14E-04 | 4.14 | 4.40E-06 | 4.95 | 1.69E-07 | 5.91 |
| 1/32 | 7.18E-06 | 3.98 | 1.33E-07 | 5.05 | 2.73E-09 | 5.95 |

**Example 2.** *(Wave propagation problem) Let $\Omega = (-2, 2) \times (-2, 2)$ and $T = 1$. We*

*consider the following wave equation with discontinuous coefficients*

$$
\begin{cases}
\dfrac{1}{\rho c^2}\partial_t u = \operatorname{div}\mathbf{q} + f, \quad \rho\partial_t\mathbf{q} = \nabla u \quad \text{in } \Omega \times (0,T), \\[2mm]
[\![u]\!] = 0, \quad [\![\mathbf{q}\cdot\mathbf{n}]\!] = 0 \quad \text{on } \Gamma \times (0,T), \\[2mm]
u = 0 \quad \text{on } \partial\Omega \times (0,T), \\[2mm]
u(\mathbf{x},0) = u_0(\mathbf{x}), \quad \mathbf{q}(\mathbf{x},0) = \mathbf{q}_0(\mathbf{x}) \quad \text{in } \Omega.
\end{cases}
\tag{6.2}
$$

*We assume the interface $\Gamma$ is the union of two closely located ellipses. We take $\Omega_1 = \{\mathbf{x} \in \Omega : \frac{(x_1-d_1)^2}{a^2} + \frac{x_2^2}{b^2} < 1 \text{ or } \frac{(x_1-d_2)^2}{a^2} + \frac{x_2^2}{b^2} < 1\}$, which is the union of two disks, and $\Omega_2 = \Omega\backslash\bar{\Omega}_1$. Here $d_1 = -0.82$, $d_2 = 0.82$, $a = 0.81$, and $b = 0.51$. The distance between two ellipses is $0.02$. We consider the wave equation (6.2) with $\rho_1 = 1/2$, $\rho_2 = 1$, $c_1 = c_2 = 1$, and the source $f$ is chosen such that the exact solution is*

$$
u(\mathbf{x},t) =
\begin{cases}
\cos(3t)\sin(r_1-1)\sin(r_2-1)\sin(3\pi x_1)\sin(3\pi x_2) & \text{in } \Omega_1, \\[3mm]
2\cos(3t)\sin(r_1-1)\sin(r_2-1)\sin(3\pi x_1)\sin(3\pi x_2) & \text{in } \Omega_2,
\end{cases}
$$

*where $r_1 = \frac{(x_1-d_1)^2}{a^2} + \frac{x_2^2}{b^2}$ $r_2 = \frac{(x_1-d_2)^2}{a^2} + \frac{x_2^2}{b^2}$. The exact solution $\mathbf{q}(\mathbf{x},t)$ is computed by (6.2) with the initial condition $\mathbf{q}_0 = 0$.*

We use the unfitted finite element method in Chen et al [7] to discretize the problem in space. Let $\mathcal{M}$ be an induced mesh which is constructed from a Cartesian partition $\mathcal{T}$ of the domain $\Omega$ with possible local refinements and hanging nodes so that the elements are large with respect to both domains $\Omega_1, \Omega_2$. Let $\mathcal{M}^\Gamma := \{K \in \mathcal{M} : K \cap \Gamma \neq \emptyset\}$ and $\mathcal{E} = \mathcal{E}^{\text{side}} \cup \mathcal{E}^\Gamma \cup \mathcal{E}^{\text{bdy}}$, where $\mathcal{E}^\Gamma := \{\Gamma_K = \Gamma \cap K : K \in \mathcal{M}\}$.

For any $K \in \mathcal{M}^\Gamma$, $i = 1,2$, let $K_i = K \cap \Omega_i$ and $K_i^h$ the polygonal approximation of $K_i$ bounded by the sides of $K$ and $\Gamma_K^h$ which is the line segment connecting two intersection points of $\Gamma_K \cap \partial K$. $K_i^h$ is the union of shape regular triangles $K_{ij}^h$, $1 \leq J_i^K \leq 3$, whose sides are the sides of $K_i^h$ and $\Gamma_K^h$. We always set $K_{i1}^h$ the element having $\Gamma_K^h$ as one of its sides. From $K_{ij}^h$ we define the curved element $\widetilde{K}_{ij}^h$ by

$$
\widetilde{K}_{i1}^h = (K_i \cap K_{i1}^h) \cup (K_i \backslash \bar{K}_{i1}^h), \quad \widetilde{K}_{ij}^h = K_i \cap K_{ij}^h, \quad j = 2, \cdots, J_i^K.
$$

Then we know that $K$ is the union of curved triangles $\widetilde{K}_{ij}^h$, $i = 1,2, j = 1, \cdots, J_i^K$.

