Multiscale convergence properties for spectral approximations of a model kinetic equation[*†]

Zheng Chen[‡], Cory D. Hauck[§]

December 21, 2017

Abstract

In this work, we prove rigorous convergence properties for a semi-discrete, moment-based approximation of a model kinetic equation in one dimension. This approximation is equivalent to a standard spectral method in the velocity variable of the kinetic distribution and, as such, is accompanied by standard algebraic estimates of the form N^{-q}, where N is the number of modes and $q > 0$ depends on the regularity of the solution. However, in the multiscale setting, the error estimate can be expressed in terms of the scaling parameter ϵ, which measures the ratio of the mean-free-path to the characteristic domain length. We show that, for isotropic initial conditions, the error in the spectral approximation is $\mathcal{O}(\epsilon^{N+1})$. More surprisingly, the coefficients of the expansion satisfy super convergence properties. In particular, the error of the ℓ^{th} coefficient of the expansion scales like $\mathcal{O}(\epsilon^{2N})$ when $\ell = 0$ and $\mathcal{O}(\epsilon^{2N+2-\ell})$ for all $1 \leq \ell \leq N$. This result is significant, because the low-order coefficients correspond to physically relevant quantities of the underlying system. All the above estimates involve constants depending on N, the time t, and the initial condition. We investigate specifically the dependence on N, in order to assess whether increasing N actually yields an additional factor of ϵ in the error. Numerical tests will also be presented to support the theoretical results.

Keywords: kinetic equation; multiscale; super convergence; spectral method; diffusion approximation

Contents

[*]This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research.

[‡]Computational and Applied Mathematics Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA. Email: chenz1@ornl.gov.

[§]Computational and Applied Mathematics Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA. Email: hauckc@ornl.gov.

1 Introduction

In this paper, we study the following linear kinetic model

$$\begin{cases} \epsilon\partial_t f(x,\mu,t) + \mu\partial_x f(x,\mu,t) + \dfrac{1}{\epsilon} f(x,\mu,t) = \dfrac{1}{\epsilon}\bar{f}(x,t), & (x,\mu,t) \in [-\pi,\pi)\times[-1,1]\times(0,\infty), \quad (1.1a) \\ f(\pi,\mu,t) = f(-\pi,\mu,t), & (\mu,t)\in[-1,1]\times(0,\infty), \quad (1.1b) \\ f(x,\mu,0) = g(x,\mu), & (x,\mu)\in[-\pi,\pi)\times[-1,1], \quad (1.1c) \end{cases}$$

where $\bar{f} = \frac{1}{2}\int_{-1}^{1} f\,d\mu$. In particular, we prove interesting convergence properties for spectral discretization with respect to the variable μ. The function f is a kinetic distribution function; the physical interpretation is that $f(x,\mu,t)$ gives the density of particles with respect to the measure $d\mu dx$ that at time t are located at position $x \in [-\pi,\pi)$ and moving with velocity $\mu \in [-1,1]$. The parameter $\epsilon > 0$ is a scaling parameter that measures the relative strength of different processes; more about this will be said below.

System (1.1) is among the most elementary examples of a kinetic model. However, despite its simplicity, it shares the basic features of many kinetic equations: particle advection (modeled by the operator $\mathcal{A}\colon f \mapsto -\mu\partial_x f$) and particle interactions (modeled by the scattering operator $\mathcal{L}\colon f \mapsto \bar{f} - f$). These basic features are found in more realistic models that describe dilute gases [9–11]; neutron [8, 12, 13, 23], photon [27, 28], and neutrino [26] radiation; charged transport in semiconductor devices [25, 30]; and ionized plasmas [5, 19]. However, connecting (1.1) to these more realistic models requires the introduction of more complicated geometries, global field equations, nonlinearities, more complex collision mechanisms, and physical boundary conditions.

Existence and uniqueness results for (1.1) follow from classical transport theory. See, for example, [12, Chapter XXI]. For data $g(x,\mu) \in L^2(d\mu dx)$, (1.1) has a unique solution $f \in C^0([0,\infty); L^2(d\mu dx))$. If further, $g \in D(\mathcal{A}) := \{u \in L^2(d\mu dx) : \mu\partial_x u \in L^2(d\mu dx)\}$, then $f \in C^1([0,\infty); L^2(d\mu dx)) \cap C^0([0,\infty); D(\mathcal{A}))$.

The scattering operator $\mathcal{L}$ is self adjoint in $L^2(d\mu)$ and satisfies

$$\int_{-1}^{1} \psi\mathcal{L}\psi\,d\mu \le 0 \quad \text{and} \quad \int_{-1}^{1} (\psi-\bar{\psi})\mathcal{L}(\psi-\bar{\psi})\,d\mu = -\int_{-1}^{1}(\psi-\bar{\psi})^2\,d\mu \tag{1.2}$$

for any function $\psi \in L^2(d\mu)$. This simple dissipative structure motivates a diffusion approximation for (1.1) when $\epsilon \ll 1$. In such cases, $f = f^{(0)} + \mathcal{O}(\epsilon)$, where $f^{(0)}$ is independent of μ and satisfies the diffusion equation [2, 4, 17, 21]

$$\partial_t f^{(0)} - \frac{1}{3}\partial_x^2 f^{(0)} = 0. \tag{1.3}$$

The diffusion approximation is useful because it removes the need for angular discretization and is therefore relatively cheap to compute; however, it does so at the expense of an $O(\epsilon)$ error. Spectral methods (see [7, 20] in general or [24, Chapter 3] for applications to kinetic transport equations), on the other hand, are more expensive but can be used to discretize (1.1) with respect to μ when ϵ is not small. A standard spectral method for (1.1) seeks an approximation

$$f^N(x,\mu,t) = \sum_{\ell=0}^{N} f_\ell^N(x,t)p_\ell(\mu), \tag{1.4}$$

such that

$$\epsilon \partial_t f^N = \mathcal{P}\mathcal{T} f^N, \qquad f^N|_{t=0} = \mathcal{P} g, \tag{1.5}$$

where $\mathcal{T} = \mathcal{A} + \epsilon^{-1}\mathcal{L}$, p_ℓ is the normalized, degree ℓ Legendre polynomial, and $\mathcal{P}$ is the orthogonal projection from $L^2(d\mu)$ onto the space $\mathbb{P}^N$ of polynomials on $[-1,1]$ with degree at most N; that is

$$\mathcal{P}\psi = \sum_{\ell=0}^{N} \psi_\ell p_\ell, \quad \text{where } \psi_\ell = \int_{-1}^{1} p_\ell \psi d\mu \tag{1.6}$$

for any $\psi \in L^2(d\mu)$. When expressed in terms of the expansion coefficients f_ℓ^N in (1.4), (1.5) takes the form of a linear, symmetric hyperbolic system of balance laws in x and t. Standard semi-group theory (see for example [6, Chapter 7] or [15, Chapter 7.4]) implies that this system has a solution in $C^0([0,\infty); [L^2(dx)]^{N+1})$ that is also in $C^1([0,\infty); [L^2(dx)]^{N+1}) \cap C^0([0,\infty); [H^1(dx)]^{N+1})$ when the expansion coefficients of $\mathcal{P}g$ are in $H^1(dx)$.

We refer to f^N as the spectral approximation or P_N solution. A straight-forward calculation shows that this approximation converges like

$$\|f(\cdot,\cdot,t) - f^N(\cdot,\cdot,t)\|_{L^2(d\mu dx)} \leq \frac{C(t)}{N^q}, \tag{1.7}$$

where q is the number of L^2 angular derivatives of f and $\partial_x f$ and the constant C depends on t but, due to the dissipative structure of $\mathcal{L}$, does not depend on ϵ in a bad way.[1]

A natural question for the spectral approximation is whether it provides an improvement over the diffusion approximation when ϵ is small. The goal of the current paper is to derive an error estimate to demonstrate that this is in fact the case. Specifically, let

$$f(x,\mu,t) = \sum_{\ell=0}^{\infty} f_\ell(x,t) p_\ell(\mu), \qquad \text{where } f_\ell(x,t) = \int_{-1}^{1} p_\ell(\mu) f(x,\mu,t) d\mu, \tag{1.8}$$

be the spectral expansion of f in $L^2(d\mu)$. For small values of ℓ, the coefficients $\{f_\ell\}$ correspond to measurable quantities and thus have physical significance. For example, f_0 is a constant multiple of the particle concentration. Thus we also derive estimates for the errors in these coefficients, respectively.

For the purposes of the current paper, we introduce the following assumption.

Assumption 1.1. *The function g is isotropic; that is, it is independent of μ. We write it as $g(x,\mu) = \frac{1}{\sqrt{2}} g_0(x)$, where the $\frac{1}{\sqrt{2}}$ is a normalization constant.*

This assumption is critical for the results in this paper, but will be removed in future work. With it, our main result is the following:

Theorem 1.2. *Suppose that $g_0 \in H^1(dx)$. Then there exists an absolute constant $\lambda_1 > 0$ such that the L^2 error of the P_N approximation satisfies*

$$\|f - f^N\|_{L^2(d\mu dx)}(t) \leq B(g) e^{-\frac{\lambda_1 t}{\epsilon^2}} + C(\partial_x g)\sqrt{t} e^{-\frac{\lambda_1 t}{\epsilon^2}} + D(g,N,t)\epsilon^{N+1}, \tag{1.9}$$

where $D(g,N,t)$ is positive and bounded for any $t > 0$ and is decreasing exponentially in t for t sufficiently large. Moreover, the L^2 error for each coefficient satisfies

$$\|f_\ell - f_\ell^N\|_{L^2(dx)}(t) \leq \begin{cases} C(\partial_x g)\sqrt{t} e^{-\frac{\lambda_1 t}{\epsilon^2}} + E(g,N,2,t)\epsilon^{2N}, & \ell = 0, \\ C(\partial_x g)\sqrt{t} e^{-\frac{\lambda_1 t}{\epsilon^2}} + E(g,N,\ell,t)\epsilon^{2N+2-\ell}, & 1 \leq \ell \leq N, \end{cases} \tag{1.10}$$

where $E(g,N,\ell,t)$ is positive and bounded for any $t > 0$ and is monotonically decreasing with respect to t.

[1] An estimate of the form (1.7) can be found in [16] when $\epsilon = O(1)$. However, a more general argument is needed to show that C can be made independent of $\epsilon \in [0,1]$. We give such an argument in the appendix.

Remark 1.3. *A formal statement of the ϵ-dependent scaling in* (1.10)*, based on a Chapman-Enskog expansion, can be found in [18]. In [22], formal asymptotic results for the SP_N equations, which are equivalent to the spectral approximation of* (1.1) *in the current setting, predict a similar scaling, at least for the coefficient f_1.*

Remark 1.4. *In our proofs, we use $\lambda_1 = 1/45$ (cf.* (3.8) *in Section 3.1). We do not believe this value is optimal; nor have we made any effort to optimize it.*

Remark 1.5. *All of the rates in* (1.9) *and* (1.10) *are observed in the numerical tests in Section 4. Moreover, we observe these rates numerically even when $\partial_x g \notin L^2(dx)$. Further discussion of this point is given in Section 4.*

Theorem 1.2 has important practical consequences for the discretization of (1.1a) in transition regimes, when ϵ is small, but not small enough to invoke the diffusion approximation. Indeed, for a fully discrete scheme in space, time, and angle, it is important to balance errors with respect to each variable. While not crucial for the solution of (1.1a), the efficiency gained from proper balancing of errors is essential for more general kinetic problems, for which the distribution function depends on six phase-space variables, plus time. Theorem 1.2 justifies the use of fewer spectral modes than the standard estimate (1.7) in transition regimes. The first statement of the theorem says that after an initial layer, the approximation of the transport solution is accurate up to $\mathcal{O}(\epsilon^{N+1})$. The second statement on the individual coefficients, which is much stronger, plays an even more important role, since it is the low-order coefficients that correspond to physically meaningful quantities. However, for more realistic applications, these estimates will ultimately need to be extended beyond the current idealized setting.

The remainder of this paper is dedicated to the proof of Theorem 1.2 and the presentation of supporting numerical results. Preliminary notation and an introduction of the modified energy are given in Section 2. Details of proofs are provided in Section 3. In Section 4, we present some numerical tests to validate the convergence rates in theory. The benefit of increasing the number of moments N is discussed in Section 5. Conclusions and future work are discussed in Section 6.

2 Preliminaries

In this section, we provide some preliminaries. We first set the notation, and then introduce the modified energy approach borrowed from [14].

2.1 Setup and Notation

The proof of Theorem 1.2 relies on estimates of expansion coefficients for functions in $L^2(d\mu dx)$.

Definition 2.1 (Legendre and Legendre-Fourier expansion)**.** *For any $u \in L^2(d\mu dx)$, the Legendre expansion of u is*

$$u(x,\mu) = \sum_{\ell=0}^{\infty} u_\ell(x) p_\ell(\mu), \qquad u_\ell(x) = \int_{-1}^{1} u(x,\mu) p_\ell(\mu) d\mu, \tag{2.1}$$

and the Legendre-Fourier expansion is

$$u(x,\mu) = \frac{1}{\sqrt{2\pi}} \sum_{\ell=0}^{\infty} \sum_{k=-\infty}^{\infty} u_{\ell,k} p_\ell(\mu) e^{ikx}, \qquad u_{\ell,k} = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} u_\ell(x) e^{-ikx} dx. \tag{2.2}$$

The coefficients u_ℓ and $u_{\ell,k}$ will be referred to as the Legendre and Legendre-Fourier coefficients.

Remark 2.2. *The definition of u_ℓ in* (2.1) *is consistent with the use of g_0 in Assumption 1.1 and the definition of f_ℓ in* (1.8)*.*

We begin by decomposing the error $e^N := f - f^N$ into the sum of two components:

$$\eta = f - \mathcal{P}f = \sum_{\ell=N+1}^{\infty} \eta_\ell(x,t) p_\ell(\mu) \quad \text{and} \quad \xi = \mathcal{P}f - f^N = \sum_{\ell=0}^{N} \xi_\ell(x,t) p_\ell(\mu), \tag{2.3}$$

where

$$\eta_\ell(x,t) = f_\ell(x,t) \quad \text{and} \quad \xi_\ell(x,t) = f_\ell(x,t) - f_\ell^N(x,t). \tag{2.4}$$

These components are orthogonal with respect to the $L^2(d\mu)$ inner product, i.e., $\int_{-1}^{1} \eta\xi d\mu = 0$.

