MATHEMATICS OF COMPUTATION
Volume 71, Number 240, Pages 1497–1528
S 0025-5718(01)01365-5
Article electronically published on December 21, 2001

COMBINED HERMITE SPECTRAL-FINITE DIFFERENCE METHOD FOR THE FOKKER-PLANCK EQUATION

JOHNSON C. M. FOK, BENYU GUO, AND TAO TANG

ABSTRACT. The convergence of a class of combined spectral-finite difference methods using Hermite basis, applied to the Fokker-Planck equation, is studied. It is shown that the Hermite based spectral methods are convergent with spectral accuracy in weighted Sobolev space. Numerical results indicating the spectral convergence rate are presented. A velocity scaling factor is used in the Hermite basis and is shown to improve the accuracy and effectiveness of the Hermite spectral approximation, with no increase in workload. Some basic analysis for the selection of the scaling factors is also presented.

1. INTRODUCTION

In the studies on Brownian motion we are principally concerned with the perpetual irregular motions exhibited by small grains or particles of colloidal size immersed in a fluid. The perpetual motions of the Brownian particles are maintained by fluctuations in the collisions with the molecules of the surrounding fluid. Under normal conditions, in a liquid, a Brownian particle will suffer about 10^{21} collisions per second and this is so frequent that we cannot really speak of separate collisions. Also, since each collision can be thought of as producing a kink in the path of the particle, it follows that we cannot hope to follow the path in any detail. The modern theory of the Brownian motion of a *free particle* (i.e., in the absence of an external field of force) generally starts with Langevin's equation

$$\frac{d\mathbf{u}}{dt} = -\beta_0 \mathbf{u} + \mathbf{A}(t), \tag{1.1}$$

where $\mathbf{u}$ denotes the velocity of the particle. According to this equation, the influence of the surrounding medium on the motion of the particle can be split into two parts: first, a systematic part $-\beta_0\mathbf{u}$ representing a dynamical friction experienced by the particle and second, a fluctuating part $\mathbf{A}(t)$ which is a characteristic of the Brownian motion.

In an analysis of the Brownian movement we regard as impracticable a detailed description of the motions of the individual particles. Instead, we emphasize the essential stochastic nature of the phenomenon and seek a description in terms of

Received by the editor December 13, 1999 and, in revised form, October 30, 2000.
2000 *Mathematics Subject Classification.* Primary 65M12, 65M70; Secondary 82C31.
Key words and phrases. Fokker-Planck equation, unbounded domain, Hermite spectral method, finite-difference method, error analysis.
This research was partially supported by FRG Grants of Hong Kong Baptist University and RGC Grants of Hong Kong Research Grants Council.

the *probability distributions* of velocity and/or position at a later time starting from given initial distributions. The Fokker-Planck equation deals with those probability distribution of position and velocity under quite general circumstances. For example, let Δt denote a time interval long compared to the periods of fluctuations of the acceleration $\mathbf{A}(t)$ occurring in the Langevin equation but short compared to intervals during which the velocity of a Brownian particle changes by appreciable amounts. In this case we should expect to derive the distribution function $W(\mathbf{u}, t+\Delta t)$ governing the probability of occurrence of $\mathbf{u}$ at time $t+\Delta t$ from the distribution $W(\mathbf{u}, t)$ at time t and a knowledge of the transition probability that $\mathbf{u}$ suffers an increment $\Delta\mathbf{u}$ in time Δt. By expanding $W(\mathbf{u}, t \pm \Delta t)$ in the form of Taylor series and passing to the limit $\Delta t \to 0$, we can obtain a special form of the Fokker-Planck equation in velocity space to describe the Brownian motion of a free particle:

$$\frac{\partial W}{\partial t} = \beta_0 \mathrm{div}_{\mathbf{u}}(\mathbf{u}W) + \beta_0 q^2 \nabla^2_{\mathbf{u}} W \tag{1.2}$$

where q is a positive constant called *thermal velocity*. Equation (1.2) is one of the simplest Fokker-Planck equations. By solving (1.2) starting with $W(\mathbf{u}, 0)$ for $t = 0$ and subject to the appropriate boundary conditions, one obtains the distribution function $W(\mathbf{u}, t)$ for all later times. Once we have found $W(\mathbf{u}, t)$, any *averaged* value of the velocity can be calculated based on simple integrations. The derivations for (1.2) and more general forms of the Fokker-Planck equations can be found in Chandrasekhar [13] and Risken [35]. The Fokker-Planck equation is now used in a number of different fields in natural science, such as solid-state physics, quantum optics, chemical physics, theoretical biology and circuit theory. The theoretical analysis for the Fokker-Planck equation can be also found in many literatures (see, e.g., Diperna and Lions [15] and Perthame [34]).

Over the past decades it has turned out that the Fokker-Planck equation provides a powerful tool with which the effects of fluctuations close to transition points can be adequately treated and that the approaches based on the Fokker-Planck equation are superior to other approaches, e.g., based on Langevin equations (1.1) (see, e.g., [25, 35]). Quite generally, the Fokker-Planck equation plays an important role in problems which involve noise, e.g., in electrical circuits. Various methods of solutions for the Fokker-Planck equation have been proposed, including transformation to Schrodinger equations, WKB methods, and matrix continued-fraction methods (see, e.g., Chapters 5 and 6 of [35]). Although analytic solutions of the Fokker-Planck equations can be found in some special cases, in general it is difficult to obtain them especially if separation of variables is impossible or if boundary values are prescribed. In general cases, numerical methods have become important in obtaining the approximate solutions of the Fokker-Planck equation. One class of works to solve the Fokker-Planck equation analytically is based on a proposal by Brinkman to expand the velocity part of the probability distribution function in Hermite functions [7]. Recursion relations for the position and time-dependent expansion coefficients follow from the recursion relations for Hermite polynomials and eigensolutions of the Fokker-Planck equation are sought. His approach has become one of the most popular methods used for solving the Fokker-Planck equation, not only for the Cauchy problems (for which some analytical solutions can be obtained) but also for the initial-boundary value problems (for which analytical solutions are in general unavailable).

Motivated by the work of Brinkman, a class of numerical methods have been proposed to solve the Fokker-Planck equation by using the Hermite functions in velocity as spectral basis (see, e.g., Boyd [4, 5] and Tang et al. [39, 40]). To better illustrate the ideas of the methods, we will consider only a special class of the Fokker-Planck equation, namely, the Kramers equation [28]. The numerical and analysis techniques used in this work can be easily extended to other types of the Fokker-Planck equations. The Kramers equation is an equation of motion for distribution functions in position and velocity space describing the Brownian motion of particles in an external field. In the one-dimensional case it has the form, $W = W(x, v, t)$,

$$\frac{\partial W}{\partial t} = -v\frac{\partial W}{\partial x} + \beta_1 \frac{\partial (vW)}{\partial v} - \frac{F_1(x)}{m}\frac{\partial W}{\partial v} + \frac{\beta_1 kT}{m}\frac{\partial^2 W}{\partial^2 v}. \tag{1.3}$$

Here W is the probability density, β_1 is the damping constant ($\tau = 1/\beta_1$ is the relaxation time), m is the mass of the particle, T is the temperature of the fluid, k is the Boltzmann's constant, and $F_1(x) = -mf'(x)$ is the external force where $mf(x)$ is the potential. For Brownian motion of particles, whose probability density W in phase space is a solution of the Fokker-Planck equation (1.3), the boundary conditions become complicated. For one-dimensional boundary value problems, it will not be well-posed if we propose boundary conditions on the left and right walls $x = x_{\min}$, $x = x_{\max}$ and $|v| < \infty$. To see this, we consider the case that there are absorbing walls at the left- and right-hand sides of the domain $x_{\min} \le x \le x_{\max}$. In this case, at the left side of the domain $x = x_{\min}$ we require that the probability current in x-direction must vanish for those particles leaving the wall into the domain, i.e., for the particles with positive velocities. Therefore, we must require that the probability density for positive velocities is zero at $x = x_{\min}$,

$$W(x_{\min}, v, t) = 0, \qquad \text{for} \qquad v > 0. \tag{1.4}$$

Similarly, since we have an absorbing wall at $x = x_{\max}$, we have

$$W(x_{\max}, v, t) = 0, \qquad \text{for} \qquad v < 0. \tag{1.5}$$

Kramers was able to derive rate expressions for various ranges of the damping constant β_1. For a piecewise parabolic potential, Blomberg [8] derived an analytical solution for strong damping in terms of parabolic cylinder functions and a numerical scheme useful towards weaker damping based on a truncated expansion in the same functions. Voigtlaender and Risken [44, 45] have performed extensive studies of other potentials by a method of matrix continued fractions. Burschka and Titulater [9, 10] calculated probability densities for the equation (1.3) with the absorbing boundary conditions (1.4)–(1.5). Several numerical methods, such as finite difference methods [12], Galerkin method [33] and mixed Hermite spectral-finite difference method [14, 40] have also been developed to solve the Fokker-Planck problems. Tang et al. [40] developed a mixed Hermite spectral-finite difference method, i.e., the Hermite spectral approximation in the velocity direction and finite-difference in the x-direction, for solving the Fokker-Planck equation with finite boundaries in space. The advantages for using Hermite basis in velocity are the following: (i) they form a complete system; (ii) they have correct natural boundary conditions in velocity space $-\infty < v < \infty$; and (iii) they lead to the tridiagonal structure of the coupling system. Spectral methods based on Hermite functions have been implemented before but were dismissed because of their poor resolution properties [19, 18]. However, recent works of Tang [39] and Holloway et al. [26, 36] suggest that with proper selection of the *scaling factors* the Hermite basis can be quite competitive when modeling functions with Gaussian-shaped profiles. In solving the Vlasov equations [18, 27], it was found that without careful velocity scaling of the Hermite functions, spectral expansions with 500 to 1500 Hermite modes are required to achieve only moderate accuracy levels. For plasma kinetics simulation, a recent paper by Schumer and Holloway [36] indicates that the Hermite based spectral methods are very efficient and extremely stable when velocity scaling and symmetric weighting are used.

While the Legendre- and Chebyshev-spectral approximations for PDEs in bounded domains have achieved great success and popularity in recent years (see, e.g., [3, 11, 16, 19]), spectral approximations for PDEs in *unbounded domains* have received only limited attention. Some earlier works on the convergence analysis of spectral methods in unbounded domains have been given by Funaro and Kavian [17], Guo [23] (on Hermite spectral approximations); by Mavriplis [32], Shen [37] (on Laguerre approximations); by Boyd [6], Grosch and Orszag [20] (on rational polynomial approximations). Although Hermite-spectral approximations have been used successfully in approximating the solutions to the Fokker-Planck equations (and also the Vlasov equations), there has been little convergence analysis for these numerical schemes. The main objective of this work is to provide a rigorous theoretical analysis for this class of spectral methods. For ease of notation, we consider the following normalized form of the Fokker-Planck equation with boundary and initial conditions:

(1.6)
$$
\begin{aligned}
&\frac{\partial W}{\partial t} + v\frac{\partial W}{\partial x} - \beta\frac{\partial}{\partial v}(vW) + F(x)\frac{\partial W}{\partial v} - \beta\mu\frac{\partial^2 W}{\partial v^2} = 0, \quad |x| < Y,\ |v| < \infty,\ t > 0,\\
&W(-Y, v, t) = b_L(v, t) \qquad \text{for} \quad v \ge 0, \quad t > 0,\\
&W(Y, v, t) = b_R(v, t) \qquad \text{for} \quad v \le 0, \quad t > 0,\\
&W(x, v, 0) = w(x, v) \qquad |x| \le Y, \quad |v| < \infty\,.
\end{aligned}
$$

Since problem (1.6) is hyperbolic-like in the x-direction, we will adopt the upwinding approximations for the term involving $\partial_x W$. Three types of combined Hermite spectral-upwinding difference schemes for (1.6) will be constructed and analyzed. Roughly speaking, the main result of this paper is the following: If the solutions of (1.6) decay exponentially to zero as $|v| \to \infty$, then the error between the exact solution W and the mixed spectral-difference solution satisfies

$$
||\mathsf{Error}|| = \mathcal{O}(\Delta t^\alpha + \Delta x) + \mathcal{O}(N^{-\gamma}) \tag{1.7}
$$

where Δt and Δx are stepsizes in the time- and x-directions respectively, N is the number of the basis functions used in the v-direction, $\alpha = 1$ or 2 depending on the order of the truncation errors associated with the finite difference approximations in time, and $\gamma > 0$ is a large number depending on the regularity of the exact solution of (1.6). In order to obtain the first part of the error bounds $\mathcal{O}(\Delta t^\alpha + \Delta x)$, we will do the following:

- Use the energy-type methods to deal with the hyperbolic system induced by the spectral approximation in velocity:
$$
\frac{\partial \mathbf{f}}{\partial t} + A\frac{\partial \mathbf{f}}{\partial x} = B\mathbf{f} + \mathbf{G}\,.
$$

With the classical energy analysis, we can show that the errors for suitable numerical approximations should be bounded by the truncation error times $e^{\|B\|_* t}$ with $\|B\|_*$ being some norm of B. The problem is that the norm for B may be proportional to N, the number of expansion terms for the spectral expansions. As a consequence, the classical results may not be applied directly. Instead some special treatment for the matrix B should be employed in order to obtain appropriate energy estimates.

- Use some new estimates developed in the next section to bound the coefficients of the Hermite expansions. In order to estimate the spectral convergence rate, one of the key ingredients is the use of the approximation theory results of Lubinsky et al. [29, 30].

The contents of this paper are organized as follows. In Section 2, we establish some results on the Hermite approximation. In Section 3, we consider several discrete hyperbolic systems and the properties of their solutions. Results in these two sections will play important roles in the error analysis. Then we construct the combined Hermite spectral-upwinding difference schemes and prove their convergence rates in Section 4. Numerical experiments are carried out in Section 5, which are used to verify the theoretical results obtained in Section 4. Some discussions on the selections of the scaling factors in Hermite functions are also included in this section. We point out that the theoretical results on the Hermite approximation and some techniques developed in this paper are also useful for analyzing other problems in unbounded domains.

