

FAST EVALUATION OF ARTIFICIAL BOUNDARY CONDITIONS FOR ADVECTION DIFFUSION EQUATIONS*

TING SUN[†], JILU WANG[‡], AND CHUNXIONG ZHENG[§]

Abstract. An artificial boundary method is developed for solving the one-dimensional advection diffusion equation in the real line. In order to construct a fully discrete fast numerical algorithm with rigorous error analysis, we start with the two-step backward difference formula for time discretization of the advection diffusion equation in the whole real line. Then, we use the discrete analogue of the Laplace transform to derive a second-order time-stepping scheme in a bounded domain equipped with a discrete artificial boundary condition (ABC). The Galerkin finite element method is used for spatial discretization. To expedite the evaluation of time convolution involved in the discrete ABC, we propose a fast algorithm based on the best rational approximation of square root function in subdomains of the complex plane. An estimate for this best rational approximation enables us to prove optimal-order convergence of the fully discrete numerical scheme (integrating the fast approximation algorithm). Several numerical examples are provided to illustrate the convergence of numerical solutions and the effectiveness of the proposed fast approximation algorithm.

Key words. artificial boundary method, fast algorithm, advection diffusion equation, rational approximation, error estimates

AMS subject classifications. 65M12, 65D30, 65N15, 65Y20

DOI. 10.1137/19M130145X

1. Introduction. Advection diffusion equations can be used to describe numerous natural and social phenomena, ranging from semiconductor simulation to financial modeling; see [28]. Among these applications, the evolution equations defined in unbounded domains are frequently used. In this paper, we are interested in the Cauchy problem of the following one-dimensional advection diffusion equation:

$$(1.1) \quad \partial_t u + \mathcal{L}u = f \quad \forall x \in \mathbb{R}, \quad \forall t > 0,$$

$$(1.2) \quad u(x, 0) = \phi(x) \quad \forall x \in \mathbb{R},$$

$$(1.3) \quad \lim_{x \rightarrow \pm\infty} u(x, t) = 0 \quad \forall t > 0,$$

where $\mathcal{L} = 2a\partial_x - \sigma\partial_x^2$ denotes the advection diffusion operator, $2a$ the advection velocity, and σ the viscosity coefficient. We assume that both the initial function ϕ and the source function f are spatially compactly supported in a bounded interval $\Omega := (x_-, x_+)$.

From the numerical point of view, many difficulties arise due to the unboundedness of the computational domain, since a direct discretization generally leads to

*Received by the editors November 21, 2019; accepted for publication (in revised form) September 18, 2020; published electronically December 18, 2020.

<https://doi.org/10.1137/19M130145X>

Funding: The work of the second author was partially supported by National Natural Science Foundation of China grant U1930402. The work of the third author was partially supported by Natural Science Foundation of Xinjiang Autonomous Region grant 2019D01C026 and National Natural Science Foundation of China grant 11771248.

[†]College of Mathematics and System Sciences, Xinjiang University, Urumqi 830046, People's Republic of China (suntsing19@163.com).

[‡]Corresponding author. Beijing Computational Science Research Center, Beijing 100193, People's Republic of China (jiluwang@csrc.ac.cn).

[§]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China; College of Mathematics and Systems Science, Xinjiang University, Urumqi 830046, People's Republic of China (czheng@mail.tsinghua.edu.cn).

an algebraic system involving an infinite number of degrees of freedom. In the past several decades, many efforts have been made to overcome these difficulties. One of the most popular approaches is the artificial boundary method, which has been extensively investigated in various disciplines; see the monograph [19] and the review papers [12, 15, 34]. In this method, the key step is the construction of suitable exact artificial/absorbing boundary conditions (ABCs), which are specified at some artificial boundaries, such that the solution of the problem in a corresponding bounded domain is equal to the original solution in the unbounded domain restricted to this bounded domain. For time-dependent problems, exact ABCs are usually nonlocal operators containing some temporal convolutions, and frequently, one needs to approximate and localize this nonlocal relation to reduce the computational cost.

There are two approaches to design approximate and localized ABCs. One approach is to directly approximate the exact ABCs for the continuous problem [7, 27], reducing these exact ABCs to localized versions. In [17], Halpern derived exact ABCs for a linear advection diffusion equation with small viscosity and approximated them by using generalized continued fractions, which leads to well-posed initial boundary value problems and produces errors that are powers of the viscosity. The relevant results have been generalized to linear advection diffusion equations with variable advection coefficient in [25]. Later, Halpern and Rauch [18] constructed and analyzed ABCs for diffusion equations with variable coefficients, curved artificial boundary, and arbitrary convection. They discretized the exact ABCs by using the geometric identification of the Dirichlet to Neumann map and the rational interpolation of square root function in the complex plane. The application of rational approximations to design ABCs for a model parabolic system was also analyzed by Hagstrom [14], where the model was chosen to display many features of the linearized compressible Navier-Stokes equations. The ABCs designed in these works are discretizations of the exact Dirichlet to Neumann map of the continuous problem in bounded domains. Jiang and Greengard [21] developed a fast algorithm for the evaluation of the exact ABCs for the Schrödinger equation, in which a Neumann to Dirichlet (NtD) integral with a kernel proportional to $1/\sqrt{t}$ is contained in the boundary conditions. The nonlocal convolution integral was approximated by using a linear combination of exponents and can be evaluated recursively. The fast algorithm in [21] required $\mathcal{O}(N \log N)$ floating-point operations and $\mathcal{O}(\log N)$ storage, where $N = T/\tau$ with T denoting the length of the time interval and τ the time step size.

Another approach is to derive exact boundary conditions for the semi- or fully discretized numerical scheme in an unbounded domain [2, 5, 26, 38]. These ABCs often contain discrete convolutions in time, while direct evaluation of these discrete convolutions is time-consuming. Therefore, it is important to develop fast algorithms with less memory requirements to reduce the computational cost. Usually, fast algorithms are carried out by utilizing the summation of exponentials to approximate the convolution kernel. The summation can be achieved, for example, by the quadrature approximation of the kernel function in the time domain [5, 30, 24, 31], the direct rational approximation of kernel symbols [1, 6, 9, 20], or the quadrature approximation of contour integrals in the Laplace domain [23, 26]. In [9], Druskin, Guettel and Knizhnerman investigated the indefinite Helmholtz problem in the half space of \mathbb{R}^{n+1} . In their works, the half space was split into $x_1 \in [0, \infty)$ and $y \in \mathbb{R}^n$. The problem was first discretized in the y -variable and then an ABC was derived, which expresses as an NtD mapping. Since an operator $A^{-1/2}$ is involved in the ABC, the rational approximation of $x^{-1/2}$ was studied. The classical Zolotarev rational approximation for $x^{-1/2}$ only holds for positive intervals. Considering the spectrum of A contains a

union of positive and negative intervals, an extension of the Zolotarev approximation was developed in [9]. The resulting NtD mapping can be expressed into a form of Stieltjes continued fraction. By introducing suitable new unknowns, the NtD ABC is equivalent to a set of finite difference relations, which were linked with the spatial finite difference discretization of the continuous PML equation. The same strategy was applied earlier to deal with the Laplace equation on a semi-infinite strip and the exterior hyperbolic problem in [20] and [6], respectively, where the difference lies in the spectrum of the operator involved in the NtD mapping. In [16], Hagstrom and Warburton derived the complete radiation boundary conditions for wave equations in the half space by using the Laplace transform. The inverse Laplace transform is performed on the complex contour $1/T_0 + i\mathbb{R}$ associated with the evolution time T_0 . Thus, their rational approximation is dependent on T_0 and the contour. There are many more works on the fast evaluation of ABCs. Interested readers are referred to [3, 4, 10, 11, 22, 32, 35, 36, 37] for more details. It is known that an inappropriate design of fast computation of the time convolution may lead to degeneracy of the convergence rate and cause difficulties in the error analysis of the overall numerical schemes. As a result, it is always an interesting and important task to design fast computational schemes whose error estimates can be proved to be optimal.

In this paper, we develop a fast and stable algorithm with rigorous analysis for solving advection diffusion equations in unbounded domains based on the second approach, which makes stability and error analysis easier for our problem. In most of the previous works, the stability and error estimates of the overall numerical schemes, combining the time-stepping methods for PDEs and fast algorithms for nonlocal convolution integrals, are hard to analyze if the continuous convolution are directly discretized. Considering this point, instead of directly approximating the ABCs of the continuous problem, we consider the two-step backward difference formula (BDF) for time discretization of problem (1.1)–(1.3) in the real line, and then derive an exact ABC for the time-discrete problem by using the \mathcal{Z} -transform. The Galerkin finite element method is used for spatial discretization. Then, we propose a fast algorithm to solve the fully discrete scheme accurately and efficiently based on a best rational approximation theory. More precisely, inspired by the work in [8], we extend the best approximation theory of the square root function in the real axis to the complex plane and prove convergence of such approximations in suitable subdomains of the complex plane. Applying these best rational approximations, we develop a fast algorithm to efficiently evaluate the convolution operator involved in the discrete ABCs with less memory requirements. If T is the length of the time interval and $\tau = T/N$ is the time step size, the fast algorithm reduces the computational cost of evaluating the discrete temporal convolution from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \ln N)$. Furthermore, we present a rigorous error analysis of the fast algorithm and establish optimal-order error estimates of the fully discrete numerical scheme for problem (1.1)–(1.3).

2. Construction of exact ABCs. The Cauchy problem (1.1)–(1.3) is defined in the infinite spatial domain \mathbb{R} . To construct the numerical solutions, it is a common practice to truncate problem (1.1)–(1.3) into the bounded domain Ω . To ensure the well-posedness of the truncated bounded domain problem, suitable ABCs should be imposed at the two artificial boundary points x_{\pm} . This can be achieved by analyzing the homogeneous residual exterior problem, as illustrated in the following.

Let $\Omega_- = (-\infty, x_-)$ and $\Omega_+ = (x_+, +\infty)$. We first consider the exterior problem on the semi-infinite interval Ω_+ :

$$(2.1) \quad \partial_t u + \mathcal{L}u = 0 \quad \forall x \in \Omega_+, \forall t > 0,$$

$$(2.2) \quad u(x, 0) = 0 \quad \forall x \in \Omega_+,$$

$$(2.3) \quad \lim_{x \rightarrow +\infty} u(x, t) = 0 \quad \forall t > 0.$$

The Laplace transform of the system (2.1)–(2.3) in time yields

$$(2.4) \quad s\widehat{u} + \mathcal{L}\widehat{u} = 0 \quad \forall x \in \Omega_+, \forall s \in \mathbb{C}_+,$$

$$(2.5) \quad \lim_{x \rightarrow +\infty} \widehat{u}(x, s) = 0 \quad \forall s \in \mathbb{C}_+,$$

where \mathbb{C}_+ denotes the right complex half plane, i.e., $\mathbb{C}_+ = \{s \in \mathbb{C} : \Re s > 0\}$. Parametrized by the complex argument s , the general solution of the second-order ordinary differential equation (2.4) is

$$\widehat{u}(x, s) = c_1(s) \exp\left(\frac{a + \sqrt{\sigma s + a^2}}{\sigma} x\right) + c_2(s) \exp\left(\frac{a - \sqrt{\sigma s + a^2}}{\sigma} x\right).$$

Due to the boundary condition (2.5) at infinity, we can see that $c_1(s) = 0$. Differentiating the above equation, we derive

$$\partial_x \widehat{u}(x, s) = \frac{a - \sqrt{\sigma s + a^2}}{\sigma} \widehat{u}(x, s) \quad \forall x \in \Omega_+, \forall s \in \mathbb{C}_+.$$

Thus, taking the inverse Laplace transform of the above equality results in an exact ABC at the right boundary point x_+ :

$$(2.6) \quad \sigma \partial_x u(x_+, t) - au(x_+, t) + \sqrt{\sigma \partial_t + a^2} u(x_+, t) = 0 \quad \forall t > 0,$$

where $\sqrt{\sigma \partial_t + a^2}$ denotes the multiplier operator (in time) associated with the symbol $\sqrt{\sigma s + a^2}$, namely,

$$\sqrt{\sigma \partial_t + a^2} v(t) := \mathcal{L}_s^{-1}[\sqrt{\sigma s + a^2} \widehat{v}(s)](t) \quad \forall t > 0,$$

with \mathcal{L}_s^{-1} denoting the inverse Laplace transform with respect to the s -variable.

An exact ABC also can be derived at the left boundary point x_- :

$$(2.7) \quad \sigma \partial_x u(x_-, t) - au(x_-, t) - \sqrt{\sigma \partial_t + a^2} u(x_-, t) = 0 \quad \forall t > 0.$$

Consequently, confined into Ω , the solution of the Cauchy problem (1.1)–(1.3) coincides with the solution of the following initial boundary value problem:

$$(2.8) \quad \partial_t u + \mathcal{L}u = f \quad \forall x \in \Omega, \forall t > 0,$$

$$(2.9) \quad u(x, 0) = \phi(x) \quad \forall x \in \Omega,$$

$$(2.10) \quad \sigma \partial_n u(x_\pm, t) \mp au(x_\pm, t) + \sqrt{\sigma \partial_t + a^2} u(x_\pm, t) = 0 \quad \forall t > 0.$$

Here, the symbol ∂_n denotes the outward normal derivative at the boundary points x_\pm .

3. Numerical schemes with discrete ABCs. In this section, we construct a fully discrete numerical method for solving problem (1.1)–(1.3). Instead of directly approximating the truncated finite domain problem (2.8)–(2.10) with exact ABCs, we first construct a BDF2 scheme for problem (1.1)–(1.3), with which and using the \mathcal{Z} -transform, a second-order scheme equipped with discrete ABCs on the bounded domain Ω is obtained. Then, we present a fully discrete numerical scheme by using a Galerkin finite element method in spatial discretization.

We first introduce the \mathcal{Z} -transform in the following subsection.