For any integers $p, q \geq 1$, the space $P^p(K)$ denotes the space of polynomials of degree at most $p$ in $K$ and $Q^{p,q}(K)$ denotes the space of polynomials of degree at most $p$ for the first variable and $q$ for the second variable in $K$. For any $K \in \mathcal{M}^\Gamma$, we define the interface finite element spaces

$$
W_p(K) = \{\varphi : \varphi|_{\widetilde{K}_{ij}^h} \in P^p(\widetilde{K}_{ij}^h), \; i = 1,2, \; j = 1, \cdots, J_i^K\},
$$

and $X_p(K) = W_p(K) \cap H^1(K_1 \cup K_2)$. Notice that the functions in $X_p(K)$ are conforming in each $K_i, i = 1,2$. Now we define the following unfitted finite element spaces

$$
X_p(\mathcal{M}) := \{v \in H^1(\Omega_1 \cup \Omega_2) : v|_K \in X_p(K) \;\; \forall K \in \mathcal{M}^\Gamma,
$$

$$v|_K \in Q^p(K) \quad \forall K \in \mathcal{M} \backslash \mathcal{M}^\Gamma\},$$
$$\mathbf{W}_p(\mathcal{M}) := \{\boldsymbol{\psi} : \boldsymbol{\psi}|_K \in [W_p(K)]^2 \quad \forall K \in \mathcal{M}^\Gamma,$$
$$\boldsymbol{\psi}|_K \in Q^{p-1,p}(K) \times Q^{p,p-1}(K) \quad \forall K \in \mathcal{M} \backslash \mathcal{M}^\Gamma\}.$$

Let $X_p^0(\mathcal{M}) = X_p(\mathcal{M}) \cap H_0^1(\Omega_1 \cup \Omega_2)$, where $H_0^1(\Omega_1 \cup \Omega_2) = \{v \in H^1(\Omega_1 \cup \Omega_2) : v = 0 \text{ on } \partial\Omega\}$.

The semi-discrete unfitted finite element method for solving (6.2) is then to find $(u_h, \mathbf{q}_h) \in X_p^0(\mathcal{M}) \times \mathbf{W}_p(\mathcal{M})$ such that for all $(\varphi_h, \boldsymbol{\psi}_h) \in X_p^0(\mathcal{M}) \times \mathbf{W}_p(\mathcal{M})$,

$$\left(\frac{1}{\rho c^2} \partial_t u_h, \varphi_h\right)_{\mathcal{M}} = -(\mathbf{q}_h, \nabla \varphi_h)_{\mathcal{M}} + \langle \mathbf{q}_h^- \cdot \mathbf{n}, [\![\varphi_h]\!]\rangle_{\mathcal{E}^\Gamma} + (f, \varphi_h)_{\mathcal{M}}, \tag{6.3}$$

$$(\rho \partial_t \mathbf{q}_h, \boldsymbol{\psi}_h)_{\mathcal{M}} = -(u_h, \mathrm{div}\boldsymbol{\psi}_h)_{\mathcal{M}} + \langle u_h^+, [\![\boldsymbol{\psi}_h]\!] \cdot \mathbf{n}\rangle_{\mathcal{E}}, \tag{6.4}$$

$$u_h(\mathbf{x}, 0) = (\mathcal{P}_h u_0)(\mathbf{x}), \quad \mathbf{q}_h(\mathbf{x}, 0) = (\boldsymbol{P}_h \mathbf{q}_0)(\mathbf{x}) \text{ in } \Omega, \tag{6.5}$$

where $\mathcal{P}_h : L^2(\Omega) \to X_p^0(\mathcal{M})$ and $\boldsymbol{P}_h : [L^2(\Omega)]^2 \to \mathbf{W}_p(\mathcal{M})$ are the standard $L^2$ projection operators.

It is shown in [7, Theorem 2.2] that the following energy error of the semi-discrete scheme

$$E_{en}(t) := (\|(u - u_h)(\cdot, t)\|_{L^2(\Omega)}^2 + \|(\mathbf{q} - \mathbf{q}_h)(\cdot, t)\|_{L^2(\Omega)}^2)^{1/2}.$$

has $p$-th order convergence. The semi-discrete problem (6.3)-(6.5) is an ODE system which is solved by the continuous time Galerkin method in this paper. By Theorem 2.1 and Theorem 2.2, we know that the energy error has $O(h^p + \tau^{r+1})$ convergence rate in the norm $\max_{0 \leq t \leq T} E_{en}(t)$ and $O(h^p + \tau^{2r})$ in the $E_{en}(t)$ at nodes $t = t_n$, $n = 1, \ldots, N$.