2. Some results on Hermite approximation

Based on the choice of the weight function $\omega(v)$ there exist several kinds of Hermite approximations. The first one is to use the standard Hermite polynomials as the base functions (see Szegö [38], Gottlieb and Orszag [19], Canuto et al. [11], Bernardi and Maday [3], and Guo [22]). In this case, $\omega(v) = e^{-v^2}$. Recently, Guo [23] established some approximation results in the corresponding weighted Sobolev space, which were successfully used in the analysis of the Hermite spectral method for some nonlinear problems. Funaro and Kavian employed the Hermite functions as the basis functions with the weight function $\omega(v) = e^{v^2/4}$. They also derived some approximation results important in the analysis of the related Hermite spectral method for differential equations. Tang et al. [39] and Tang [40] considered orthogonal systems with the weight functions $\omega(v) = e^{\alpha^2 v^2}$ and $\omega(v) \equiv 1$, respectively. The orthogonal systems have been used for numerical simulations on certain differential equations. In general, the choice of $\omega(v)$ depends on the asymptotic behaviour of the solutions of the considered problems. In many problems arising in quantum mechanics and statistical physics, the solutions decay exponentially as $|v| \to \infty$. In this case, it is reasonable to take the basis functions as those used in Funaro and Kavian [17], or as in Tang et al. [40].

In this paper, *we will confine our work to the case* $\omega(v) = e^{\alpha^2 v^2}$ *with* $\alpha > 0$. We begin by introducing some notation. Let

$$L^2_\omega(\mathbf{R}) = \Big\{ u \mid u \text{ is measurable and } \|u\|_{\omega,\mathbf{R}} < \infty \Big\}$$

be equipped with the norm

$$\|u\|_{\omega,\mathbf{R}} = \left(\int_{\mathbf{R}} u^2(v)\omega(v)dv\right)^{\frac{1}{2}}.$$

The associated inner product is

$$(u,w)_{\omega,\mathbf{R}} = \int_{\mathbf{R}} u(v)w(v)\omega(v)dv.$$

For any nonnegative integer m,

$$H^m_\omega(\mathbf{R}) = \left\{u \;\middle|\; \frac{d^k u}{dv^k} \in L^2_\omega(\mathbf{R}), \quad 0 \le k \le m\right\}$$

with the following semi-norm and norm,

$$|u|_{m,\omega,\mathbf{R}} = \left\|\frac{\partial^m u}{\partial v^m}\right\|_{\omega,\mathbf{R}}, \qquad \|u\|_{m,\omega,\mathbf{R}} = \left(\sum_{k=0}^{m} |u|^2_{k,\omega,\mathbf{R}}\right)^{\frac{1}{2}}.$$

In particular, $\|u\|_{0,\omega,\mathbf{R}} = \|u\|_{\omega,\mathbf{R}}$. Moreover, for any $r > 0$, we define the space $H^r_\omega(\mathbf{R})$ and its norm $\|u\|_{r,\omega,\mathbf{R}}$ by space interpolation as in Adams [2].

Now let $H_n(v)$ be the Hermite polynomial of degree n,

$$H_n(v) = (-1)^n e^{v^2}\frac{d^n}{dv^n}(e^{-v^2}). \tag{2.8}$$

The generalized Hermite function $\widetilde{H}_n(v)$ is given by

$$\widetilde{H}_n(v) = d_n H_n(\alpha v)e^{-\alpha^2 v^2}, \qquad \alpha > 0, \qquad n \ge 0, \tag{2.9}$$

where $d_n = 1/\sqrt{2^n n!}$. The function $\widetilde{H}_n(v)$ is the n-th eigenfunction of the following singular Liouville problem:

$$\frac{d}{dv}\left(e^{-\alpha^2 v^2}\frac{d}{dv}\left(e^{\alpha^2 v^2}u(v)\right)\right) + \lambda u(v) = 0, \qquad v \in \mathbf{R}. \tag{2.10}$$

The corresponding eigenvalues are $\lambda_n = 2\alpha^2 n$. It can be shown that, for all $n \ge 0$,

$$\begin{aligned}
&\alpha v\widetilde{H}_n(v) = \sqrt{\frac{n+1}{2}}\widetilde{H}_{n+1}(v) + \sqrt{\frac{n}{2}}\widetilde{H}_{n-1}(v),\\
&\frac{d\widetilde{H}_n(v)}{dv} = -\alpha\sqrt{2(n+1)}\widetilde{H}_{n+1}(v),\\
&v\frac{d\widetilde{H}(v)}{dv} = -\sqrt{(n+1)(n+2)}\widetilde{H}_{n+2}(v) - (n+1)\widetilde{H}_n(v),\\
&\frac{d^2\widetilde{H}_n(v)}{dv^2} = 2\alpha^2\sqrt{(n+1)(n+2)}\widetilde{H}_{n+2}(v),
\end{aligned} \tag{2.11}$$

where $\widetilde{H}_j(v) \equiv 0$ for $j < 0$. The set of functions $\widetilde{H}_n(v)$ is the $L^2_\omega(\mathbf{R})$-orthogonal system, namely,

$$\int_{\mathbf{R}} \widetilde{H}_m(v)\widetilde{H}_n(v)\omega(v)dv = \frac{\sqrt{\pi}}{\alpha}\delta_{m,n} \tag{2.12}$$

where $\delta_{m,n}$ is the Kronecker delta. By the second recurrence relation in (2.11), the set of derivatives for $\widetilde{H}_n(v)$ is also an orthogonal system, i.e.,

$$\int_{\mathbf{R}} \frac{d\widetilde{H}_m(v)}{dv}\frac{d\widetilde{H}_n(v)}{dv}\omega(v)dv = 2\alpha(n+1)\sqrt{\pi}\delta_{m,n}. \tag{2.13}$$

For any $u \in L^2_\omega(\mathbf{R})$, we can expand u in the following form:

$$u(v) = \sum_{n=0}^{\infty} \widehat{u}_n \widetilde{H}_n(v) \tag{2.14}$$

with the Hermite coefficients

$$\widehat{u}_n = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbf{R}} u(v)\widetilde{H}_n(v)\omega(v)dv, \qquad n \geq 0.$$

We now turn to the approximation theory of the Hermite approximation. Let N be any positive integer and $\mathbb{P}_N$ be the set of polynomials of degree at most N. Define

$$V_N := \left\{ q(v)e^{-\alpha^2 v^2} \;\middle|\; q(v) \in \mathbb{P}_N \right\}. \tag{2.15}$$

To analyze the spectral convergence property for the Hermite spectral method, the following inverse inequalities and imbedding inequalities are needed.

Lemma 2.1. *For any $\phi \in V_N$,*

$$|\phi|_{1,\omega,\mathbf{R}} \leq \alpha\sqrt{2(N+1)}\|\phi\|_{\omega,\mathbf{R}}.$$

Proof. Clearly $\phi \in L^2_\omega(\mathbf{R})$. Since $\phi \in V_N$, there exist coefficients $\widehat{\phi}_n$ such that

$$\phi(v) = \sum_{n=0}^{N} \widehat{\phi}_n \widetilde{H}_n(v).$$

This, together with (2.12)–(2.13), yields

$$|\phi|^2_{1,\omega,\mathbf{R}} = 2\alpha\sqrt{\pi}\sum_{n=0}^{N}(n+1)\widehat{\phi}_n^2 \leq 2\alpha\sqrt{\pi}(N+1)\sum_{n=0}^{N}\widehat{\phi}_n^2 = 2\alpha^2(N+1)\|\phi\|^2_{\omega,\mathbf{R}}$$

which implies the desired result. □

Lemma 2.2. *For any $u \in H^1_\omega(\mathbf{R})$,*

$$\begin{aligned}\|u\|_{\omega,\mathbf{R}} &\leq \frac{\sqrt{2}}{\alpha}|u|_{1,\omega,\mathbf{R}},\\ \|vu\|_{\omega,\mathbf{R}} &\leq \frac{1}{\alpha^2}|u|_{1,\omega,\mathbf{R}}.\end{aligned}$$

Proof. Since $u \in H^1_\omega(\mathbf{R})$, $u(v)\omega^{1/2}(v) \to 0$ as $|v| \to \infty$. By integration by parts,

$$\begin{aligned}\left|\int_{\mathbf{R}} vu^2(v)\omega(v)dv\right| &= \frac{1}{2\alpha^2}\left|\int_{\mathbf{R}} u^2(v)d\omega(v)\right|\\ &= \frac{1}{\alpha^2}\left|\int_{\mathbf{R}} u(v)\frac{du(v)}{dv}\omega(v)dv\right| \leq \frac{1}{\alpha^2}\|u\|_{\omega,\mathbf{R}}|u|_{1,\omega,\mathbf{R}}.\end{aligned} \tag{2.16}$$

The boundedness of the last term implies that $vu^2(v)\omega(v) \to 0$ as $|v| \to \infty$. As a result, using integration by parts gives

$$\begin{aligned}\|vu\|^2_{\omega,\mathbf{R}} &= \frac{1}{2\alpha^2}\int_{\mathbf{R}} vu^2(v)d\omega(v)\\ &= -\frac{1}{2\alpha^2}\int_{\mathbf{R}} u^2(v)\omega(v)dv - \frac{1}{\alpha^2}\int_{\mathbf{R}} vu(v)\frac{du(v)}{dv}\omega(v)dv,\end{aligned}$$

whence

$$\|vu\|^2_{\omega,\mathbf{R}} + \frac{1}{2\alpha^2}\|u\|^2_{\omega,\mathbf{R}} \leq \frac{1}{\alpha^2}\|vu\|_{\omega,\mathbf{R}}|u|_{1,\omega,\mathbf{R}}.$$

Then the desired results follow from the above inequality and (2.16). □

Lemma 2.3. *For any $u \in H^1_\omega(\mathbf{R})$ and all $v \in \mathbf{R}$,*

$$|u(v)| \le \sqrt{2} e^{-\alpha|v|} \|u\|_{1,\omega,\mathbf{R}}.$$

Proof. By Lemma 2.2,

$$\begin{aligned} u^2(v)e^{\alpha^2 v^2} &= \int_{-\infty}^{v} \frac{d}{dv}\Big(u^2(v)e^{\alpha^2 v^2}\Big) dv \\ &= 2\int_{-\infty}^{v} u\frac{du}{dv} e^{\alpha^2 v^2} dv + 2\alpha^2 \int_{-\infty}^{v} vu^2 e^{\alpha^2 v^2} dv \\ &\le \|u\|^2_{1,\omega,\mathbf{R}} + \alpha^4 \|vu\|^2_{\omega,\mathbf{R}} + \|u\|^2_{\omega,\mathbf{R}} \\ &\le 2\|u\|^2_{1,\omega,\mathbf{R}}. \end{aligned}$$

This completes the proof of this lemma. ☐

Next we consider the orthogonal projections. The $L^2_\omega(\mathbf{R})$-orthogonal projection $P_N : L^2_\omega(\mathbf{R}) \to V_N$ is a mapping such that, for any $u \in L^2_\omega(\mathbf{R})$,

$$(u - P_N u, \phi)_{\omega,\mathbf{R}} = 0, \qquad \forall \phi \in V_N.$$

Equivalently,

$$P_N u(v) = \sum_{n=0}^{N} \widehat{u}_n \widetilde{H}_n(v).$$

We also introduce the operator A as

$$Au(v) = -\frac{d}{dv}\left(e^{-\alpha^2 v^2}\frac{d}{dv}\Big(u(v)e^{\alpha^2 v^2}\Big)\right). \tag{2.17}$$

It follows from Lemma 2.2 that A is a continuous mapping from $H^2_\omega(\mathbf{R})$ to $L^2_\omega(\mathbf{R})$. Let c be a generic positive constant independent of N, which may be different in different places.

Theorem 2.1. *For any $u \in H^r_\omega(\mathbf{R})$ and $r \ge 0$,*

$$\|u - P_N u\|_{\omega,\mathbf{R}} \le c(\alpha^2 N)^{-\frac{r}{2}} \|u\|_{r,\omega,\mathbf{R}}. \tag{2.18}$$

Proof. It follows from the orthogonal relation (2.12) and the Hermite expansion (2.14) that

$$\|u - P_N u\|^2_{\omega,\mathbf{R}} = \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} \widehat{u}_N^2. \tag{2.19}$$

We first consider any even integer r. Then by the singular Liouville equation (2.10) and integration by parts,

$$\begin{aligned} \int_{\mathbf{R}} u(v)\widetilde{H}_n(v)\omega(v)dv &= \frac{1}{2\alpha^2 n}\int_{\mathbf{R}} A\widetilde{H}_n(v)u(v)\omega(v)dv \\ &= -\frac{1}{2\alpha^2 n}\int_{\mathbf{R}} \frac{d}{dv}\Big(u(v)\omega(v)\Big)\frac{d}{dv}\Big(\widetilde{H}_n(v)\omega(v)\Big)\omega^{-1}(v)dv \\ &= \frac{1}{2\alpha^2 n}\int_{\mathbf{R}} Au(v)\widetilde{H}_n(v)\omega(v)dv \\ &= \cdots \\ &= (2\alpha^2 n)^{-\frac{r}{2}}\int_{\mathbf{R}} A^{\frac{r}{2}}u(v)\widetilde{H}_n(v)\omega(v)dv. \end{aligned} \tag{2.20}$$

Consequently

$$|\widehat{u}_n| = \frac{\alpha}{\sqrt{\pi}}(2\alpha^2 n)^{-\frac{r}{2}} \left| \int_{\mathbf{R}} A^{\frac{r}{2}} u(v) \widetilde{H}_n(v)\omega(v)dv \right| . \tag{2.21}$$

Furthermore

$$\begin{aligned} \|v - P_N v\|^2_{\omega,\mathbf{R}} &\leq c(\alpha^2 n)^{-r} \sum_{n=N+1}^{\infty} \left| \int_{\mathbf{R}} A^{\frac{r}{2}} u(v) \widetilde{H}_n(v)\omega(v)dv \right| \\ &\leq c(\alpha^2 n)^{-r} \|A^{\frac{r}{2}} u\|^2_{\omega,\mathbf{R}} \leq c(\alpha^2 n)^{-r} \|u\|^2_{r,\omega,\mathbf{R}} \, . \end{aligned} \tag{2.22}$$

Next, let r be any odd integer. Using Liouville equation (2.10) and arguments similar to above gives

$$\begin{aligned} \int_{\mathbf{R}} u(v)\widetilde{H}_n(v)\omega(v)dv &= (2\alpha^2 n)^{-\frac{r-1}{2}} \int_{\mathbf{R}} A^{\frac{r-1}{2}} u(v)\widetilde{H}_n(v)\omega(v)dv \\ &= -(2\alpha^2 n)^{-\frac{r+1}{2}} \int_{\mathbf{R}} \frac{d}{dv}\Big(A^{\frac{r-1}{2}} u(v)\omega(v)\Big) \frac{d}{dv}\Big(\widetilde{H}_n(v)\omega(v)\Big)\omega^{-1}(v)dv \, . \end{aligned} \tag{2.23}$$