3.1. \mathcal{Z} -transform of a sequence of functions. Given a Hilbert space \mathcal{H} , the space $\ell^2(\mathcal{H})$ of semi-infinite sequences is defined by

$$\ell^2(\mathcal{H}) = \left\{ g = \{g^n\}_{n=0}^\infty : g^n \in \mathcal{H}, \|g\|_{\ell^2(\mathcal{H})} = \left(\sum_{n=0}^\infty \|g^n\|_{\mathcal{H}}^2 \right)^{\frac{1}{2}} < \infty \right\}.$$

We introduce the subspace $\ell_0^2(\mathcal{H}) = \{g = \{g^n\}_{n=0}^\infty \in \ell^2(\mathcal{H}) : g^0 = 0\}$. The linear space $\ell^2(\mathcal{H})$ is a Hilbert space with the inner product

$$(f, g)_{\ell^2(\mathcal{H})} \equiv \sum_{n=0}^\infty (f^n, g^n)_{\mathcal{H}} \quad \forall f, g \in \ell^2(\mathcal{H}).$$

For any $g = \{g^n\}_{n=0}^\infty \in \ell^2(\mathcal{H})$, its \mathcal{Z} -transform is defined as $\tilde{g}(z) = \sum_{n=0}^\infty g^n z^n$, which is an \mathcal{H} -valued function, holomorphic in the unit disk \mathbb{D} . The limit $\tilde{g}(z) = \lim_{r \nearrow 1} \tilde{g}(rz)$ exists in $L^2(\partial\mathbb{D}; \mathcal{H})$, and the following Parseval identity holds:

$$(3.1) \quad (f, g)_{\ell^2(\mathcal{H})} = \int_{\partial\mathbb{D}} (\tilde{f}(z), \tilde{g}(z))_{\mathcal{H}} \mu(dz) \quad \forall f, g \in \ell^2(\mathcal{H}),$$

where $\mu(dz) = \frac{1}{2\pi} d\theta$ is defined through the change of variable $z = e^{i\theta}$ with $\theta \in [-\pi, \pi]$.

3.2. The BDF2 time discretization with discrete ABCs. Let T denote the length of the time interval, and let $t_n = n\tau$, $n = 0, 1, 2, \dots, N$, with $\tau = \frac{T}{N}$ denoting the step size for time discretization. We discretize the Cauchy problem (1.1)–(1.3) by the standard BDF2 scheme:

$$(3.2) \quad \frac{3U^n - 4U^{n-1} + U^{n-2}}{2\tau} + \mathcal{L}U^n = f^n \quad \forall x \in \mathbb{R},$$

$$(3.3) \quad \lim_{x \rightarrow \pm\infty} U^n(x) = 0,$$

for $n \geq 2$, where $U^n(x) \approx u(x, t_n)$ and $f^n(x) = f(x, t_n)$. The starting approximations are set as

$$(3.4) \quad U^0 = \phi,$$

$$(3.5) \quad U^1 = \psi := \phi + \tau(f^0 - \mathcal{L}\phi),$$

for all $x \in \mathbb{R}$, where U^1 is obtained by the backward Euler method. Note that both U^0 and U^1 are compactly supported into the bounded domain Ω .

Similarly as in section 2, we confine the above problem into the exterior domain Ω_+ . Thus, the semidiscrete problem (3.2)–(3.5) reduces to

$$(3.6) \quad \frac{3U^n - 4U^{n-1} + U^{n-2}}{2\tau} + \mathcal{L}U^n = 0 \quad \forall x \in \Omega_+,$$

$$(3.7) \quad U^0(x) = U^1(x) = 0 \quad \forall x \in \Omega_+,$$

$$(3.8) \quad \lim_{x \rightarrow +\infty} U^n(x) = 0$$

for $n \geq 2$. With the notation introduced in section 3.1, let $\tilde{U}(x, z)$ denote the \mathcal{Z} -transform of the sequence $\{U^n(x)\}_{n=0}^{+\infty}$. Applying the \mathcal{Z} -transform to the above system, we obtain

$$\begin{aligned} \frac{3 - 4z + z^2}{2\tau} \tilde{U}(x, z) + \mathcal{L}\tilde{U}(x, z) &= 0 & \forall x \in \Omega_+, \forall z \in \mathbb{D}, \\ \lim_{x \rightarrow +\infty} \tilde{U}(x, z) &= 0 & \forall z \in \mathbb{D}. \end{aligned}$$

Clearly, the general solution of the above second-order ordinary differential equation is

$$\begin{aligned} \tilde{U}(x, z) &= c_1^+(z) \exp\left(\frac{a + \sqrt{a^2 + \frac{\sigma}{2\tau}(3 - 4z + z^2)}}{\sigma} x\right) \\ &\quad + c_2^+(z) \exp\left(\frac{a - \sqrt{a^2 + \frac{\sigma}{2\tau}(3 - 4z + z^2)}}{\sigma} x\right). \end{aligned}$$

Since $\lim_{x \rightarrow +\infty} \tilde{U}(x, z) = 0$, we have $c_1^+(z) = 0$. By differentiating the above equation, we obtain

$$(3.9) \quad \partial_x \tilde{U}(x, z) = \frac{a - \sqrt{a^2 + \frac{\sigma}{2\tau}(3 - 4z + z^2)}}{\sigma} \tilde{U}(x, z) \quad \forall x \in \Omega_+, \forall z \in \mathbb{D}.$$

Note that $\sqrt{a^2 + \frac{\sigma}{2\tau}(3 - 4z + z^2)}$ is analytic at $z = 0$. Then, it has a power series expansion:

$$(3.10) \quad \tilde{J}(z) := \sqrt{a^2 + \frac{\sigma}{2\tau}(3 - 4z + z^2)} = \sum_{j=0}^{\infty} \lambda_j z^j \quad \forall z \in \mathbb{D}.$$

With the above result and noting that $\tilde{U}(x, z) = \sum_{n=0}^{\infty} U^n(x) z^n$, we obtain an exact ABC from (3.9) at the right artificial boundary point $x = x_+$:

$$(3.11) \quad \sigma \partial_x U^n(x_+) = a U^n(x_+) - \sum_{j=0}^n \lambda_j U^{n-j}(x_+)$$

for $n \geq 2$. Applying the same technique, we can also get an exact ABC at the left artificial boundary point $x = x_-$:

$$(3.12) \quad \sigma \partial_x U^n(x_-) = a U^n(x_-) + \sum_{j=0}^n \lambda_j U^{n-j}(x_-)$$

for $n \geq 2$.

Hence, instead of directly discretizing the bounded domain problem (2.8)–(2.10), we approximate problem (1.1)–(1.3) in time and obtain a BDF2 scheme with discrete ABCs in Ω :

$$(3.13) \quad \frac{3U^n - 4U^{n-1} + U^{n-2}}{2\tau} + \mathcal{L}U^n = f^n \quad \forall x \in \Omega, n \geq 2,$$

$$(3.14) \quad U^0(x) = \phi(x), \quad U^1(x) = \psi(x) \quad \forall x \in \Omega,$$

$$(3.15) \quad \sigma \partial_n U^n(x_{\pm}) \mp a U^n(x_{\pm}) + (\mathcal{J} * U)^n(x_{\pm}) = 0 \quad n \geq 2,$$

where $\mathcal{J}*$ is the convolution operator corresponding to the symbol $\tilde{J}(z)$, namely, for any sequence $v = \{v^n\}_{n=0}^{+\infty}$, we define

$$(3.16) \quad (\mathcal{J} * v)^n := \sum_{j=0}^n \lambda_j v^{n-j}.$$

Furthermore, we use a similar notation to define

$$(3.17) \quad (\mathcal{J} * v)(t_n) := \sum_{j=0}^n \lambda_j v(t_{n-j})$$

for a time-dependent function $v(t)$.

In the following two lemmas, we prove stability of the convolution operator and give an error estimate for $\mathcal{J}*$. These results will be used in the next sections. For simplicity of notation, we denote by C a generic positive constant, which is independent of n , the time step size τ , and the spatial mesh size h .

3.3. Stability and error estimate of time discretization.

LEMMA 3.1. *The discrete convolution operator $\mathcal{J}*$ is stable, in the sense that for any fixed $n > 0$ and $v = \{v^m\}_{m=0}^n$, it holds that*

$$(3.18) \quad \sum_{m=0}^n (\mathcal{J} * v)^m v^m \geq |a| \sum_{m=0}^n (v^m)^2.$$

Proof. For any fixed $n > 0$, we assume $v^m = 0$ for all $m > n$. Then, we have

$$\begin{aligned} \sum_{m=0}^n (\mathcal{J} * v)^m v^m &= \sum_{m=0}^{+\infty} (\mathcal{J} * v)^m v^m = \Re \int_{\partial\mathbb{D}} \widetilde{\mathcal{J} * v}(z) \overline{\tilde{v}(z)} \mu(dz) \\ &= \int_{\partial\mathbb{D}} \Re \widetilde{\mathcal{J}}(z) |\tilde{v}(z)|^2 \mu(dz) \\ &\geq |a| \int_{\partial\mathbb{D}} |\tilde{v}(z)|^2 \mu(dz) = |a| \sum_{m=0}^{+\infty} (v^m)^2 = |a| \sum_{m=0}^n (v^m)^2. \end{aligned}$$

This finishes the proof. \square

LEMMA 3.2. *Let $v = v(t)$ be a smooth function for $t \in [0, T]$ with $v \in H^4(0, T)$ and $v(0) = v'(0) = v''(0) = v'''(0) = 0$. Then*

$$(3.19) \quad |(\mathcal{J} * v)(t_n) - \sqrt{\sigma \partial_t + a^2} v(t_n)| \leq C \sigma \tau^2,$$

where the constant C is independent of τ .

Proof. For $v \in H^4(0, T)$, there exists an extension $v^* \in H^4(\mathbb{R}_+)$ of v such that $v^*(t) = v(t)$ for $t \in [0, T]$ and $\|v^*\|_{H^4(\mathbb{R}_+)} \leq C \|v\|_{H^4(0, T)}$ [29, Theorem 5, p. 181]. Since $v(0) = v'(0) = v''(0) = v'''(0) = 0$, we further extend $v^*(t)$ to be zero on $t \in (-\infty, 0]$ and obtain a function $v^*(t) \in H^4(\mathbb{R})$. Then, we define

$$(3.20) \quad (\mathcal{J} * v^*)(t) := \sum_{j=0}^{\infty} \lambda_j v^*(t - j\tau) \quad \forall t \in \mathbb{R},$$

which is consistent with the definition in (3.16) and (3.17). Taking the Fourier transform of (3.20) in time yields

$$\begin{aligned} \mathcal{F}[(\mathcal{J} * v^*)(t)](\xi) &= \int_{\mathbb{R}} (\mathcal{J} * v^*)(t) e^{-it\xi} dt \\ &= \sum_{j=0}^{\infty} \int_{\mathbb{R}} \lambda_j v^*(t - j\tau) e^{-it\xi} dt \\ &= \widetilde{\mathcal{J}}(e^{-i\tau\xi}) \mathcal{F}v^*(\xi) \end{aligned}$$

$$\begin{aligned}
&= \sqrt{a^2 + \sigma i \xi} \mathcal{F}^* v(\xi) + \left(\tilde{J}(e^{-i\tau\xi}) - \sqrt{a^2 + \sigma i \xi} \right) \mathcal{F} v^*(\xi) \\
&= \mathcal{F} \left[\sqrt{a^2 + \sigma \partial_t} v^*(t) \right] (\xi) + \left(\tilde{J}(e^{-i\tau\xi}) - \sqrt{a^2 + \sigma i \xi} \right) \mathcal{F} v^*(\xi).
\end{aligned}$$

By using Taylor expansion, we can easily get

$$|\tilde{J}(e^{-i\tau\xi}) - \sqrt{a^2 + \sigma i \xi}| \leq C\sigma\tau^2 |\xi|^3.$$

With the above estimates, we have

$$\begin{aligned}
|(\mathcal{J}^* v^*)(t) - \sqrt{a^2 + \sigma \partial_t} v^*(t)| &= \left| \mathcal{F}^{-1} \left[(\tilde{J}(e^{-i\tau\xi}) - \sqrt{a^2 + \sigma i \xi}) \mathcal{F} v^*(\xi) \right] (t) \right| \\
&\leq \int_{\mathbb{R}} \left| \tilde{J}(e^{-i\tau\xi}) - \sqrt{a^2 + \sigma i \xi} \right| |\mathcal{F} v^*(\xi)| d\xi \\
&\leq C\sigma\tau^2 \int_{\mathbb{R}} |\xi|^3 |\mathcal{F} v^*(\xi)| d\xi \\
&\leq C\sigma\tau^2 \int_{\mathbb{R}} \frac{1}{1+|\xi|} (1+|\xi|^4) |\mathcal{F} v^*(\xi)| d\xi \\
&\leq C\sigma\tau^2 \left(\int_{\mathbb{R}} \frac{1}{(1+|\xi|)^2} d\xi \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}} (1+|\xi|^4)^2 |\mathcal{F} v^*(\xi)|^2 d\xi \right)^{\frac{1}{2}} \\
&\leq C\sigma\tau^2 \left(\int_0^\infty |v^*(t)| + |\partial_t^4 v^*(t)|^2 dt \right)^{\frac{1}{2}} \\
&\leq C\sigma\tau^2 \left(\int_0^T |v(t)| + |\partial_t^4 v(t)|^2 dt \right)^{\frac{1}{2}}
\end{aligned}$$

for $t \in [0, T]$. Thus, taking $t = t_n$ in the above estimate and noting that $v^*(t_n) = v(t_n)$ results in the inequality (3.19). This completes the proof of Lemma 3.2. \square

3.4. Spatial discretization. In this section, we discretize the BDF2 scheme (3.13)–(3.15) by the Galerkin finite element method to obtain a fully discrete scheme. Let M be a positive integer and $h = (x_+ - x_-)/M$ the mesh size. Let S_h^r denote the finite element space of continuous piecewise polynomials of degree at most $r \geq 1$. Over the finite element space S_h^r , we define the Ritz projection $R_h : H^1(\Omega) \rightarrow S_h^r$ by

$$(\partial_x(R_h\varphi - \varphi), \partial_x v_h) = 0 \quad \forall v_h \in S_h^r$$

with $\int_{\Omega} (\varphi - R_h\varphi) dx = 0$. It is well known that the Ritz projection satisfies the following standard error estimates [33]:

(3.21)

$$\|R_h\varphi - \varphi\|_{L^2} + h\|R_h\varphi - \varphi\|_{H^1} \leq Ch^{r+1}\|\varphi\|_{H^{r+1}(\Omega)} \quad \forall \varphi \in H^{r+1}(\Omega),$$

(3.22)

$$\|R_h\varphi - \varphi\|_{L^\infty} + h\|R_h\varphi - \varphi\|_{W^{1,\infty}} \leq C\ell_h h^{r+1}\|\varphi\|_{W^{r+1,\infty}(\Omega)} \quad \forall \varphi \in W^{r+1,\infty}(\Omega),$$

where

$$(3.23) \quad \ell_h = \begin{cases} 1 + |\ln h| & \text{if } r = 1, \\ 1 & \text{if } r \geq 2. \end{cases}$$

The following inverse inequality will be frequently used in this paper:

$$(3.24) \quad \|v_h\|_{H^1} \leq Ch^{-1}\|v_h\|_{L^2} \quad \forall v_h \in S_h^r.$$

With the above notation, the finite element approximation of the BDF2 scheme (3.13)–(3.15) is to find $u_h^n \in S_h^r$ such that

$$(3.25) \quad \left(\frac{3u_h^n - 4u_h^{n-1} + u_h^{n-2}}{2\tau}, v_h \right) + A(u_h^n, v_h) + \left[(\mathcal{J} * u_h)^n(x_{\pm}) \right] v_h(x_{\pm}) = (f^n, v_h)$$

for all $v_h \in S_h^r$ and $n = 2, 3, \dots, N$, with $u_h^0 = \Pi_h \phi$ and $u_h^1 = \Pi_h \psi$, where

$$(3.26) \quad A(u, v) := (\sigma \partial_x u, \partial_x v) + (a \partial_x u, v) - (au, \partial_x v)$$

and Π_h denotes the standard Lagrange interpolation operator; the analogue of estimate (3.21) also holds for Π_h . By convention, we define

$$[(\mathcal{J} * u_h)^n(x_{\pm})] v_h(x_{\pm}) := [(\mathcal{J} * u_h)^n(x_+)] v_h(x_+) + [(\mathcal{J} * u_h)^n(x_-)] v_h(x_-).$$

In the following, we will use the same expression for simplicity of notation.