Note that these two ellipses are close but not tangent. To resolve the interface $\Gamma$ well, we locally refine the mesh near the interface such that the interface deviation $\eta_K \leq \eta_0 = 0.05$ for all $K \in \mathcal{M}^\Gamma$. For the concept of the interface deviation we refer to [7, Definition 2.2], see also Chen et al [6, Definition 2.2]. As an illustration, we show the computational mesh for $h = 1/4$ in Figure 1.

In this example, we test the accuracy of the error in the $\|\cdot\|_{L^\infty(0,T;L^2(\Omega))}$ norm and $E_{en}(T)$. As in Example 1, to test the accuracy at nodes, we set $\tau = \frac{p}{2r}$, thus the wall time of using Algorithm 3.1 for the sequential computation is proportion to $Tr\nu^{\frac{1}{r}}$ which minimizes at $r = \ln \nu$ for $r > 0$, where $\nu = h^{-\frac{p}{2}}$. Table 3 shows the error $E_{en}(T)$ at the terminal time when $r = \lfloor \ln \nu \rfloor + 1$.

To test the accuracy in the $\|\cdot\|_{L^\infty(0,T;L^2(\Omega))}$ norm, we set $\tau = h^{\frac{p}{r+1}}$ and thus $N = T/\tau = T\mu^{\frac{1}{r+1}}$, where $\mu = h^{-p}$. As in Example 1, the wall time of using Algorithm 3.1 and Algorithm 3.2 for the parallel computation is proportion to $N = T\mu^{\frac{1}{r+1}}$ which decreases in $r$. However, since the wave equation is not dissipative, we cannot use Algorithm 4.1 for the sequential computation. In this case, the wall time of using Algorithm 3.1 and Algorithm 3.2 for the sequential computation is proportion to $rN + r^2 N = Tr(r+1)\mu^{\frac{1}{r+1}}$ which is increasing if $\ln \mu \leq 6$ and minimizes at $r^{**} = [-(3 - \ln \mu) + \sqrt{(3 - \ln \mu)^2 - 8}]/4$ if $\ln \mu \geq 6$. Since $r^{**} \geq 1$ is equivalent to $\ln \mu \geq 6$, the optimal choice of the order for minimizing the computation wall time is $r = \max(1, \lfloor r^{**} \rfloor + 1)$. Table 3 shows the error

$$\max_{0 \leq n \leq N-1, 1 \leq k \leq 10} E_{en}(t_n + 0.1k\tau_n)$$

31

as the approximation of $\max_{0 \leq t \leq T} E_{en}(t)$ when $r = \max(1, \lfloor r^{**} \rfloor + 1)$. We clearly observe the optimal $p$-th order convergence and the superior performance of high order methods from Tables 3-4.
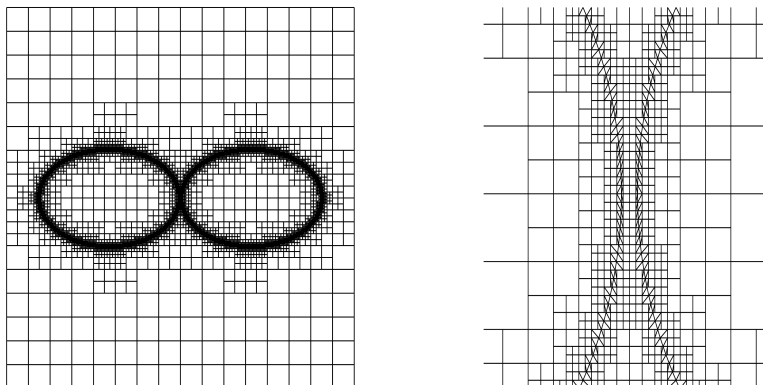


Figure 1: Illustration of the computational domain and the mesh (left) and the corresponding zoomed local mesh (right) with $h = 1/4$ in Example 2.

Table 3: Example 2: numerical errors of $E_{en}(T)$ and orders.

|  | $p = 3$ | | $p = 4$ | | $p = 5$ | |
|---|---|---|---|---|---|---|
| $h$ | error | order | error | order | error | order |
| 1/4 | 4.22E-01 | – | 2.21E-01 | – | 1.25E-01 | – |
| 1/8 | 8.97E-02 | 2.24 | 2.64E-02 | 3.06 | 6.20E-03 | 4.33 |
| 1/16 | 1.32E-02 | 2.76 | 1.81E-03 | 3.87 | 3.82E-05 | 4.96 |
| 1/32 | 1.66E-03 | 2.99 | 1.13E-04 | 4.00 | 1.19E-06 | 5.00 |

Table 4: Example 2: numerical errors of $\max_{0 \leq t \leq T} E_{en}(t)$ and orders.

|  | $p = 3$ | | $p = 4$ | | $p = 5$ | |
|---|---|---|---|---|---|---|
| $h$ | error | order | error | order | error | order |
| 1/4 | 5.23E-01 | – | 2.95E-01 | – | 1.69E-01 | – |
| 1/8 | 1.26E-01 | 2.05 | 3.70E-02 | 3.00 | 8.87E-03 | 4.25 |
| 1/16 | 1.90E-02 | 2.72 | 2.58E-03 | 3.84 | 2.93E-04 | 4.92 |
| 1/32 | 2.45E-03 | 2.96 | 1.66E-04 | 3.95 | 9.21E-06 | 4.99 |

# References

[1] G. Akrivis, C. Makridakis, and R. H. Nochetto, Galerkin and Runge-Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence, *Numer. Math.*, 118: 429–456, 2011.