By virtue of the first two recurrence equations in (2.11), we have

$$\begin{aligned} \omega(v)^{-1}\frac{d}{dv}(\widetilde{H}_n(v)\omega(v)) &= 2\alpha^2 v\widetilde{H}_n(v) - \alpha\sqrt{2(n+1)}\widetilde{H}_{n+1}(v) \\ &= \alpha\sqrt{2n}\widetilde{H}_{n-1}(v), \\ \omega(v)^{-1}\frac{d}{dv}\Big(A^{\frac{r-1}{2}} u(v)\omega(v)\Big) &= \frac{d}{dv}A^{\frac{r-1}{2}} u(v) + 2\alpha^2 v A^{\frac{r-1}{2}} u(v) \, . \end{aligned} \tag{2.24}$$

Substituting the above two results into (2.23) and using Lemma 2.2 leads to the same result as (2.22) for the case that r is odd. Therefore, we have proved the inequality (2.18) when r is an integer. The inequality (2.18) can be established for any $r \geq 0$ by using space interpolation. $\square$

Theorem 2.2. *For any $u \in H^r_\omega(\mathbf{R})$ and $0 \leq \mu \leq r$,*

$$\|u - P_N u\|_{\mu,\omega,\mathbf{R}} \leq c(\alpha^2 N)^{\frac{\mu}{2}-\frac{r}{2}} \|u\|_{r,\omega,\mathbf{R}}.$$

Proof. By space interpolation, we only have to use induction to prove the conclusion for any integer μ. Obviously Theorem 2.1 implies the desired result with $\mu = 0$. Now assume it is true for $\mu - 1$, which yields

$$\left\| \frac{du}{dv} - P_N\frac{du}{dv} \right\|_{\mu-1,\omega,\mathbf{R}} \leq c(\alpha^2 N)^{\frac{\mu-r}{2}} \left\| \frac{du}{dv} \right\|_{r-1,\omega,\mathbf{R}} \leq c(\alpha^2 N)^{\frac{\mu-r}{2}} \|u\|_{r,\omega,\mathbf{R}}. \tag{2.25}$$

It follows from the triangular inequality that

$$\begin{aligned} \|u - P_N u\|_{\mu,\omega,\mathbf{R}} \leq{} & \left\| \frac{du}{dv} - P_N\frac{du}{dv} \right\|_{\mu-1,\omega,\mathbf{R}} \\ & + \left\| P_N\frac{du}{dv} - \frac{d}{dv}(P_N u) \right\|_{\mu-1,\omega,\mathbf{R}} + \|u - P_N u\|_{\omega,\mathbf{R}}. \end{aligned} \tag{2.26}$$

Using the second equation of (2.11) gives

$$P_N \frac{du}{dv}(v) = -\alpha \sum_{n=1}^{N} \sqrt{2n}\widehat{u}_{n-1}\widetilde{H}_n(v),$$

$$\frac{d}{dv}(P_N u(v)) = -\alpha \sum_{n=1}^{N+1} \sqrt{2n}\widehat{u}_{n-1}\widetilde{H}_n(v)\,.$$

As a consequence,

$$P_N \frac{du}{dv}(v) - \frac{d}{dv}(P_N u(v)) = \alpha\sqrt{2(N+1)}\widehat{u}_N \widetilde{H}_{N+1}(v). \tag{2.27}$$

Next, using Theorem 2.1 yields

$$\frac{\sqrt{\pi}}{\alpha}|\widehat{u}_N|^2 \le \frac{\sqrt{\pi}}{\alpha}\sum_{n=N}^{\infty}|\widehat{u}_N|^2 \le \|u - P_{N-1}u\|^2_{\omega,\mathbf{R}} \le c(\alpha^2 N)^{-r}\|u\|^2_{r,\omega,\mathbf{R}}. \tag{2.28}$$

Moreover, it follows from Lemma 2.1 and (2.12) that

$$\|\widetilde{H}_{N+1}\|^2_{\mu-1,\omega,\mathbf{R}} \le c(\alpha^2 N)^{\mu-1}\|\widetilde{H}_{N+1}\|^2_{\omega,\mathbf{R}} \le \frac{c}{\alpha}(\alpha^2 N)^{\mu-1}. \tag{2.29}$$

Combining the results (2.27)–(2.29) leads to

$$\left\| P_N \frac{du}{dv} - \frac{d}{dv}(P_N u)\right\|^2_{\mu-1,\omega,\mathbf{R}} \le c(\alpha^2 N)^{\mu-r}\|u\|^2_{r,\omega,\mathbf{R}}. \tag{2.30}$$

By inserting (2.25) and (2.30) into (2.26), and also by applying Theorem 2.1 to the last term of (2.26), we complete procedure for the induction. $\square$

In order to obtain the optimal error estimation of Hermite approximation to differential equations, we often need to compare the numerical solutions with the $H^n_\omega(\mathbf{R})$-orthogonal projections of the exact solutions. The $H^m_\omega(\mathbf{R})$-orthogonal projection $P^m_N : H^m_\omega(\mathbf{R}) \to V_N$ is a mapping such that, for any $u \in H^1_\omega(\mathbf{R})$,

$$\left(\frac{d^m}{dv^m}(u - P^1_N u), \frac{d^m}{dv^m}\phi\right)_{\omega,\mathbf{R}} = 0, \qquad \forall \phi \in V_N. \tag{2.31}$$

Now, let $m = 1$ and assume that

$$P^1_N u(v) = \sum_{n=0}^{N} a_n \widetilde{H}_n(v)\,.$$

By the second equation of (2.11),

$$\frac{d}{dv}P^1_N u(v) = -\alpha\sum_{n=0}^{N}\sqrt{2(n+1)}a_n\widetilde{H}_{n+1}(v)\,.$$

Similarly

$$\frac{d}{dv}u(v) = -\alpha\sum_{n=0}^{\infty}\sqrt{2(n+1)}\widehat{u}_n\widetilde{H}_{n+1}(v).$$

By substituting the above two equalities into (2.31) and taking $\phi = \widetilde{H}_n(v)$, $0 \le n \le N$, we know from the second equation of (2.11) and (2.12) that $a_n = \hat{u}_n$ for all $0 \le n \le N$. It means that P^1_N is exactly the same as P_N. It is also true for P^m_N. So it suffices to compare the numerical solutions with the $L^2_\omega(\mathbf{R})$-orthogonal projections of the exact solutions in the numerical analysis of Hermite spectral approximation to differential equations of any order. This feature is one of advantages of the Hermite function (2.9).

Remark 2.1. If $\alpha = \frac{1}{2}$, then the $\widetilde{H}_n(v)$ becomes the Hermite function as discussed in Funaro and Kavian [17]. Theorem 2.1 generalizes the corresponding results in [17], while other results in this section are new, which make the use of the method of Funaro and Kavian possible for more general problems.

3. Some results on discrete hyperbolic systems

In this section, we investigate some discrete hyperbolic systems arising in the combined Hermite spectral-upwinding difference schemes for (1.6). Without loss of generality, we assume that the solution interval in x is $[-1,1]$, i.e., $Y = 1$ in (1.6). To begin with, we introduce some notation useful in our error analysis. Let $I = (-1,1)$ and $h = 1/M$, with M a fixed positive integer. Let

$$I_h = \left\{ x = jh \;\middle|\; -M+1 \le j \le M-1 \right\} \quad \text{and} \quad \bar{I}_h = I_h \cup \{-1,1\}.$$

For any scalar functions $u, w \in C(\bar{I})$, the discrete inner product and the discrete norm are defined by

$$(u,w)_h = h \sum_{x \in I_h} u(x)v(x), \qquad \|u\|_h = (u,u)_h^{\frac{1}{2}}.$$

For any vector functions $\mathbf{u} = [u_0, \ldots, u_N]^T$ and $\mathbf{w} = [w_0, \ldots, w_N]^T$

$$(\mathbf{u},\mathbf{w})_h = h \sum_{x \in I_h} \mathbf{u}^T(x)\mathbf{w}(x), \qquad \|\mathbf{u}\|_h = (\mathbf{u},\mathbf{u})_h^{\frac{1}{2}}.$$

Next, let τ be the mesh size of the variable t,

$$Q_\tau = \left\{ t = k\tau \;\middle|\; 1 \le k \le \left[\frac{T}{\tau}\right] \right\} \quad \text{and} \quad \bar{Q}_\tau = Q_\tau \cup \{0\}.$$

We shall use the following notation:

$$\begin{aligned}
\Delta_x \mathbf{u}(x,t) &= \frac{1}{h}(\mathbf{u}(x+h,t) - \mathbf{u}(x,t)), \qquad & \nabla_x \mathbf{u}(x,t) &= \Delta_x \mathbf{u}(x-h,t),\\
\Delta_t \mathbf{u}(x,t) &= \frac{1}{\tau}(\mathbf{u}(x,t+\tau) - \mathbf{u}(x,t)), \qquad & \nabla_t \mathbf{u}(x,t) &= \Delta_t \mathbf{u}(x,t-\tau),\\
\bar{u}(x,t) &= \frac{1}{2}(\mathbf{u}(x,t) + \mathbf{u}(x,t+\tau)), & &
\end{aligned} \tag{3.32}$$

where the first four represent the usual forward or backward difference quotients, and the last one is the average in time. Let λ_n be $(N+1)$ distinct real numbers, arranged as

$$\lambda_0 < \lambda_1 < \cdots < \lambda_q < 0 < \lambda_{q+1} < \cdots < \lambda_N.$$

We then split the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_0, \lambda_1, \cdots, \lambda_N)$ into the positive and nagative parts respectively, i.e., $\Lambda = \Lambda^+ + \Lambda^-$, where

$$\Lambda^\pm = \mathrm{diag}\left(\lambda_0^\pm, \lambda_1^\pm, \cdots, \lambda_N^\pm\right)$$

with $\lambda^+ = \max(\lambda, 0)$ and $\lambda^- = \min(\lambda, 0)$. We also denote D_1 and D_2 as the following constant matrices:

(3.33)
$$D_1 = \mathrm{diag}[\underbrace{1,\ldots,1}_{q+1},0,\ldots,0], \qquad D_2 = \mathrm{diag}[0,\ldots,0,\underbrace{1,\ldots,1}_{N-q}].$$

The main purpose of this section is to provide an energy-type analysis for the numerical approximations of the hyperbolic system of the following type:

$$\frac{\partial \mathbf{f}}{\partial t} + A\frac{\partial \mathbf{f}}{\partial x} = B\mathbf{f} + \mathbf{G}\,. \tag{3.34}$$

In dealing with the above equation, the matrix B requires some special attention *since the matrix norm of B may be dependent of N,* the number of terms for the Hermite spectral expansion. We make two assumptions on B:

- (H_1): The 2-norm of the coefficient matrix B is uniformly bounded with respect to N. Namely, there exists a constant C_1, *independent of* N, such that for any vector $\mathbf{u} \in \mathbf{R}^{(N+1)}$
$$(\mathbf{u}, B\mathbf{u})_h \le C_1 \|\mathbf{u}\|_h^2\,.$$
- (H_2): The 2-norm of the coefficient matrix B is not uniformly bounded with respect to N, but instead there exist a constant C_2, *independent of* N, and a constant d_N, such that for any vectors $\mathbf{u}, \mathbf{v} \in \mathbf{R}^{(N+1)}$
$$(\mathbf{u}, (d_N I + B)\mathbf{v})_h \le \frac{d_N}{2}\Big(\|\mathbf{u}\|_h^2 + \|\mathbf{v}\|_h^2\Big) + C_2\Big(\|\mathbf{u}\|_h^2 + \|\mathbf{v}\|_h^2\Big).$$

For implicit schemes below, it can be verified that the assumption (H_1) is satisfied, i.e., the standard 2-norm of the matrix B is bounded. However, for an explicit scheme, the 2-norm of the matrix B is no longer bounded, which gives some difficulties in the energy-method analysis. This is the reason that we propose the assumption (H_2). It will be verified in next section that the matrix B associated with the explicit scheme satisfies the assumption (H_2).

3.1. **Implicit scheme I.** Let $\mathbf{f}(x,t)$ be a vector function with the components $f_n(x,t)$, $0 \le n \le N$, defined on $\bar{I}_h \times \bar{Q}_\tau$. The first discrete hyperbolic system is as follows:

(3.35)
$$\begin{cases} \Delta_t \mathbf{f}(x,t) + \dfrac{1}{\alpha}\Lambda^- \Delta_x \bar{\mathbf{f}}(x,t) + \dfrac{1}{\alpha}\Lambda^+ \nabla_x \bar{\mathbf{f}}(x,t) = B(x)\bar{\mathbf{f}}(x,t) + \bar{\mathbf{G}}(x,t), \\ \hfill x \in I_h, t \in \bar{Q}_\tau, \\ D_1\mathbf{f}(1,t) = \mathbf{g}_1(t), \qquad D_2\mathbf{f}(-1,t) = \mathbf{g}_2(t), \quad t \in Q_\tau, \\ \mathbf{f}(x,0) = \mathbf{f}_0(x), \hfill x \in \bar{I}_h, \end{cases}$$

where $\alpha > 0$ is a positive constant, $B(x)$ is a given matrix dependent on x, $\mathbf{G}(x,t)$ is a given source term, $\bar{\mathbf{f}}$ and $\bar{\mathbf{G}}$ are the averages defined by (3.32), and $\mathbf{g}_1(t)$ and $\mathbf{g}_2(t)$ are given vector functions with the following form:

$$\mathbf{g}_1(t) = [g_{1,0},\ldots,g_{1,q},0,\ldots,0]^T,$$
$$\mathbf{g}_2(t) = [0,\ldots,0,g_{2,q+1},\ldots,g_{2,N}]^T.$$

Clearly (3.35) is only a usual upwinding scheme which is implicit in time. There are many existing results concerning the continuous dependence of $\|\mathbf{f}(t)\|_h$ on $\|\mathbf{B}\|_h$,

$\|\mathbf{G}(t)\|_h$, $\|\mathbf{g}_1(t)\|_h$, $\|\mathbf{g}_2(t)\|_h$ and $\|\mathbf{f}_0(t)\|_h$ (see, e.g., Thomée [42] and Guo [21]). It is easy to establish the following result.