In the numerical implementation of (3.25), one needs to evaluate two discrete convolutions at each time step and solve an algebraic system with constant coefficient matrix of bandwidth $2r + 1$. Even though the algebraic system can be solved within linear complexity, the convolution operations would be very costly for a large number of time steps, in both memory and computation. In the next section, we intend to develop an approximation method for the convolution operator such that both the computational cost at a single step and the total memory cost will be essentially independent of the number of time steps.

4. Fast convolution approximation. In this section, we present an efficient algorithm for approximating the convolution quadrature operator $\mathcal{J} *$ by using a best rational approximation of the square root function.

4.1. Implementation algorithm. Recalling the definition (3.16) of the convolution quadrature operator $\mathcal{J} *$ and (3.10), we can easily get

$$(4.1) \quad (\mathcal{J} * v)^n = \mathcal{Z}^{-1} \left\{ \sqrt{a^2 + \frac{\sigma}{2\tau}(3 - 4z + z^2)} \tilde{v}(z) \right\} = \mathcal{Z}^{-1} \left\{ \sqrt{s(z)} \tilde{v}(z) \right\},$$

where $\tilde{v}(z) = \sum_{n=0}^{\infty} v^n z^n$ denotes the \mathcal{Z} -transform for any $v = \{v^n\}_{n=0}^{\infty}$ and \mathcal{Z}^{-1} denotes the inverse \mathcal{Z} -transform, and

$$(4.2) \quad s(z) = C_1 + C_2 z + C_3 z^2, \quad C_1 = a^2 + \frac{3\sigma}{2\tau}, \quad C_2 = -\frac{2\sigma}{\tau}, \quad C_3 = \frac{\sigma}{2\tau}.$$

The basic idea of fast approximation for $\mathcal{J} *$ is to efficiently approximate the square root function \sqrt{s} by the following type of rational function:

$$(4.3) \quad \Phi(s) = \sum_{m=1}^p \frac{w_m}{s - q_m} + w_{p+1}s + w_{p+2}$$

with $q_m < 0$ for all $m = 1, 2, \dots, p$, where q_m denote the poles of the rational function. Details on the procedure to construct such rational approximations can be found in subsections 4.2 and 4.3. Here, we first explain how we reduce the computational cost of the convolution operator by using the above rational approximation. For this purpose, we approximate $\mathcal{J} *$ by the convolution operator $\mathcal{K} *$ defined by

$$(4.4) \quad (\mathcal{K} * v)^n := \mathcal{Z}^{-1} \{ \Phi(s(z)) \tilde{v}(z) \}.$$

Next, we will show that the new convolution operator $\mathcal{K}*$ can be evaluated efficiently. To this end, we introduce the functions

$$(4.5) \quad \tilde{v}_m(z) := \frac{w_m \tilde{v}(z)}{s(z) - q_m}$$

for all $m = 1, 2, \dots, p$. The inverse \mathcal{Z} -transform of $\tilde{v}_m(z)$ is denoted by $\{v_m^n\}_{n=0}^{+\infty}$. Together with (4.3), we can rewrite (4.4) as

$$(4.6) \quad \begin{aligned} (\mathcal{K}*_v)^n &= \mathcal{Z}^{-1} \left\{ \sum_{m=1}^p \frac{w_m \tilde{v}(z)}{s(z) - q_m} + w_{p+1} s(z) \tilde{v}(z) + w_{p+2} \tilde{v}(z) \right\} \\ &= \mathcal{Z}^{-1} \left\{ \sum_{m=1}^p \tilde{v}_m(z) + w_{p+1} s(z) \tilde{v}(z) + w_{p+2} \tilde{v}(z) \right\} \\ &= \sum_{m=1}^p v_m^n + w_{p+1} (C_1 v^n + C_2 v^{n-1} + C_3 v^{n-2}) + w_{p+2} v^n. \end{aligned}$$

In view of (4.5), we obtain

$$(C_1 - q_m) v_m^n + C_2 v_m^{n-1} + C_3 v_m^{n-2} = w_m v^n,$$

which leads to

$$(4.7) \quad v_m^n = \frac{w_m v^n - C_2 v_m^{n-1} - C_3 v_m^{n-2}}{C_1 - q_m}$$

for $n = 0, 1, 2, \dots, N$ and $m = 1, 2, \dots, p$, with $v_m^0 = v_m^1 = 0$. Substituting the above result into (4.6), we have

$$(4.8) \quad \begin{aligned} (\mathcal{K}*_v)^n &= \left(w_{p+1} C_1 + w_{p+2} + \sum_{m=1}^p \frac{w_m}{C_1 - q_m} \right) v^n + w_{p+1} (C_2 v^{n-1} + C_3 v^{n-2}) \\ &\quad - \sum_{m=1}^p \frac{C_2 v_m^{n-1} + C_3 v_m^{n-2}}{C_1 - q_m}. \end{aligned}$$

Clearly, to evaluate the convolution $(\mathcal{K}*_v)^n$ recursively, only two history elements of the sequence $\{v^n\}_{n=0}^{+\infty}$ are required to be stored and the computational cost is $\mathcal{O}(p)$ at each time step. If p is relatively small, both the memory cost and the computational complexity are significantly reduced, especially for large number of time steps. In the following sections, as analyzed in Lemma 4.1, Corollary 4.5, section 4.3, and Theorem 5.1, it can be seen that choosing $p = \mathcal{O}(\ln(\tau^{-1}))$ in numerical implementation is sufficient to obtain the optimal-order convergence of the fully discrete numerical scheme; see Remark 5.3 for more details. Thus, the fast algorithm reduces the computational cost of evaluating the discrete temporal convolution from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \ln N)$.

Therefore, instead of solving system (3.25), we replace the convolution operator $\mathcal{J}*$ in (3.25) by $\mathcal{K}*$ and construct a new numerical method for solving problem (1.1)–(1.3), which is to find $u_h^n \in S_h^r$ such that

$$(4.9) \quad \left(\frac{3U_h^n - 4U_h^{n-1} + U_h^{n-2}}{2\tau}, v_h \right) + A(U_h^n, v_h) + \left[(\mathcal{K}*_U_h)^n(x_\pm) \right] v_h(x_\pm) = (f^n, v_h)$$

for all $v_h \in S_h^r$ and $n = 2, 3, \dots, N$, with $U_h^0 = \Pi_h \phi$ and $U_h^1 = \Pi_h \psi$, where the operator A is defined in (3.26). Compared with the scheme (3.25), one only needs to solve an algebraic system (4.9) with the same coefficient matrix at each time step. If the computational cost of solving a specific algebraic system is $\mathcal{O}(M)$, the overall computational cost is $\mathcal{O}(N(M + p))$.

For the convolution operator $\mathcal{K}*$, we have the following estimates.

LEMMA 4.1. *If $\max_{s \in \Theta} |\Phi(s) - \sqrt{s}| \leq \epsilon$ with $\Theta = \{s(z) | z \in \bar{\mathbb{D}}\}$ (see (4.2)), then for all $n > 0$, $\{v^m\}_{m=0}^n$ and $\{\omega^m\}_{m=0}^n$, it holds that*

$$\left| \sum_{m=0}^n [(\mathcal{K} * v)^m - (\mathcal{J} * v)^m] \omega^m \right| \leq \epsilon \left(\sum_{m=0}^n |v^m|^2 \right)^{\frac{1}{2}} \left(\sum_{m=0}^n |\omega^m|^2 \right)^{\frac{1}{2}} \quad \forall n \geq 0.$$

Proof. For any fixed $n \geq 0$, we assume $v^m = 0$ and $\omega^m = 0$ for all $m > n$. Then, we have

$$\begin{aligned} & \left| \sum_{m=0}^n [(\mathcal{K} * v)^m - (\mathcal{J} * v)^m] \omega^m \right| \\ &= \left| \sum_{m=0}^{+\infty} [(\mathcal{K} * v)^m - (\mathcal{J} * v)^m] \omega^m \right| = \left| \int_{\partial\mathbb{D}} [\widetilde{\mathcal{K} * v}(z) - \widetilde{\mathcal{J} * v}(z)] \overline{\widetilde{\omega}(z)} \mu(dz) \right| \\ &= \left| \int_{\partial\mathbb{D}} [\Phi(s(z)) - \sqrt{s(z)}] \widetilde{v}(z) \overline{\widetilde{\omega}(z)} \mu(dz) \right| \leq \epsilon \int_{\partial\mathbb{D}} |\widetilde{v}(z)| |\widetilde{\omega}(z)| \mu(dz) \\ &\leq \epsilon \left(\int_{\partial\mathbb{D}} |\widetilde{v}(z)|^2 \mu(dz) \right)^{\frac{1}{2}} \left(\int_{\partial\mathbb{D}} |\widetilde{\omega}(z)|^2 \mu(dz) \right)^{\frac{1}{2}} \\ &= \epsilon \left(\sum_{m=0}^{+\infty} |v^m|^2 \right)^{\frac{1}{2}} \left(\sum_{m=0}^{+\infty} |\omega^m|^2 \right)^{\frac{1}{2}} = \epsilon \left(\sum_{m=0}^n |v^m|^2 \right)^{\frac{1}{2}} \left(\sum_{m=0}^n |\omega^m|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

This ends the proof. \square

4.2. Rational approximation of \sqrt{s} for complex variables. In [8], a sequence of rational approximations of the function \sqrt{x} was proposed on a prescribed closed interval $[\alpha^2, \beta^2] \subset \mathbb{R}^+$. Such rational approximations are the best in the sense of uniform relative error. Thus, we call them the best relative Chebyshev approximations. To begin with, we first recall the recurrence relation between these best rational approximations. In [8], $\nu_k(x)$ with $k \geq 1$ was introduced to denote the best relative Chebyshev approximation of degree $(k, k-1)$ for \sqrt{x} in $[\alpha^2, \beta^2]$. Then, the uniform relative error

$$(4.10) \quad e_{k,k-1} := \sup_{x \in [\alpha^2, \beta^2]} \left| x^{-\frac{1}{2}} \nu_k(x) - 1 \right|$$

is the minimal, compared to any other rational approximations of the same degree. Let $\omega_k := \frac{1}{\sqrt{1-\epsilon_{k,k-1}^2}} \nu_k$. The application of one step of Heron's algorithm yields a new rational function $\omega_{2k}(x)$ of degree $(2k, 2k-1)$ [8, pp. 45–46], i.e., $\omega_{2k}(x) = \frac{1}{2}(\omega_k(x) + \frac{x}{\omega_k(x)})$. It turns out that the function ω_{2k} is equal to ν_{2k} up to a multiplicative constant. More precisely, it holds that

$$\begin{aligned}
\nu_{2k}(x) &= \frac{2\sqrt{1-e_{k,k-1}^2}}{1+\sqrt{1-e_{k,k-1}^2}} \cdot \omega_{2k}(x) = \frac{2\sqrt{1-e_{k,k-1}^2}}{1+\sqrt{1-e_{k,k-1}^2}} \cdot \frac{1}{2} \left(\omega_k(x) + \frac{x}{\omega_k(x)} \right) \\
(4.11) \quad &= \frac{\sqrt{1-e_{k,k-1}^2}}{1+\sqrt{1-e_{k,k-1}^2}} \left(\frac{\nu_k}{\sqrt{1-e_{k,k-1}^2}} + \frac{x\sqrt{1-e_{k,k-1}^2}}{\nu_k} \right).
\end{aligned}$$

A starting approximation for this recursive procedure is the best constant function. It can be verified that the best constant approximation of \sqrt{x} in $[\alpha^2, \beta^2]$ and the relative error are

$$\nu_0(x) := \frac{2\alpha\beta}{\alpha+\beta}, \quad e_{0,0} := \frac{\beta-\alpha}{\beta+\alpha}.$$

In this paper, for simplicity of notation, we denote the best rational approximations of \sqrt{x} in $[\alpha^2, \beta^2]$ by

$$(4.12) \quad \Phi_{n,[\alpha^2,\beta^2]}(x) := \begin{cases} \nu_0(x) & \text{if } n = 0, \\ \nu_{2^{n-1}}(x) & \text{if } n \geq 1, \end{cases}$$

and let

$$(4.13) \quad E_n := E_{n,[\alpha^2,\beta^2]} := \sup_{x \in [\alpha^2, \beta^2]} \left| x^{-\frac{1}{2}} \Phi_{n,[\alpha^2,\beta^2]}(x) - 1 \right|.$$

By (4.11), we have

$$\begin{aligned}
(4.14) \quad \Phi_{n+1,[\alpha^2,\beta^2]}(x) &= \frac{\sqrt{1-E_n^2}}{1+\sqrt{1-E_n^2}} \left(\frac{\Phi_{n,[\alpha^2,\beta^2]}(x)}{\sqrt{1-E_n^2}} + \frac{x\sqrt{1-E_n^2}}{\Phi_{n,[\alpha^2,\beta^2]}(x)} \right), \\
&\Phi_{0,[\alpha^2,\beta^2]}(x) = \frac{2\alpha\beta}{\alpha+\beta},
\end{aligned}$$

and the minimal relative error satisfies [8, see (12)]

$$(4.15) \quad E_{n+1} = \frac{E_n^2}{(1+\sqrt{1-E_n^2})^2}, \quad E_0 = \frac{\beta-\alpha}{\beta+\alpha}.$$

In [8], the following error estimate was proved.