Equations for the expansion coefficients $\{f_\ell\}_{\ell=0}^{\infty}$ are derived using the three-term recurrence relation for the Legendre polynomials:

$$\mu p_\ell(\mu) = a_\ell p_{\ell+1}(\mu) + a_{\ell-1} p_{\ell-1}(\mu), \tag{2.5}$$

where

$$\frac{1}{2} < a_\ell = \frac{\ell+1}{\sqrt{(2\ell+1)(2\ell+3)}} \leq \frac{1}{\sqrt{3}}. \tag{2.6}$$

By taking the $L^2(d\mu)$ inner product of (1.1a) with p_ℓ, $\ell = 0, \ldots, \infty$, and invoking (2.5), one arrives at an infinite system of equations for the expansion coefficients $\{f_\ell\}_{\ell=0}^{\infty}$:

$$\begin{cases} \epsilon\partial_t f_0 + a_0 \partial_x f_1 = 0, & \ell = 0, \\ \epsilon\partial_t f_\ell + a_\ell \partial_x f_{\ell+1} + a_{\ell-1}\partial_x f_{\ell-1} + \frac{1}{\epsilon} f_\ell = 0, & \ell \geq 1. \end{cases} \tag{2.7}$$

When applied to (1.5), the same procedure yields a similar set of equations for the coefficients $\{f_\ell^N\}_{\ell=0}^{N}$:

$$\begin{cases} \epsilon\partial_t f_0^N + a_0 \partial_x f_1^N = 0, & \ell = 0, \\ \epsilon\partial_t f_\ell^N + a_\ell \partial_x f_{\ell+1}^N + a_{\ell-1}\partial_x f_{\ell-1}^N + \frac{1}{\epsilon} f_\ell^N = 0, & 1 \leq \ell \leq N-1, \\ \epsilon\partial_t f_N^N + a_{N-1}\partial_x f_{N-1}^N + \frac{1}{\epsilon} f_N^N = 0, & \ell = N, \end{cases} \tag{2.8}$$

with initial condition $f_\ell^N(\cdot, 0) = f_\ell(\cdot, 0)$, for $\ell = 0, \ldots, N$. We refer to this system as the P_N system. The equations in (2.8) differ in form from the first $N+1$ equations of (2.7) only when $\ell = N$; it is this difference that is the origin of the error between Pf and f^N. Subtracting (2.8) from (2.7) yields a system equations for $\{\xi\}_{\ell=0}^{N}$, with an additional source term in the last equation:

$$\begin{cases} \epsilon\partial_t \xi_0 + a_0 \partial_x \xi_1 = 0, & \ell = 0, \\ \epsilon\partial_t \xi_\ell + a_\ell \partial_x \xi_{\ell+1} + a_{\ell-1}\partial_x \xi_{\ell-1} + \frac{1}{\epsilon} \xi_\ell = 0, & 1 \leq \ell \leq N-1, \\ \epsilon\partial_t \xi_N + a_{N-1}\partial_x \xi_{N-1} + \frac{1}{\epsilon} \xi_N = -a_N \partial_x f_{N+1}, & \ell = N. \end{cases} \tag{2.9}$$

with initial condition $\xi_\ell(\cdot, 0) = 0$, for $\ell = 0, \ldots, N$.

By taking Fourier transforms in x, we can write (2.7) in terms of the Legendre-Fourier coefficients of f.

$$\begin{cases} \epsilon\partial_t f_{0,k} + a_0 ik f_{1,k} = 0, & \ell = 0, \\ \epsilon\partial_t f_{\ell,k} + a_\ell ik f_{\ell+1,k} + a_{\ell-1} ik f_{\ell-1,k} + \frac{1}{\epsilon} f_{\ell,k} = 0, & \ell \geq 1. \end{cases} \tag{2.10}$$

Similarly, the Legendre-Fourier coefficients of f^N satisfy

$$\begin{cases} \epsilon\partial_t f_{0,k}^N + a_0 ik f_{1,k}^N = 0, & \ell = 0, \\ \epsilon\partial_t f_{\ell,k}^N + a_\ell ik f_{\ell+1,k}^N + a_{\ell-1} ik f_{\ell-1,k}^N + \frac{1}{\epsilon} f_{\ell,k}^N = 0, & 1 \leq \ell \leq N-1, \\ \epsilon\partial_t f_{N,k}^N + a_{N-1} ik f_{N-1,k}^N + \frac{1}{\epsilon} f_{N,k}^N = 0, & \ell = N, \end{cases} \tag{2.11}$$

and the Legendre-Fourier coefficients of ξ satisfy

$$\begin{cases} \epsilon\partial_t \xi_{0,k} + a_0 ik \xi_{1,k} = 0, & \ell = 0, \\ \epsilon\partial_t \xi_{\ell,k} + a_\ell ik \xi_{\ell+1,k} + a_{\ell-1} ik \xi_{\ell-1,k} + \frac{1}{\epsilon} \xi_{\ell,k} = 0, & 1 \leq \ell \leq N-1, \\ \epsilon\partial_t \xi_{N,k} + a_{N-1} ik \xi_{N-1,k} + \frac{1}{\epsilon} \xi_{N,k} = -a_N ik f_{N+1,k}, & \ell = N. \end{cases} \tag{2.12}$$

It turns out the behavior of the Legendre-Fourier coefficients $f_{\ell,k}$ and $\xi_{\ell,k}$ depends on the wave number k, with the long-time behavior being dominated by the low frequency parts. We therefore separate the coefficients into high and low frequency terms.

Definition 2.3 (High and low frequency parts)**.** *Let $\epsilon > 0$ be given and let $u \in L^2(d\mu dx)$ have Legendre and Legendre-Fourier coefficients as defined in Definition 2.1. Then u_ℓ can be decomposed into a high frequency part u_ℓ^{high} and a low frequency part u_ℓ^{low}, given by*

$$u_\ell^{\text{high}}(x) := \sum_{|k|\epsilon > \frac{1}{2}} u_{\ell,k} e^{ikx}, \quad \textit{and} \quad u_\ell^{\text{low}}(x) := \sum_{|k|\epsilon \leq \frac{1}{2}} u_{\ell,k} e^{ikx}, \tag{2.13}$$

respectively. Similarly, u can be decomposed into a high frequency part u^{high} and a low frequency part u^{low}, given by

$$u^{\text{high}}(x,\mu) := \sum_{\ell=0}^{\infty} u_\ell^{\text{high}}(x)p_\ell(\mu), \quad \textit{and} \quad u^{\text{low}}(x,\mu) := \sum_{\ell=0}^{\infty} u_\ell^{\text{low}}(x)p_\ell(\mu), \tag{2.14}$$

respectively.

2.2 Modified energy method

One may conclude from (1.2) that solutions of (1.1) dissipate the energy functional $\mathcal{H}\colon L^2(d\mu dx) \to \mathbb{R}$, given by

$$\mathcal{H}(u) = \|u\|^2_{L^2(d\mu dx)}. \tag{2.15}$$

A key tool in the proof of Theorem 1.2 is the spectral decomposition of $\mathcal{H}$.

Definition 2.4. *Given $\epsilon > 0$ and any $u \in L^2(d\mu dx)$ with Legendre-Fourier expansion in* (2.2)*, let*

$$\mathcal{H}_k^j(u) := \frac{1}{2}\sum_{\ell=j}^{\infty} |u_{\ell,k}|^2. \tag{2.16}$$

A direct consequence of Definitions 2.3 and 2.4 is that

$$\sum_{|k|\epsilon>\frac{1}{2}} \mathcal{H}_k^0(u) = \frac{1}{2}\|u^{\text{high}}\|^2_{L^2(d\mu dx)} \quad \text{and} \quad \sum_{|k|\epsilon\leq\frac{1}{2}} \mathcal{H}_k^0(u) = \frac{1}{2}\|u^{\text{low}}\|^2_{L^2(d\mu dx)}. \tag{2.17}$$

Since f satisfies (2.10), it follows that

$$\partial_t \mathcal{H}_k^0(f) + \frac{2}{\epsilon^2}\mathcal{H}_k^1(f) + ik\sum_{\ell=0}^{\infty} a_\ell \left(f^*_{\ell,k}f_{\ell+1,k} + f^*_{\ell+1,k}f_{\ell,k}\right) = 0. \tag{2.18}$$

For real-valued f, the real part of (2.18) gives

$$\partial_t \mathcal{H}_k^0(f) + \frac{2}{\epsilon^2}\mathcal{H}_k^1(f) = 0. \tag{2.19}$$

Thus $\mathcal{H}_k^0(f)$ is a non-increasing function of time. However, this is not enough to prove that $\mathcal{H}_k^0(f)$ decays to zero or how. In a similar calculation, (2.12) implies

$$\partial_t \mathcal{H}_k^0(\xi) + \frac{2}{\epsilon^2}\mathcal{H}_k^1(\xi) \leq \frac{a_N|k|}{\epsilon}|\xi_{N,k}||f_{N+1,k}| \leq \frac{1}{2\epsilon^2}|\xi_{N,k}|^2 + \frac{k^2}{6}|f_{N+1,k}|^2, \tag{2.20}$$

where the third expression is a direct consequence of Young's inequality and the bound on a_N from (2.6).

In order to estimate the decay rate of the $\mathcal{H}_k^0(f)$ or $\mathcal{H}_k^0(\xi)$, the energy needs to be modified. Thus following [14], we modify the energy by adding a compensating function.

Definition 2.5 (Compensating function)**.** *Given any $u \in L^2(d\mu dx)$ with Legendre-Fourier expansion in* (2.2)*, a compensating function for $\mathcal{H}_k^0(u)$ in* (2.16) *is a real-valued function*

$$h_k^\gamma(u) = -\frac{\gamma}{4a_0}\operatorname{Im}(u_{0,k}u^*_{1,k}), \tag{2.21}$$

where $\gamma \in \mathbb{R}$ is a positive scalar parameter to be determined and a_0 is the constant defined in (2.6)*.*

The role of the compensating function is elucidated by the following lemma

Lemma 2.6. *Let $u \in L^2(d\mu dx)$ have Legendre-Fourier coefficients that satisfy*

$$\begin{cases} \epsilon \partial_t u_{0,k} + a_0 ik u_{1,k} = 0, \\ \epsilon \partial_t u_{1,k} + a_1 ik u_{2,k} + a_0 ik u_{0,k} + \frac{1}{\epsilon} u_{1,k} = 0. \end{cases} \tag{2.22}$$

Then

$$(1 - \frac{\gamma}{2})\mathcal{H}_k^0(u) \leq (\mathcal{H}_k^0 + h_k^\gamma)(u) \leq (1 + \frac{\gamma}{2})\mathcal{H}_k^0(u); \tag{2.23}$$

and, for positive k, the time derivative is bounded by

$$\partial_t h_k^\gamma(u) \leq -\gamma \left(\frac{k}{16\epsilon} |u_{0,k}|^2 - \left(\frac{k}{4\epsilon} + \frac{3}{8\epsilon^3 k} \right) |u_{1,k}|^2 - \frac{k}{5\epsilon} |u_{2,k}|^2 \right). \tag{2.24}$$

Proof. From the definition of h_k^γ in (2.21) and the fact that $a_0 = 1/\sqrt{3}$, it follows that

$$|h_k^\gamma(u)| \leq \frac{\gamma}{2} |u_{0,k}||u_{1,k}| \leq \frac{\gamma}{2} \left(\frac{1}{2} |u_{0,k}|^2 + \frac{1}{2} |u_{1,k}|^2 \right) \leq \frac{\gamma}{2} \mathcal{H}_k^0(u), \tag{2.25}$$

which immediately implies (2.23). To derive (2.24), we differentiate (2.21) in time and use (2.22) to conclude that

$$\partial_t h_k^\gamma(u) = \frac{\gamma}{4} \left(-\frac{k}{\epsilon} |u_{0,k}|^2 + \frac{k}{\epsilon} |u_{1,k}|^2 + \frac{1}{a_0 \epsilon^2} \operatorname{Im}(u_{0,k} u_{1,k}^*) - \frac{a_1 k}{a_0 \epsilon} \operatorname{Re}(u_{0,k} u_{2,k}^*) \right). \tag{2.26}$$

Using Young's inequality, we compute bounds for the last two terms in (2.26):

$$\frac{1}{a_0 \epsilon^2} \operatorname{Im}(u_{0,k} u_{1,k}^*) \leq \frac{k}{2\epsilon} |u_{0,k}|^2 + \frac{1}{2 a_0^2 \epsilon^3 k} |u_{1,k}|^2 = \frac{k}{2\epsilon} |u_{0,k}|^2 + \frac{3}{2\epsilon^3 k} |u_{1,k}|^2 \tag{2.27}$$

and

$$-\frac{a_1 k}{a_0 \epsilon} \operatorname{Re}(u_{0,k} u_{2,k}^*) \leq \frac{k}{4\epsilon} |u_{0,k}|^2 + \frac{a_1^2 k}{a_0^2 \epsilon} |u_{2,k}|^2 = \frac{k}{4\epsilon} |u_{0,k}|^2 + \frac{4k}{5\epsilon} |u_{2,k}|^2. \tag{2.28}$$

These bounds, when substituted into (2.26), give (2.24). □

3 Proofs

This section is dedicated to the proof of Theorem 1.2, which proceeds in 4 steps. First, in Section 3.1, we determine bounds on the coefficients $f_{\ell,k}$ for $\ell = 0, \ldots, \infty$ and $k = -\infty, \ldots, \infty$. Second, in Section 3.2, we use the bounds on $f_{\ell,k}$ to estimate η. Third, in Section 3.3, we use the bound on $f_{N+1,k}$ to estimate ξ. Fourth, in Section 3.4, we compute finer estimates on ξ_ℓ^{low} for $\ell = 0, \ldots, N$. In Section 3.5, the results of these four steps are combined to prove Theorem 1.2. More specifically, the first three steps are used to establish the spectral error in (2.8), while the last is required to establish the moment errors given in (2.9).