Theorem 3.1. *Let $\mathbf{f}(x,t)$ be the solution of* (3.35) *and time step τ be sufficiently small. If the matrix B satisfies the assumption* (H_1)*, then, for all $t \in \bar{Q}_\tau$,*

$$\|\mathbf{f}(t)\|_h^2 \le ce^{ct}\mathcal{G}_{h,\tau}(t),$$

where c is a positive constant independent of h, τ and N, and $\mathcal{G}_{h,\tau}(t)$ is defined by

$$\mathcal{G}_{h,\tau}(t) = \|\mathbf{f}_0\|_h^2 + \sum_{\substack{\eta\in\bar{Q}_\tau\\ \eta\le t}} \Big(\|\mathbf{G}(\eta)\|_h^2 + \|\Lambda^-\mathbf{g}_1(\eta)\|^2 + \|\Lambda^+\mathbf{g}_2(\eta)\|^2\Big). \tag{3.36}$$

3.2. **Implicit scheme II.** The second finite-difference system to be considered is of the following form:

$$\begin{cases} \nabla_t\mathbf{f}(x,t) + \dfrac{1}{\alpha}\Lambda^-\Delta_x\mathbf{f}(x,t) + \dfrac{1}{\alpha}\Lambda^+\nabla_x\mathbf{f}(x,t) = B(x)\mathbf{f}(x,t) + \mathbf{G}(x,t), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x \in I_h, t\in Q_\tau, \\ D_1\mathbf{f}(1,t) = \mathbf{g}_1(t), \qquad D_2\mathbf{f}(-1,t) = \mathbf{g}_2(t), \quad t\in Q_\tau, \\ \mathbf{f}(x,0) = \mathbf{f}_0(x), \qquad\qquad\qquad\qquad\qquad x\in \bar{I}_h. \end{cases} \tag{3.37}$$

In other words, this is a finite difference approximation with *backward* Euler in time and *upwinding* in space. Again with the standard energy estimates we have the following result.

Theorem 3.2. *Let $\mathbf{f}(x,t)$ be the solution of* (3.37) *and time step τ be sufficiently small. If the matrix B satisfies the assumption* (H_1)*, then for all $t \in \bar{Q}_\tau$*

$$\|\mathbf{f}(t)\|_h^2 \le ce^{ct}\mathcal{G}_{h,\tau}(t)$$

where $\mathcal{G}_{h,\tau}$ is defined by (3.36).

3.3. **Explicit scheme.** In this subsection we consider the following explicit system

$$\begin{cases} \Delta_t\mathbf{f}(x,t) + \dfrac{1}{\alpha}\Lambda^-\Delta_x\mathbf{f}(x,t) + \dfrac{1}{\alpha}\Lambda^+\nabla_x\mathbf{f}(x,t) = B(x)\mathbf{f}(x,t) + \mathbf{G}(x,t), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad x \in I_h, t\in \bar{Q}_\tau, \\ D_1\mathbf{f}(1,t) = \mathbf{g}_1(t), \qquad D_2\mathbf{f}(-1,t) = \mathbf{g}_2(t), \quad t\in Q_\tau, \\ \mathbf{f}(x,0) = \mathbf{f}_0(x), \qquad\qquad\qquad\qquad\qquad x\in \bar{I}_h. \end{cases} \tag{3.38}$$

This is the standard *forward* Euler approximation in time and *upwinding* in space. For any $(x,t) \in I_h\times\bar{Q}_\tau$, it follows from (3.38) that

$$\begin{aligned} \mathbf{f}(x,t+\tau) &= -\frac{\tau}{\alpha h}\Lambda^-\mathbf{f}(x+h,t) + \Big(I + \frac{\tau}{\alpha h}\Lambda^- - \frac{\tau}{\alpha h}\Lambda^+ + \tau B\Big)\mathbf{f}(x,t) \\ &\quad + \frac{\tau}{h\alpha}\Lambda^+\mathbf{f}(x-h,t) + \tau G(x,t)\,. \end{aligned} \tag{3.39}$$

Assume that (H_2) is satisfied. Then there exist d_N and C_2 such that

$$\begin{aligned} &\Big(\mathbf{f}(t+\tau), (d_N I + B)\mathbf{f}(t)\Big)_h \\ &\le \frac{d_N}{2}\Big(\|\mathbf{f}(t+\tau)\|_h^2 + \|\mathbf{f}(t)\|_h^2\Big) + C_2\Big(\|\mathbf{f}(t+\tau)\|_h^2 + \|\mathbf{f}(t)\|_h^2\Big). \end{aligned} \tag{3.40}$$

If we further assume that the generalized CFL condition

$$\frac{\tau}{\alpha h}\max_{0\le j\le N}|\lambda_j|+\tau d_N\le 1 \tag{3.41}$$

is satisfied, then we have

$$\begin{aligned}
&\Big(\mathbf{f}(t+\tau),(I+\frac{\tau}{\alpha h}\Lambda^- -\frac{\tau}{\alpha h}\Lambda^+ -\tau d_N I)\mathbf{f}(t)\Big)_h\\
\le\;& \frac{1}{2}\Big(\mathbf{f}(t+\tau),(I+\frac{\tau}{\alpha h}\Lambda^- -\frac{\tau}{\alpha h}\Lambda^+ -\tau d_N I)\mathbf{f}(t+\tau)\Big)_h\\
&+\frac{1}{2}\Big(\mathbf{f}(t),(I+\frac{\tau}{\alpha h}\Lambda^- -\frac{\tau}{\alpha h}\Lambda^+ -\tau d_N I)\mathbf{f}(t)\Big)_h .
\end{aligned}\tag{3.42}$$

It is also observed that

$$\begin{aligned}
&-\Big(\mathbf{f}(t+\tau),\Lambda^-\mathbf{f}(\bullet+h,t)\Big)_h\\
\le\;& -\frac{1}{2}\Big(\mathbf{f}(t+\tau),\Lambda^-\mathbf{f}(t+\tau)\Big)_h-\frac{1}{2}\Big(\mathbf{f}(t),\Lambda^-\mathbf{f}(t)\Big)_h+h\|\Lambda^-\bar{\mathbf{g}}_1(t)\|^2 ,
\end{aligned}\tag{3.43}$$

$$\begin{aligned}
&\Big(\mathbf{f}(t+\tau),\Lambda^+\mathbf{f}(\bullet-h,t)\Big)_h\\
\le\;& \frac{1}{2}\Big(\mathbf{f}(t+\tau),\Lambda^+\mathbf{f}(t+\tau)\Big)_h+\frac{1}{2}\Big(\mathbf{f}(t),\Lambda^+\mathbf{f}(t)\Big)_h+h\|\Lambda^+\bar{\mathbf{g}}_2(t)\|^2 .
\end{aligned}\tag{3.44}$$

Take the discrete inner product for (3.39) by multiplying it with $\mathbf{f}(x,t+\tau)$. Then by using (3.40), together with the estimates (3.42)–(3.44) we obtain

$$\begin{aligned}
\|\mathbf{f}(t+\tau)\|_h^2\le&\left(\frac{1}{2}+C_2\tau+\frac{\tau}{2}\right)\|\mathbf{f}(t+\tau)\|_h^2+\left(\frac{1}{2}+C_2\tau\right)\|\mathbf{f}(t)\|_h^2\\
&+\frac{\tau}{2}\|\mathbf{G}(t)\|_h^2+\frac{\tau}{\alpha}\Big(\|\Lambda^-\mathbf{g}_1(t)\|_h^2+\|\Lambda^+\mathbf{g}_2(t)\|_h^2\Big) .
\end{aligned}$$

We obtain the following result from the above Gronwall type inequality.

Theorem 3.3. *Assume that* (H_2) *and the generalized CFL condition* (3.41) *are satisfied. Then for all* $t\in\bar{Q}_\tau$

$$\|\mathbf{f}(t)\|_h^2\le ce^{ct}\mathcal{G}_{h,\tau}(t)$$

where $\mathcal{G}_{h,\tau}$ *is defined by* (3.36).

Remark 3.1. If $\lambda_0<\lambda_1<\cdots<\lambda_{q-1}<\lambda_q=0<\lambda_{q+1}<\cdots<\lambda_N$, then by slightly modifying the definitions of $D_1, D_2, \mathbf{g}_1$ and $\mathbf{g}_2$, we can recover all the results obtained in this section.

4. The Hermite spectral-finite difference schemes

In this section, we consider the Hermite spectral-finite difference schemes for (1.6) and their error analysis. We begin by introducing some notation. Let $\partial_t=\frac{\partial}{\partial t}, \partial_x=\frac{\partial}{\partial x}$ and

(4.1)

$$V_\omega^{m,r}=C^m\Big(I;H_\omega^r(\mathbf{R})\Big)$$

$$\text{with norm}\quad \|u(\bullet,\bullet,t)\|_{m,r,\omega}=\max_{0\le k\le m}\max_{x\in I}\|\partial_x^k u(x,\bullet,t)\|_{r,\omega,\mathbf{R}} .$$

We further define

$$
(4.2)\qquad
\begin{aligned}
\|u\|_{p,m,r,\omega} &= \max_{0\le k\le p}\max_{0\le t\le T}\|\partial_t^k u(\bullet,\bullet,t)\|_{m,r,\omega},\\
\|u\|_{\Delta,r,\omega} &= \max_{0\le t\le T}\Big(h\sum_{x\in I_h}\|u(x,\bullet,t)\|^2_{r,\omega}\Big)^{\frac12}.
\end{aligned}
$$

For $W\in L^2_\omega(\mathbf{R})$, we can expand it in Hermite functions

$$
(4.3)\qquad W(x,v,t)=\sum_{n=0}^{\infty}\widehat{W}_n(x,t)\widetilde{H}_n(v).
$$

Its truncated expansion is

$$
(4.4)\qquad W_N(x,v,t)=\sum_{n=0}^{N}\widehat{W}_n(x,t)\widetilde{H}_n(v).
$$

If $W\in C\Big(0,T;V_\omega^{0,r}\Big)$, then by Theorem 2.2 we have, for all $x\in I$, $0\le t\le T$ and $0\le\mu\le r$,

$$
(4.5)\qquad \|W(x,\bullet,t)-W_N(x,\bullet,t)\|_{\mu,\omega,\mathbf{R}}\le C(\alpha^2N)^{\frac{\mu-r}{2}}\|W(x,\bullet,t)\|_{r,\omega,\mathbf{R}}.
$$

Furthermore, it follows from (2.28) that

$$
(4.6)\qquad |\widehat{W}_n(x,t)|\le C\alpha^{1/2-r}2^{-r/2}n^{-r/2}\|W(x,\bullet,t)\|_{r,\omega,\mathbf{R}}.
$$

4.1. **Hermite spectral expansion.** By substituting the expansion (4.3) into (1.6), we obtain from the recurrence relation (2.11) that

(4.7)

$$
\begin{aligned}
&\frac{\partial\widehat{W}_n}{\partial t}+\frac{1}{\alpha}\left[\sqrt{\frac{n}{2}}\frac{\partial\widehat{W}_{n-1}}{\partial x}+\sqrt{\frac{n+1}{2}}\frac{\partial\widehat{W}_{n+1}}{\partial x}\right]=-\beta n\widehat{W}_n+\alpha\sqrt{2n}F(x)\widehat{W}_{n-1}\\
&\qquad+\beta(2\alpha^2\mu-1)\sqrt{n(n-1)}\widehat{W}_{n-2},\qquad x\in I,\ 0\le t\le T, n\ge 0,\\
&\sum_{n=0}^{\infty}\widehat{W}_n(-1,t)\widetilde{H}_n(v)=b_L(v,t),\quad \text{for}\quad v\ge 0,\ t\in(0,T],\\
&\sum_{n=0}^{\infty}\widehat{W}_n(1,t)\widetilde{H}_n(v)=b_R(v,t),\quad \text{for}\quad v\le 0,\ t\in(0,T],\\
&\widehat{W}_n(x,0)=\widehat{W}_{0,n}(x),\qquad x\in\bar{I}, n\ge 0.
\end{aligned}
$$

where $\widehat{W}_{-1}=\widehat{W}_{-2}=0$. The system (4.7) is an infinite hyperbolic system and in order to solve it we have to ignore some terms in (4.7). More precisely, let

$$
\mathcal{F}(x,t)=[\mathcal{F}_0,\mathcal{F}_1,\cdots,\mathcal{F}_N]^T;
$$

then the Hermite spectral method is to solve the following finite system of hyperbolic equations:

(4.8)

$$\begin{aligned}
&\frac{\partial \mathcal{F}_n}{\partial t} + \frac{1}{\alpha}\left[\sqrt{\frac{n}{2}}\frac{\partial \mathcal{F}_{n-1}}{\partial x} + \sqrt{\frac{n+1}{2}}\frac{\partial \mathcal{F}_{n+1}}{\partial x}\right] = -\beta n \mathcal{F}_n + \alpha\sqrt{2n}F(x)\mathcal{F}_{n-1} \\
&\qquad + \beta(2\alpha^2\mu - 1)\sqrt{n(n-1)}\mathcal{F}_{n-2}, \qquad x \in I,\ 0 \le t \le T, 0 \le n \le N, \\
&\sum_{n=0}^{N} \mathcal{F}_n(-1,t)\widetilde{H}_n(v) = b_L(v,t), \quad \text{for} \quad v \ge 0,\ t \in (0,T], \\
&\sum_{n=0}^{N} \mathcal{F}_n(1,t)\widetilde{H}_n(v) = b_R(v,t), \quad \text{for} \quad v \le 0,\ t \in (0,T], \\
&\mathcal{F}_n(x,0) = \widehat{W}_{0,n}(x), \qquad x \in \bar{I}, 0 \le n \le N.
\end{aligned}$$

where $\mathcal{F}_{-1} = \mathcal{F}_{-2} = 0, \mathcal{F}_{N+1} \equiv 0$. We are now in a position to specify the boundary conditions for $\mathcal{F}_n(\pm 1, t)$. We will do so by using the collocation idea to the second and third equations in (4.8). To this end, we first denote by λ_k the zeros of the Hermite polynomial $H_{N+1}(\lambda)$. By Szegö [38] and Timan [43], they are distinct real numbers, situated around the origin symmetrically, arranged as

$$\lambda_0 < \lambda_1 < \cdots < \lambda_{N-1} < \lambda_N, \quad \lambda_N = -\lambda_{N-n}.$$

For simplicity, we assume that N *is an odd integer*. Then $\lambda_n < 0$ for $0 \le n \le N_1 := (N-1)/2$ and $\lambda_n > 0$ for $N_1 + 1 \le n \le N$. Letting $v = \lambda_k/\alpha$ in the third equation of (4.8) gives

$$\sum_{n=0}^{N} c_k d_n H_n(\lambda_k)\mathcal{F}_n(1,t) = c_k b_R(\lambda_k/\alpha, t)e^{\lambda_k^2}, \quad 0 \le k \le N_1 \tag{4.9}$$

where

$$c_k = \left(\sum_{n=0}^{N} d_n^2 H_n^2(\lambda_k)\right)^{-\frac{1}{2}}, \qquad 0 \le k \le N. \tag{4.10}$$