THEOREM 4.2 (see [8, Theorem 3.3]). *Let $\kappa = \frac{\alpha}{\beta}$; then, we have*

$$(4.16) \quad E_n \leq \frac{4}{\omega^{2^n}}$$

for $n = 1, 2, 3, \dots$, with $\omega = \exp[\pi L(\kappa)]$ and $L(\kappa) = \frac{K(\kappa)}{K(\sqrt{1-\kappa^2})}$, where $K(\kappa)$ denotes the complete elliptic integral defined by $K(\kappa) = \int_0^{\frac{\pi}{2}} \frac{d\theta}{\sqrt{1-\kappa^2 \sin^2 \theta}}$.

Remark 4.3. Theorem 4.2 shows that the relative error E_n depends very weakly on the relative size of the approximation interval $[\alpha^2, \beta^2]$, i.e., the ratio $\frac{\alpha}{\beta}$. More precisely, according to the expressions 8.113 in [13], asymptotically we have $K(\kappa) \sim \frac{\pi}{2}(1 + \mathcal{O}(\kappa^2))$ and $K(\sqrt{1-\kappa^2}) \sim \ln \frac{4}{\kappa}(1 + o(1))$, as $\kappa \rightarrow 0^+$, which further yields

$$(4.17) \quad L(\kappa) \sim \frac{\frac{\pi}{2}}{\ln \frac{4}{\kappa}}(1 + o(1))$$

as $\kappa \rightarrow 0^+$. The above result implies that $L(\kappa)$ tends to 0, and thus ω in (4.16) to 1, fairly slowly as $\kappa \rightarrow 0^+$. Consequently, even for a large interval $[\alpha^2, \beta^2]$, as n gets moderately large, the relative error E_n of the rational approximation $\Phi_{n,[\alpha^2,\beta^2]}(x)$ will become very small.

Theorem 4.2 enables us to conjecture that the rational function $\Phi_{n,[\alpha^2,\beta^2]}(s)$ with respect to a complex variable s would also be a good approximation of the function \sqrt{s} , at least when s is suitably close to $[\alpha^2, \beta^2]$. In the following theorem, we will prove that our conjecture is correct to some extent. To this end, let $\mathcal{A}_{d,\varphi}$ denote the sectorial ring-shaped domain defined by

$$\mathcal{A}_{d,\varphi} = \{r \exp(i\theta) : r \in [d^{-1}, d] \text{ and } \theta \in [-\varphi, \varphi]\} \subset \mathbb{C}$$

for $d > 1$ and $\varphi \in [0, \pi/2)$. Here, we introduce the following function directly related to the rational function $\Phi_{n,[c^{-1},c]}(\cdot)$:

$$\zeta_{n,[c^{-1},c]}(s) := s^{-\frac{1}{2}} \Phi_{n,[c^{-1},c]}(s)$$

for all $s \in \mathbb{C}^+$ and $n = 0, 1, 2, \dots$ with $c \in (1, \infty)$. Clearly, to analyze the approximating error of $\Phi_{n,[c^{-1},c]}(s)$ in the complex plane, it suffices to estimate the distance between $\zeta_{n,[c^{-1},c]}(s)$ and 1. In our work, we obtain the following error estimates for the rational approximation of \sqrt{s} for complex variables.

THEOREM 4.4. *There exists a positive constant $\delta < 1$, such that for $c > 3/2$ and $m \geq 1$, by setting $N_1 = \lceil \log_2 \frac{L(2/3)}{L(c^{-1})} \rceil + 2$, we have*

$$(4.18) \quad \sup_{s \in \mathcal{A}_{2c,\pi/2}} |\zeta_{N_1+m,[c^{-1},c]}(s) - 1| \leq C\delta^{2^m}.$$

The proof of Theorem 4.4 is presented in the appendix. By the above theorem, we are able to obtain an asymptotic estimate for the degree of rational approximation and the proof can be also found in the appendix.

COROLLARY 4.5. *Let $c > 3/2$ and $\varepsilon \in (0, 1)$. There exists a positive integer N_* satisfying*

$$(4.19) \quad 2^{N_*} = \mathcal{O}(\ln c \ln \varepsilon^{-1})$$

such that for all $n \geq N_*$, it holds that

$$(4.20) \quad \sup_{s \in \mathcal{A}_{2c,\pi/2}} |\zeta_{n,[c^{-1},c]}(s) - 1| \leq \varepsilon.$$

4.3. Factorized rational approximations (4.3). In the above subsection, (4.14) and Theorem 4.4 give a guideline for the construction of rational approximations in prescribed approximate domain with the error bound (4.18). More precisely, if we need to approximate \sqrt{s} in the prescribed domain $\Theta \subset \mathbb{C}^+$, we first find a sectorial ring-shaped domain $\mathcal{A}_{2c,\pi/2}$ that contains Θ . Then we use the analytical continuation of $\Phi_{n,[c^{-1},c]}(x)$, which is still denoted as $\Phi_{n,[c^{-1},c]}(s)$ for $s \in \mathbb{C}^+$, as our approximation. The degree n can be determined via (4.19) and the estimate (4.20).

In our work, we need to factorize the rational function $\Phi_{n,[c^{-1},c]}(x)$ of degree $(2^{n-1}, 2^{n-1}-1)$ into the simple form of (4.3) to perform the fast algorithm introduced in subsection 4.1. Or equivalently, it suffices to find the poles of the rational function $\Phi_{n,[c^{-1},c]}(x)$. Although the best relative Chebyshev approximation $\Phi_{n,[c^{-1},c]}(x)$ can

be computed recursively via the relation (4.14), such a formula is not suitable to compute the poles recursively. The reason is that the simple form (4.3) of rational approximations cannot be preserved by the relation (4.14). Thus, instead of using (4.14), we use the following formula [8, Lemma 3.1] to recursively generate the rational approximations $\Phi_{n,[c^{-1},c]}(x)$ of \sqrt{s} , which maintains the structure of (4.3):

$$(4.21) \quad \Phi_{n,[\alpha^2,\beta^2]}(x) = r(x)\Phi_{n-1,\left[\frac{4}{(\alpha+\beta)^2}, \frac{1}{\alpha\beta}\right]}(\xi(x)),$$

where $\xi(x) = \frac{x}{r^2(x)}$ and $r(x) = \frac{x+\alpha\beta}{2}$. This formula indicates the relation between the best rational approximation of degree $(2^{n-1}, 2^{n-1} - 1)$ and the one of degree $(2^{n-2}, 2^{n-2} - 1)$.

In the following, we present a more detailed process of obtaining the sum-of-poles form of $\Phi_{n,[\alpha_n^2,\beta_n^2]}(x)$, i.e., (4.3), by using the formula (4.21). Here, we assume that the degree n of the rational approximation $\Phi_{n,[\alpha_n^2,\beta_n^2]}(x)$ is fixed which can be determined via (4.19), and we set $\alpha_n = c^{-1/2}$ and $\beta_n = c^{1/2}$. Note that the poles and weights of rational functions are dependent on the domains. Thus, before introducing the detailed process of obtaining the sum-of-poles form of rational functions, we first apply (4.21) to compute the domains of the rational functions. To this end, we define two functions g_L and g_R by

$$g_L(\xi_1, \xi_2) = \frac{2}{\xi_1 + \xi_2}, \quad g_R(\xi_1, \xi_2) = \frac{1}{\sqrt{\xi_1\xi_2}}.$$

Then, the domain of $\Phi_{n-1,[\alpha_{n-1}^2,\beta_{n-1}^2]}(x)$ can be easily obtained since we have $\alpha_{n-1} = g_L(\alpha_n, \beta_n)$ and $\beta_{n-1} = g_R(\alpha_n, \beta_n)$ by (4.21). We repeat this process and get the domains of $\Phi_{k,[\alpha_k^2,\beta_k^2]}(x)$ for $k = n-2, n-3, \dots, 0$, recursively.

As the domains are known, we are able to derive the sum-of-poles form of $\Phi_{n,[\alpha_n^2,\beta_n^2]}(x)$. First, it is straightforward to get the explicit expression of $\Phi_{0,[\alpha_0^2,\beta_0^2]}(x)$ by (4.14) as α_0 and β_0 are obtained in the above recursive process. Assume that we have already factorized $\Phi_{k-1,[\alpha_{k-1}^2,\beta_{k-1}^2]}(x)$ of the following form:

$$\sum_{m=1}^{l_{k-1}} \frac{w_m}{x - q_m} + w_{l_{k-1}+1}x + w_{l_{k-1}+2}$$

for $k = 1, 2, \dots, n$. By (4.21), we get

$$\Phi_{k,[\alpha_k^2,\beta_k^2]}(x) = r(x) \left(\sum_{m=1}^{l_{k-1}} \frac{w_m}{\xi(x) - q_m} + w_{l_{k-1}+1}\xi(x) + w_{l_{k-1}+2} \right).$$

Now, the task of factorizing $\Phi_{k,[\alpha_k^2,\beta_k^2]}(x)$ is reduced to factorizing several simple rational functions of the form

$$(4.22) \quad \frac{r(x)}{\xi(x) - q_m}, \quad r(x)\xi(x), \quad r(x),$$

where q_m , $m = 1, 2, \dots, l_{k-1}$, are the poles of $\Phi_{k-1,[\alpha_{k-1}^2,\beta_{k-1}^2]}(\xi)$. Note that the rational functions in (4.22) are of degree $(3, 2)$, $(1, 1)$, and $(1, 0)$, respectively, and it is quite easy to compute their poles and weights. Consequently, the sum-of-poles form of $\Phi_{k,[\alpha_k^2,\beta_k^2]}(x)$ follows. We repeat this process and obtain the factorized rational approximation $\Phi_{n,[\alpha_n^2,\beta_n^2]}(x)$ of the form (4.3) for \sqrt{s} , recursively. And again, the degree n can be determined via (4.19) and the estimate (4.20).

5. Error estimates of numerical solutions. In this section, we prove the following error estimates for the fully discrete scheme (4.9).

THEOREM 5.1. *Suppose that the solution u of the PDE problem (2.8)–(2.10) is sufficiently smooth or, equivalently, the solution of the original problem (1.1)–(1.3) is sufficiently smooth. Let $\{U_h^m\}_{m=0}^N$ be the numerical solution of the discrete problem (4.9) with $\epsilon = \mathcal{O}(\tau^2)$ specified in Lemma 4.1. Then, there exists a positive constant τ_0 such that for $\tau \leq \tau_0$, the numerical solution given by (4.9) satisfies the following error estimates:*

$$(5.1) \quad \max_{1 \leq m \leq N} \|u^m - U_h^m\|_{L^2} \leq C(\tau^2 + \ell_h h^{r+1}),$$

$$(5.2) \quad \left(\tau \sum_{m=1}^N \|u^m - U_h^m\|_{H^1}^2 \right)^{\frac{1}{2}} \leq C(\tau^2 + \ell_h h^r),$$

where ℓ_h is defined in (3.23), and C is a positive constant independent of τ and h .

Proof. Clearly, the exact solution $u^n := u(x, t_n)$ satisfies the following weak formulation:

$$(5.3) \quad \begin{aligned} & \left(\frac{3u^n - 4u^{n-1} + u^{n-2}}{2\tau}, v \right) + A(u^n, v) + \left[(\mathcal{J}*u)^n(x_{\pm}) \right] v(x_{\pm}) \\ &= (f^n, v) + \left(\frac{3u^n - 4u^{n-1} + u^{n-2}}{2\tau} - \partial_t u^n, v \right) \\ & \quad + \left[(\mathcal{J}*u)^n(x_{\pm}) - \sqrt{\sigma \partial_t + a^2} u^n(x_{\pm}) \right] v(x_{\pm}) \end{aligned}$$

for all $v \in H^1(\Omega)$, and $n = 2, 3, \dots, N$.

Let

$$e_h^n = R_h u^n - U_h^n$$

for $n = 0, 1, \dots, N$. Subtracting (5.3) from the fully discrete scheme (4.9) results in the following error equation:

$$(5.4) \quad \begin{aligned} & \left(\frac{3e_h^n - 4e_h^{n-1} + e_h^{n-2}}{2\tau}, v_h \right) + A(e_h^n, v_h) + \left[(\mathcal{J}*e_h)^n(x_{\pm}) \right] v_h(x_{\pm}) \\ &= \left(\frac{3u^n - 4u^{n-1} + u^{n-2}}{2\tau} - \partial_t u^n, v_h \right) \\ & \quad + \left(R_h \frac{3u^n - 4u^{n-1} + u^{n-2}}{2\tau} - \frac{3u^n - 4u^{n-1} + u^{n-2}}{2\tau}, v_h \right) \\ & \quad + \left(a \partial_x (R_h u^n - u^n), v_h \right) - \left(a (R_h u^n - u^n), \partial_x v_h \right) \\ & \quad + \left[(\mathcal{J}*u)^n(x_{\pm}) - \sqrt{\sigma \partial_t + a^2} u^n(x_{\pm}) \right] v_h(x_{\pm}) \\ & \quad + \left[(\mathcal{J}*(R_h u - u))^n(x_{\pm}) \right] v_h(x_{\pm}) + \left[((\mathcal{K} - \mathcal{J})*U_h)^n(x_{\pm}) \right] v_h(x_{\pm}) \\ &:= \sum_{j=1}^7 I_j(v_h) \end{aligned}$$

for all $v_h \in S_h^r$ and $n = 2, 3, \dots, N$.