In many cases, the proofs below rely on the decomposition of functions into high- and low-frequency components, as prescribed in Definition 2.3. Since we consider only real-valued functions $u \in L^2(d\mu dx)$, $u_{\ell,k}^* = u_{\ell,-k}$. Therefore $|u_{\ell,k}| = |u_{\ell,-k}|$, which means it is sufficient to consider only non-negative components of the Fourier spectrum, i.e., wave numbers $k \geq 0$.

3.1 Bounding the coefficients of f

In this section, we first use the method of modified energy to bound $\mathcal{H}_k^0(f)$ in Lemma 3.1. With such bounds and method of induction, we find bounds on $f_{\ell,k}$ in Lemma 3.2.

Lemma 3.1. *For any $g_0 \in L^2(dx)$,*

$$\mathcal{H}_k^0(f)(t) \leq \begin{cases} 6e^{-\frac{2\lambda_1 t}{\epsilon^2}} \mathcal{H}_k^0(g), & |k|\epsilon > \frac{1}{2}, \quad (3.1) \\ 6e^{-2\lambda_2 k^2 t} \mathcal{H}_k^0(g), & |k|\epsilon \leq \frac{1}{2}, \quad (3.2) \end{cases}$$

with $\lambda_1 = \frac{1}{45}$ (cf. (3.8)) and $\lambda_2 = \frac{4}{45}$ (cf. (3.12)).

Proof. We set $u = f$ in (2.24), add the result to (2.19), and use the fact that $\mathcal{H}_k^1(f) = \mathcal{H}_k^3(f) + \frac{1}{2}|f_{1,k}|^2 + \frac{1}{2}|f_{2,k}|^2$. This gives

$$\partial_t \left(\mathcal{H}_k^0(f) + h_k^\gamma(f)\right) \leq -\frac{2}{\epsilon^2}\mathcal{H}_k^3(f) - \sum_{\ell=0}^{2} c_{\gamma,\ell}|f_{\ell,k}|^2, \tag{3.3}$$

where

$$c_{\gamma,0} = \frac{\gamma k}{16\epsilon}, \qquad c_{\gamma,1} = \frac{1}{\epsilon^2} - \frac{\gamma k}{4\epsilon} - \frac{3\gamma}{8\epsilon^3 k}, \qquad c_{\gamma,2} = \frac{1}{\epsilon^2} - \frac{\gamma k}{5\epsilon}. \tag{3.4}$$

We next separate the frequency spectrum into high-frequency terms, when $k\epsilon > 1/2$, and low-frequency terms, when $0 \leq k\epsilon \leq 1/2$. The choice of γ and the subsequent estimates will depend on which part of the spectrum is being considered.

(i) **High frequency.** For $k\epsilon > 1/2$, we set

$$\gamma = \gamma^{\text{high}} := \frac{16}{29}\frac{1}{k\epsilon} < \frac{32}{29} \tag{3.5}$$

so that

$$c_{\gamma,0} = \frac{1}{29\epsilon^2}, \qquad c_{\gamma,1} = \left(\frac{1}{\epsilon^2} - \frac{4}{29\epsilon^2} - \frac{6}{29k^2\epsilon^4}\right) > \frac{1}{29\epsilon^2}, \qquad c_{\gamma,2} = \left(\frac{1}{\epsilon^2} - \frac{16}{145\epsilon^2}\right) > \frac{1}{29\epsilon^2}. \tag{3.6}$$

By substituting these bounds into (3.3), we find that

$$\partial_t \left(\mathcal{H}_k^0(f) + h_k^\gamma(f)\right) \leq -\frac{2}{\epsilon^2}\mathcal{H}_k^3(f) - \frac{1}{29\epsilon^2}\sum_{\ell=0}^{2}|f_{\ell,k}|^2 \leq -\frac{2}{29\epsilon^2}\mathcal{H}_k^0(f) \leq -\frac{2}{45\epsilon^2}\left(\mathcal{H}_k^0(f) + h_k^\gamma(f)\right), \tag{3.7}$$

where the last inequality uses the upper bound on $\mathcal{H}_k^0(f) + h_k^\gamma(f)$ in (2.23) and the upper bound on γ^{high} in (3.5). We integrate the inequality in (3.7) and apply the bounds in (2.23), using the fact that $\frac{1}{3} < \frac{13}{29} < 1 - \frac{\gamma}{2} < 1 + \frac{\gamma}{2} < \frac{45}{29} < 2$. This gives

$$\frac{1}{3}\mathcal{H}_k^0(f)(t) < \left(\mathcal{H}_k^0(f) + h_k^\gamma(f)\right)(t) \leq e^{-\frac{2\lambda_1 t}{\epsilon^2}}\left(\mathcal{H}_k^0(g) + h_k^\gamma(g)\right) < 2e^{-\frac{2\lambda_1 t}{\epsilon^2}}\mathcal{H}_k^0(g), \quad \lambda_1 = \frac{1}{45}, \tag{3.8}$$

from which (3.1) follows.

(ii) **Low frequency.** For $k = 0$, the result in (3.2) follows trivially from (2.22). For $0 < k\epsilon \leq \frac{1}{2}$, we let

$$\gamma = \gamma^{\text{low}} := \frac{64}{29}k\epsilon \leq \frac{32}{29} \tag{3.9}$$

so that

$$c_{\gamma,0} = \frac{4k^2}{29}, \qquad c_{\gamma,1} = \left(\frac{1}{\epsilon^2} - \frac{16k^2}{29} - \frac{24}{29\epsilon^2}\right) \geq \frac{4k^2}{29}, \qquad c_{\gamma,2} = \left(\frac{1}{\epsilon^2} - \frac{64}{145\epsilon^2}\right) > \frac{4k^2}{29}. \tag{3.10}$$

By substituting these bounds into (3.3), we find that

$$\partial_t \left(\mathcal{H}_k^0(f) + h_k^\gamma(f)\right) \leq -\frac{2}{\epsilon^2}\mathcal{H}_k^3(f) - \frac{4}{29}k^2\sum_{\ell=0}^{2}|f_{\ell,k}|^2 \leq -\frac{8}{29}k^2\mathcal{H}_k^0(f) \leq -\frac{8}{45}k^2\left(\mathcal{H}_k^0(f) + h_k^\gamma(f)\right), \tag{3.11}$$

where the last inequality uses the upper bound on H in (2.23) and the upper bound on γ^{low} in (3.9). We integrate the inequality in (3.11) and apply the bounds in (2.23), using the fact that $\frac{1}{3} < \frac{13}{29} \leq 1 - \frac{\gamma}{2} < 1 + \frac{\gamma}{2} \leq \frac{45}{29} < 2$. This gives

$$\frac{1}{3}\mathcal{H}_k^0(f)(t) < \left(\mathcal{H}_k^0(f) + h_k^\gamma(f)\right)(t) \leq e^{-2\lambda_2 k^2 t}\left(\mathcal{H}_k^0(g) + h_k^\gamma(g)\right) < 2e^{-2\lambda_2 k^2 t}\mathcal{H}_k^0(g), \quad \lambda_2 = \frac{4}{45}, \tag{3.12}$$

from which (3.2) follows.

□

Lemma 3.2. *Let $g_0 \in L^2(dx)$ be given. For $|k|\epsilon > 1/2$,*

$$|f_{\ell,k}|(t) \le \sqrt{12\mathcal{H}_k^0(g)}\, e^{-\frac{\lambda_1 t}{\epsilon^2}}, \quad \ell = 0, 1, 2, \ldots. \tag{3.13}$$

As a result,

$$\|f_\ell^{\text{high}}\|_{L^2(dx)}(t) \le \|f^{\text{high}}\|_{L^2(d\mu dx)}(t) \le \sqrt{6}\, \|g^{\text{high}}\|_{L^2(d\mu dx)} e^{-\frac{\lambda_1 t}{\epsilon^2}}, \quad \ell = 0, 1, 2, \ldots, \tag{3.14}$$

For $|k|\epsilon \le 1/2$,

$$|f_{\ell,k}|(t) \le C_\ell^k \epsilon^\ell k^\ell e^{-\lambda_2 k^2 t}, \quad \ell = 0, 1, 2, \ldots, \tag{3.15}$$

with

$$C_\ell^k(g) = \sqrt{12\mathcal{H}_k^0(g)} A^\ell \quad \textit{and} \quad A = \frac{2}{\sqrt{3}(1 - \lambda_2/4)} \simeq 1.2. \tag{3.16}$$

As a result,

$$\|f_\ell^{\text{low}}\|_{L^2(dx)}(t) \le F(g, \ell, t)\epsilon^\ell, \quad \ell = 0, 1, 2, \ldots \tag{3.17}$$

where

$$\begin{aligned} F(g, \ell, t) &= \left[2 \max_{k>0} (C_\ell^k)^2 \sum_{k>0} k^{2\ell} e^{-2\lambda_2 k^2 t} + \delta_{\ell,0} |g_{0,0}|^2 \right]^{\frac{1}{2}} \\ &= \left[24 \max_{k>0} \mathcal{H}_k^0(g) \sum_{k>0} (Ak)^{2\ell} e^{-2\lambda_2 k^2 t} + \delta_{\ell,0} |g_{0,0}|^2 \right]^{\frac{1}{2}} \end{aligned} \tag{3.18}$$

is positive, bounded for any $t > 0$, independent of k or ϵ, and monotonically decreasing with respect to t. As a result,

$$\|f_\ell\|_{L^2(dx)}(t) \le \sqrt{6} \|g^{\text{high}}\|_{L^2(d\mu dx)} e^{-\frac{\lambda_1 t}{\epsilon^2}} + F(g, \ell, t)\epsilon^\ell. \tag{3.19}$$

Proof. We again consider high and low frequencies separately.

(i) **High frequency.** For $|k|\epsilon > 1/2$, the definition of $\mathcal{H}_k^0$ in (2.16), along with bound in (3.1), implies that

$$|f_{\ell,k}|^2(t) \le 2\mathcal{H}_k^0(f)(t) \le 12\mathcal{H}_k^0(g) e^{-\frac{2\lambda_1 t}{\epsilon^2}}. \tag{3.20}$$

Taking square roots gives (3.13). We sum (3.20) over all k such that $|k|\epsilon > 1/2$ and use the definition of f_ℓ^{high} in (2.13) and the expression for $\mathcal{H}_k^0$ in (2.17) to conclude that

$$\|f_\ell^{\text{high}}\|_{L^2(dx)}^2(t) = \sum_{|k|\epsilon > \frac{1}{2}} |f_{\ell,k}|^2(t) \le \sum_{|k|\epsilon > \frac{1}{2}} 12\mathcal{H}_k^0(g) e^{-\frac{2\lambda_1 t}{\epsilon^2}} = 6\|g^{\text{high}}\|_{L^2(d\mu dx)}^2 e^{-\frac{2\lambda_1 t}{\epsilon^2}}. \tag{3.21}$$

Taking square roots gives (3.14).

(ii) **Low frequency.** To establish (3.15) for $0 \le k\epsilon \le 1/2$, we consider three cases, the first two of which are rather specific.

- Case 1: $\ell = 0$, $k = 0$. In this case, direct inspection of (2.10) shows that $f_{0,0}(t) = g_{0,0}$ is constant w.r.t. t. The assumption that g is isotropic implies that $g_{\ell,0}(t) = 0$ for $\ell \ge 1$ and all $t \ge 0$. Hence $\mathcal{H}_0^0(g) = \frac{1}{2}|g_{0,0}|^2$, whereby $C_0^0 = \sqrt{6}|g_{0,0}|$. Thus the bound in (3.15) is satisfied.
- Case 2: $\ell \ge 1$, $k = 0$. In this case, (2.10) implies that $f_{\ell,0}(t) = e^{-\frac{t}{\epsilon^2}} f_{\ell,0}(0)$. Hence, with the isotropic assumption on g, $f_{\ell,0}(t) = 0$. Thus the bound in (3.15) holds.

– Case 3: $\ell \geq 0$, $0 < k\epsilon \leq 1/2$. In this case, we actually prove the stronger statement

$$|f_{n,k}|(t) \leq C_\ell^k \epsilon^\ell k^\ell e^{-\lambda_2 k^2 t}, \qquad \ell \geq 0, \quad n \geq \ell, \tag{3.22}$$

with C_ℓ^k defined in (3.16). The result in (3.15) then follows by setting $n = \ell$ in (3.22). We proceed by induction on ℓ. According to the definition of $\mathcal{H}_k^0$ in (2.16) and the bound in (3.2),

$$|f_{n,k}|^2(t) \leq 2\mathcal{H}_k^0(f)(t) \leq 12\mathcal{H}_k^0(g)\, e^{-2\lambda_2 k^2 t}, \qquad \lambda_2 = \frac{4}{45}, \qquad n \geq 0. \tag{3.23}$$

Taking square roots in (3.23) recovers (3.22) for the case $\ell = 0$. Next, assume that (3.22) holds for $\ell = \ell_*$ for some $\ell_* \geq 0$ fixed. Using (2.10), the estimate (3.22) with $\ell = \ell_*$, and the fact that $a_n \leq 1/\sqrt{3}$ for all $n \geq 0$, we arrive at the following estimate for $|f_{n,k}|$ for all $n \geq \ell_* + 1$:

$$\partial_t |f_{n,k}| + \frac{1}{\epsilon^2}|f_{n,k}| \leq \frac{k}{\epsilon}\left(a_n|f_{n+1,k}| + a_{n-1}|f_{n-1,k}|\right) \leq \frac{2k}{\sqrt{3}\epsilon}\left(C_{\ell_*}^k \epsilon^{\ell_*} k^{\ell_*} e^{-\lambda_2 k^2 t}\right). \tag{3.24}$$

Thus integration of (3.24) in time (with an integrating factor on the left-hand side) gives

$$\begin{aligned}
|f_{n,k}|(t) &\leq e^{-\frac{t}{\epsilon^2}}|g_{n,k}| + \frac{2k}{\sqrt{3}\epsilon}\int_0^t e^{-\frac{t-s}{\epsilon^2}}\left(C_{\ell_*}^k \epsilon^{\ell_*} k^{\ell_*} e^{-\lambda_2 k^2 s}\right) ds \\
&= \frac{2k}{\sqrt{3}\epsilon}\int_0^t e^{-\frac{t-s}{\epsilon^2}}\left(C_{\ell_*}^k \epsilon^{\ell_*} k^{\ell_*} e^{-\lambda_2 k^2 s}\right) ds \\
&= \frac{2}{\sqrt{3}}\frac{C_{\ell_*}^k}{1-\lambda_2\epsilon^2 k}\epsilon^{\ell_*+1}k^{\ell_*+1}\left(e^{-\lambda_2 k^2 t} - e^{-\frac{t}{\epsilon^2}}\right) \\
&\leq AC_{\ell_*}^k \epsilon^{\ell_*+1}k^{\ell_*+1}e^{-\lambda_2 k^2 t} \\
&= C_{\ell_*+1}^k \epsilon^{\ell_*+1}k^{\ell_*+1}e^{-\lambda_2 k^2 t},
\end{aligned} \tag{3.25}$$

where $C_{\ell_*+1}^k$ is given in (3.16) and we have again used the fact that $g_{n,k} = 0$ for $n \geq 1$. This proves (3.22) and hence (3.15).