It follows from Christoffel-Darboux formula (see, e.g., Abramowitz and Stegun [1]) and L'Hospital's rule that

$$c_k = \Big((N+1)d_N H_N(\lambda_k)\Big)^{-1}.$$

The above formula is useful in computation. Similarly we derive that

(4.11)

$$\sum_{n=0}^{N} c_k d_n H_n(\lambda_k)\mathcal{F}_n(-1,t) = c_k b_L(\lambda_k/\alpha, t)e^{\lambda_k^2}, \quad N_1 < k \le N.$$

Furthermore, we define the matrix U in the following way:

$$\mathbf{u}_k = [u_{0,k}, \cdots, u_{N,k}]^T\ ,\ U = [\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_N]. \tag{4.12}$$

where

$$u_{n,k} = c_k d_n H_n(\lambda_k).$$

Using the definition of U, we make a linear transformation for $\mathcal{F}$ to get $\mathbf{F} := U^T\mathcal{F}$. Now we want to obtain governing equations and boundary conditions for $\mathbf{F}$, which are also the equations for the numerical computations. Clearly, it follows from (4.9) and (4.11) that the boundary conditions for $\mathbf{F}$ are

(4.13)
$$D_1\mathbf{F}(1,t) = \Big[c_0 b_R(\lambda_0/\alpha,t)e^{\lambda_0^2}, \cdots, c_{N_1} b_R(\lambda_{N_1}/\alpha,t)e^{\lambda_{N_1}^2}, 0, \cdots, 0\Big]^T,$$
$$D_2\mathbf{F}(-1,t) = \Big[0, \cdots, 0, c_{N_1+1} b_L(\lambda_{N_1+1}/\alpha,t)e^{\lambda_{N_1+1}^2}, \cdots, c_N b_L(\lambda_N/\alpha,t)e^{\lambda_N^2}\Big]^T.$$

where the matrices D_1, D_2 are defined by (3.33). We now need to obtain governing equations for $\mathbf{F}$. Let

(4.14)
$$R = \begin{bmatrix} 0 & \alpha_1 & & & \\ \alpha_1 & 0 & \alpha_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \alpha_{N-1} & 0 & \alpha_N \\ & & & \alpha_N & 0 \end{bmatrix}, \quad S = \begin{bmatrix} s_0 & & & & \\ \gamma_1 & s_1 & & & \\ \delta_2 & \gamma_2 & s_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \delta_N & \gamma_N & s_N \end{bmatrix}$$

where

(4.15)
$$\alpha_n = \sqrt{\frac{n}{2}}, \quad s_n = -\beta n, \quad \gamma_n = \alpha\sqrt{2n}F(x), \quad \delta_n = \beta(2\alpha^2\mu - 1)\sqrt{n(n-1)}.$$

Lemma 4.1. *The matrix R satisfies the following properties:*

(i) *The eigenvalues of R are λ_k, the zeros of the Hermite polynomial $H_{N+1}(\lambda)$;*
(ii) *The eigenvectors of R corresponding to the eigenvalue λ_k are U_k, defined by* (4.12)*;*
(iii) *U is an orthogonal matrix.*

Proof. The above results can be obtained in a way similar to that provided in Tang et al. [40]. ☐

It follows from the first equations of (4.8) that

$$\frac{\partial \mathcal{F}}{\partial t} + \frac{1}{\alpha}R\frac{\partial \mathcal{F}}{\partial x} = S\mathcal{F}. \tag{4.16}$$

Using Lemma 4.1 gives

$$U^T R U = \Lambda = \Lambda^+ + \Lambda^-$$

where

$$\Lambda^- = \mathrm{diag}\Big(\lambda_0, \cdots, \lambda_{N_1}, 0, \cdots, 0\Big), \quad \Lambda^+ = \mathrm{diag}\Big(0, \cdots, 0, \lambda_{N_1+1}, \cdots, \lambda_N\Big).$$

By premultiplying (4.16) by U^T, we obtain

$$\frac{\partial \mathbf{F}}{\partial t} + \frac{1}{\alpha}\Lambda\frac{\partial \mathbf{F}}{\partial x} = B\mathbf{F}, \tag{4.17}$$

where $\mathbf{F} = U^T\mathcal{F}$ and $B = U^T S U$. Combining (4.13) and (4.17) gives the following system for $\mathbf{F}$, which will be used for numerical computation:

$$\begin{aligned} &\frac{\partial \mathbf{F}}{\partial t} + \frac{1}{\alpha}\Lambda\frac{\partial \mathbf{F}}{\partial x} = B\mathbf{F}, \\ &D_1\mathbf{F}(1,t) = V_R, \qquad D_2\mathbf{F}(-1,t) = V_L, \\ &\mathbf{F}(x,0) = U^T W^{(0)}(x), \end{aligned} \tag{4.18}$$

where

$$V_R := \left[c_0 b_R(\lambda_0/\alpha, t)e^{\lambda_0^2}, \cdots, c_{N_1} b_R(\lambda_{N_1}/\alpha, t)e^{\lambda_{N_1}^2}, 0, \cdots, 0\right]^T, \tag{4.19}$$

$$V_L := \left[0, \cdots, 0, c_{N_1+1} b_L(\lambda_{N_1+1}/\alpha, t)e^{\lambda_{N_1+1}^2}, \cdots, c_N b_L(\lambda_N/\alpha, t)e^{\lambda_N^2}\right]^T, \tag{4.20}$$

$$W^{(0)}(x) := \left[\widehat{W}_{0,0}(x), \cdots, \widehat{W}_{0,N}(x)\right]^T. \tag{4.21}$$

Here $W^{(0)}(x)$ are the coefficients of the Hermite expansion for $W(x, v, 0)$.

4.2. **Combined spectral-difference schemes.** We are now in a position to solve (4.18) numerically. Using the upwinding method introduced in the last section, the numerical solution $\mathbf{f}(x,t)$, which is the approximation for $\mathbf{F}$ in (4.17), can be determined by one of the following discrete hyperbolic systems

(4.22) Second-order in time

$$\Delta_t \mathbf{f}(x,t) + \frac{1}{\alpha}\Lambda^- \Delta_x \bar{\mathbf{f}}(x,t) + \frac{1}{\alpha}\Lambda^+ \nabla_x \bar{\mathbf{f}}(x,t) = B(x)\bar{\mathbf{f}}(x,t), \qquad x \in I_h, t \in Q_\tau,$$

(4.23) Backward Euler

$$\nabla_t \mathbf{f}(x,t) + \frac{1}{\alpha}\Lambda^- \Delta_x \mathbf{f}(x,t) + \frac{1}{\alpha}\Lambda^+ \nabla_x \mathbf{f}(x,t) = B(x)\mathbf{f}(x,t), \qquad x \in I_h, t \in Q_\tau,$$

(4.24) Forward Euler

$$\Delta_t \mathbf{f}(x,t) + \frac{1}{\alpha}\Lambda^- \Delta_x \mathbf{f}(x,t) + \frac{1}{\alpha}\Lambda^+ \nabla_x \mathbf{f}(x,t) = B(x)\mathbf{f}(x,t), \qquad x \in I_h, t \in Q_\tau.$$

In all cases, the boundary and initial conditions are the same:

$$\begin{aligned} &D_1\mathbf{f}(1,t) = V_R, \quad D_2\mathbf{f}(-1,t) = V_L, \quad t \in Q_\tau, \\ &\mathbf{f}(x,0) = U^T W^{(0)}(x), \quad x \in \bar{I}_h, \end{aligned}$$

where V_L, V_R and $W^{(0)}(x)$ are defined by (4.19)–(4.21). Then the numerical approximation of (1.6) is given by

$$W_\Delta(x,v,t) = \sum_{n=0}^{N} \mathcal{F}_{\Delta,n}(x,t)\widetilde{H}_n(v), \tag{4.25}$$

where

$$\mathcal{F}_\Delta = [\mathcal{F}_{\Delta,0}, \mathcal{F}_{\Delta,1}, \cdots, \mathcal{F}_{\Delta,N}]^T := U\mathbf{f}.$$

4.3. **Error analysis.** We now turn to the error analysis. Let

$$\begin{aligned} &\mathcal{W} := [\widehat{W}_0(x,t), \cdots, \widehat{W}_N(x,t)]^T, \\ &\mathbf{A} = \left[0, \cdots, 0, -(\sqrt{2}\alpha)^{-1}\partial_x\sqrt{N+1}\,\widehat{W}_{N+1}(x,t)\right]^T, \\ &\mathbf{g}_1 = [\sigma_1, \cdots, \sigma_{N_1}, 0, \cdots, 0]^T, \\ &\mathbf{g}_2 = [0, 0, \cdots, 0, \sigma_{N_1+1}, \cdots, \sigma_N]^T, \end{aligned} \tag{4.26}$$

where $\widehat{W}_n$ are the Hermite expansion coefficients for W, given by (4.3), and

$$\begin{aligned}\sigma_k &= -c_k \sum_{n=N+1}^{\infty} d_n H_n(\lambda_k)\widehat{W}_n(1,t), \qquad 0 \le n \le N_1,\\ \sigma_k &= -c_k \sum_{n=N+1}^{\infty} d_n H_n(\lambda_k)\widehat{W}_n(-1,t), \qquad N_1+1 \le k \le N.\end{aligned} \tag{4.27}$$

Having the above notation, we obtain from (4.7) that

$$\begin{aligned}&\frac{\partial \mathcal{W}}{\partial t} + \frac{1}{\alpha} R \frac{\partial \mathcal{W}}{\partial x} = S\mathcal{W} + U^T \mathbf{A}\,, \quad x \in I, t \in (0,T],\\ &D_1 U^T \mathcal{W}(1,t) = V_R + \mathbf{g}_1 \quad 0 < t \le T,\\ &D_2 U^T \mathcal{W}(-1,t) = V_L + \mathbf{g}_2 \quad 0 < t \le T,\\ &\mathcal{W}(x,0) = W^{(0)}(x)\,.\end{aligned} \tag{4.28}$$

Now let $\mathbf{W} = U^T \mathcal{W}$. By premultiplying (4.28) by U^T, we obtain

$$\begin{aligned}&\frac{\partial \mathbf{W}}{\partial t} + \frac{1}{\alpha} \Lambda \frac{\partial \mathbf{W}}{\partial x} = B\mathbf{W} + U^T \mathbf{A}\,, \quad x \in I, t \in (0,T],\\ &D_1 \mathbf{W}(1,t) = V_R + \mathbf{g}_1 \quad 0 < t \le T,\\ &D_2 \mathbf{W}(-1,t) = V_L + \mathbf{g}_2 \quad 0 < t \le T,\\ &\mathbf{W}(x,0) = U^T W^{(0)}(x)\,,\end{aligned} \tag{4.29}$$

where as before $B = U^T S U$.

Now at the grid points $(x,t) \in \bar{I}_h \times \bar{Q}_\tau$ we let the error between $\mathbf{F}$ and $\mathbf{W}$ be $\mathbf{e}$, i.e., $\mathbf{e} := \mathbf{W} - \mathbf{F}$, and we will estimate the error $\mathbf{e}$. We first consider the second-order (in time) scheme (4.22). It satisfies

(4.30)

$$\Delta_t \mathbf{e}(x,t) + \frac{1}{\alpha}\Lambda^- \Delta_x \bar{\mathbf{e}}(x,t) + \frac{1}{\alpha}\Lambda^+ \nabla_x \bar{\mathbf{e}}(x,t) = B\bar{\mathbf{e}} + U^T \mathbf{A} + \mathcal{T}_\Delta(x,t)\,, \qquad x \in I_h, t \in \bar{Q}_\tau,$$

$$D_1 \mathbf{e}(1,t) = \mathbf{g}_1\,, \quad D_2 \mathbf{e}(-1,t) = \mathbf{g}_2 \quad t \in \bar{Q}_\tau,$$

$$\mathbf{e}(x,0) = 0\,,$$

where $\mathcal{T}_\Delta$ are the truncation errors induced by finite difference approximations in (4.22):

$$\mathcal{T}_\Delta(x_j,t_k) \sim \tau^2 \partial_t^3 \mathbf{W}(x_j,t_k^*) + \alpha^{-1} h \Lambda \partial_x^2 \mathbf{W}(x_j^*,t_k),$$

where t_k^*, x_j^* are some intermediate values.

Lemma 4.2. *If the solution of* (1.6) *satisfies* $W \in C^3(0,T;V_\omega^{0,0}) \cap C^0(0,T;V_\omega^{2,1})$, *then for* $0 < t_k \le T$

$$\|\mathcal{T}_\Delta(\bullet,t_k)\|_h \le C\tau^2 + Ch. \tag{4.31}$$

Proof. Due to the orthogonality of U, we have

$$\|\partial_t^3 \mathbf{W}(\bullet,t)\|_h = \|\partial_t^3 \mathcal{W}(\bullet,t)\|_h\,.$$

Since the solution of (1.6) satisfies $W \in C^3(0,T;V_\omega^{0,0})$, we obtain, by observing that $\mathcal{W}$ is a vector of the Hermite expansion coefficients of W, that

$$\|\partial_t^3 \mathcal{W}(\bullet,t)\|_h \le C \max_{x\in I} \|\partial_t^3 W(x,\bullet,t)\|_{0,\omega,\mathbf{R}}\,.$$

By the notations (4.1) and (4.2), we have for any $0 < t \le T$ that

(4.32)
$$\|\partial_t^3 \mathbf{W}(\bullet, t)\|_h \le C \max_{x\in I} \|\partial_t^3 W(x, \bullet, t)\|_{0,\omega,\mathbf{R}} \le C\|\partial_t^3 W(\bullet, \bullet, t)\|_{0,0,\omega} \le C\|W\|_{3,0,0,\omega}.$$

Furthermore, using the facts that $\mathbf{W} = U^T \mathcal{W}$ and $U^T R U = \Lambda$ we have

$$\begin{aligned}(\Lambda\partial_x^2\mathbf{W})^T(\Lambda\partial_x^2\mathbf{W}) &= (\partial_x^2\mathbf{W})^T\Lambda^T\Lambda\partial_x^2\mathbf{W}\\ &= (\partial_x^2\mathcal{W})^T U\Lambda^T U^T U\Lambda U^T\partial_x^2\mathcal{W} = (R\partial_x^2\mathcal{W})^T(R\partial_x^2\mathcal{W}).\end{aligned} \tag{4.33}$$