We begin with the error analysis at the starting approximations. From (3.4)–(3.5), it is easy to see that

$$\begin{aligned}\|\psi - u^1\|_{L^2} &\leq C\tau^2 \|u_{tt}\|_{L^2}, \\ \|\psi - u^1\|_{H^1} &\leq C\tau^2 \|u_{tt}\|_{H^1}.\end{aligned}$$

Since $U_h^0 = \Pi_h \phi$ and $U_h^1 = \Pi_h \psi$, combining (3.21) and the above results, we have

$$(5.5) \quad \|u^0 - U_h^0\|_{L^2} \leq C(\tau^2 + h^{r+1}),$$

$$(5.6) \quad \|u^0 - U_h^0\|_{H^1} \leq C(\tau^2 + h^r),$$

$$(5.7) \quad \|u^1 - U_h^1\|_{L^2} \leq \|u^1 - \psi\|_{L^2} + \|\psi - \Pi_h \psi\|_{L^2} \leq C(\tau^2 + h^{r+1}),$$

$$(5.8) \quad \|u^1 - U_h^1\|_{H^1} \leq \|u^1 - \psi\|_{H^1} + \|\psi - \Pi_h \psi\|_{H^1} \leq C(\tau^2 + h^r).$$

In the following, we prove that (5.1)–(5.2) hold also for $m = n$. Substituting $v_h = e_h^n$ into the error equation (5.4) yields

(5.9)

$$\begin{aligned}\frac{1}{4\tau} \|e_h^n\|_{L^2}^2 - \frac{1}{4\tau} \|e_h^{n-1}\|_{L^2}^2 + \frac{1}{4\tau} \|2e_h^n - e_h^{n-1}\|_{L^2}^2 - \frac{1}{4\tau} \|2e_h^{n-1} - e_h^{n-2}\|_{L^2}^2 + \sigma \|\partial_x e_h^n\|_{L^2}^2 \\ + (\mathcal{J} * e_h)^n(x_\pm) e_h^n(x_\pm) \leq \sum_{j=1}^7 I_j(e_h^n),\end{aligned}$$

where we have used the following telescopic formula:

$$\begin{aligned}\left(\frac{3e_h^n - 4e_h^{n-1} + e_h^{n-2}}{2\tau}, e_h^n \right) &= \frac{1}{4\tau} \|e_h^n\|_{L^2}^2 - \frac{1}{4\tau} \|e_h^{n-1}\|_{L^2}^2 + \frac{1}{4\tau} \|2e_h^n - e_h^{n-1}\|_{L^2}^2 \\ &\quad - \frac{1}{4\tau} \|2e_h^{n-1} - e_h^{n-2}\|_{L^2}^2 + \frac{1}{4\tau} \|e_h^n - 2e_h^{n-1} + e_h^{n-2}\|_{L^2}^2.\end{aligned}$$

Now, we estimate the right-hand side of (5.9). By Taylor expansion, it is clear that

$$I_1(e_h^n) \leq \left\| \frac{3u^n - 4u^{n-1} + u^{n-2}}{2\tau} - \partial_t u^n \right\|_{L^2} \|e_h^n\|_{L^2} \leq C\tau^2 \|e_h^n\|_{L^2} \leq C\tau^4 + C\|e_h^n\|_{L^2}^2.$$

System (1.1)–(1.3) implies that both $u(x_\pm, t)$ and its time derivatives vanish at $t = 0$. As a result, we can use Lemma 3.2 and get

$$\begin{aligned}I_5(e_h^n) &\leq \left| (\mathcal{J} * u)^n(x_\pm) - \sqrt{\sigma \partial_t + a^2} u^n(x_\pm) \right| |e_h^n(x_\pm)| \\ &\leq C\sigma\tau^2 |e_h^n(x_\pm)| \leq C\tau^4 + \frac{|a|}{16} |e_h^n(x_\pm)|^2.\end{aligned}$$

By the projection error estimates (3.21)–(3.22) and applying integration by parts, we further obtain

$$\begin{aligned}I_2(e_h^n) &\leq \left\| R_h \frac{3u^n - 4u^{n-1} + u^{n-2}}{2\tau} - \frac{3u^n - 4u^{n-1} + u^{n-2}}{2\tau} \right\|_{L^2} \|e_h^n\|_{L^2} \\ &\leq Ch^{r+1} \|e_h^n\|_{L^2} \leq Ch^{2(r+1)} + C\|e_h^n\|_{L^2}^2, \\ I_3(e_h^n) &= - \left(a(R_h u^n - u^n), \partial_x e_h^n \right) + a(R_h u^n(x_+) - u^n(x_+)) e_h^n(x_+) \\ &\quad - a(R_h u^n(x_-) - u^n(x_-)) e_h^n(x_-) \\ &\leq C\|R_h u^n - u^n\|_{L^2} \|\partial_x e_h^n\|_{L^2} + |a| \|R_h u^n - u^n\|_{L^\infty} |e_h^n(x_+)| \\ &\quad + |a| \|R_h u^n - u^n\|_{L^\infty} |e_h^n(x_-)|\end{aligned}$$

$$\begin{aligned}
&\leq Ch^{r+1}\|\partial_x e_h^n\|_{L^2} + C\ell_h h^{r+1}|e_h^n(x_\pm)| \\
&\leq C\ell_h^2 h^{2(r+1)} + \frac{\sigma}{4}\|\partial_x e_h^n\|_{L^2}^2 + \frac{|a|}{16}|e_h^n(x_\pm)|^2, \\
I_4(e_h^n) &\leq C\|R_h u^n - u^n\|_{L^2}\|\partial_x e_h^n\|_{L^2} \leq Ch^{r+1}\|\partial_x e_h^n\|_{L^2} \leq Ch^{2(r+1)} + \frac{\sigma}{4}\|\partial_x e_h^n\|_{L^2}^2, \\
I_6(e_h^n) &\leq \left\| \mathcal{J}*(R_h u^n - u^n) \right\|_{L^\infty} |e_h^n(x_\pm)| = \left\| R_h(\mathcal{J}*u)^n - (\mathcal{J}*u)^n \right\|_{L^\infty} |e_h^n(x_\pm)| \\
&\leq C\ell_h h^{r+1}|e_h^n(x_\pm)| \leq C\ell_h^2 h^{2(r+1)} + \frac{|a|}{16}|e_h^n(x_\pm)|^2,
\end{aligned}$$

where by convention we define $|e_h^n(x_\pm)| := |e_h^n(x_+)| + |e_h^n(x_-)|$. Summing up (5.9) and using the above estimates of $I_i(e_h^n)$, $i = 1, 2, \dots, 6$, gives

$$\begin{aligned}
(5.10) \quad & \frac{1}{4}\|e_h^n\|_{L^2}^2 + \frac{1}{4}\|2e_h^n - e_h^{n-1}\|_{L^2}^2 + \tau\sigma \sum_{m=2}^n \|\partial_x e_h^m\|_{L^2}^2 + \tau \sum_{m=2}^n \left[(\mathcal{J}*e_h)^m(x_\pm) \right] e_h^m(x_\pm) \\
&\leq C \left(\tau^4 + \ell_h^2 h^{2(r+1)} + \tau \sum_{m=2}^n \|e_h^m\|_{L^2}^2 \right) + \tau \sum_{m=2}^n \left(\frac{3\sigma}{4}\|\partial_x e_h^m\|_{L^2}^2 + \frac{|a|}{4}|e_h^m(x_\pm)|^2 \right) \\
&\quad - \tau \sum_{m=2}^n \left[((\mathcal{K} - \mathcal{J})*e_h)^m(x_\pm) \right] e_h^m(x_\pm) + \tau \sum_{m=2}^n \left[((\mathcal{K} - \mathcal{J})*R_h u)^m(x_\pm) \right] e_h^m(x_\pm).
\end{aligned}$$

Equivalently, we can rewrite the above estimate as follows:

$$\begin{aligned}
(5.11) \quad & \frac{1}{4}\|e_h^n\|_{L^2}^2 + \frac{1}{4}\|2e_h^n - e_h^{n-1}\|_{L^2}^2 + \tau\sigma \sum_{m=2}^n \|\partial_x e_h^m\|_{L^2}^2 + \tau \sum_{m=0}^n \left[(\mathcal{J}*e_h)^m(x_\pm) \right] e_h^m(x_\pm) \\
&\leq C \left(\tau^4 + \ell_h^2 h^{2(r+1)} + \tau \sum_{m=2}^n \|e_h^m\|_{L^2}^2 \right) + \tau \sum_{m=2}^n \left(\frac{3\sigma}{4}\|\partial_x e_h^m\|_{L^2}^2 + \frac{|a|}{4}|e_h^m(x_\pm)|^2 \right) \\
&\quad - \tau \sum_{m=0}^n \left[((\mathcal{K} - \mathcal{J})*e_h)^m(x_\pm) \right] e_h^m(x_\pm) + \tau \sum_{m=0}^1 \left[((\mathcal{K} - \mathcal{J})*e_h)^m(x_\pm) \right] e_h^m(x_\pm) \\
&\quad + \tau \sum_{m=0}^n \left[((\mathcal{K} - \mathcal{J})*R_h u)^m(x_\pm) \right] e_h^m(x_\pm) - \tau \sum_{m=0}^1 \left[((\mathcal{K} - \mathcal{J})*R_h u)^m(x_\pm) \right] e_h^m(x_\pm) \\
&\quad + \tau \sum_{m=0}^1 \left[(\mathcal{J}*e_h)^m(x_\pm) \right] e_h^m(x_\pm) \\
&= C \left(\tau^4 + \ell_h^2 h^{2(r+1)} + \tau \sum_{m=2}^n \|e_h^m\|_{L^2}^2 \right) + \tau \sum_{m=2}^n \left(\frac{3\sigma}{4}\|\partial_x e_h^m\|_{L^2}^2 + \frac{|a|}{4}|e_h^m(x_\pm)|^2 \right) \\
&\quad - \tau \sum_{m=0}^n \left[((\mathcal{K} - \mathcal{J})*e_h)^m(x_\pm) \right] e_h^m(x_\pm) + \tau \sum_{m=0}^n \left[((\mathcal{K} - \mathcal{J})*R_h u)^m(x_\pm) \right] e_h^m(x_\pm) \\
&\quad - \tau \sum_{m=0}^1 \left[((\mathcal{K} - \mathcal{J})*U_h)^m(x_\pm) \right] e_h^m(x_\pm) + \tau \sum_{m=0}^1 \left[(\mathcal{J}*e_h)^m(x_\pm) \right] e_h^m(x_\pm).
\end{aligned}$$

Clearly, by Lemma 4.1, we have

$$\begin{aligned}
(5.12) \quad & \left| \tau \sum_{m=0}^n \left[((\mathcal{K} - \mathcal{J}) * e_h)^m(x_\pm) \right] e_h^m(x_\pm) \right| + \left| \tau \sum_{m=0}^n \left[((\mathcal{K} - \mathcal{J}) * R_h u)^m(x_\pm) \right] e_h^m(x_\pm) \right| \\
& \leq \epsilon \tau \sum_{m=0}^n |e_h^n(x_\pm)|^2 + \epsilon \left(\tau \sum_{m=0}^n |R_h u^m(x_\pm)|^2 \right)^{\frac{1}{2}} \left(\tau \sum_{m=0}^n |e_h^m(x_\pm)|^2 \right)^{\frac{1}{2}} \\
& \leq \epsilon \tau \sum_{m=0}^n |e_h^n(x_\pm)|^2 + \epsilon \left(\tau \sum_{m=0}^n |e_h^m(x_\pm)|^2 \right)^{\frac{1}{2}} \\
& \leq \epsilon \tau \sum_{m=0}^n |e_h^n(x_\pm)|^2 + C\epsilon^2 + \frac{\tau|a|}{4} \sum_{m=0}^n |e_h^m(x_\pm)|^2.
\end{aligned}$$

Now, it suffices to estimate the last two terms in the right-hand side of (5.11). Since $U_h^0 = \Pi_h \phi$, $U_h^1 = \Pi_h \psi$, we can easily see that

$$\begin{aligned}
|e_h^0(x_\pm)| & \leq |R_h \phi(x_\pm) - \phi(x_\pm)| + |\phi(x_\pm) - \Pi_h \phi(x_\pm)| \leq C\ell_h h^{r+1}, \\
|e_h^1(x_\pm)| & \leq |R_h \psi(x_\pm) - \psi(x_\pm)| + |\psi(x_\pm) - \Pi_h \psi(x_\pm)| \leq C\ell_h h^{r+1}.
\end{aligned}$$

Together with Lemma 4.1, we have

$$\begin{aligned}
(5.13) \quad & \left| \tau \sum_{m=0}^1 \left[((\mathcal{K} - \mathcal{J}) * U_h)^m(x_\pm) \right] e_h^m(x_\pm) \right| \leq \epsilon \left(\tau \sum_{m=0}^1 |U_h^m(x_\pm)|^2 \right)^{\frac{1}{2}} \left(\tau \sum_{m=0}^1 |e_h^m(x_\pm)|^2 \right)^{\frac{1}{2}} \\
& \leq \epsilon \tau \left(|e_h^0(x_\pm)|^2 + |e_h^1(x_\pm)|^2 \right)^{\frac{1}{2}} \\
& \leq \epsilon^2 \tau^2 + C\ell_h^2 h^{2(r+1)}.
\end{aligned}$$

From (3.10), we can obtain $\lambda_0 = \mathcal{O}(\tau^{-\frac{1}{2}})$ and $\lambda_1 = \mathcal{O}(\tau^{-\frac{1}{2}})$ by Taylor expansion. Then, we have

$$\begin{aligned}
(5.14) \quad & \tau \sum_{m=0}^1 \left[(\mathcal{J} * e_h)^m(x_\pm) \right] e_h^m(x_\pm) \\
& = \tau \left[(\mathcal{J} * e_h)^0(x_\pm) \right] e_h^0(x_\pm) + \tau \left[(\mathcal{J} * e_h)^1(x_\pm) \right] e_h^1(x_\pm) \\
& = \tau \lambda_0 e_h^0(x_\pm) e_h^0(x_\pm) + \tau \left[\lambda_0 e_h^1(x_\pm) + \lambda_1 e_h^0(x_\pm) \right] e_h^1(x_\pm) \\
& \leq C\tau^{\frac{1}{2}} |e_h^0(x_\pm)|^2 + C\tau^{\frac{1}{2}} |e_h^1(x_\pm)|^2 + C\tau^{\frac{1}{2}} |e_h^0(x_\pm)| |e_h^1(x_\pm)| \\
& \leq C\ell_h^2 h^{2(r+1)}.
\end{aligned}$$