To show (3.17), we sum (3.15) over all low frequency values of k:

$$\begin{aligned}
\|f_\ell^{\text{low}}\|_{L^2(dx)}^2(t) &= \sum_{|k|\epsilon \leq \frac{1}{2}} |f_{\ell,k}|^2(t) \leq |f_{\ell,0}|^2(t) + 2\sum_{0<k\epsilon\leq\frac{1}{2}} |f_{\ell,k}|^2(t) \\
&\leq \delta_{\ell,0}|g_{0,0}|^2 + 2\sum_{0<k\epsilon\leq\frac{1}{2}}\left\{(C_\ell^k)^2\epsilon^{2\ell}k^{2\ell}e^{-2\lambda_2 k^2 t}\right\} \\
&\leq \delta_{\ell,0}|g_{0,0}|^2 + 2\epsilon^{2\ell}\max_{0<k\epsilon\leq\frac{1}{2}}(C_\ell^k)^2\sum_{0<k\epsilon\leq\frac{1}{2}} k^{2\ell}e^{-2\lambda_2 k^2 t} \\
&\leq F(g,\ell,t)^2\epsilon^{2\ell},
\end{aligned} \tag{3.26}$$

where $F(g,\ell,t)$ is given in (3.18). This proves (3.17).

□

Remark 3.3. *While $F(g,\ell,t)$ is independent of ϵ, it depends on ℓ. A more careful examination of this dependence is provided in Section 5.*

Remark 3.4. *The assumption that g is isotropic is critical to the proof above. More specifically, it is needed in order to ignore the contribution of the initial condition in the first line of (3.25). If $g_{n,k}$ is not zero, then the estimates are quite different and the proofs are much more complicated. We leave the analysis for anisotropic initial conditions to future work.*

Remark 3.5. *The proof above also works for the coefficients f_ℓ^N of the P_N system. Hence the estimates on f_ℓ in Lemma 3.2 also apply to f_ℓ^N. In Section 4, we use f_ℓ^N as a proxy for f_ℓ in order to verify these estimates numerically.*

3.2 Estimating η

In this section, we use the bounds on f_ℓ to bound η.

Lemma 3.6. *Let $g_0 \in L^2(dx)$ be given.*

$$\|\eta^{\text{high}}\|_{L^2(d\mu dx)}(t) \le \sqrt{6}\|g^{\text{high}}\|_{L^2(d\mu dx)} e^{-\frac{\lambda_1 t}{\epsilon^2}}, \tag{3.27}$$

and

$$\|\eta^{\text{low}}\|_{L^2(d\mu dx)}(t) \le \sqrt{2}F(g, N+1, t)\epsilon^{N+1}, \tag{3.28}$$

where F is given in (3.18) *with $\ell = N+1$. As a result,*

$$\|\eta\|_{L^2(d\mu dx)}(t) \le \sqrt{6}\|g^{\text{high}}\|_{L^2(d\mu dx)} e^{-\frac{\lambda_1 t}{\epsilon^2}} + \sqrt{2}F(g, N+1, t)\epsilon^{N+1}. \tag{3.29}$$

Proof. We prove (3.27) and (3.28), which combine to give (3.29).

(i) **High frequencies.** We first recall the high-frequency definitions in (2.13) and (2.14) and the definition of η in (2.3). We then apply (3.14). This gives

$$\|\eta^{\text{high}}\|_{L^2(d\mu dx)}(t) \le \|f^{\text{high}}\|_{L^2(d\mu dx)}(t) \le \sqrt{6}\|g^{\text{high}}\|_{L^2(d\mu dx)} e^{-\frac{\lambda_1 t}{\epsilon^2}}. \tag{3.30}$$

(ii) **Low frequencies.** We recall the low-frequency definitions in (2.13) and (2.14) and then apply the bound in (3.17). This gives

$$\|\eta^{\text{low}}\|^2_{L^2(d\mu dx)}(t) = \sum_{\ell=N+1}^{\infty} \|f_\ell^{\text{low}}\|^2_{L^2(dx)}(t) \le \sum_{\ell=N+1}^{\infty} \left[F(g,\ell,t)\epsilon^\ell\right]^2 \tag{3.31}$$

Using the definition of $F(g,\ell,t)$ in (3.18), we have

$$\begin{aligned}
\sum_{\ell=N+1}^{\infty} \left[F(g,\ell,t)\epsilon^\ell\right]^2 &= \sum_{\ell=N+1}^{\infty} \left(24 \max_{k>0} \mathcal{H}_k^0(g) \sum_{k>0} (Ak)^{2\ell} e^{-2\lambda_2 k^2 t} \epsilon^{2\ell} \right) && (3.32)\\
&= \epsilon^{2(N+1)} 24 \max_{k>0} \mathcal{H}_k^0(g) \sum_{k>0} (Ak)^{2(N+1)} e^{-2\lambda_2 k^2 t} \sum_{\ell=0}^{\infty} (Ak\epsilon)^{2\ell} && (3.33)\\
&= \left[F(g,N+1,t)\right]^2 \epsilon^{2(N+1)} \sum_{\ell=0}^{\infty} (Ak\epsilon)^{2\ell}. && (3.34)
\end{aligned}$$

Recall that $0 < A < 1.2$. Therefore, $(Ak\epsilon) < 0.6 < 1$ and

$$\sum_{\ell=0}^{\infty} (Ak\epsilon)^{2\ell} = \frac{1}{1-(Ak\epsilon)^2} < 2. \tag{3.35}$$

□

3.3 Estimating ξ

With bounds on $f_{N+1,k}$ in (3.13), we use method of modified energy to bound $\mathcal{H}_k^0(\xi)$ and then estimate ξ.

Lemma 3.7. *Let $g_0 \in H^1(dx)$, then*

$$\mathcal{H}_k^0(\xi)(t) \le \begin{cases} 6te^{-\frac{2\lambda_1 t}{\epsilon^2}} \mathcal{H}_k^0(\partial_x g), & |k|\epsilon > \frac{1}{2}, \quad (3.36)\\ \frac{t}{2}[C_{N+1}^k]^2 k^{2(N+2)} e^{-2\lambda_2 k^2 t} \epsilon^{2(N+1)}, & |k|\epsilon \le \frac{1}{2}, \quad (3.37) \end{cases}$$

with $\lambda_1 = \frac{1}{45}$, $\lambda_2 = \frac{4}{45}$ *and* C^k_{N+1} *defined in* (3.16). *Hence*

$$\|\xi^{\text{high}}\|_{L^2(d\mu dx)}(t) \leq \sqrt{6t} e^{-\frac{\lambda_1 t}{\epsilon^2}} \|\partial_x g^{\text{high}}\|_{L^2(d\mu dx)} \tag{3.38}$$

and

$$\|\xi^{\text{low}}\|_{L^2(d\mu dx)}(t) \leq \frac{\sqrt{t}}{A} F(g, N+2, t)\epsilon^{N+1}, \tag{3.39}$$

where F *is given in* (3.18) *with* $\ell = N+2$. *As a result,*

$$\|\xi\|_{L^2(d\mu dx)}(t) \leq \sqrt{6t}\|\partial_x g\|_{L^2(d\mu dx)} e^{-\frac{\lambda_1 t}{\epsilon^2}} + \frac{\sqrt{t}}{A} F(g, N+2, t)\epsilon^{N+1}. \tag{3.40}$$

Proof. The proof relies on the same calculations as Lemma 3.1, but must incorporate the presence of a source term in the energy equation. (Compare (2.19) to (2.20).) We set $u = \xi$ in (2.24), add the result to (2.20), and use the fact that $\mathcal{H}^1_k(\xi) = \mathcal{H}^3_k(\xi) + \frac{1}{2}|\xi_{1,k}|^2 + \frac{1}{2}|\xi_{2,k}|^2$. This gives

$$\begin{aligned}
\partial_t \left(\mathcal{H}^0_k(\xi) + h^\gamma_k(\xi)\right) &\leq -\frac{2}{\epsilon^2}\mathcal{H}^3_k(\xi) - \sum_{\ell=0}^{2} c_{\gamma,\ell}|\xi_{\ell,k}|^2 + \frac{1}{2\epsilon^2}|\xi_{N,k}|^2 + \frac{k^2}{6}|f_{N+1,k}|^2 \\
&\leq -\frac{1}{\epsilon^2}\mathcal{H}^3_k(\xi) - \sum_{\ell=0}^{2} c_{\gamma,\ell}|\xi_{\ell,k}|^2 + \frac{k^2}{6}|f_{N+1,k}|^2,
\end{aligned} \tag{3.41}$$

where the coefficients $c_{\gamma,0}$, $c_{\gamma,1}$, and $c_{\gamma,2}$ are defined in (3.4) and the term $\frac{1}{2\epsilon^2}|\xi_{N,k}|^2$ in the first line has been absorbed by $-\frac{2}{\epsilon^2}\mathcal{H}^3_k(\xi)$.[2] As in the proof of Lemma 3.1, we separate the frequency spectrum of ξ into high frequency and low-frequency parts, and choose γ appropriately in each case.

(i) **High frequency.** For $k\epsilon > 1/2$, we set $\gamma = \gamma^{\text{high}}$ (defined in (3.5)) into (3.41) and repeat the arguments in part (i) of the proof of Lemma 3.1. This gives

$$\partial_t \left(\mathcal{H}^0_k(\xi) + h^\gamma_k(\xi)\right) \leq -\frac{2\lambda_1}{\epsilon^2}(\mathcal{H}^0_k(\xi) + h^\gamma_k(\xi)) + \frac{k^2}{6}|f_{N+1,k}|^2. \tag{3.42}$$

We integrate (3.42) in time. Using (3.13) to evaluate $|f_{N+1,k}|$ and the fact that $\left(\mathcal{H}^0_k(\xi) + h^\gamma_k(\xi)\right)(0) = 0$, we find that

$$\left(\mathcal{H}^0_k(\xi) + h^\gamma_k(\xi)\right)(t) \leq \frac{k^2}{6}\int_0^t e^{-\frac{2\lambda_1(t-s)}{\epsilon^2}}|f_{N+1,k}|^2(s)ds \leq 2te^{-\frac{2\lambda_1 t}{\epsilon^2}}k^2\mathcal{H}^0_k(g). \tag{3.43}$$

Applying the left bound from (2.23), we find

$$\mathcal{H}^0_k(\xi)(t) \leq 6te^{-\frac{2\lambda_1 t}{\epsilon^2}}k^2\mathcal{H}^0_k(g) = 6te^{-\frac{2\lambda_1 t}{\epsilon^2}}\mathcal{H}^0_k(\partial_x g), \tag{3.44}$$

which is (3.36). Then (3.38) is recovered by summing over all high frequency values of k.

(ii) **Low frequency.** When $k = 0$, (2.12) implies that $\xi_{\ell,k} = 0$ (since the initial condition is zero by definition). For $0 < k\epsilon \leq 1/2$, we set $\gamma = \gamma^{\text{low}}$ (defined in (3.9)) into (3.41) and repeat the arguments in part (ii) of the proof of Lemma 3.1. This gives

$$\partial_t \left(\mathcal{H}^0_k(\xi) + h^\gamma_k(\xi)\right) \leq -2\lambda_2 k^2 \left(\mathcal{H}^0_k(\xi) + h^\gamma_k(\xi)\right) + \frac{k^2}{6}|f_{N+1,k}|^2. \tag{3.45}$$

[2]The cost of combining these two terms is that the coefficient of $\mathcal{H}^3_k(\xi)$ in (3.41) is only half the coefficient of $\mathcal{H}^3_k(f)$ in (3.3). However, the bound with respect to these coefficients is very loose. Hence the estimates in the proof of Lemma 3.1 follow, except for the source term.

We integrate (3.45) in time, using the fact $\left(\mathcal{H}_k^0(\xi)+h_k^\gamma(\xi)\right)(0)=0$ and the estimate in (3.15) for $|f_{N+1,k}|$. This gives

$$\begin{aligned}\left(\mathcal{H}_k^0(\xi)+h_k^\gamma(\xi)\right)(t) &\leq \frac{k^2}{6}e^{-2\lambda_2 k^2 t}\int_0^t e^{2\lambda_2 k^2 s}|f_{N+1,k}|^2\,ds \\ &\leq \frac{t}{6}(C_{N+1}^k)^2 k^{2(N+2)}e^{-2\lambda_2 k^2 t}\epsilon^{2(N+1)}. \end{aligned} \tag{3.46}$$

To arrive at (3.37) from (3.46), we use the fact that $\mathcal{H}_k^0(\xi)\leq 3(\mathcal{H}_k^0(\xi)+h_k^\gamma(\xi))$ (cf. (2.23)). Then to establish (3.39), we sum (3.37) over the low frequency values of k:

$$\begin{aligned}\|\xi^{\text{low}}\|^2_{L^2(d\mu dx)}(t) &\leq \sum_{|k|\epsilon\leq\frac{1}{2}}\frac{12t}{A^2}(Ak)^{2(N+2)}e^{-2\lambda_2 k^2 t}\mathcal{H}_k^0(g)\epsilon^{2(N+1)} \\ &\leq \frac{24t}{A^2}\epsilon^{2(N+1)}\max_{0<k\epsilon\leq\frac{1}{2}}\mathcal{H}_k^0(g)\sum_{0<k\epsilon\leq\frac{1}{2}}(Ak)^{2(N+2)}e^{-2\lambda_2 k^2 t} \\ &\leq \frac{t}{A^2}\epsilon^{2(N+1)}\left[F(g,N+2,t)\right]^2, \end{aligned} \tag{3.47}$$

with F defined in (3.18).