[2] A.K. Aziz and P. Monk, Continuous finite elements in space and time for the heat equation, *Math. Comp.*, 52: 255-274, 1989.

[3] C. Bernardi and Y. Maday, Spectral methods, *Handbook of Numerical Analysis*, 5: 209–485, 1997.

[4] W.C. Brown, *Matrices over Commutative Rings,* Marcel Dekker, New York, 1993.

[5] J. Chen, Z. Chen, T. Cui, and L. Zhang, An adaptive finite element method for the eddy current model with circuit/field couplings, *SIAM J. Sci. Comput.*, 32: 1020–1042, 2010.

[6] Z. Chen, K. Li, and X. Xiang, An adaptive high-order unfitted finite element method for elliptic interface problems, *Numer. Math.*, 149: 507-548, 2021.

[7] Z. Chen, Y. Liu, and X. Xiang, A high order explicit time finite element method for the acoustic wave equation with discontinuous coefficients, arXiv:2112.02867v2.

[8] Y. Cheng, X. Meng, and Q. Zhang, Application of generalized Gauss-Radau projections for the local discontinuous Galerkin method for linear convection-diffusion equations, *Math. Comp.*, 86: 1233–1267, 2017.

[9] B. Cockburn and C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems, *SIAM J. Numer. Anal.*, 35: 2440–2463, 1998.

[10] J. Douglas Jr., T. Dupont, and M.F. Wheeler, A quasi projection analysis of Galerkin methods for parabolic and hyperbolic equations, *Math. Comp.*, 32: 345-362, 1978.

[11] D. A. French and T. E. Peterson, A continuous space-time finite element method for the wave equation, *Math. Comp.,* 65: 491–506, 1996.

[12] E. Gallopoulos, and Y. Saad, On the parallel solution of parabolic equations, *Proceedings of the 3rd international conference on Supercomputing,* 17–28, 1989.

[13] M.J. Gander, 50 years of time parallel time integration, *in Multiple shooting and time domain decomposition,* T. Carraro, M. Geiger, S. Körkel, and R. Rannacher, eds., Springer-Verlag, Berlin, 2015, pp. 69–114.

[14] G.H. Golub and C.F. Van Loan, *Matrix Computations,* third edition, The John Hopkins University Press, Baltimore, Maryland, 1996

[15] R. Griesmaier, and P. Monk, Discretization of the Wave Equation Using Continuous Elements in Time and a Hybridizable Discontinuous Galerkin Method in Space, *J. Sci. Comput.,* 58: 472–498, 2014.

[16] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II,* Springer-Verlag, Berlin, Heidelberg, 1996.

[17] B.L. Hulme, One-step piecewise polynomial Galerkin methods for initial value problems, *Math. Comp.*, 26: 415–426, 1972.

[18] O. Perron, *Algebra, Volume II*, Guschens Bücherei, Berlin, 1933.

[19] T. Richter, A. Springer, and Vexler, Efficient numerical realization of discontinuous Galerkin methods for temporal discretization of parabolic problems, *Numer. Math.*, 124: 151–182, 2013.

[20] E. B. Saff and R. S. Varga, On the zeros and Poles of Padé Approximation to $e^z$, *Numer. Math.*, 25: 1–14, 1975.

[21] D. Schötzau and Ch. Schwab, *p- and hp- Finite Element Methods,* Oxford Science Publications, New York, 1998.

[22] D. Schötzau and Ch. Schwab, The time discretization of parabolic problems by the hp-version of the discontinuous Galerkin finite element method, *SIAM J. Numer. Anal.*, 38: 837–875, 2000.

[23] B.N. Southworth, O. Krzysik, W. Pazner, and H. De Sterck, Fast solution of fully implicit Runge-Kutta and discontinuous Galerkin in time for numerical PDEs, Part I: The linear setting, *SIAM J. Sci. Comput.*, 44: A416–A443, 2022.

[24] G. Szegö, *Orthogonal Polynomials*, American Mathematical Society, New York, 1939.

[25] Y. Wu and Y. Bai, Error analysis of energy-preserving mixed finite element methods for the Hodge wave equation, *SIAM J. Numer. Anal.*, 59: 1433–1454, 2021.