Let $\mathbf{W}(x,t) = [w_0, w_1, \cdots, w_N]^T$. By the definition of the tri-diagonal matrix R, we obtain from (4.33) that

$$\begin{aligned}& h\sum_j\sum_{n=0}^N \lambda_n^2\partial_x^2 w_n(x_j^*, t)^2\\ &\quad = h\sum_j\sum_{n=0}^N\Big(\alpha_{n-1}\partial_x^2\widehat{W}_{n-1}(x_j^*, t) + \alpha_{n+1}\partial_x^2\widehat{W}_{n+1}(x_j^*, t)\Big)^2\\ &\quad \le Ch\sum_j\sum_{n=0}^N n\partial_x^2\widehat{W}_n(x_j^*, t)^2\end{aligned}$$

where $\alpha_n = 0$ except $\alpha_n = \sqrt{n/2}$ for $0 \le n \le N$. Since $W \in C^0(0, T; V_\omega^{2,1})$, we have from the above estimate and the second equation of (2.11) that, for any $t \in (0, T]$,

(4.34)
$$h\sum_j\sum_{n=0}^N\lambda_n^2\partial_x^2 w_n(x_j^*, t)^2 \le Ch\sum_j\|\partial_x^2 W(x_j^*, \bullet, t)\|_{1,\omega,\mathbf{R}}^2 \le C\|W\|_{0,2,1,\omega}^2.$$

Combining (4.32) and (4.34) we obtain the desired result. □

Lemma 4.3. *If the solution of* (1.6) *satisfies* $W \in C^0(0, T; V_\omega^{0,r_0})$, *where* $r_0 \ge 2$, *then*

$$\|\Lambda^-\mathbf{g}_1\| \le CN^{11/12-r_0/2}, \qquad \|\Lambda^+\mathbf{g}_2\| \le CN^{11/12-r_0/2}. \tag{4.35}$$

Proof. We will prove the first inequality in (4.35); the second one can be obtained in a similar way. Consider $|\lambda_k\sigma_k|$, $0 \le k \le N_1 = (N-1)/2$, where σ_k are defined by (4.27). By the recurrence formula for $H_n(v)$, we have

(4.36)
$$\lambda_k\sigma_k(t) = -\frac{c_k}{2}\sum_{n=N+1}^\infty d_n\widehat{W}_n(1,t)\Big(H_{n+1}(\lambda_k) + 2nH_{n-1}(\lambda_k)\Big).$$

It is known that for large n (see, e.g., Abramowitz and Stegun [1])

$$H_n(v) \sim e^{v^2/2}\frac{n!}{(\frac{n}{2})!}\cos\Big(\sqrt{(2n+1)v} - \frac{1}{2}n\pi\Big). \tag{4.37}$$

Due to the above asymptotic formula and the Stirling formula,

$$n! \sim \sqrt{2\pi n}\, n^{n+\frac{1}{2}}e^{-n}, \tag{4.38}$$

we obtain

$$d_n|H_{n+1}(\lambda_k)| \le Ce^{\lambda_k^2/2}\frac{(n+1)}{(\frac{n+1}{2})!}\sqrt{\frac{n!}{2^n}} \le Ce^{\lambda_k^2/2}\,. \tag{4.39}$$

Similarly

$$2nd_n|H_{n-1}(\lambda_k)| \le Ce^{\lambda_k^2/2}\,. \tag{4.40}$$

Furthermore, by Lemma 2.5 of Lubinsky and Moricz [30], we have, for all $N \ge 1$ and $|v| \le \sqrt{2N+2}$,

$$\left(\sum_{n=0}^{N} d_n^2 H_n^2(v)\right)^{-1} \sim \sqrt{\frac{2}{\pi(N+1)}}\, e^{\lambda_k^2} \max\Big((N+1)^{-2/3}, 1-|v|(2N+2)^{-1/2}\Big)^{-1/2}. \tag{4.41}$$

Using the following fact (see Levin and Lubinsky [29])

$$|\lambda_k| \le \sqrt{2N+2}\Big(1-N^{-\frac{2}{3}}\Big), \qquad 0 \le k \le N\,,$$

we obtain from (4.41) and the definition of c_k, (4.10), that

$$c_k \le CN^{-1/12}e^{-\lambda_k^2/2}\,. \tag{4.42}$$

For $W \in C^0(0,T;V_\omega^{0,r_0})$ and a constant $1 < q < r_0$, we use Cauchy inequality to obtain

$$\begin{aligned} |\lambda_k\sigma_k(t)| &\le CN^{-1/12}\sum_{n=N+1}^{\infty}|\widehat{W}_n(1,t)| \\ &\le CN^{-1/6}\sum_{n=N+1}^{\infty} n^{-q} + \sum_{n=N+1}^{\infty} n^q|\widehat{W}_n(1,t)|^2 \\ &\le CN^{5/6-q} + CN^{q-r_0}\|W(1,\bullet,t)\|_{r_0,\omega,\mathbf{R}}\,. \end{aligned} \tag{4.43}$$

Choose $q = r_0/2 + 5/12$. It follows from $r_0 \ge 2$ that $1 < q < r_0$. The result (4.43) implies

$$|\lambda_k\sigma_k(t)| \le CN^{5/12-r_0/2}\,.$$

Therefore, by the definition of $\mathbf{g}_1$, namely (4.26), we have

$$\big\|\Lambda^-\mathbf{g}_1(t)\big\| \le CN^{11/12-r_0/2}\,. \tag{4.44}$$

This completes the proof of Lemma 4.3. □

We further observe that if $W \in C^0(0,T;V_\omega^{1,r_1})$, then

$$\big\|U^T\mathbf{A}(t)\big\|_h = (\sqrt{2}\alpha)^{-1}\sqrt{N+1}\|\partial_x\widehat{W}_{N+1}(\bullet,t)\|_h \le CN^{1/2-r_1/2}\,, \tag{4.45}$$

where in the last step we have used the inequality (4.6).

In order to obtain convergence-rate estimates for numerical scheme (4.22) by using the energy estimate in Theorem 3.1, we need to verify the assumption (H_1).

Lemma 4.4. *Let $B = U^TSU$, where U is an orthogonal matrix and S is defined by (4.14). If μ in (1.6) satisfies $0 < \alpha \le \mu^{-1/2}$, then there exists a constant C_1, such that for any $\mathbf{u} \in \mathbf{R}^{N+1}$,*

$$(\mathbf{u}, B\mathbf{u})_h \le C_1\|\mathbf{u}\|_h\,.$$

Proof. Let $\mathbf{y} = [y_0, y_1, \cdots, y_N]^T := U\mathbf{u}$. Since $B = U^T SU$ and U is orthogonal, it can be verified that

$$\begin{aligned}(\mathbf{u}, B\mathbf{u})_h &= (U\mathbf{u}, SU\mathbf{u})_h = (\mathbf{y}, S\mathbf{y})_h \\ &= \sum_{x\in I_h}\Big(\sum_{n=0}^N s_n y_n(x)^2 + \sum_{n=1}^N \gamma_n y_{n-1}(x)y_n(x) + \sum_{n=2}^N \delta_{n-1}y_{n-2}y_n(x)\Big) \\ &\le \sum_{x\in I_h}\sum_{n=0}^N \sigma_n y_n^2(x) \qquad (4.46)\end{aligned}$$

where, for $n \ge 2$,

$$\begin{aligned}\sigma_n &= \frac{1}{2}\Big(- 2\beta n + \alpha\sqrt{2n}\|F\|_\infty + \alpha\sqrt{2n+2}\|F\|_\infty \\ &\qquad + \beta|2\alpha^2\mu - 1|\sqrt{(n-1)(n-2)} + \beta|2\alpha^2\mu - 1|\sqrt{n(n+1)}\Big).\end{aligned}$$

For large n,

$$\sigma_n \sim n\beta\Big(- 1 + |2\alpha^2\mu - 1|\Big).$$

Therefore, if $0 < \alpha \le \mu^{-1/2}$, then $\sigma_n \le 0$ provided that n is sufficiently large. This proves that $\sigma_n \le C$ for all $n \ge 0$. This result, together with (4.46), yields

$$(\mathbf{u}, B\mathbf{u})_h \le C\|\mathbf{y}\|_h^2 = C\|\mathbf{u}\|_h^2\,.$$

This completes the proof of Lemma 4.4. □

We are now ready to state and prove the convergence result for the numerical scheme (4.22).

Theorem 4.1. *Let W be the solution of* (1.6) *and W_Δ be the numerical approximation given by* (4.25) *with $\mathbf{f}$ being computed by scheme* (4.22). *If $0 < \alpha \le \mu^{-1/2}$ and $W \in C^3(0,T; V_\omega^{0,0}) \cap C^0(0,T; V_\omega^{0,r_0} \cap V_\omega^{1,r_1} \cap V_\omega^{2,1})$, then, for all $t \in Q_\tau$,*

$$\Big\|W(\bullet,\bullet,t) - W_\Delta(\bullet,\bullet,t)\Big\|_{h,0,\omega} \le C\Big(\tau^2 + h + N^{11/12 - r_0/2} + N^{1/2 - r_1/2}\Big).$$

Proof. It follows from (4.5) that

$$\|W(\bullet,\bullet,t) - W_\Delta(\bullet,\bullet,t)\|_{h,0,\omega} \le CN^{-r_0}\|W(\bullet,\bullet,t)\|_{h,0,\omega}. \qquad (4.47)$$

By (4.25), noting $\mathcal{F}_\Delta = U\mathbf{f}, \mathcal{F} = U\mathbf{W}$ and U is orthogonal we have

$$\begin{aligned}&\|W_N(\bullet,\bullet,t) - W_\Delta(\bullet,\bullet,t)\|_{h,0,\omega} = C\|\mathcal{F}_\Delta(\bullet,t) - \mathcal{F}(\bullet,t)\|_h \\ &= C\|\mathbf{f} - \mathbf{W}\|_h = C\|\mathbf{e}(\bullet,t)\|_h\end{aligned}$$

where $\mathbf{e}$ satisfies (4.30). It follows from Theorem (3.1), Lemmas 4.2–4.4 and (4.45) that

$$\|\mathbf{e}(\bullet,t)\|_h \le C\Big(\tau^2 + h + N^{11/12 - r_0/2} + N^{1/2 - r_1/2}\Big). \qquad (4.48)$$

Using the triangular inequality for (4.47) and (4.48) we obtain the desired result. □

Similarly, we can prove the following result for the backward Euler method (4.23).

Theorem 4.2. *Let W be the solution of* (1.6) *and W_Δ be the numerical approximation given by* (4.25) *with* $\mathbf{f}$ *being computed by scheme* (4.23). *If $0 < \alpha \le \mu^{-1/2}$ and $W \in C^2(0,T;V_\omega^{0,0}) \cap C^0(0,T;V_\omega^{0,r_0} \cap V_\omega^{1,r_1} \cap V_\omega^{2,1})$, then, for all $t \in Q_\tau$,*

$$\Big\| W(\bullet,\bullet,t) - W_\Delta(\bullet,\bullet,t) \Big\|_{h,0,\omega} \le C\Big(\tau + h + N^{11/12-r_0/2} + N^{1/2-r_1/2}\Big).$$

Finally, we will bound the errors introduced by the spectral-difference approximation (4.24). In order to apply the energy estimate in Theorem 3.3, we need to verify that the matrix $B = U^T SU$, where S is given by (4.14), satisfies (H_2).

Lemma 4.5. *Let $B = U^T SU$, where U is an orthogonal matrix and S is defined by* (4.14). *If μ in* (1.6) *satisfies $0 < \alpha \le \mu^{-1/2}$, then there exist a constant C_2 (independent of N) and a constant $d_N = \beta N$ (β is given in* (1.6)), *such that for any vectors $\mathbf{u}, \mathbf{v} \in \mathbf{R}^{(N+1)}$*

$$(\mathbf{u}, (d_N I + B)\mathbf{v})_h \le \frac{d_N}{2}\Big(\|\mathbf{u}\|_h^2 + \|\mathbf{v}\|_h^2\Big) + C_2\Big(\|\mathbf{u}\|_h^2 + \|\mathbf{v}\|_h^2\Big).$$

Proof. Let $\mathbf{y} = [y_0, y_1, \cdots, y_N]^T := U\mathbf{u}$ and $\mathbf{z} = [z_0, z_1, \cdots, z_N]^T := U\mathbf{u}$. Since $B = U^T SU$ and U is orthogonal, it can be verified that

$$(\mathbf{u}, (d_N I + B)\mathbf{v})_h = (\mathbf{y}, (d_N I + S)\mathbf{z})_h \,.$$

Since the diagonal elements of the matrix $d_N I + S$ are positive, we can use the definition of S to obtain

$$(\mathbf{y}, (d_N I + S)\mathbf{z})_h = \sum_{x \in I_h} \Big(\sum_{n=0}^{N} (d_N + s_n) y_n z_n + \sum_{n=1}^{N} \gamma_n y_{n-1} z_n + \sum_{n=2}^{N} \delta_{n-1} y_{n-2} z_n \Big)$$

$$\le \quad \sum_{x \in I_h} \sum_{n=0}^{N} \Big(\frac{d_N}{2}(y_n^2 + z_n^2) + p_n y_n^2 + q_n z_n^2 \Big)$$

where for $n \ge 2$

$$p_n = \frac{1}{2}\Big(-\beta n + \alpha\sqrt{2n+2}\|F\|_\infty + \beta|2\alpha^2\mu - 1|\sqrt{n(n+1)}\Big)\,,$$
$$q_n = \frac{1}{2}\Big(-\beta n + \alpha\sqrt{2n}\|F\|_\infty + \beta|2\alpha^2\mu - 1|\sqrt{(n-1)(n-2)}\Big)\,.$$

Similar to the proof for Lemma 4.4, we have $p_n \le C$ and $q_n \le C$ for all $n \ge 0$, provided that $0 < \alpha \le \mu^{-1/2}$. These results, together with $\|\mathbf{y}\|_h = \|\mathbf{u}\|_h$ and $\|\mathbf{z}\|_h = \|\mathbf{u}\|_h$, yield the desired inequality. ☐

The error estimate below follows from Theorem 3.3 and the above lemma.

Theorem 4.3. *Let W be the solution of* (1.6) *and W_Δ be the numerical approximation given by* (4.25) *with* $\mathbf{f}$ *being computed by scheme* (4.24). *If $W \in C^2(0,T;V_\omega^{0,0}) \cap C^0(0,T;V_\omega^{0,r_0} \cap V_\omega^{1,r_1} \cap V_\omega^{2,1})$ and if the generalized CFL condition*

$$\frac{\tau}{\alpha h} \max_{0 \le j \le N} |\lambda_j| + \beta N \tau \le 1 \tag{4.49}$$

is satisfied, then, for all $t \in Q_\tau$,

$$\Big\| W(\bullet,\bullet,t) - W_\Delta(\bullet,\bullet,t) \Big\|_{h,0,\omega} \le C\Big(\tau + h + N^{11/12-r_0/2} + N^{1/2-r_1/2}\Big).$$

Remark 4.1. Since $\max|\lambda_j| \sim \sqrt{2N}$, Theorem 4.3 implies that the time step used in the explicit scheme (4.24) is of the order $\tau \sim \alpha h/(\sqrt{2N} + \alpha\beta N h)$.