□

3.4 Finer estimate on ξ_ℓ^{low}

In Lemma 3.7, we proved an ϵ-dependent estimate for ξ^{low}. In this section, we combine the method of induction with a reduced version of the modified energy to refine the estimate for ξ_ℓ^{low}.

Lemma 3.8. *Let $g_0\in L^2(dx)$ be given. For $|k|\epsilon\leq\frac{1}{2}$,*

$$|\xi_{\ell,k}|(t)\leq\begin{cases}\tilde{C}_{N-1}^k(t)\epsilon^{2N}k^{3N}e^{-\lambda_2 k^2 t}, & \ell=0, \quad (3.48)\\ \tilde{C}_{N+1-\ell}^k(t)\epsilon^{2N+2-\ell}k^{3N+4-2\ell}e^{-\lambda_2 k^2 t}, & 1\leq\ell\leq N, \quad (3.49)\end{cases}$$

where

$$\tilde{C}_{N+1-\ell}^k(t)=M(t)\tilde{C}_{N-\ell}^k(t),\quad \textit{and}\quad \tilde{C}_1^k(t)=\max\{1,t^{1/2}\}\hat{C}(t)C_{N+1}^k, \tag{3.50}$$

with

$$M(t)=\frac{2\max\{1,t^{1/2}\}}{\sqrt{3}(1-\lambda_2/4)}\quad \textit{and}\quad \hat{C}(t)=\frac{2t^{1/2}+k^{-1}}{\sqrt{3}(1-\lambda_2/4)}, \tag{3.51}$$

and C_{N+1}^k defined in (3.16).

Proof. For $k=0$, (3.48) and (3.49) hold, since direct inspection of (2.12) shows that $\xi_{\ell,0}(t)=0$ for all $\ell=0,1,\cdots,N$. We therefore consider only $0<k\leq\frac{1}{2\epsilon}$. In this case, we establish (3.49) and (3.48) by proving the stronger statements:

(1) for $2\leq\ell\leq N$,

$$|\xi_{n,k}|(t)\leq\tilde{C}_{N+1-\ell}^k(t)\epsilon^{2N+2-\ell}k^{3N+4-2\ell}e^{-\lambda_2 k^2 t},\quad 0\leq n\leq\ell; \tag{3.52}$$

(2) for $\ell=1$,

$$|\xi_{1,k}|(t)\leq\tilde{C}_N^k(t)\epsilon^{2N+1}k^{3N+1}e^{-\lambda_2 k^2 t}. \tag{3.53}$$

When $\ell=0$, (3.48) is recovered by setting $\ell=2$ and $n=0$ in (3.52). When $\ell=1$, (3.49) is recovered by (3.53). When $2\leq\ell\leq N$, (3.49) is recovered by setting $n=\ell$ in (3.52).

(1) To prove (3.52), we use the method of induction, starting with N and working backward.

(a) For the initial step in the induction, we need to show that for $\ell = N$,

$$|\xi_{n,k}|(t) \le \tilde{C}_1^k(t)\epsilon^{N+2}k^{N+4}e^{-\lambda_2 k^2 t}, \quad n = 0, 1, \ldots, N. \tag{3.54}$$

We prove (3.54) in two sub-steps.

(i) The first sub-step is to show (3.54) for $n = 1, \ldots, N$. We combine the last two equations of (2.12). This gives

$$\partial_t|\xi_{n,k}| + \frac{1}{\epsilon^2}|\xi_{n,k}| \le \frac{k}{\epsilon}\left(a_{n-1}|\xi_{n-1,k}| + (1-\delta_{n,N})a_n|\xi_{n+1,k}| + \delta_{n,N}a_N|f_{N+1,k}|\right), \tag{3.55}$$

for $1 \le n \le N$. It follows from (3.37) that

$$|\xi_{\ell,k}|(t) \le t^{1/2}C_{N+1}^k\epsilon^{N+1}k^{N+2}e^{-\lambda_2 k^2 t}, \quad \ell = 0, 1, \cdots, N, \quad 0 < k \le \frac{1}{2\epsilon}, \tag{3.56}$$

where C_{N+1}^k is defined in (3.16). We use (3.56) to estimate $\xi_{n-1,k}$ and $\xi_{n+1,k}$ and (3.15) to estimate $f_{N+1,k}$. Then (3.55) reduces to

$$\partial_t|\xi_{n,k}| + \frac{1}{\epsilon^2}|\xi_{n,k}| \le \frac{1}{\sqrt{3}}\left(2t^{1/2} + \frac{1}{k}\right)C_{N+1}^k\epsilon^N k^{N+3}e^{-\lambda_2 k^2 t}, \quad 1 \le n \le N. \tag{3.57}$$

We then integrate in time, using the zero initial condition for $\xi_{n,k}$ to find

$$|\xi_{n,k}|(t) \le \hat{C}(t)C_{N+1}^k\epsilon^{N+2}k^{N+3}e^{-\lambda_2 k^2 t}, \quad 1 \le n \le N, \tag{3.58}$$

with $\hat{C}(t)$ defined in (3.51). Since $\hat{C}C_{N+1}^k \le \tilde{C}_1^k$, (3.58) verifies (3.54) for $n = 1, \ldots, N$. [3]

(ii) The second sub-step is to use the result of (i) to show that (3.54) holds for $n = 0$. Since using (2.12) directly will result in order reduction by one power of ϵ, we instead consider the smaller system for $\{\xi_{n,k}\}_{n=0}^{N-1}$ and treat $\xi_{N,k}$ as a source term:

$$\begin{cases} \epsilon\partial_t\xi_{0,k} + a_0 ik\xi_{1,k} = 0, & n = 0; \\ \epsilon\partial_t\xi_{n,k} + a_n ik\xi_{n+1,k} + a_{n-1}ik\xi_{n-1,k} + \frac{1}{\epsilon}\xi_{n,k} = 0, & 1 \le n \le N-2; \\ \epsilon\partial_t\xi_{N-1,k} + a_{N-2}ik\xi_{N-2,k} + \frac{1}{\epsilon}\xi_{N-1,k} = -a_{N-1}ik\xi_{N,k}, & n = N-1. \end{cases} \tag{3.59}$$

We then repeat the arguments used to establish (3.37) using the estimate for $\xi_{N,k}$ in (3.58) instead of the estimate for $f_{N+1,k}$. This procedure requires the introduction of a new functional

$$\mathcal{H}_k^{j,i} \colon u \mapsto \frac{1}{2}\sum_{n=j}^{i}|u_{n,k}|^2, \quad u \in L^2(d\mu dx), \tag{3.60}$$

which is defined such that $\mathcal{H}_k^{j,N} = \mathcal{H}_k^j$ (cf. (2.16)). With $\mathcal{H}_k^{0,N-1}(\xi)$ and compensating function $h_k^\gamma(\xi)$, defined in (2.21), one can first derive a differential inequality analogous to (3.41) and then follow the arguments in part (ii) of the proof of Lemma 3.7. The result is

$$\mathcal{H}_k^{0,N-1}(\xi)(t) \le \frac{t}{2}\hat{C}(t)^2(C_{N+1}^k)^2\epsilon^{2(N+2)}k^{2(N+4)}e^{-2\lambda_2 k^2 t}, \tag{3.61}$$

and then

$$|\xi_{n,k}|(t) \le t^{1/2}\hat{C}(t)C_{N+1}^k\epsilon^{N+2}k^{N+4}e^{-\lambda_2 k^2 t}, \quad 0 \le n \le N-1. \tag{3.62}$$

As compared to (3.37), the extra powers of ϵ and k in (3.62) come from the higher powers in the estimate for $\xi_{N,k}$ in (3.58) when compared to the estimate for $f_{N+1,k}$ in (3.15). Since $t^{1/2}\hat{C}C_{N+1}^k \le \tilde{C}_1^k$, (3.62) verifies (3.54) for $n = 0$.

[3] Note that power of k in (3.58) is $N+3$, which is actually better than the estimate in (3.54). However in the second substep, an additional power of k is needed (cf. (3.62)) in order to gain an additional factor of ϵ in the estimate for $\xi_{0,k}$.

(b) For the next step of the induction, we assume that for some $3 \le \ell_* \le N$ fixed, (3.52) holds for $\ell = \ell_*$:

$$|\xi_{n,k}|(t) \le \tilde{C}^k_{N+1-\ell_*}(t)\epsilon^{2N+2-\ell_*}k^{3N+4-2\ell_*}e^{-\lambda_2 k^2 t}, \quad 0 \le n \le \ell_*. \tag{3.63}$$

We would like to show

$$|\xi_{n,k}|(t) \le \tilde{C}^k_{N+1-(\ell_*-1)}(t)\epsilon^{2N+2-(\ell_*-1)}k^{3N+4-2(\ell_*-1)}e^{-\lambda_2 k^2 t}, \quad 0 \le n \le \ell_* - 1. \tag{3.64}$$

(i) We first show (3.64) for $1 \le n \le \ell_* - 1$. Using (2.12), the estimate (3.63) for $\xi_{n-1,k}$ and $\xi_{n+1,k}$ and the fact that $a_{n-1} \le 1/\sqrt{3}$ and $a_n \le 1/\sqrt{3}$, we derive the following estimate for $|\xi_{n,k}|$ for $1 \le n \le \ell_* - 1$:

$$\begin{aligned}\partial_t|\xi_{n,k}| + \frac{1}{\epsilon^2}|\xi_{n,k}| &\le \frac{k}{\epsilon}\left(a_{n-1}|\xi_{n-1,k}| + a_n|\xi_{n+1,k}|\right)\\ &\le \frac{2}{\sqrt{3}}\tilde{C}^k_{N+1-\ell_*}(t)\epsilon^{2N+1-\ell_*}k^{3N+5-2\ell_*}e^{-\lambda_2 k^2 t}.\end{aligned} \tag{3.65}$$

Integration of (3.65) in time gives

$$|\xi_{n,k}|(t) \le \frac{2}{\sqrt{3}(1-\lambda_2/4)}\tilde{C}^k_{N+1-\ell_*}(t)\epsilon^{2N+3-\ell_*}k^{3N+5-2\ell_*}e^{-\lambda_2 k^2 t}, \tag{3.66}$$

which recovers (3.64), using the definition of $\tilde{C}^k_{N+2-\ell_*}$ in (3.50).

(ii) We next show (3.64) for $n = 0$, repeating the argument from step (a)(ii). We consider the smaller system for $\{\xi_{n,k}\}_{n=0}^{\ell_*-2}$ and treat $\xi_{\ell_*-1,k}$ as a source term:

$$\begin{cases} \epsilon\partial_t\xi_{0,k} + a_0 ik\xi_{1,k} = 0, & n = 0;\\ \epsilon\partial_t\xi_{n,k} + a_n ik\xi_{n+1,k} + a_{n-1}ik\xi_{n-1,k} + \frac{1}{\epsilon}\xi_{n,k} = 0, & 1 \le n \le \ell_* - 3;\\ \epsilon\partial_t\xi_{\ell_*-2,k} + a_{\ell_*-3}ik\xi_{\ell_*-3,k} + \frac{1}{\epsilon}\xi_{\ell_*-2,k} = -a_{\ell_*-2}ik\xi_{\ell_*-1,k}, & n = \ell_* - 2.\end{cases} \tag{3.67}$$

Using the energy $\mathcal{H}_k^{0,\ell_*-2}(\xi)$ defined in (3.60) and compensating function $h_k^\gamma(\xi)$, defined in (2.21), we have

$$\mathcal{H}_k^{0,\ell_*-2}(\xi)(t) \le \frac{2t}{\left(\sqrt{3}(1-\lambda_2/4)\right)^2}\left(\tilde{C}^k_{N+1-\ell_*}(t)\right)^2\epsilon^{2(2N+3-\ell_*)}k^{2(3N+6-2\ell_*)}e^{-2\lambda_2 k^2 t}, \tag{3.68}$$

and then

$$|\xi_{n,k}|(t) \le \frac{2t^{1/2}}{\sqrt{3}(1-\lambda_2/4)}\tilde{C}^k_{N+1-\ell_*}(t)\epsilon^{2N+3-\ell_*}k^{3N+6-2\ell_*}e^{-\lambda_2 k^2 t}, \quad 0 \le n \le \ell_* - 2. \tag{3.69}$$

This estimate is analogous to (3.62).

(2) To prove (3.53), one just need to repeat the argument in (b)(i) with $\ell_* = 2$.

□

The coefficients $\tilde{C}^k_{N-1}(t)$ and $\tilde{C}^k_{N+1-\ell}(t)$ in the estimates for $\xi_{\ell,k}(t)$ can be replaced by some time independent coefficients, at the cost of a reduced decay rate in the error.

Lemma 3.9. *Let $g_0 \in L^2(dx)$ be given. For $|k|\epsilon \le \frac{1}{2}$,*

$$|\xi_{\ell,k}|(t) \le \begin{cases}\bar{C}(N,2)C^k_{N+1}\epsilon^{2N}k^{3N}e^{-\frac{\lambda_2}{2}k^2 t}, & \ell = 0, \quad (3.70)\\ \bar{C}(N,\ell)C^k_{N+1}\epsilon^{2N+2-\ell}k^{3N+4-2\ell}e^{-\frac{\lambda_2}{2}k^2 t}, & 1 \le \ell \le N, \quad (3.71)\end{cases}$$

where

$$\bar{C}(N,\ell) = 2A^{N-\ell+1}\left(\frac{N-\ell+2}{\lambda_2}\right)^{\frac{N-\ell+2}{2}}e^{-\frac{N-\ell}{2}+\frac{\lambda_2}{2}-1} \tag{3.72}$$

and C^k_{N+1} is defined in (3.16). *Hence*

$$\|\xi^{\text{low}}_\ell\|_{L^2(dx)}(t) \le \begin{cases} E(g,N,2,t)\epsilon^{2N}, & \ell = 0, \quad (3.73) \\ E(g,N,\ell,t)\epsilon^{2N+2-\ell}, & 1 \le \ell \le N, \quad (3.74) \end{cases}$$

where

$$E(g,N,\ell,t) = \bar{C}(N,\ell)A^{-2N-3+2\ell}F(g,3N+4-2\ell,t/2) \tag{3.75}$$

and F is defined in (3.18).