Thus, by Lemma 3.1 and with the estimates (5.12)–(5.14), (5.11) reduces to

$$\begin{aligned}
(5.15) \quad & \frac{1}{4} \|e_h^n\|_{L^2}^2 + \frac{1}{4} \|2e_h^n - e_h^{n-1}\|_{L^2}^2 + \tau \sigma \sum_{m=2}^n \|\partial_x e_h^m\|_{L^2}^2 + |a| \tau \sum_{m=0}^n |e_h^m(x_\pm)|^2 \\
& \leq C \left(\tau^4 + \ell_h^2 h^{2(r+1)} + \tau \sum_{m=2}^n \|e_h^n\|_{L^2}^2 \right) + \tau \sum_{m=2}^n \left(\frac{3\sigma}{4} \|\partial_x e_h^m\|_{L^2}^2 + \frac{|a|}{4} |e_h^m(x_\pm)|^2 \right) \\
& + \epsilon \tau \sum_{m=0}^n |e_h^n(x_\pm)|^2 + C\epsilon^2 + \frac{\tau|a|}{4} \sum_{m=0}^n |e_h^m(x_\pm)|^2 + \epsilon^2 \tau^2 + C\ell_h^2 h^{2(r+1)}.
\end{aligned}$$

Since $\epsilon = \mathcal{O}(\tau^2)$, in view of the above result and using the estimates (5.5)–(5.8) and Gronwall's inequality, we infer that there exists $\tau_0 > 0$ such that

$$(5.16) \quad \max_{2 \leq n \leq N} \|e_h^n\|_{L^2}^2 + \tau \sum_{n=2}^N \|\partial_x e_h^n\|_{L^2}^2 \leq C(\tau^2 + \ell_h h^{r+1})^2,$$

when $\tau \leq \tau_0$. Now, (5.1)–(5.2) follow immediately combining the estimates (3.21) and (5.5)–(5.8). The proof of Theorem 5.1 is complete. \square

Remark 5.2. If we consider the advection-dominated problem, i.e., the viscosity coefficient σ in (1.1) becomes very small, we need to reestimate the terms $I_3(e_h^n)$ and $I_4(e_h^n)$ on the right-hand side of (5.9). To this end, using (3.21) and the inverse inequality (3.24), we have

$$\begin{aligned} I_3(e_h^n) &\leq C \|\partial_x(R_h u^n - u^n)\|_{L^2} \|e_h^n\|_{L^2} \leq Ch^r \|e_h^n\|_{L^2} \leq Ch^{2r} + C \|e_h^n\|_{L^2}^2, \\ I_4(e_h^n) &\leq C \|R_h u^n - u^n\|_{L^2} \|\partial_x e_h^n\|_{L^2} \leq Ch^{r+1} h^{-1} \|e_h^n\|_{L^2} \leq Ch^{2r} + C \|e_h^n\|_{L^2}^2. \end{aligned}$$

Then, applying the same technique used in the proof of Theorem 5.1, we can easily obtain the following error estimates for the advection-dominated problem:

$$(5.17) \quad \max_{1 \leq m \leq N} \|u^m - U_h^m\|_{L^2} + \left(\tau \sum_{m=1}^N \sigma \|u^m - U_h^m\|_{H^1}^2 \right)^{\frac{1}{2}} \leq C(\tau^2 + h^r).$$

In such a case, the constant C of the above estimate is independent of the viscosity coefficient σ .

Remark 5.3. To obtain optimal error estimates for the fully discrete scheme (4.9), Theorem 5.1 requires $\epsilon = \mathcal{O}(\tau^2)$ specified in Lemma 4.1. To this end, we also need to choose $\varepsilon = \mathcal{O}(\tau^2)$ in Corollary 4.5. As a result, if we approximate \sqrt{s} by a rational function of degree $(2^{N_*-1}, 2^{N_*-1} - 1)$ with N_* satisfying $2^{N_*} = \mathcal{O}(\ln c \ln \tau^{-2})$ in Corollary 4.5, then optimal error estimates for the fully discrete scheme (4.9) can be proved.

6. Numerical examples. In this section, we present two numerical examples to illustrate our theoretical analysis.

Example 6.1. We first consider a Cauchy problem of an advection diffusion equation

$$(6.1) \quad \begin{aligned} \frac{\partial u}{\partial t} + 2a \frac{\partial u}{\partial x} &= \sigma \frac{\partial^2 u}{\partial x^2}, \\ u(x, 0) &= \exp \left(-\frac{(x + 0.5)^2}{0.00125} \right), \end{aligned}$$

for $x \in \mathbb{R}$ and $t \in [0, T]$ with $a = 0.5$, $\sigma = 0.01$, $T = 1.3$, where the initial data is a Gaussian profile. In the above equation, the exact solution can be expressed analytically as

$$u(x, t) = \frac{0.025}{\sqrt{0.00625 + 2\sigma t}} \exp \left(-\frac{(x + 0.5 - 2at)^2}{0.00125 + 4\sigma t} \right).$$

Here, we are interested in the spatial computation domain $\Omega_c = [x_-, x_+] = [-1.5, 0.8]$ such that the initial data is negligibly small outside of Ω_c .

We discretize the above problem by the fully discrete scheme (4.9). To investigate the temporal convergence rate, we first set $\tau = T/N$ with $N = 32, 64, 128, 256, 512$ and choose P1 elements with a sufficiently small spatial mesh size $h = 10^{-4}$ such that the spatial discretization error can be relatively negligible. Here, the error of the rational approximation ϵ is chosen to be of $\mathcal{O}(\tau^2)$ in both Examples 6.1 and 6.2 to satisfy the requirement of Theorem 5.1 on ϵ . Then, as shown in the left part of Figure 1, second-order convergence in both L^2 -norm and H^1 -norm is observed by comparing it with the second-order slope, which is consistent with the analysis in Theorem 5.1.

Second, we solve problem (6.1) by using the fully discrete scheme (4.9) and taking the mesh size $h = (x_+ - x_-)/M$ with $M = 100, 200, 400, 800, 1600$ for both P1 and P2 elements. Here, a sufficiently small temporal step size $\tau = 2 \times 10^{-5}$ is used in order to focus on the spatial discretization error. Clearly, the results in Figure 2 show that the spatial convergence rates of P1 and P2 elements are $\mathcal{O}(h^2)$ and $\mathcal{O}(h^3)$, respectively, when measured in L^2 -norm and $\mathcal{O}(h)$ and $\mathcal{O}(h^2)$ when measured in H^1 -norm.

To further test the efficiency of the fast algorithm proposed in this paper, we compare the CPU time (in seconds) of the fast method with that of the direct method in the right part of Figure 1. Here, we choose P1 elements with $h = 5 \times 10^{-3}$ and $\tau = T/N$ with $N = 2^{13}, 2^{14}, 2^{15}, 2^{16}, 2^{17}$. We observe that the computational cost is reduced to be linearly dependent on N by employing the fast algorithm.

Example 6.2. We next consider the advection diffusion equation with 0 initial condition and left boundary condition defined as follows:

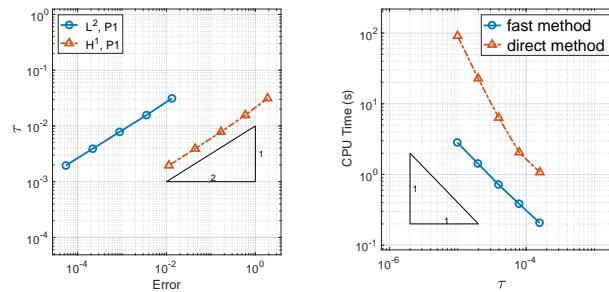


FIG. 1. Left: Temporal convergence in Example 6.1. Right: CPU time of the fast method and the direct method in Example 6.1.

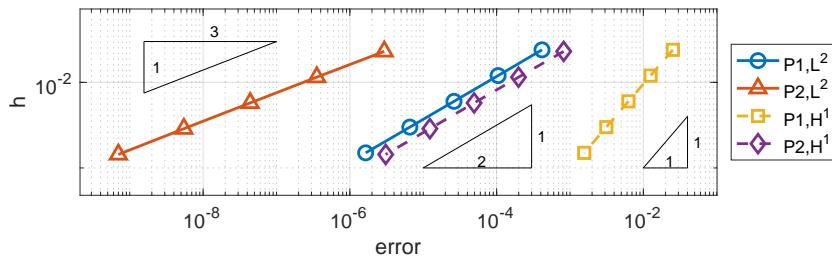


FIG. 2. Spatial convergence in Example 6.1.

$$(6.2) \quad \begin{aligned} \frac{\partial u}{\partial t} + 2a \frac{\partial u}{\partial x} &= \sigma \frac{\partial^2 u}{\partial x^2}, \\ u(0, t) &= \frac{\sin^3(t)}{\sqrt{1+t^2}}, \\ u(x, 0) &= 0, \end{aligned}$$

for $x \in [0, \infty)$ and $t \in [0, T]$, with $a = 0.5$, $\sigma = 0.01$, and $T = 10$. Here, the spatial computation domain that we are interested in is $\Omega_c = [0, 1]$. Note that in this example, we do not have the analytic expression of the exact solution. Thus, a reference solution is computed by the proposed scheme in a larger domain $[0, 8]$ with P2 elements and sufficiently small time step $\tau = 10^{-6}$ and spatial mesh size $h = 2^{-12}$.

Similarly as in the previous example, we first solve problem (6.2) by the fully discrete scheme (4.9) with P1 elements by taking $h = 10^{-4}$ and $\tau = T/N$ with $N = 32, 64, 128, 256, 512$ to investigate the temporal convergence rate. In the left part of Figure 3, the convergence rate in time is shown to be $\mathcal{O}(\tau^2)$ in both L^2 -norm and H^1 -norm. In the right part of Figure 3, the efficiency of the fast algorithm for problem (6.2) is also presented and the numerical results are obtained by using P2 elements with $h = 5 \times 10^{-3}$ and $\tau = T/N$ with $N = 2^{13}, 2^{14}, 2^{15}, 2^{16}, 2^{17}$.

For spatial convergence, we set $h = 1/M$ with $M = 10, 20, 40, 80, 160$ and $\tau = 5 \times 10^{-5}$. Then, we apply the fully discrete scheme (4.9) with both P1 and P2 elements to solve problem (6.2). Again, optimal spatial convergence rates can be observed in Figure 4.

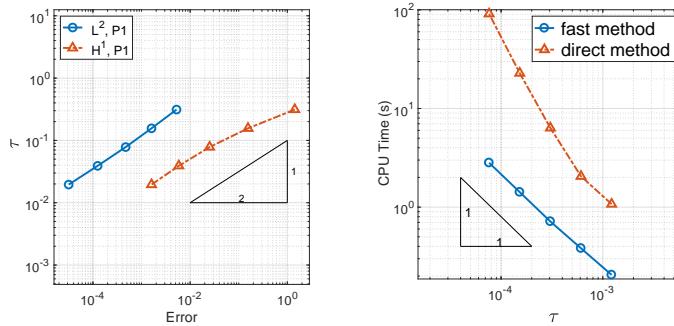


FIG. 3. Left: Temporal convergence in Example 6.2. Right: CPU time of the fast method and the direct method in Example 6.2.

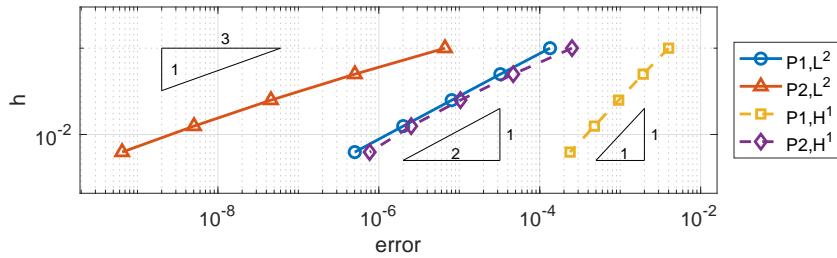


FIG. 4. Spatial convergence in Example 6.2.

7. Conclusion. In this paper, we constructed a fully discrete BDF2 Galerkin finite element scheme for solving advection diffusion equations in unbounded domains. Based on the best rational approximation of the square root function in subdomains of the complex plane, a fast algorithm was developed to approximate the nonlocal convolution integral involved in the discrete ABCs. Then, we presented a framework of error estimation for fully discrete numerical scheme with fast algorithms for the problem in the real line. Our method and analysis can be generalized to the special case of flow in an infinite tunnel (where the velocity is parallel to the tunnel and homogeneous in the section). In the more general cases of multidimensional unbounded domains, it is still challenging to extend our analysis to obtain error estimates of the overall schemes combining the time-stepping methods for PDEs and fast approximations for convolution integrals. The main difficulty is due to the fact that the exact ABCs cannot be derived by the characteristic decomposition. Further efforts must be made to overcome this difficulty. In the case that the convection term is missing, things become easier. We refer readers to [24, 31] for numerical approximations with fast algorithms for the heat equation in multidimensional infinite domains by a Fourier-based method and Green's functions.

Furthermore, in our work, the A-stability of BDF2 is crucial since it is used to obtain the stability estimate of the discrete convolution operator \mathcal{J}^* in (3.18). Higher-order BDF methods are not A-stable and therefore do not have the property above. As a result, the current error analysis cannot be directly generalized to higher-order BDF methods. Intuitively, it is possible to extend our work to high-order A-stable Runge–Kutta methods, while the structure of these Runge–Kutta methods is different from the BDF2 analyzed in this paper. Rigorous error analysis for high-order A-stable Runge–Kutta methods (with fast algorithms for the ABCs) still deserves further investigation in the future.

Appendix A. Proof of Theorem 4.4 and Corollary 4.5. We first present several lemmas, which are useful in the proof of Theorem 4.4.