5. Numerical results

In this section, we will consider some issues for the numerical implementation of the numerical methods considered in this work. First, we will discuss the use of the scaling factor, α, which is important in applying the Hermite spectral methods (see, e.g., [5, 39]). Secondly, we will test our Hermite expansion methods for a simplified Fokker-Planck equations where the distribution function W depends on v only. Finally, we will use the combined spectral-difference schemes to solve a test problem, in order to verify our convergence theory. Particular attention has been paid to the implementation of the Hermite spectral methods.

5.1. **Scaling factor.** Although the Hermite methods presented above enjoy a theoretical spectral convergence rate, the actual error decays considerably slower than the Chebyshev or Legendre method for similar problems in finite intervals. The poor resolution property of Hermite functions, which was pointed out by Gottlieb and Orszag in [19], is one of the main reasons why Hermite functions are rarely used in practice. However, the resolution of Hermite functions can be greatly improved by using a proper scaling factor [39]. We will extend the theory developed in [39] to deal with the spectral approximations for the Fokker-Planck equation (see subection 5.3). In this subsection, we will discuss how to choose the scaling factor for a given Gaussian type function. If the initial condition for the Fokker-Planck equation is of Gaussian type, then its solution can be bounded by a Gaussian type function and its stationary solution is of the form of Gaussian type also. It is therefore a basic requirement that the expansion methods should approximate function $\exp(-sv^2)$ accurately and efficiently for any given (positive) values of s. To analyze the effectiveness of the Hermite expansion, we expand

$$\exp\left(-sv^2\right) = \sum_{n=0}^{\infty} b_n \widetilde{H}_n(v), \tag{5.1}$$

where

$$\widetilde{H}_n(v) = \frac{1}{\sqrt{2^n n!}} H_n(\alpha v) e^{-\alpha^2 v^2}.$$

It is seen that the basis functions $\widetilde{H}_n(v)$ involve a parameter α which should be chosen with some caution. We can re-write (5.1) into the following form:

$$\exp\left(-s_1 v^2\right) = \sum_{n=0}^{\infty} b_n \frac{1}{\sqrt{2^n n!}} H_n(v) e^{-v^2}, \quad s_1 := \frac{s}{\alpha^2}. \tag{5.2}$$

A direct calculation gives that $b_{2k+1} \equiv 0$ and

$$b_{2k} = \frac{1}{\sqrt{2^{2k}(2k)! s_1}} \left(\frac{1-s_1}{s_1}\right)^k \frac{(2k)!}{k!}, \tag{5.3}$$

for $k \geq 0$. An application of the Sterling's formula yields

$$b_{2k} \sim \frac{1}{\left(\pi k s_1^2\right)^{1/4}} \left(\frac{1-s_1}{s_1}\right)^k, \qquad k \gg 1. \tag{5.4}$$

The above equation indicates that the Hermite expansion cannot produce any reasonable approximations in the case $s_1 < 0.5$. Further, since $\phi_{2k}(v) = \mathcal{O}(k^{-1/4})$ the $2k$-th term of (5.1) is of order $O(k^{-1/2})$ in the case $s_1 = 0.5$. This implies that spectral accuracy cannot be observed in the case $s_1 = 0.5$.

TABLE 1. Maximum error obtained by using the Hermite expansion (5.1) with $\alpha = 1$ for $f(v) = \exp(-sv^2)$.

N	$s = 0.4$	$s = 0.5$	$s = 0.6$	$s = 1.5$
10	1.7D+0	1.7D–1	1.6D–2	3.7D–4
20	9.4D+0	1.2D–1	1.5D–3	1.1D–6
30	5.9D+1	1.0D–1	1.7D–4	3.9D–9
40	3.9D+2	8.8D–2	1.9D–5	1.4D–11

The above analysis suggests that for a Gaussion function $\exp(-sv^2)$, the scaling factor α in the basis function $\widetilde{H}_n(v)$ must satisfy $\alpha < \sqrt{2s}$. To give a quantitative understanding of this statement, we approximate the Gaussian distribution function $f(v) = \exp(-sv^2)$ using the Hermite expansion (5.1) without a scaling factor, i.e., $\alpha = 1$. We present in Table 1 the maximum errors, defined by $\max_{v\in[-3,3]} |f(v) - f_N(v)|$. We choose $s = 0.4, 0.5, 0.6$ and 1.5 in the numerical experiments, and the numerical results in Table 1 suggest that the truncated series is divergent when $\alpha = 0.4$. In the case $\alpha = 0.5$ spectral accuracy cannot be observed. These observations support our earlier analysis that α must be less than $\sqrt{2s}$.

5.2. **Application to a simplified Fokker-Planck equation.** One of the simplest FP equations is of the form (1.2). In 1-D, it is given by

$$\frac{\partial W}{\partial t} = \gamma \frac{\partial(vW)}{\partial v} + \gamma\beta \frac{\partial^2 W}{\partial v^2}, \tag{5.5}$$

where $W(v,t)$ is the distributive function, $v \in (-\infty, \infty)$ is the particle velocity, γ^{-1} the particle relaxation time and $\sqrt{\beta}$ the thermal velocity. By solving (5.5), together with the initial distribution $W(v,0)$, one may obtain the distributive function $W(v,t)$ for all later times.

Let the exact solution W of (5.5) be approximated by

$$W_N(v,t) = \sum_{n=0}^{N} a_n(t)\widetilde{H}_n(v), \tag{5.6}$$

where again we set the scaling factor α in $\widetilde{H}_n(v)$ as 1. We want to show that with the constant scaling factor some Gaussian type solutions cannot be well approximated. Using the Hermite expansion methods we obtain from (5.5) and (5.6) that

$$\frac{da_0(t)}{dt} = 0, \tag{5.7}$$

$$\frac{da_n(t)}{dt} = (2\beta - 1)\gamma\sqrt{(n-1)n}\,a_{n-2}(t) - n\gamma a_n(t), \tag{5.8}$$

for $n = 1, \cdots, N$, with $a_n(t) \equiv 0$ whenever $n < 0$ or $n > N$.

To see the performance of the Hermite spectral methods with velocity scaling, we consider the following test problem.

Example 5.1. Consider (5.5) with $\gamma = 0.01$ (which corresponds to a relaxation time of 100), $\beta = 0.5$ and the following initial condition:

$$W(v,0) = (1 + \sin(\pi v))\exp\left(-sv^2\right), \tag{5.9}$$

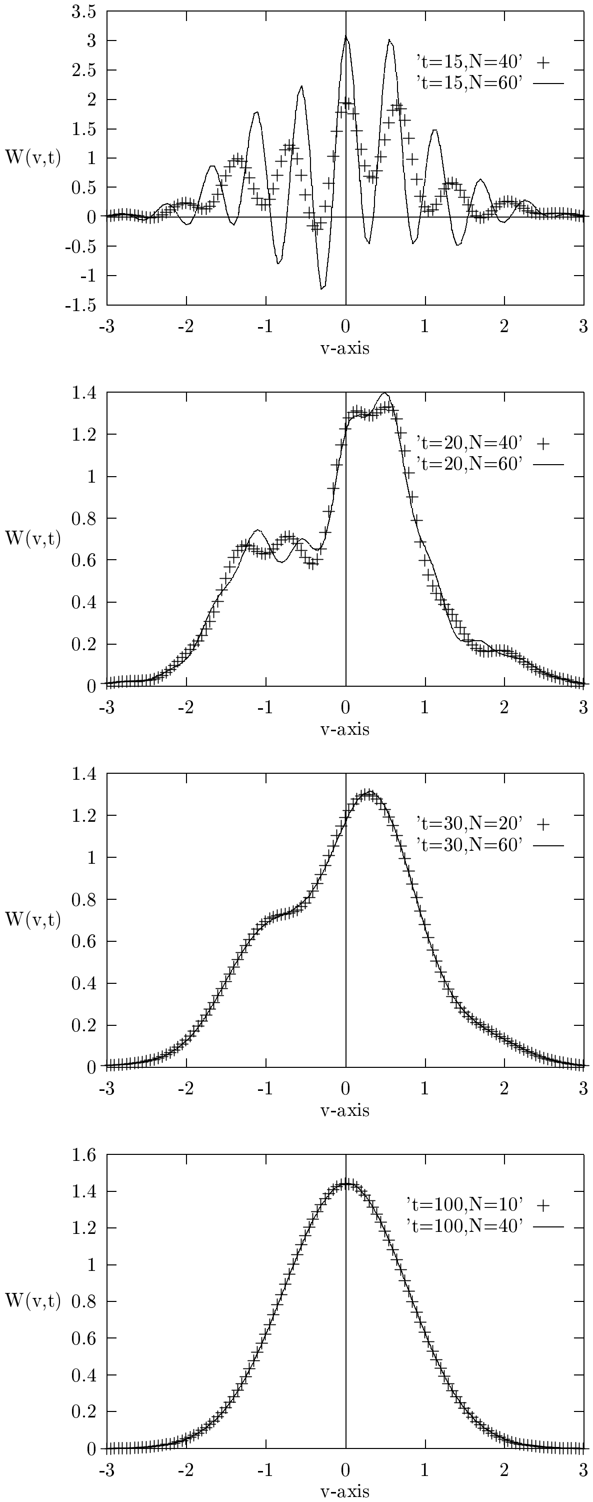

FIGURE 1. Numerical solution at different time levels for Example 5.1 with initial function (5.11).

where s is a positive constant. The stationary solution of (5.5) and (5.9) is

$$W(v,t) \sim \sqrt{\frac{1}{s}}\exp\left(-v^2\right), \qquad t \gg 0. \tag{5.10}$$

As in the last subsection, we test the above problem by using $s = 0.4, 0.5, 0.6$ and 1.5. It is observed that for $s > 0.5$ very accurate numerical approximations are obtained by using about 20 expansion terms (i.e., $N = 20$). However, for $s = 0.4$ convergent results cannot be obtained before the solution reaches the stationary sate. Figure 1 shows the numerical results with various values of N for

$$W(v,0) = (1+\sin(\pi v))\exp\left(-0.4v^2\right). \tag{5.11}$$

Since the initial data has the power constant $s = 0.4$, it is expected from the experience of the last subsection that the Hermite expansion with the scaling factor $\alpha = 1$ will not be convergent, at least for small values of the time t. However, it is seen that the stationary solution (5.10) has the power constant $s = 1$, and as a result it is expected that when t becomes large the Hermite spectral methods with scaling factor $\alpha = 1$ will lead to accurate approximations. These theoretical predictions are well verified in the plots of Figure 1.

5.3. **Application of the Hermite spectral-finite difference methods.** In this section, we consider a numerical example by using the combined spectral-difference schemes (4.22)–(4.24). We would verify the Theorem 4.1–4.3; i.e., the numerical schemes are of spectral accuracy in v-direction, first order in x-direction and first/second order in t-direction. To this end, we consider the following test problem.

Example 5.2. Consider the Fokker-Planck equation

$$\frac{\partial w}{\partial t} = -v\frac{\partial w}{\partial x} + \frac{\partial (vw)}{\partial v} + \frac{\partial^2 w}{\partial v^2}, \qquad |x| \le 1,\ |v| < \infty, \tag{5.12}$$

with the initial and boundary conditions (for $x = -1, v \ge 0$ and $x = 1, v \le 0$) such that the exact solution is

$$w(x,v,t) = \frac{1}{2}\left[1 + \cos\left(\frac{\pi}{2}(x - (1-e^{-t})v)\right) q(t)\right] \exp\left(-v^2/2\right), \tag{5.13}$$

where

$$q(t) = \exp\left[-\frac{\pi^2}{4}\left(t + 2e^{-t} - \frac{1}{2}e^{-2t} - \frac{3}{2}\right)\right].$$

The main reason for choosing the above test example is that its analytic solution can be found which enables us to test the accuracy of the numerical schemes. In the following numerical calculations, the length of the x-interval is chosen to be 0.4 and the scaling factor α (see (2.9)) is chosen as $1/\sqrt{2}$. The effects of using other values of the scaling factor α will be also investigated (see Figure 5). Although the results reported below are obtained by using the numerical scheme (4.22), similar results supporting Theorems 4.2 and 4.3 have been also computed. Due to the limitation of the space, we will not include them here.

Convergence rate for the spectral approximation. In Figure 2, we plot the l^2-errors for the mixed finite-difference-spectral method (4.22), with parameters $N = 9, h = \tau = 0.001$. The step sizes in both x and t directions are chosen to be very

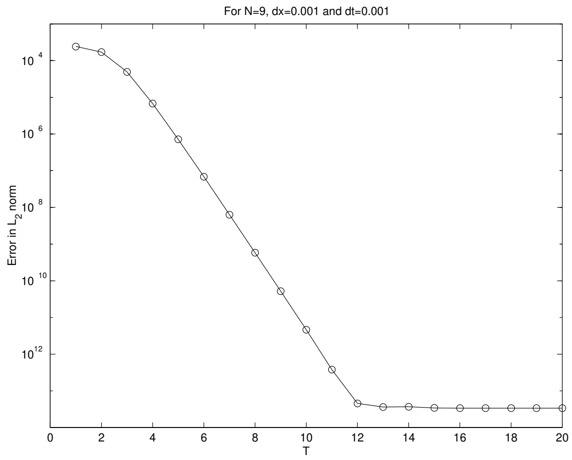

FIGURE 2. The l^2-error in time for Example 5.2.

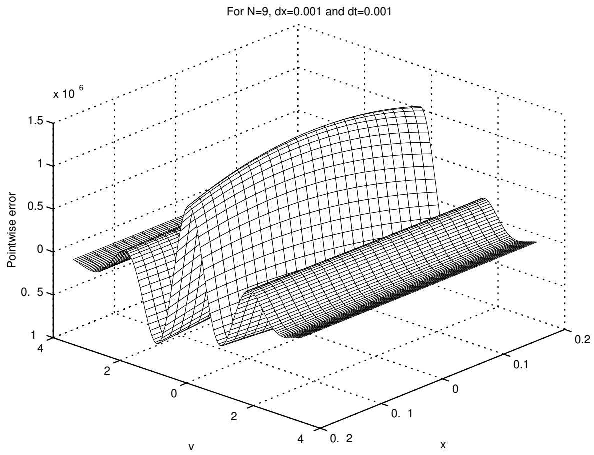

FIGURE 3. The pointwise error for Example 5.2 at $t = 5$.

small so that we can verify the exponential rate of convergence for the spectral approximations in the v direction. Indeed it is observed in Figure 2 that spectral convergence is achieved. For $t \geq 12$, the l^2-error remains the order of about $\mathcal{O}(10^{-14})$, which is about the machine accuracy in double precision. It is seen from the exact solution of this problem that the asymptotic solution for large t is $\frac{1}{2}\exp(-v^2/2)$, i.e., a pure Gaussian type function. As a result the spectral approach in v direction will give very accurate approximation for large time solution as seen in Figure 1. Pointwise error at $t = 5$ is plotted in Figure 3. As expected, the largest errors occur at the zero axis for v (see also the similar observation in [40]).