Proof. The strategy is simple: use part of the exponentially decaying term in (3.48) and (3.49) to control powers of t in the other coefficients. Since $\max\{1,t^{1/2}\} \le (t+1)^{1/2}$, (3.50) and (3.51) imply the following bound:

$$\begin{aligned} \tilde{C}^k_{N+1-\ell}(t)e^{-\lambda_2 k^2 t} &\le A^{N-\ell+1}(t+1)^{(N-\ell+1)/2}\left(t^{1/2}+\frac{1}{2k}\right)C^k_{N+1}e^{-\lambda_2 k^2 t} \\ &\le 2A^{N-\ell+1}(t+1)^{(N-\ell+2)/2}e^{-\frac{\lambda_2}{2}t}C^k_{N+1}e^{-\frac{\lambda_2}{2}k^2 t}, \quad k>1, \end{aligned} \tag{3.76}$$

The product $2A^{N-\ell+1}(t+1)^{(N-\ell+2)/2}e^{-\frac{\lambda_2}{2}t}$ takes its maximum value $\bar{C}(N,\ell)$ at $t = (N-\ell+2)/\lambda_2 - 1$. This proves (3.70) and (3.71).

To establish (3.73) and (3.74), we sum (3.70) and (3.71), respectively, over all low frequency values of k. For example, summing (3.71) gives

$$\begin{aligned} \|\xi^{\text{low}}_\ell\|^2_{L^2(dx)}(t) &= \sum_{|k|\epsilon\le\frac{1}{2}} |\xi_{\ell,k}|^2(t) \le 2\sum_{0<k\epsilon\le\frac{1}{2}}\left(\bar{C}(N,\ell)C^k_{N+1}\epsilon^{2N+2-\ell}k^{3N+4-2\ell}e^{-\frac{\lambda_2}{2}k^2 t}\right)^2 \\ &\le (\bar{C}(N,\ell))^2\epsilon^{2(2N+2-\ell)}2\max_{0<k\epsilon\le\frac{1}{2}}(C^k_{N+1})^2\sum_{0<k\epsilon\le\frac{1}{2}}k^{2(3N+4-2\ell)}e^{-\lambda_2 k^2 t} \\ &= (\bar{C}(N,\ell))^2\epsilon^{2(2N+2-\ell)}24\max_{0<k\epsilon\le\frac{1}{2}}\mathcal{H}^0_k(g)A^{2(N+1)}\sum_{0<k\epsilon\le\frac{1}{2}}k^{2(3N+4-2\ell)}e^{-\lambda_2 k^2 t}, \end{aligned} \tag{3.77}$$

which yields (3.74). Then (3.73) is derived similarly.

□

Remark 3.10. *One could easily prove bounds of the form* (3.73) *and* (3.74) *by using* (3.48) *and* (3.49) *directly, with the coefficient E replaced by $\tilde{E}(g,N,\ell,t) = \max\{1,t^{1/2}\}\hat{C}(t)M(t)^{N-\ell}A^{-2N-3+2\ell}F(g,3N+4-2\ell,t)$. Although $\tilde{E}$ decays more quickly for large t (due to the difference in the third argument in F), the time-dependent factor $\max\{1,t^{1/2}\}\hat{C}(t)M(t)^{N-\ell}$ makes the analysis of $\tilde{E}$ as a function of N more difficult. We instead use E in* (3.75) *because it is easier to bound its growth with respect to N. This fact will be useful in Section 5.2.*

3.5 Proof of Theorem 1.2

We first prove the error estimate for f^N. Since $f - f^N = \eta + \xi$, we simply apply the triangle inequality and combine the estimates for η in (3.29) and ξ in (3.40), and get

$$\begin{aligned} \|f-f^N\|_{L^2(d\mu dx)}(t) \le& \sqrt{6}\|g^{\text{high}}\|_{L^2(d\mu dx)}e^{-\frac{\lambda_1 t}{\epsilon^2}} + \sqrt{2}F(g,N+1,t)\epsilon^{N+1} \\ &+ \sqrt{6t}\|\partial_x g\|_{L^2(d\mu dx)}e^{-\frac{\lambda_1 t}{\epsilon^2}} + \frac{\sqrt{t}}{A}F(g,N+2,t)\epsilon^{N+1}. \end{aligned} \tag{3.78}$$

This establishes (1.9) with constants

$$B(g) = \sqrt{6}\|g\|_{L^2(d\mu dx)}, \quad C(\partial_x g) = \sqrt{6}\|\partial_x g\|_{L^2(d\mu dx)}, \quad D(g,N,t) = \sqrt{2}F(g,N+1,t) + \frac{\sqrt{t}}{A}F(g,N+2,t). \tag{3.79}$$

We next prove the error estimate for f_ℓ^N. Since $f_\ell - f_\ell^N = \xi_\ell = \xi_\ell^{\text{high}} + \xi_\ell^{\text{low}}$, we combine the estimate (3.38) with (3.73) and (3.74). After some additional trivial bounds ($\|\xi_\ell^{\text{high}}\|_{L^2(dx)} \le \|\xi^{\text{high}}\|_{L^2(d\mu dx)}$ and $\|\partial_x g^{\text{high}}\|_{L^2(d\mu dx)} \le \|\partial_x g\|_{L^2(d\mu dx)}$), we arrive at

$$\|f_\ell - f_\ell^N\|_{L^2(dx)}(t) \le \begin{cases} \sqrt{6t}e^{-\frac{\lambda_1 t}{\epsilon^2}}\|\partial_x g\|_{L^2(d\mu dx)} + E(g,N,2,t)\epsilon^{2N}, & \ell = 0, \quad (3.80) \\ \sqrt{6t}e^{-\frac{\lambda_1 t}{\epsilon^2}}\|\partial_x g\|_{L^2(d\mu dx)} + E(g,N,\ell,t)\epsilon^{2N+2-\ell}, & 1 \le \ell \le N, \quad (3.81) \end{cases}$$

with E defined in (3.75). This establishes (1.10) and completes the proof.

4 Numerical Examples

In this section, we perform numerical tests to demonstrate the theoretical results, by exploring different values of ϵ, N, and the initial condition g. All calculations are based on the P_N system (2.8). Since the exact solution of f is not readily available, we use f^N with $N = 65$ as a reference solution in order to calculate L^2 errors, and as discussed in Remark 3.5, we use f_ℓ^N as a proxy for the coefficients f_ℓ of the exact solution in order to test the asymptotic estimates in Lemma 3.2.

For the spatial discretization of (2.8), we use a Fourier-Galerkin method, typically with 100 Fourier modes, although more modes are added as needed to ensure that the spatial error neglected can be neglected. The method is implemented with a fast Fourier transform (FFT) algorithm. What remains is an ODE for the Fourier-Galerkin coefficients that can be solved exactly (up to machine precision). In some situations, the size of the coefficients differs by many orders of magnitude. Thus in order to rule out the effect of cumulative round-off error that this discrepancy may create, we use the Multiprecision Computing Toolbox for MATLAB by Advanpix LLC. [1] with 250 digits.

Example 4.1. We start with the kinetic equation (1.1) with three different initial conditions:

$$\begin{aligned} g(x,\mu) &= g^{(1)}(x) = 1 + 1_{[-\frac{\pi}{2},\frac{\pi}{2}]}(x), \\ g(x,\mu) &= g^{(2)}(x) = 1 + \cos(x)1_{[-\frac{\pi}{2},\frac{\pi}{2}]}(x), \\ g(x,\mu) &= g^{(3)}(x) = 1 + \cos(x). \end{aligned} \quad (4.1)$$

Simple calculations with Fourier analysis imply $g^{(1)} \in H^q(dx)$ for $q < \frac{1}{2}$ and $g^{(2)} \in H^q(dx)$ for $q < \frac{3}{2}$, while $g^{(3)}$ is a smooth function. However, $g^{(1)}$ does not satisfy the regularity assumption in Theorem 1.2 which is required for the high-frequency bound (3.36) in Lemma 3.7.

We solve the P_N system (2.8) with $\epsilon = 2 \cdot 4^{-m}$, with $m = 1, \dots, 5$, and $N = 1, 2, \cdots, 5$. For each ϵ and N, $L^2(d\mu dx)$ errors with respect to the reference solution are listed in Table 1. The convergence rates of the coefficients f_ℓ^N for P_4 and P_5 are listed in Tables 2 and 3. The convergence rates of the errors ξ_ℓ for f_ℓ^N are listed in Tables 4 and 5.

For all three initial conditions, we observe the convergence rates for f^N indicated by (1.9) in Table 1, the convergence rates for ξ_ℓ indicated by (1.10) in Tables 4 and 5, and the covergence rates for f_ℓ^N indicated by Remark 3.5 in Tables 2 and 3. We observe these rates even for $g^{(1)}$, which is not in $H^1(dx)$ and thus does not satisfy the hypothesis used to prove these estimates. This discrepancy may be due to the fact that the Fourier-Galerkin method uses a finite number of waves and thus the numerical approximation of $g^{(1)}$ is in $H^1(dx)$. However, even with $10,000$ Fourier modes, the results do not change. Thus while $g^{(1)} \in H^1(dx)$ may be a necessary condition, it may be impossible to verify it numerically in this example.

5 The benefit of increasing N

The goal of any apriori error estimate is to provide an indication of how the accuracy of an approximation will improve as a given parameter varies. For example, the spectral estimate in (1.7) suggests that $e^N = f - f^N$ behaves like

$$\frac{\|e^{N+1}\|}{\|e^N\|} \sim \left(\frac{N}{N+1}\right)^q, \quad (5.1)$$

$g^{(1)}$

ϵ	P_1 error	order	P_2 error	order	P_3 error	order	P_4 error	order	P_5 error	order
1/2	6.34E-02		3.27E-02		2.07E-02		1.48E-02		1.17E-02	
1/8	2.60E-03	2.30	1.71E-04	3.79	1.97E-05	5.02	3.39E-06	6.05	6.39E-07	7.08
1/32	1.60E-04	2.01	2.50E-06	3.05	7.09E-08	4.06	3.00E-09	5.07	1.39E-10	6.08
1/128	9.99E-06	2.00	3.89E-08	3.00	2.76E-10	4.00	2.91E-12	5.00	3.38E-14	6.01
1/512	6.24E-07	2.00	6.07E-10	3.00	1.08E-12	4.00	2.84E-15	5.00	8.24E-18	6.00

$g^{(2)}$

ϵ	P_1 error	order	P_2 error	order	P_3 error	order	P_4 error	order	P_5 error	order
1/2	4.39E-02		1.59E-02		6.63E-03		3.44E-03		2.21E-03	
1/8	2.41E-03	2.09	1.76E-04	3.25	1.88E-05	4.23	2.27E-06	5.28	2.90E-07	6.45
1/32	1.49E-04	2.01	2.65E-06	3.03	7.04E-08	4.03	2.11E-09	5.04	6.64E-11	6.05
1/128	9.31E-06	2.00	4.14E-08	3.00	2.74E-10	4.00	2.05E-12	5.00	1.62E-14	6.00
1/512	5.82E-07	2.00	6.46E-10	3.00	1.07E-12	4.00	2.00E-15	5.00	3.94E-18	6.00

$g^{(3)}$

ϵ	P_1 error	order	P_2 error	order	P_3 error	order	P_4 error	order	P_5 error	order
1/2	5.78E-02		1.24E-02		2.37E-03		3.94E-04		5.79E-05	
1/8	3.51E-03	2.02	2.14E-04	2.93	1.35E-05	3.73	8.50E-07	4.43	5.34E-08	5.04
1/32	2.18E-04	2.00	3.31E-06	3.01	5.21E-08	4.01	8.18E-10	5.01	1.28E-11	6.01
1/128	1.36E-05	2.00	5.17E-08	3.00	2.03E-10	4.00	7.98E-13	5.00	3.13E-15	6.00
1/512	8.50E-07	2.00	8.07E-10	3.00	7.94E-13	4.00	7.80E-16	5.00	7.64E-19	6.00

Table 1: Errors and convergence rates for the P_N solutions in Example 4.1. According to Theorem 1.2, the theoretical order of convergence is $N + 1$.

$g^{(1)}$

ϵ	f_0^N	order	f_1^N	order	f_2^N	order	f_3^N	order	f_4^N	order
1/2	2.16E+00		1.43E-01		3.84E-02		1.48E-02		1.35E-02	
1/8	2.16E+00	0.00	3.32E-02	1.06	2.19E-03	2.07	1.62E-04	3.25	1.96E-05	4.72
1/32	2.16E+00	0.00	8.25E-03	1.00	1.36E-04	2.01	2.49E-06	3.01	7.09E-08	4.05
1/128	2.16E+00	0.00	2.06E-03	1.00	8.48E-06	2.00	3.89E-08	3.00	2.76E-10	4.00
1/512	2.16E+00	0.00	5.16E-04	1.00	5.30E-07	2.00	6.07E-10	3.00	1.08E-12	4.00

$g^{(2)}$

ϵ	f_0^N	order	f_1^N	order	f_2^N	order	f_3^N	order	f_4^N	order
1/2	1.91E+00		1.21E-01		3.65E-02		1.28E-02		5.90E-03	
1/8	1.90E+00	0.00	2.73E-02	1.07	1.99E-03	2.10	1.73E-04	3.10	1.88E-05	4.15
1/32	1.90E+00	0.00	6.78E-03	1.00	1.23E-04	2.01	2.65E-06	3.01	7.04E-08	4.03
1/128	1.90E+00	0.00	1.69E-03	1.00	7.69E-06	2.00	4.14E-08	3.00	2.74E-10	4.00
1/512	1.90E+00	0.00	4.23E-04	1.00	4.81E-07	2.00	6.46E-10	3.00	1.07E-12	4.00

$g^{(3)}$

ϵ	f_0^N	order	f_1^N	order	f_2^N	order	f_3^N	order	f_4^N	order
1/2	1.61E+00		2.24E-01		5.50E-02		1.20E-02		2.36E-03	
1/8	1.59E+00	0.01	5.20E-02	1.05	3.36E-03	2.02	2.13E-04	2.91	1.35E-05	3.72
1/32	1.59E+00	0.00	1.29E-02	1.00	2.09E-04	2.00	3.31E-06	3.01	5.21E-08	4.01
1/128	1.59E+00	0.00	3.23E-03	1.00	1.30E-05	2.00	5.17E-08	3.00	2.03E-10	4.00
1/512	1.59E+00	0.00	8.08E-04	1.00	8.15E-07	2.00	8.07E-10	3.00	7.94E-13	4.00

Table 2: Convergence rates of the coefficients f_ℓ^N for the P_4 solution in Example 4.1. According to Lemma 3.2, the theoretical order of convergence is ℓ.