LEMMA A.1. *For $c > 1$, there holds*

$$(A.1) \quad \zeta_{n,[c^{-1},c]}(s) = \zeta_{n-1,[t^{-1},t]}(g(s,c))$$

for all $s \in \mathbb{C}^+$ and $n = 1, 2, \dots$, where $t = \frac{1+c}{2\sqrt{c}}$ and $g(s,c) = \frac{4st}{(s+1)^2}$.

Proof. Instead of considering (A.1) with respect to $s \in \mathbb{C}^+$, we start proving the following relation:

$$(A.2) \quad \zeta_{n,[c^{-1},c]}(y) = \zeta_{n-1,[t^{-1},t]}(g(y,c))$$

for all $y \in \mathbb{R}^+$ and $n = 1, 2, \dots$.

By the definition (4.13) of E_n , it is easy to see that

$$E_n = \sup_{x \in [\alpha^2, \beta^2]} \left| x^{-\frac{1}{2}} \Phi_{n,[\alpha^2, \beta^2]}(x) - 1 \right| = \sup_{y \in [\frac{\alpha}{\beta}, \frac{\beta}{\alpha}]} \left| y^{-\frac{1}{2}} (\alpha\beta)^{-\frac{1}{2}} \Phi_{n,[\frac{\alpha}{\beta}, \frac{\beta}{\alpha}]}(\alpha\beta y) - 1 \right|,$$

where we have let $x = \alpha\beta y$ and used change of variables. The above result further shows that $(\alpha\beta)^{-\frac{1}{2}} \Phi_{n,[\frac{\alpha}{\beta}, \frac{\beta}{\alpha}]}(\alpha\beta y)$ is the best rational approximation of \sqrt{y} in $[\frac{\alpha}{\beta}, \frac{\beta}{\alpha}]$, i.e.,

$$\Phi_{n,[\frac{\alpha}{\beta}, \frac{\beta}{\alpha}]}(y) = (\alpha\beta)^{-\frac{1}{2}} \Phi_{n,[\frac{\alpha}{\beta}, \frac{\beta}{\alpha}]}(\alpha\beta y).$$

As a result, if the interval $[\alpha^2, \beta^2]$ is transformed to $[\frac{\alpha}{\beta}, \frac{\beta}{\alpha}]$, the above equality implies that the best rational approximation satisfies

$$(A.3) \quad \Phi_{n,[\alpha^2,\beta^2]}(x) = (\alpha\beta)^{\frac{1}{2}} \Phi_{n,[\frac{\alpha}{\beta},\frac{\beta}{\alpha}]}(y).$$

Furthermore, Lemma 3.1 in [8] yields the relation (4.21) between the best rational approximation of degree $(2^n, 2^n - 1)$ and that of degree $(2^{n-1}, 2^{n-1} - 1)$. Applying the transformation (A.3) on both sides of (4.21) leads to

$$\begin{aligned} (\alpha\beta)^{\frac{1}{2}} \Phi_{n+1,[c^{-1},c]}(y) &= (\alpha\beta)^{\frac{1}{2}} \Phi_{n+1,[\frac{\alpha}{\beta},\frac{\beta}{\alpha}]}(y) = \Phi_{n+1,[\alpha^2,\beta^2]}(x) = r(x) \Phi_{n,[\frac{4}{(\alpha+\beta)^2},\frac{1}{\alpha\beta}]}(\xi) \\ &= r(x) \sqrt{\frac{2}{\alpha+\beta} \cdot \frac{1}{\sqrt{\alpha\beta}}} \Phi_{n,[\frac{2\sqrt{\alpha\beta}}{\alpha+\beta},\frac{\alpha+\beta}{2\sqrt{\alpha\beta}}]}(\eta) \\ &= \frac{\alpha\beta y + \alpha\beta}{2} \sqrt{\frac{2}{\alpha+\beta} \cdot \frac{1}{\sqrt{\alpha\beta}}} \Phi_{n,[t^{-1},t]}(\eta) \end{aligned}$$

with $c = \frac{\beta}{\alpha}$, $y = \frac{x}{\alpha\beta}$, $\eta = \frac{(\alpha+\beta)\sqrt{\alpha\beta}}{2}\xi$, $t = \frac{\alpha+\beta}{2\sqrt{\alpha\beta}}$, and we further have

$$\eta = \frac{(\alpha+\beta)\sqrt{\alpha\beta}}{2} \cdot \frac{x}{r^2(x)} = \frac{(\alpha+\beta)\sqrt{\alpha\beta}}{2} \cdot \frac{4\alpha\beta y}{(\alpha\beta y + \alpha\beta)^2} = \frac{\alpha+\beta}{2\sqrt{\alpha\beta}} \cdot \frac{4y}{(y+1)^2} = \frac{4yt}{(y+1)^2}.$$

Consequently, we obtain

$$\begin{aligned} (A.4) \quad \Phi_{n+1,[c^{-1},c]}(y) &= (\alpha\beta)^{-\frac{1}{2}} \cdot \frac{\alpha\beta y + \alpha\beta}{2} \sqrt{\frac{2}{\alpha+\beta} \cdot \frac{1}{\sqrt{\alpha\beta}}} \Phi_{n,[t^{-1},t]}(\eta) \\ &= t^{-\frac{1}{2}} \frac{y+1}{2} \Phi_{n,[t^{-1},t]}(\eta). \end{aligned}$$

Dividing both sides of (A.4) by \sqrt{y} results in (A.2) for all $y \in \mathbb{R}^+$. Because both $\zeta_{n,[c^{-1},c]}(\cdot)$ and $\zeta_{n-1,[t^{-1},t]}(\cdot)$ are rational functions, (A.1) follows immediately for all $s \in \mathbb{C}^+$. The proof of Lemma A.1 is complete. \square

Based on Lemma A.1, we have the following result.

LEMMA A.2. *For $c > 1$, there holds*

$$(A.5) \quad \zeta_{n,[c^{-1},c]}(\mathcal{A}_{2c,\pi/2}) \subset \zeta_{0,[c_n^{-1},c_n]}(\mathcal{A}_{2c_n,\varphi_n})$$

for $n = 1, 2, \dots$, with $c_0 = c$, $c_n = \frac{1+c_{n-1}}{2\sqrt{c_{n-1}}}$, $\varphi_0 = \frac{\pi}{2}$, $\varphi_n = \left(\frac{4}{\pi} \tan^{-1}(2c_{n-1}) - 1\right) \varphi_{n-1}$.

Proof. We first prove that for $c > 1$, $\varphi \in [0, \pi/2)$, and $n = 1, 2, \dots$, we have

$$(A.6) \quad \zeta_{n,[c^{-1},c]}(\mathcal{A}_{2c,\varphi}) \subset \zeta_{n-1,[t^{-1},t]}(\mathcal{A}_{2t,\varphi^*}),$$

where $t = \frac{1+c}{2\sqrt{c}}$ and $\varphi^* = \left(\frac{4}{\pi} \tan^{-1}(2c) - 1\right) \varphi$. By Lemma A.1, it suffices to prove that for $c > 1$ and $s = r \exp(i\theta) \in \mathcal{A}_{2c,\varphi}$, we have

$$(A.7) \quad g(s, c) = \frac{4r \exp(i\theta)t}{(r \exp(i\theta) + 1)^2} \in \mathcal{A}_{2t,\varphi^*}.$$

Actually, if $r \in [1, 2c]$ and $\theta \in (-\varphi, \varphi)$, we have

$$|\operatorname{Arg} g(s, c)| = \left| \theta - 2 \tan^{-1} \frac{r \sin \theta}{r \cos \theta + 1} \right| < 2 \tan^{-1} \frac{2c \sin \varphi}{2c \cos \varphi + 1} - \varphi =: \eta(\varphi).$$

Since $g(s, c) = g(s^{-1}, c)$, this inequality holds also for $r \in [(2c)^{-1}, 1]$ and $\theta \in (-\varphi, \varphi)$. It is easy to verify that $\eta(\varphi)$ is an increasing and convex function in $[0, \pi/2]$ for all $c > 1$. As a result, we get

$$\eta(\varphi) \leq \left(\frac{4}{\pi} \tan^{-1}(2c) - 1 \right) \varphi =: \varphi^* < \varphi.$$

On the other hand, since $1 + r^2 \leq 1 + r^2 + 2r \cos \theta \leq (1 + r)^2$ for all $\theta \in (-\varphi, \varphi)$, we derive

$$\frac{1}{2t} \leq \frac{8ct}{(1+2c)^2} \leq \frac{4rt}{(1+r)^2} \leq |g(s, c)| = \frac{4rt}{r^2 + 2r \cos \theta + 1} \leq \frac{4rt}{1+r^2} \leq 2t.$$

Combining the above two inequalities yields (A.7) for all $s \in \mathcal{A}_{2c, \varphi}$, which further implies (A.6). By applying (A.6) recursively, we immediately derive (A.5). This ends the proof. \square

Now we start to estimate the distance between $\zeta_{n, [c^{-1}, c]}(s)$ and 1 in the complex plane. Here, we first give a rough estimate for the error $|\zeta_{n, [c^{-1}, c]}(s) - 1|$, with which we will present the proof of Theorem 4.4.

LEMMA A.3. *For $c > \frac{3}{2}$, let $N_1 = \lceil \log_2 \frac{L(2/3)}{L(c^{-1})} \rceil + 2$; then for all $s \in \mathcal{A}_{2c, \pi/2}$, we have*

$$(A.8) \quad |\zeta_{N_1, [c^{-1}, c]}(s) - 1| \leq \frac{17\sqrt{3}}{32} < 1.$$

Proof. First, we estimate the sequence $\{c_n\}$ in Lemma A.2. From Lemma A.2, we have $c_n = \frac{1+c_{n-1}}{2\sqrt{c_{n-1}}}$ for $n \geq 1$. Let $\kappa_n = \frac{1}{c_n}$; then $\kappa_n = \frac{2\sqrt{\kappa_{n-1}}}{1+\kappa_{n-1}}$. By Lemma 9.1 in [8] (see (54)), we get

$$L(\kappa_n) = 2L(\kappa_{n-1}) = 2^n L(\kappa_0) = 2^n L(c^{-1}).$$

For $N_1 = \lceil \log_2 \frac{L(\frac{2}{3})}{L(c^{-1})} \rceil + 2$, we derive

$$2^{N_1-2} L(c^{-1}) \geq L\left(\frac{2}{3}\right),$$

or equivalently,

$$L(\kappa_{N_1-2}) \geq L\left(\frac{2}{3}\right).$$

Considering $L(\kappa)$ is an increasing function in $(0, 1)$, the above result implies $\kappa_{N_1-2} \geq \frac{2}{3}$. Noting that $\kappa_n = \frac{1}{c_n}$, we then derive $1 < c_{N_1-2} \leq \frac{3}{2}$, which leads to

$$(A.9) \quad 1 < c_{N_1} < c_{N_1-1} < c_{N_1-2} < \frac{3}{2},$$

since $\{c_n\}$ is decreasing.

On the other hand, because $\varphi_n = \left(\frac{4}{\pi} \tan^{-1}(2c_{n-1}) - 1 \right) \varphi_{n-1}$, applying (A.9) recursively yields

$$(A.10) \quad \varphi_{N_1} = \prod_{n=0}^{N_1-1} \left(\frac{4}{\pi} \tan^{-1}(2c_{n-1}) - 1 \right) \varphi_0 < \left(\frac{4}{\pi} \tan^{-1}(3) - 1 \right)^2 \cdot \frac{\pi}{2} =: \gamma \cdot \frac{\pi}{2}.$$

Together with (A.9) and (A.10), we can see that

$$\mathcal{A}_{2c_{N_1},\varphi_{N_1}} \subset \mathcal{A}_{3,\frac{\gamma\pi}{2}}.$$

Since $(\mathcal{A}_{2c_{N_1},\varphi_{N_1}})^{\frac{1}{2}} = (\mathcal{A}_{2c_{N_1},\varphi_{N_1}})^{-\frac{1}{2}}$, by applying the above result, it can be directly verified that

$$(\mathcal{A}_{2c_{N_1},\varphi_{N_1}})^{-\frac{1}{2}} = (\mathcal{A}_{2c_{N_1},\varphi_{N_1}})^{\frac{1}{2}} \subset \left(\mathcal{A}_{3,\frac{\gamma\pi}{2}}\right)^{\frac{1}{2}} \subset B_1(\sqrt{3}/2),$$

where $B_1(\sqrt{3}/2)$ denotes a disk of radius $\frac{\sqrt{3}}{2}$ centered at 1. Then, for all $s \in \mathcal{A}_{2c_{N_1},\varphi_{N_1}}$, we obtain $|s^{-\frac{1}{2}} - 1| \leq \frac{\sqrt{3}}{2}$ and

$$\begin{aligned} |\zeta_{0,[c_{N_1}^{-1},c_{N_1}]}(s) - 1| &= \left| \frac{2s^{-\frac{1}{2}}}{c_{N_1}^{\frac{1}{2}} + c_{N_1}^{-\frac{1}{2}}} - 1 \right| \\ &\leq \left| \frac{2}{c_{N_1}^{\frac{1}{2}} + c_{N_1}^{-\frac{1}{2}}} - 1 \right| |s^{-\frac{1}{2}} - 1| + \left| \frac{2}{c_{N_1}^{\frac{1}{2}} + c_{N_1}^{-\frac{1}{2}}} - 1 \right| + |s^{-\frac{1}{2}} - 1| \\ &\leq \frac{\sqrt{3}}{64} \frac{\sqrt{3}}{2} + \frac{\sqrt{3}}{64} + \frac{\sqrt{3}}{2} < \frac{17\sqrt{3}}{32} < 1, \end{aligned}$$

where we have used the inequality $|\frac{2}{(\frac{3}{2})^{\frac{1}{2}} + (\frac{3}{2})^{-\frac{1}{2}}} - 1| < \frac{\sqrt{3}}{64}$. Applying Lemma A.2 leads to (A.8). This ends the proof. \square

Lemma A.3 gives an estimate of n to ensure that $|\zeta_{n,[c^{-1},c]}(s) - 1|$ is uniformly bounded by a number less than 1. In the following, we will use new techniques to show that the error function $|\zeta_{n,[c^{-1},c]}(s) - 1|$ tends to 0 fairly fast. To this end, we introduce a new function $R_n(s)$ defined by

$$(A.11) \quad R_n(s) = \frac{\zeta_{n,[c^{-1},c]}(s)}{1 + E_n} = \frac{\Phi_{n,[c^{-1},c]}(s)}{(1 + E_n)\sqrt{s}}$$

for $s \in \mathcal{A}_{2c,\pi/2}$. It is straightforward to deduce from (4.14) and (4.15) that

$$R_{n+1}(s) = \frac{1}{2} \left((1 + E_n)R_n(s) + \frac{1 - E_n}{R_n(s)} \right).$$

By applying the Möbius transformation,

$$(A.12) \quad \mu_n(s) = \frac{R_n(s) - 1}{R_n(s) + 1},$$

we derive

$$\begin{aligned} \mu_{n+1}(s) &= \frac{R_{n+1}(s) - 1}{R_{n+1}(s) + 1} = \frac{(1 + E_n)R_n^2(s) + 1 - E_n - 2R_n(s)}{(1 + E_n)R_n^2(s) + 1 - E_n + 2R_n(s)} \\ &= \frac{(R_n(s) - 1 + E_n(R_n(s) + 1))(R_n(s) - 1)}{(R_n(s) + 1 + E_n(R_n(s) - 1))(R_n(s) + 1)} \\ &= \mu_n(s) \frac{\mu_n(s) + E_n}{1 + E_n\mu_n(s)}. \end{aligned}$$

Note that $\zeta_n(s) \rightarrow 1$ is equivalent to $\mu_n(s) \rightarrow 0$ as $n \rightarrow \infty$. With the above results, we can now prove Theorem 4.4.