Convergence rates for the finite-difference approximation. Our theoretical predictions in last section indicate that there will have only first-order spatial and temporal accuracy for the schemes (4.22)–(4.24), except for (4.22) for which $\mathcal{O}(h+\tau^2)$ can be achieved. This result is well understood in finite-difference theory, so we just simply plot the l^2-errors as a function of N with $h = 0.01, 0.02$ and 0.04 in Figure 4. A first-order convergence rate in x-direction is observed, which is in

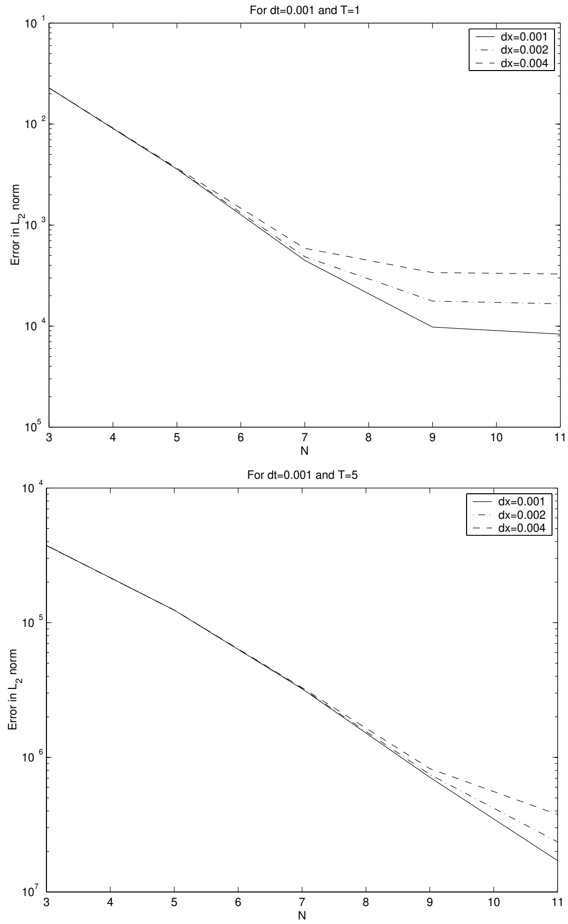

FIGURE 4. The l^2-error as a function of N with different values of h. The top picture is for $t = 1$ and the bottom one is for $t = 5$.

agreement with the theoretical predications. Similarly, a second-order convergence rate in time, i.e., $\mathcal{O}(\tau^2)$, have been also observed in our numerical computations.

Variation with the scaling factor. Finally, we investigate the role of the scaling factor α for the test problem Example 5.2. In Figure 5, we plot the l^2-errors obtained by the scheme (4.22), with the use of the parameters $N = 7, h = 0.005$ and $\tau = 0.001$. Using the discussions in subsection 5.2, together with the exact solution (5.13), we can conclude that the optimal choice of α is $1/\sqrt{2} \approx 0.7071$ for Example 5.2. Indeed this predication is verified by our computational results given in Figure 5.

In practice, the scaling factors may not be a constant with respect to time and therefore some adaptive computation for the scaling factor should be used during the time integration to enhance spectral accuracy. We will not give further discussion on this issue due to the limitation of space, but just point out a recent paper of Schumer and Holloway [36] where a *variable scaling factor* for the Hermite basis was constructed for solving the nonlinear Vlasov-Poisson equations. The

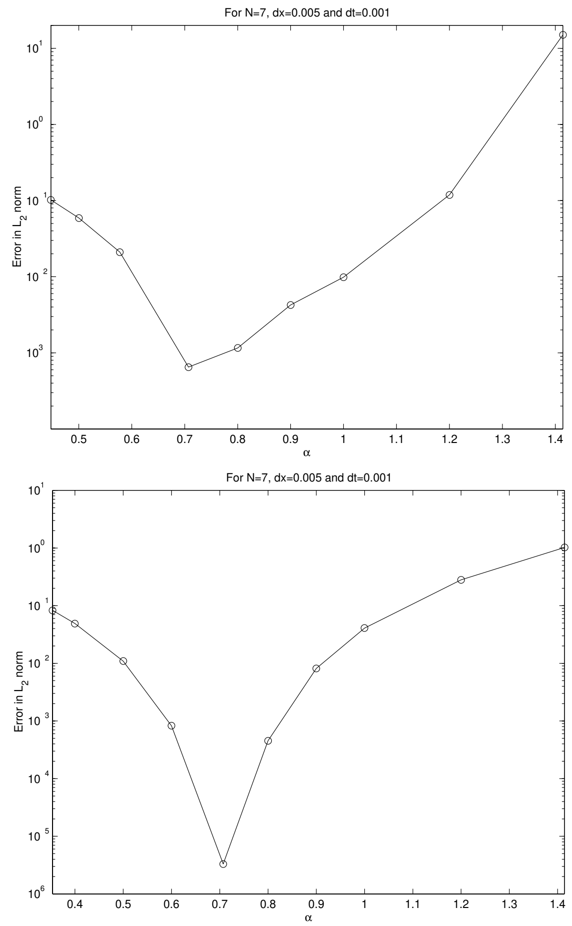

FIGURE 5. The l^2-error as a function of α, for $t = 1$ (top) and $t = 5$ (bottom).

principal ideas in [36] are also useful for Hermite spectral approximations to the Fokker-Planck equations. Scaling factors are also used in a recent work of Shen [37] for the Laguerre spectral approximations, which also greatly enhance the resolution capacities of the Laguerre functions.

References

[1] M. Abramowitz and I. R. Stegun, *Handbook of Mathematical Functions*, Dover, 1972. MR **94b:**00012

[2] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975. MR **56:**9247

[3] C. Bernardi, and Y. Maday, Spectral methods, in *Handbook of Numerical Analysis*, 209-486, ed. by Ciarlet, P.G. and Lions, J.L., Elsevier, Amsterdam, 1997. CMP 98:01

[4] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, (Springer–Verlag Berlin, Heidelberg, 1989).

[5] J. P. Boyd, Asymptotic coefficients of Hermite function series, *J. Comput. Phys.*, **54** (1984), p. 382. MR **86c:**41012

[6] J. P. Boyd, Orthogonal rational functions on a semi-infinite interval, *J. Comput. Phys.*, **70** (1987), 63-88. MR **88d:**65034

[7] H. C. Brinkman, Brownian motion in a field of force and the diffusion theory of chemical reactions. *Physica*, **22** (1956), 29-34.

[8] C. Blomberg, The Brownian motion theory of chemical transition rates, *Physica*, **86A** (1977), 49–66.

[9] M. A. Burschka and U. M. Titulaer, The kinetic boundary layer for the Fokker-Planck equation: a Brownian particle in an unbounded field, *Physica*, **112A**, (1982), 315–330.

[10] M. A. Burschka and U. M. Titulaer, The kinetic boundary layer for the Fokker-Planck equation: selectively absorbing boundaries, *J. Stat. Phys.* **26**, (1981) 59–71. MR **83b:**82108

[11] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988. MR **89m:**76004

[12] B. Cartling, Kinetics of activated processes from nonstationary solutions of the Fokker-Planck equation for a bistable potential., *J. Chem. Phys.*, **87** (1987), 2638-2648.

[13] S. Chandrasekhar, Stochastic problems in physics and astronomy, *Rev. Mod. Phys.*, **15** (1943), 1-89.

[14] T. Chen, A theoretical and numerical study for the Fokker-Planck equation, MSc Thesis, Dept of Math., Simon Fraser University, B. C., Canada, 1992.

[15] R. J. DiPerna and P. L. Lions, On the Fokker-Planck-Boltzmann equation, *Commun. Math. Phys.*, **120** (1988), 1-23. MR **90b:**35203

[16] D. Funaro, *Polynomial Approximation of Differential Equations*, Springer-Verlag, Berlin, 1992. MR **94c:**65078

[17] D. Funaro and O. Kavian, Approximation of some diffusion evolution equations in unbounded domain by Hermite functions, *Math. Comp.*, **57** (1991), 597-619. MR **92k:**35156

[18] R. R. J. Gagne and M. M. Shoucri, A splitting scheme for the numerical solution of a one-dimensional Vlasov equation, *J. Comput. Phys.*, **24** (1977), 445.

[19] D. Gottlieb and S. Orszag, *Numerical Analysis of Spectral Methods, Theory and Applications*, CBMS-NSF Regional Conference Series in Applied Mathematics, 26, SIAM , Philadelphia, 1997. MR **58:**24983

[20] C. E. Grosch and S. A. Orszag, Numerical solution of problems in unbounded regions: Coordinates transfroms, *J. Comput. Phys.*, **25** (1977), 273-295. MR **58:**8372

[21] B. Guo, *Finite Difference Methods for Partial Differential Equations*, Science Press, Beijing, 1988.

[22] B. Guo, *Spectral Methods and Their Applications*, World Scientific, River Edge, NJ, 1998. MR **2000b:**65194

[23] B. Guo, Error estimation of Hermite spectral method for nonlinear partial differential equations, *Math. Comp.*, **68**, (1999), 1067-1078. MR **99i:**65111

[24] B. Y. Guo and J. Shen, Laguerre-Galerkin method for nonlinear partial differential equations on a semi-infinite interval, *Numer. Math.*, **86**, 2000, 635–654. MR **2001h:**65152

[25] D. W. Heerman, *Computer Simulation Methods in Theoretical Physics*, Springer Verlag, Berlin, 1986.

[26] J. P. HOLLOWAY, Spectral velocity discretizations for the Vlasov-Maxwell equations, *Trans. Theory and Stat. Phys.*, **25** (1996), 1–32. MR **96k:**82071
[27] G. JOYCE, G. KNORR AND H. K. MEIER, Numerical integration methods of the Vlasov equation. *J. Comput. Phys.*, **8** (1971), 53–63.
[28] H. A. KRAMERS, Brownian motion in a field force and the diffusion of chemical reactions, *Physica*, **7** (1940), 284–304. MR **2:**140d
[29] A. L. LEVIN AND D. S. LUBINSKY, Christoffel functions, orthogonal polynomials, and Nevaiś conjecture for Freud weights, *Constr. Approx.*, **8** (1992), 461-535. MR **94f:**42030
[30] D. S. LUBINSKY AND F. MORITZ, The weighted L_p-norms of orthogonal polynomials for Freud weights, *J. Approx. Theory*, **77** (1994), 42-50.
[31] Y. MADAY, B. PERNAUD-THOMAS, AND H. VANDEVEN, Une réhabilitation des méthodes spèctrales de type Laguerre, *Rech. Aérospat.*, **6**, 1985, 353–375. MR **88b:**65135
[32] C. MAVRIPLIS, Laguerre polynomials for infinite-domain spectral elements, *J. Comp. Phys.*, **80** (1989), 480-488. MR **90f:**65228
[33] P. MOORE AND J. FLAHERTY, Adaptive local overlapping grid methods for parabolic systems in two space dimensions, *J. Comp. Phys.*, **98** (1992), 54-63.
[34] B. PERTHAME, Higher moments for kinetic equations, The Vlasov-Poisson and Fokkel-Planck cases, *Math. Meth. App. Sci.*, **13** (1990), 441-452. MR **91j:**82044
[35] H. RISKEN, *The Fokker-Planck equation: Methods of solution and applications,* 2nd ed., Springer-Verlag, Berlin, 1989. MR **90a:**82002
[36] J. W. SCHUMER AND J. P. HOLLOWAY, Vlasov simulations using velocity-scaled Hermite representations. *J. Comp. Phys*, **144**, (1998), 626-661.
[37] J. SHEN, Stable and efficient spectral methods in unbounded domains using Laguerre functions, *SIAM J. Numer. Anal.,* **38** (2000), 1113-1133. MR **2001g:**65165
[38] G. SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc., New York, 1959. MR **21:**5029
[39] T. TANG, The Hermite spectral method for Gaussian type funcitons, *SIAM J. Sci. Comp.*, **14**, (1993), 594-606. MR **93m:**65026
[40] T. TANG, S. MCKEE, AND M. W. REEKS, A spectral method for the numerical solutions of a kinetic equation describing the dspersion of small particles in a turbulent flow, *J. Comp. Phys*, **103**, (1991), 222-230.
[41] R. TÉMAN, *Analysis Numérique*, Presses Universitaires de France, Paris, 1970.
[42] V. THOMÉE, Difference methods for two-dimensional mixed problems for hyperbolic first order systems, *Arch. Rat. Mech. Anal.*, **8** (1961), 68-88. MR **23:**B2591
[43] A. F. TIMAN, *Theory of Approximation of Functions of a Real Variable*, Pergamon Press, Oxford, 1963. MR **33:**465
[44] K. VOIGTLAENDER AND H. RISEN, Eigenvalues of the Fokker-Planck and BGK operators for a double-well potential, *Chem. Phys. Lett.*, **105** (1984), 506-510.
[45] K. VOIGTLAENDER AND H. RISEN, Solutions of the Fokker-Planker equation for a double-well potential in terms of matrix continued fractions, *J. Stat. Phys.*, **40** (1985), 397-429.
[46] J. A. C. WEIDEMAN, The eigenvalues of Hermite and rational spectral differentiation matrices, *Numer. Math.*, **61** (1992), 409-431. MR **92k:**65071

DEPARTMENT OF MATHEMATICS, THE HONG KONG BAPTIST UNIVERSITY, KOWLOON TONG, HONG KONG
E-mail address: `cmfok@math.hkbu.edu.hk`

DEPARTMENT OF MATHEMATICS, SHANGHAI NORMAL UNIVERSITY, SHANGHAI 200234, P. R. CHINA
E-mail address: `byguo@guomai.sh.cn`

DEPARTMENT OF MATHEMATICS, THE HONG KONG BAPTIST UNIVERSITY, KOWLOON TONG, HONG KONG.
E-mail address: `ttang@math.hkbu.edu.hk`