$g^{(1)}$

ϵ	f_0^N	order	f_1^N	order	f_2^N	order	f_3^N	order	f_4^N	order	f_5^N	order
1/2	2.16E+00		1.43E-01		3.80E-02		1.54E-02		1.00E-02		8.71E-03	
1/8	2.16E+00	0.00	3.32E-02	1.06	2.19E-03	2.06	1.62E-04	3.28	1.90E-05	4.52	3.39E-06	5.66
1/32	2.16E+00	0.00	8.25E-03	1.00	1.36E-04	2.01	2.49E-06	3.01	7.08E-08	4.03	3.00E-09	5.07
1/128	2.16E+00	0.00	2.06E-03	1.00	8.48E-06	2.00	3.89E-08	3.00	2.76E-10	4.00	2.91E-12	5.00
1/512	2.16E+00	0.00	5.16E-04	1.00	5.30E-07	2.00	6.07E-10	3.00	1.08E-12	4.00	2.84E-15	5.00

$g^{(2)}$

ϵ	f_0^N	order	f_1^N	order	f_2^N	order	f_3^N	order	f_4^N	order	f_5^N	order
1/2	1.91E+00		1.21E-01		3.65E-02		1.30E-02		4.91E-03		2.85E-03	
1/8	1.90E+00	0.00	2.73E-02	1.07	1.99E-03	2.10	1.73E-04	3.11	1.85E-05	4.02	2.27E-06	5.15
1/32	1.90E+00	0.00	6.78E-03	1.00	1.23E-04	2.01	2.65E-06	3.01	7.03E-08	4.02	2.11E-09	5.04
1/128	1.90E+00	0.00	1.69E-03	1.00	7.69E-06	2.00	4.14E-08	3.00	2.74E-10	4.00	2.05E-12	5.00
1/512	1.90E+00	0.00	4.23E-04	1.00	4.81E-07	2.00	6.46E-10	3.00	1.07E-12	4.00	2.00E-15	5.00

$g^{(3)}$

ϵ	f_0^N	order	f_1^N	order	f_2^N	order	f_3^N	order	f_4^N	order	f_5^N	order
1/2	1.61E+00		2.24E-01		5.50E-02		1.20E-02		2.30E-03		3.93E-04	
1/8	1.59E+00	0.01	5.20E-02	1.05	3.36E-03	2.02	2.13E-04	2.91	1.35E-05	3.71	8.50E-07	4.43
1/32	1.59E+00	0.00	1.29E-02	1.00	2.09E-04	2.00	3.31E-06	3.01	5.21E-08	4.01	8.18E-10	5.01
1/128	1.59E+00	0.00	3.23E-03	1.00	1.30E-05	2.00	5.17E-08	3.00	2.03E-10	4.00	7.98E-13	5.00
1/512	1.59E+00	0.00	8.08E-04	1.00	8.15E-07	2.00	8.07E-10	3.00	7.94E-13	4.00	7.80E-16	5.00

Table 3: Convergence rates of the coefficients f_ℓ^N for the P_5 solution in Example 4.1. According to Lemma 3.2, the theoretical order of convergence is ℓ.

$g^{(1)}$

ϵ	ξ_0	order	ξ_1	order	ξ_2	order	ξ_3	order	ξ_4	order
1/2	5.93E-03		3.62E-03		4.78E-03		4.71E-03		5.87E-03	
1/8	6.71E-08	8.22	1.35E-08	9.02	2.62E-08	8.74	1.18E-07	7.64	6.17E-07	6.61
1/32	9.34E-13	8.07	4.45E-14	9.11	3.20E-13	8.16	6.59E-12	7.06	1.39E-10	6.06
1/128	1.42E-17	8.00	1.68E-19	9.01	4.81E-18	8.01	4.00E-16	7.00	3.38E-14	6.00
1/512	2.16E-22	8.00	6.42E-25	9.00	7.33E-23	8.00	2.44E-20	7.00	8.25E-18	6.00

$g^{(2)}$

ϵ	ξ_0	order	ξ_1	order	ξ_2	order	ξ_3	order	ξ_4	order
1/2	8.37E-04		4.47E-04		7.31E-04		8.59E-04		1.35E-03	
1/8	2.28E-08	7.58	5.15E-09	8.20	7.69E-09	8.27	3.89E-08	7.21	2.85E-07	6.11
1/32	3.01E-13	8.10	1.71E-14	9.10	8.96E-14	8.19	2.25E-12	7.04	6.65E-11	6.03
1/128	4.55E-18	8.01	6.46E-20	9.01	1.35E-18	8.01	1.37E-16	7.00	1.62E-14	6.00
1/512	6.94E-23	8.00	2.46E-25	9.00	2.05E-23	8.00	8.34E-21	7.00	3.95E-18	6.00

$g^{(3)}$

ϵ	ξ_0	order	ξ_1	order	ξ_2	order	ξ_3	order	ξ_4	order
1/2	1.62E-08		9.57E-08		8.87E-07		7.49E-06		5.70E-05	
1/8	5.52E-11	4.10	9.91E-12	6.62	2.13E-10	6.01	3.36E-09	5.56	5.32E-08	5.03
1/32	9.48E-16	7.91	3.47E-17	9.06	3.21E-15	8.01	2.02E-13	7.01	1.28E-11	6.01
1/128	1.46E-20	8.00	1.32E-22	9.00	4.89E-20	8.00	1.23E-17	7.00	3.13E-15	6.00
1/512	2.22E-25	8.00	5.02E-28	9.00	7.46E-25	8.00	7.53E-22	7.00	7.65E-19	6.00

Table 4: Convergence rates of the error ξ_ℓ in the coefficients f_ℓ^N for the P_4 solution in Example 4.1. According to (1.10) in Theorem 1.2, the theoretical order of convergence is $2N = 8$ for f_0^N and $2N + 2 - \ell = 10 - \ell$ for f_ℓ^N, $\ell = 1, \ldots, 4$.

$g^{(1)}$

ϵ	ξ_0	order	ξ_1	order	ξ_2	order	ξ_3	order	ξ_4	order	ξ_5	order
1/2	4.06E-03		2.96E-03		3.32E-03		3.28E-03		3.62E-03		4.55E-03	
1/8	4.02E-09	9.97	1.13E-09	10.66	1.23E-09	10.68	4.42E-09	9.75	2.27E-08	8.64	1.18E-07	7.62
1/32	3.13E-15	10.15	2.10E-16	11.18	7.85E-16	10.29	1.52E-14	9.08	3.12E-13	8.07	6.56E-12	7.07
1/128	2.95E-21	10.01	4.93E-23	11.01	7.31E-22	10.02	5.75E-20	9.01	4.74E-18	8.00	3.98E-16	7.00
1/512	2.81E-27	10.00	1.17E-29	11.00	6.96E-28	10.00	2.19E-25	9.00	7.22E-23	8.00	2.43E-20	7.00

$g^{(2)}$

ϵ	ξ_0	order	ξ_1	order	ξ_2	order	ξ_3	order	ξ_4	order	ξ_5	order
1/2	5.07E-04		2.72E-04		4.34E-04		4.69E-04		5.59E-04		8.59E-04	
1/8	1.42E-09	9.22	3.26E-10	9.84	4.12E-10	10.00	1.27E-09	9.25	6.31E-09	8.22	3.89E-08	7.22
1/32	1.15E-15	10.12	6.76E-17	11.10	2.54E-16	10.32	4.28E-15	9.09	8.73E-14	8.07	2.23E-12	7.04
1/128	1.09E-21	10.01	1.60E-23	11.01	2.35E-22	10.02	1.62E-20	9.01	1.32E-18	8.00	1.36E-16	7.00
1/512	1.03E-27	10.00	3.80E-30	11.00	2.24E-28	10.00	6.17E-26	9.00	2.02E-23	8.00	8.29E-21	7.00

$g^{(3)}$

ϵ	ξ_0	order	ξ_1	order	ξ_2	order	ξ_3	order	ξ_4	order	ξ_5	order
1/2	1.09E-10		8.15E-10		9.02E-09		9.28E-08		8.74E-07		7.45E-06	
1/8	2.10E-13	4.51	3.97E-14	7.16	8.41E-13	6.69	1.32E-11	6.39	2.10E-10	6.01	3.34E-09	5.56
1/32	2.33E-19	9.89	8.53E-21	11.07	7.89E-19	10.01	4.98E-17	9.01	3.16E-15	8.01	2.01E-13	7.01
1/128	2.24E-25	9.99	2.02E-27	11.00	7.51E-25	10.00	1.90E-22	9.00	4.82E-20	8.00	1.23E-17	7.00
1/512	2.14E-31	10.00	4.82E-34	11.00	7.16E-31	10.00	7.23E-28	9.00	7.35E-25	8.00	7.49E-22	7.00

Table 5: Convergence rates of the error ξ_ℓ in the coefficients f_ℓ^N for the P_5 solution in Example 4.1. According to (1.10) in Theorem 1.2, the theoretical order of convergence is $2N = 10$ for f_0^N and $2N + 2 - \ell = 12 - \ell$ for f_ℓ^N, $\ell = 1, \ldots, 5$.

where q is related to the regularity of f. Thus the gain realized by increasing N to $N+1$ is not expected to be large, especially when q is small. On the other hand, the estimate in (1.9) of Theorem 1.2 suggests that by increasing N, we gain an additional factor of ϵ in the error estimate; that is,

$$\frac{\|e^{N+1}\|}{\|e^N\|} \sim \epsilon. \tag{5.2}$$

Similarly, (1.10) of Theorem 1.2 suggests that we gain an additional factor of ϵ^2 in the error estimate for $e_\ell^N := \xi_\ell = f_\ell - f_\ell^N$; that is

$$\frac{\|e_\ell^{N+1}\|}{\|e_\ell^N\|} \sim \epsilon^2. \tag{5.3}$$

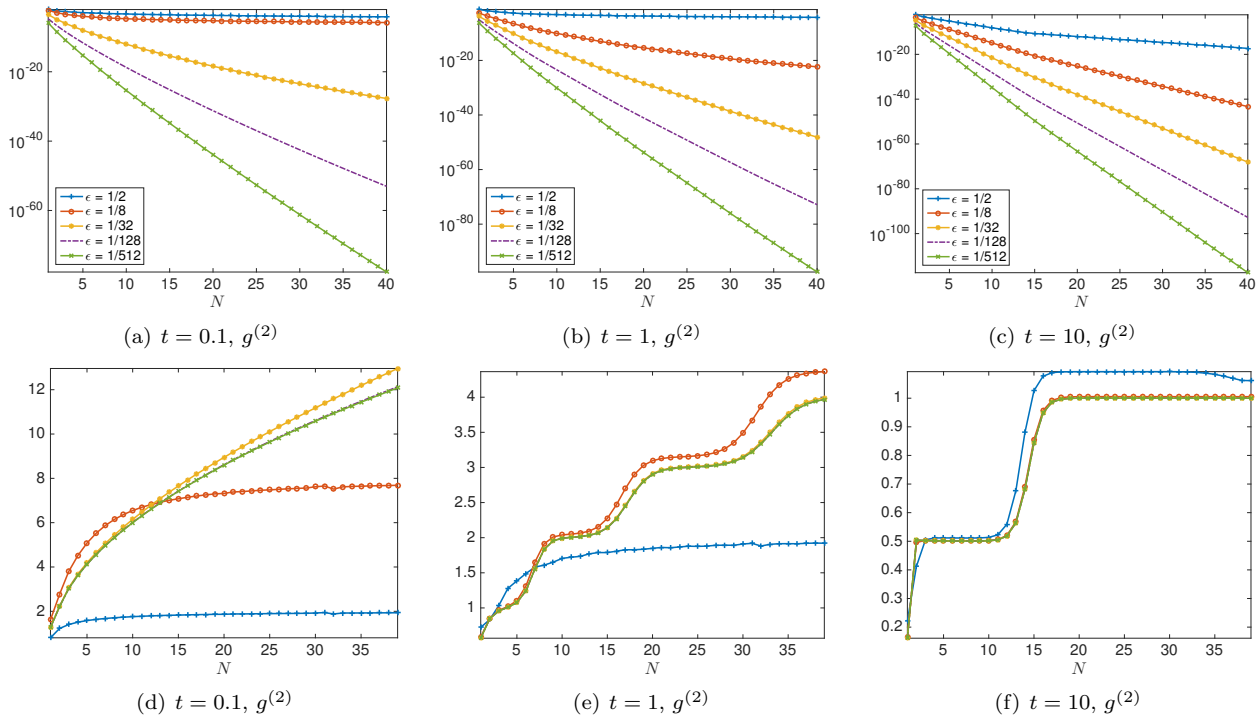
However, the statements in (5.2) and (5.3) are misleading since, unlike the spectral estimate in (1.7), the ϵ-dependent estimates in (1.9) and (1.10) include coefficients that depend on N and ℓ. We begin by exploring this question numerically.