Proof of Theorem 4.4. Because $N_1 = \lceil \log_2 \frac{L(2/3)}{L(c^{-1})} \rceil + 2$, by Theorem 4.2, it is clear that

$$\omega^{2^{N_1}} = \exp(2^{N_1} \pi L(c^{-1})) \geq \exp(4\pi L(2/3)),$$

which leads to

$$E_{N_1} \leq 4 \exp(-4\pi L(2/3)) =: c_2.$$

Recalling the definitions of $R_n(s)$ and $\mu_n(s)$, we have

$$|\mu_{N_1}(s)| = \left| \frac{\zeta_{N_1, [c^{-1}, c]}(s) - (1 + E_{N_1})}{\zeta_{N_1, [c^{-1}, c]}(s) + (1 + E_{N_1})} \right| \leq \frac{c_1 + c_2}{2 - c_1 - c_2}$$

for $s \in \mathcal{A}_{2c, \pi/2}$, where we used Lemma A.3 and set $c_1 := \frac{17\sqrt{3}}{32}$. Let $F_n = \max_{s \in \mathcal{A}_{2c, \pi/2}} |\mu_n(s)|$ for all $n \geq 0$; then the above estimate further implies

$$F_{N_1} \leq \frac{c_1 + c_2}{2 - c_1 - c_2} =: c_3 < 1.$$

It can be directly verified that

$$(A.13) \quad E_{N_1} + F_{N_1} \leq c_2 + c_3 =: \delta < 1.$$

As mentioned above, $\zeta_n(s) \rightarrow 1$ is equivalent to $\mu_n(s) \rightarrow 0$ as $n \rightarrow \infty$. In the following, we will prove that for $m \geq 1$,

$$(A.14) \quad F_{N_1+m} \leq \delta^{2^m}.$$

We start from the estimate of F_n for $n = 0, 1, 2, \dots$. It is straightforward to verify that $F_0 < 1$. We assume $F_n < 1$ and then prove that this estimate also holds for F_{n+1} by mathematical induction. Let $\mu_n(s) = re^{i\theta}$; clearly, we have $r \leq F_n < 1$ and

$$\begin{aligned} \left| \frac{\mu_n(s) + E_n}{1 + E_n \mu_n(s)} \right|^2 &= \left| \frac{re^{i\theta} + E_n}{1 + E_n re^{i\theta}} \right|^2 = \frac{(r \cos \theta + E_n)^2 + (r \sin \theta)^2}{(E_n r \cos \theta + 1)^2 + (E_n r \sin \theta)^2} \\ &= \frac{r^2 + E_n^2 + 2E_n r \cos \theta}{1 + E_n^2 r^2 + 2E_n r \cos \theta} = \frac{r^2 + E_n^2 - 1 - E_n^2 r^2}{1 + E_n^2 r^2 + 2E_n r \cos \theta} + 1 \\ &= 1 - \frac{(E_n^2 - 1)(r^2 - 1)}{1 + E_n^2 r^2 + 2E_n r \cos \theta} \leq \left(\frac{r + E_n}{1 + r E_n} \right)^2 \\ &\leq \left(\frac{F_n + E_n}{1 + F_n E_n} \right)^2 < 1. \end{aligned}$$

Therefore, we obtain

$$(A.15) \quad F_{n+1} \leq F_n \frac{F_n + E_n}{1 + E_n F_n} < F_n < 1.$$

The induction is closed and we have $F_n < 1$ for all $n \geq 0$. Estimate (A.15) implies also that

$$F_{n+1} \leq F_n(E_n + F_n).$$

Using the fact that $E_{n+1} \leq E_n^2$ for all $n \geq 1$ and (A.13), we obtain

$$(E_{N_1+1} + F_{N_1+1}) \leq (E_{N_1} + F_{N_1})^2 \leq \delta^2,$$

and thus, for $m \geq 1$,

$$E_{N_1+m} + F_{N_1+m} \leq \delta^{2^m},$$

which implies (A.14).

Now, we present the estimate of $\zeta_{n,[c^{-1},c]}$ by using (A.14). Because

$$\zeta_{n,[c^{-1},c]}(s) - 1 = E_n + \frac{(2+2E_n)\mu_n(s)}{1-\mu_n(s)} = \frac{E_n + E_n\mu_n(s) + 2\mu_n(s)}{1-\mu_n(s)},$$

we have

$$\sup_{s \in \mathcal{A}_{2c,\pi/2}} |\zeta_{N_1+m,[c^{-1},c]}(s) - 1| \leq \frac{4F_{N_1+m}}{1-F_{N_1+m}} \leq \frac{4}{1-\delta} \delta^{2^m} \leq C \delta^{2^m}.$$

The proof of Theorem 4.4 is complete. \square

By Theorem 4.4, we are able to prove an asymptotic estimate for the degree of the rational approximation given in Corollary 4.5.

Proof of Corollary 4.5. By Theorem 4.4, we let

$$N_1 = \left\lceil \log_2 \frac{L(2/3)}{L(c^{-1})} \right\rceil + 2, \quad N_2 = \left\lceil \log_2 \frac{\ln \varepsilon - \ln C}{\ln \delta} \right\rceil$$

and define $N_* := N_1 + N_2$; then

$$\sup_{s \in \mathcal{A}_{2c,\pi/2}} |\zeta_{n,[c^{-1},c]}(s) - 1| \leq \varepsilon \quad \forall n \geq N_*.$$

Furthermore, by (4.17), we can see that

$$2^{N_*} = \mathcal{O}\left(\frac{\ln \varepsilon^{-1}}{L(c^{-1})}\right) = \mathcal{O}(\ln c \ln \varepsilon^{-1}).$$

This ends the proof of Corollary 4.5. \square

REFERENCES

- [1] B. ALERT, L. GREENGARD, AND T. HAGSTROM, *Rapid evaluation of nonreflecting boundary kernels for time-domain wave propagation*, SIAM J. Numer. Anal., 37 (2000), pp. 1138–1164.
- [2] X. ANTOINE AND C. BESSE, *Unconditionally stable discretization schemes of non-reflecting boundary conditions for the one-dimensional Schrödinger equation*, J. Comput. Phys., 188 (2003), pp. 157–175.
- [3] X. ANTOINE, C. BESSE, AND P. KLEIN, *Absorbing boundary conditions for general nonlinear Schrödinger equations*, SIAM J. Sci. Comput., 33 (2011), pp. 1008–1033.
- [4] X. ANTOINE, Q. TANG, AND J. ZHANG, *On the numerical solution and dynamical laws of nonlinear fractional Schrödinger/Gross-Pitaevskii equations*, Int. J. Comput. Math., 95 (2018), pp. 1423–1443.
- [5] A. ARNOLD, M. EHRHARDT, AND I. SOFRONOV, *Discrete transparent boundary conditions for the Schrödinger equation: fast calculation, approximation, and stability*, Commun. Math. Sci., 1 (2003), pp. 501–556.
- [6] S. ASVADUROV, V. DRUSKIN, M. GUDDATI, AND L. KNIZHNERMAN, *On optimal finite-difference approximation of PML*, SIAM J. Numer. Anal., 41 (2003), pp. 287–305.
- [7] V. BASKAKOV AND A. POPOV, *Implementation of transparent boundaries for numerical solution of the Schrödinger equation*, Wave Motion, 14 (1991), pp. 123–128.
- [8] D. BRAESS AND W. HACKBUSCH, *On the efficient computation of high-dimensional integrals and the approximation by exponential sums*, in Multiscale, Nonlinear and Adaptive Approximation, R. DeVore and A. Kunoth, eds., Springer, Berlin, 2009, pp. 39–74.
- [9] V. DRUSKIN, S. GUETTEL, AND L. KNIZHNERMAN, *Near-optimal perfectly matched layers for indefinite Helmholtz problems*, SIAM Rev., 58 (2016), pp. 90–116.

- [10] M. EHRHARDT, H. HAN, AND C. ZHENG, *Numerical simulation of waves in periodic structures*, Comm. Comput. Phys., 5 (2009), pp. 849–870.
- [11] M. EHRHARDT AND C. ZHENG, *Fast numerical methods for waves in periodic media*, in Progress in Computational Physics: Wave Propagation in Periodic Media, M. Ehrhardt, ed., Springer, Berlin, 2010, pp. 135–166.
- [12] D. GIVOLI, *High-order local non-reflecting boundary conditions: A review*, Wave Motion, 39 (2004), pp. 319–326.
- [13] I. GRADSHTEYN AND I. RYZHIK, *Table of Integrals, Series, and Products*, Academic Press, Boston, 2014.
- [14] T. HAGSTROM, *Open boundary conditions for a parabolic system*, Math. Comp. Model., 20 (1994), pp. 55–68.
- [15] T. HAGSTROM, *New results on absorbing layers and radiation boundary conditions*, in Topics in Computational Wave Propagation, Lect. Notes Comput. Sci. Eng. 31, M. Ainsworth, P. Davies, D. Duncan, B. Rynne, and P. Martin, eds., Springer, Berlin, 2003, pp. 1–42.
- [16] T. HAGSTROM AND T. WARBURTON, *Complete radiation boundary conditions: Minimizing the long time error growth of local methods*, SIAM J. Numer. Anal., 47 (2009), pp. 3678–3704.
- [17] L. HALPERN, *Artificial boundary conditions for the linear advection diffusion equation*, Math. Comp., 46 (1986), pp. 425–438.
- [18] L. HALPERN AND J. RAUCH, *Absorbing boundary conditions for diffusion equations*, Numer. Math., 71 (1995), pp. 185–224.
- [19] H. HAN AND X. WU, *Artificial Boundary Method*, Springer, Berlin, 2013.
- [20] D. INGERMAN, V. DRUSKIN, AND L. KNIZHNERMAN, *Optimal finite difference grids and rational approximations of the square root I. Elliptic problems*, Comm. Pure Appl. Math., 53 (2000), pp. 1039–1066.
- [21] S. JIANG AND L. GREENGARD, *Fast evaluation of nonreflecting boundary conditions for the Schrödinger equation in one dimension*, Comput. Math. Appl., 47 (2002), pp. 955–966.
- [22] S. JIANG, J. ZHANG, Q. ZHANG, AND Z. ZHANG, *Fast evaluation of the Caputo fractional derivative and its applications to fractional diffusion equations*, Commun. Comput. Phys., 21 (2017), pp. 650–678.
- [23] B. LI, J. ZHANG, AND C. ZHENG, *Stability and error analysis for a second order fast approximation of the 1D Schrödinger equation under absorbing boundary conditions*, SIAM J. Sci. Comput., 40 (2018), pp. A4083–A4104.
- [24] J. LI AND L. GREENGARD, *On the numerical solution of the heat equation I: Fast solvers in free space*, J. Comput. Phys., 22 (2007), pp. 1891–1901.
- [25] J. P. LOHÉAC, *An artificial boundary condition for an advection-diffusion equation*, Math. Methods Appl. Sci., 14 (1991), pp. 155–175.
- [26] C. LUBICH AND A. SCHÄDLE, *Fast convolution for nonreflecting boundary conditions*, SIAM J. Sci. Comput., 24 (2002), pp. 161–182.
- [27] B. MAYFIELD, *Non-Local Boundary Conditions for the Schrödinger Equation*, Ph.D. thesis, University of Rhode Island, Providence, RI, 1989.
- [28] K. W. MORTON, *Numerical Solution of Convection-Diffusion Problems*, Chapman and Hall, London, 1996.
- [29] E. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
- [30] F. STENGER, *Approximations via Whittaker's cardinal function*, J. Approx. Theory., 17 (1976), pp. 222–240.
- [31] A. SUHOV AND A. DITKOWSKI, *Artificial boundary conditions for the simulation of the heat equation in an infinite domain*, SIAM J. Sci. Comput., 33 (2011), pp. 1765–1784.
- [32] J. SZETFEL, *Design of absorbing boundary conditions for Schrödinger equations in \mathbb{R}^d* , SIAM J. Numer. Anal., 42 (2004), pp. 1527–1551.
- [33] V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer, Berlin, 2006.
- [34] S. V. TSYRKOV, *Numerical solution of problems on unbounded domains: A review*, Appl. Numer. Math., 27 (1998), pp. 465–532.
- [35] X. WU AND J. ZHANG, *High-order local absorbing boundary conditions for heat equation in unbounded domains*, J. Comput. Math., 29 (2011), pp. 74–90.
- [36] X. YANG AND J. ZHANG, *Computation of the Schrödinger equation in the semiclassical regime on an unbounded domain*, SIAM J. Numer. Anal., 52 (2014), pp. 808–831.
- [37] J. ZHANG, Z. SUN, AND D. WANG, *Analysis of high-order absorbing boundary conditions for the Schrödinger equation*, Commun. Comput. Phys., 10 (2011), pp. 742–766.
- [38] A. ZISOWSKY AND M. EHRHARDT, *Discrete transparent boundary conditions for parabolic systems*, Math. Comp. Model., 43 (2006), pp. 294–309.