5.1 Numerical experiments

Example 5.1. We return to Example 4.1 from Section 4 with initial condition $g^{(2)}$. We again use the P_{65} solution as a reference. We compute P_N solutions with $\epsilon = 2 \cdot 4^{-m}$, $m = 1, \ldots, 5$, and values of N up to 40, which in practice is quite large. We examine the solutions at times $t = 0.1, 1, 10$.

As before the spatial discretization is a Fourier-Galerkin method that uses fast Fourier transforms (FFT) for implementation. For most cases, the spatial grid has 100 points. However, for smaller t or larger ϵ, gradients in x become larger; in such cases, more points are needed to ensure that the spacial discretization error can be neglected. Specifically, for $t = 0.1$ and $\epsilon = 1/8$, 1000 points are used; for $t = 0.1$ and $\epsilon = 1/2$, 2500 points are used; for $t = 1$ and $\epsilon = 1/2$, 1000 points are used.

In Figure 1, the ratio in (5.2) (normalized by ϵ) is plotted as a function of N. In Figures 2–4, the ratio in (5.3) (normalized by ϵ^2) is plotted for $\ell = 0, 1, 2$. We observe the following trends:



(a) $t = 0.1$, $g^{(2)}$ (b) $t = 1$, $g^{(2)}$ (c) $t = 10$, $g^{(2)}$

(d) $t = 0.1$, $g^{(2)}$ (e) $t = 1$, $g^{(2)}$ (f) $t = 10$, $g^{(2)}$

Figure 1: Results from Example 5.1. Top figures: $\|e^N\|$. Bottom figures: $\frac{\|e^{N+1}\|}{\|e^N\|\epsilon}$.

1) Larger values of t lead to smaller error ratios. Numerically, we find that for $1 \le N \le 40$,

$$\frac{\|e^{N+1}\|}{\|e^N\|} \le G_1(N,t)\epsilon, \quad \text{where} \quad G_1(N,t) \le \begin{cases} 13, & t = 0.1, \\ 4.5, & t = 1, \\ 1.1, & t = 10, \end{cases} \tag{5.4}$$

and for $1 \le N \le 40$ and $0 \le \ell \le 2$,

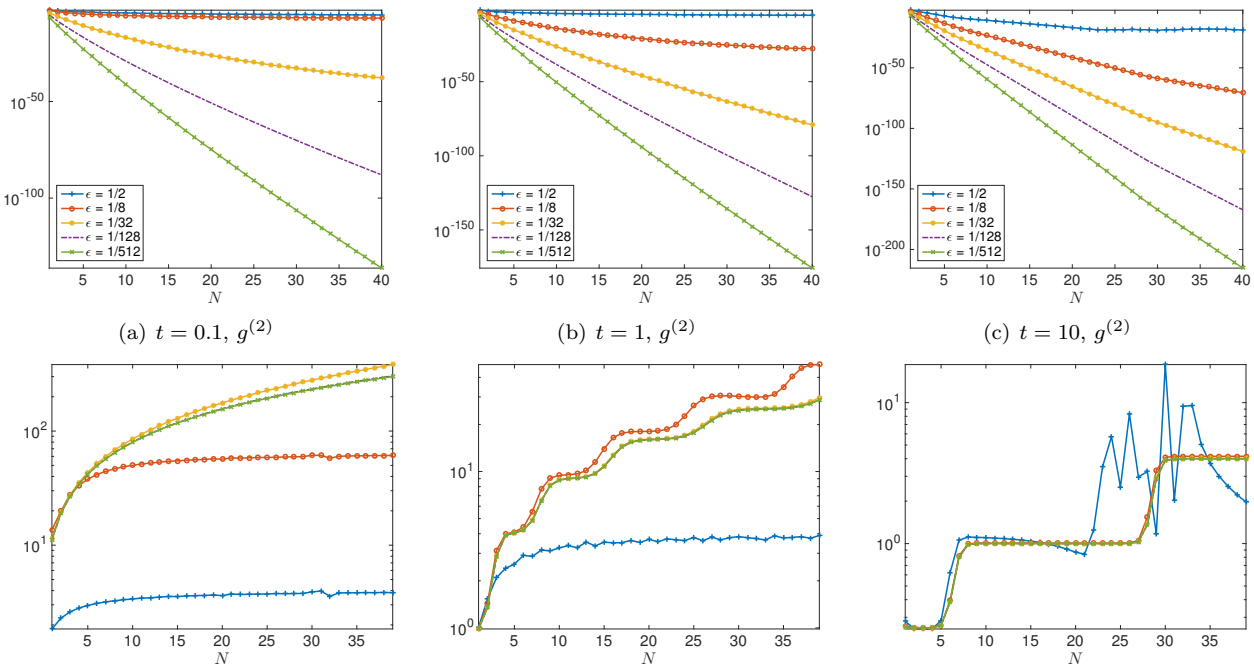
$$\frac{\|e_\ell^{N+1}\|}{\|e_\ell^N\|} \le G_2(N,t)\epsilon^2, \quad \text{where} \quad G_2(N,t) \le \begin{cases} 400, & t = 0.1, \\ 50, & t = 1, \\ 20, & t = 10. \end{cases} \tag{5.5}$$

2) For fixed t, the solution profiles of the normalized error ratios appear to convergence at ϵ decreases.

3) As N varies, the solution profiles of the normalized error ratios exhibit plateaus with sharp transitions in between. We do not yet understand the origin of this behavior.

Example 5.2. We repeat the previous test, this time using the initial condition $g^{(3)}$ from Example 4.1. Because $g^{(3)}$ is smooth, 20 grid points are sufficient to ensure that the spatial error in the Fourier-Galerkin discretization is negligible. For large values of N, the errors are so small that 300 digits are used.

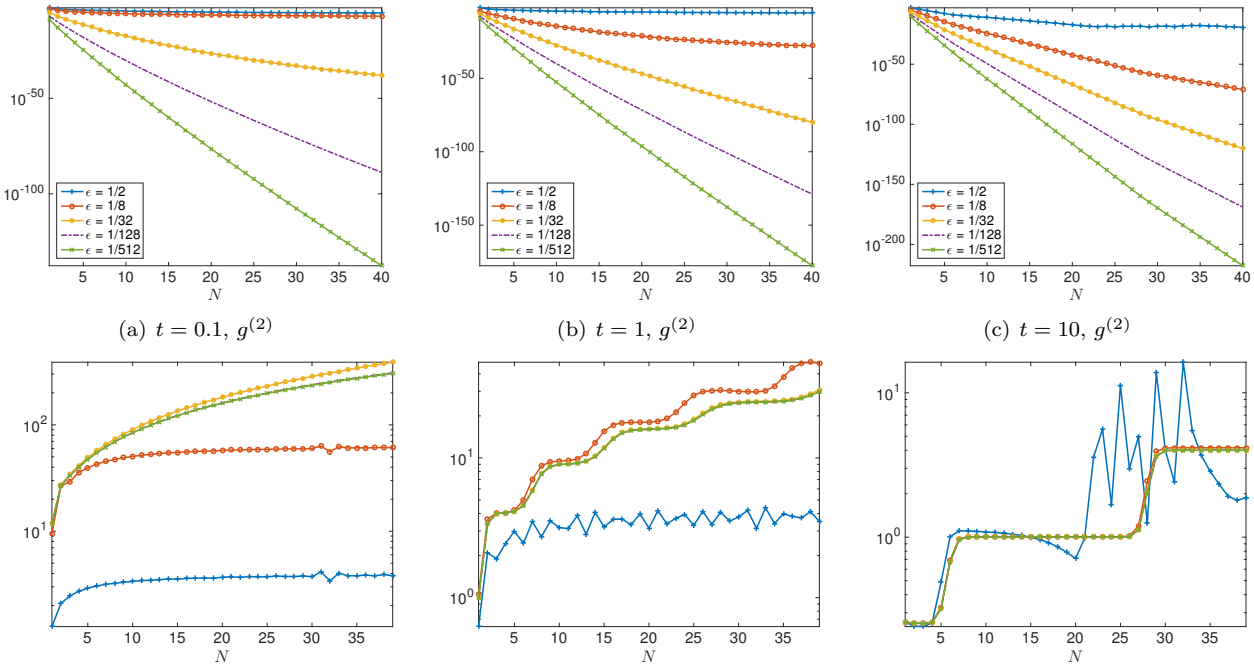
In Figure 5, the ratio in (5.2) (normalized by ϵ) is plotted as a function of N. In Figures 6–8, the ratio in (5.3) (normalized by ϵ^2) is plotted for $\ell = 0, 1, 2$ as a function of N. As in the previous example, profiles of the normalized error ratios appear to convergence at ϵ decreases. However, unlike the previous example, the ratios do not appear to decay significantly as time increases. Indeed, they are already less than one for $t = 0.1$. Numerically, we see that $e^{N+1} \le 0.5\epsilon e^N$ and $e_\ell^{N+1} \le 0.25\epsilon^2 e_\ell^N$ $(\ell = 0, 1, 2)$ for all three tested value of t. Also, we do not observe the plateaus and transitions seen in the previous example.



(a) $t = 0.1$, $g^{(2)}$ (b) $t = 1$, $g^{(2)}$ (c) $t = 10$, $g^{(2)}$

(d) $t = 0.1$, $g^{(2)}$ (e) $t = 1$, $g^{(2)}$ (f) $t = 10$, $g^{(2)}$

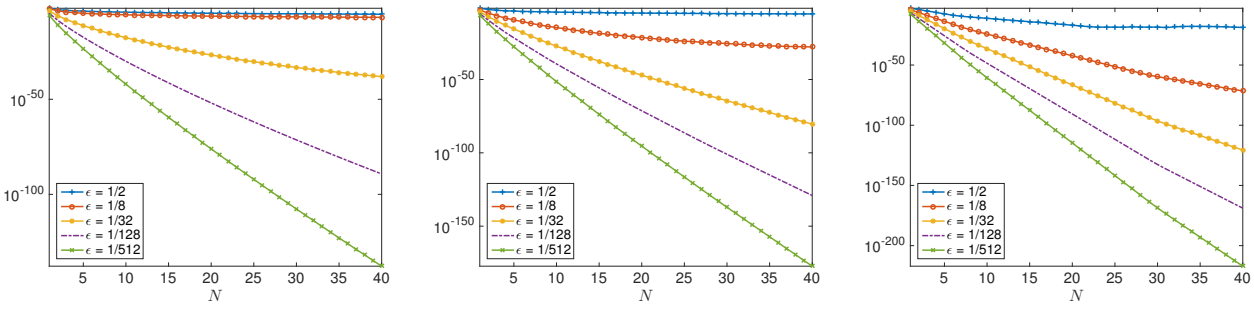
Figure 2: Results from Example 5.1. Top figures: $\|e_0^N\|$. Bottom figures: $\frac{\|e_0^{N+1}\|}{\|e_0^N\|\epsilon^2}$.



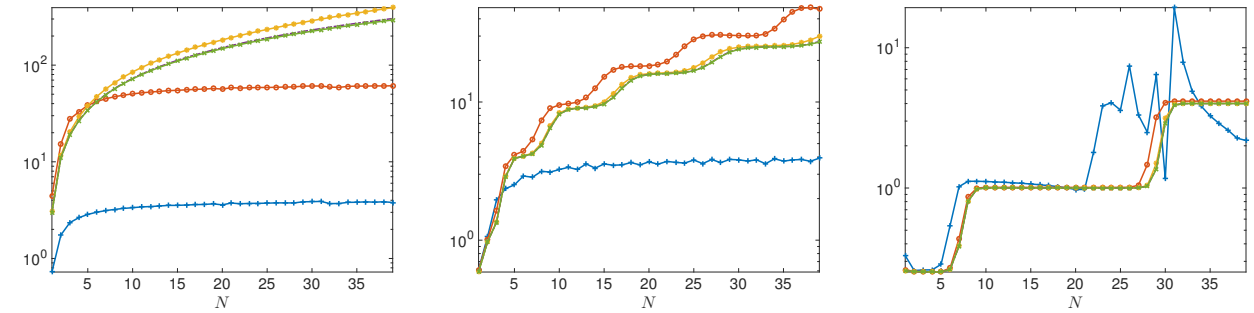
(a) $t = 0.1$, $g^{(2)}$ (b) $t = 1$, $g^{(2)}$ (c) $t = 10$, $g^{(2)}$

(d) $t = 0.1$, $g^{(2)}$ (e) $t = 1$, $g^{(2)}$ (f) $t = 10$, $g^{(2)}$

Figure 3: Results from Example 5.1. Top figures: $\|e_1^N\|$. Bottom figures: $\frac{\|e_1^{N+1}\|}{\|e_1^N\|\epsilon^2}$.

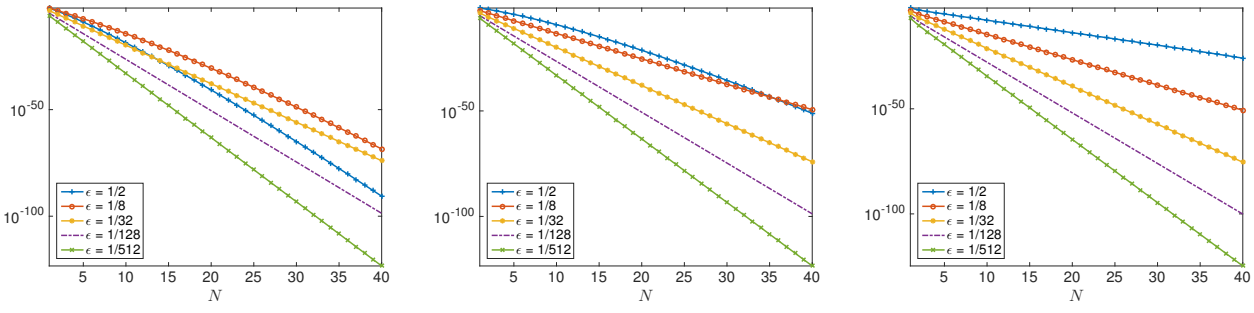


(a) $t = 0.1$, $g^{(2)}$ (b) $t = 1$, $g^{(2)}$ (c) $t = 10$, $g^{(2)}$

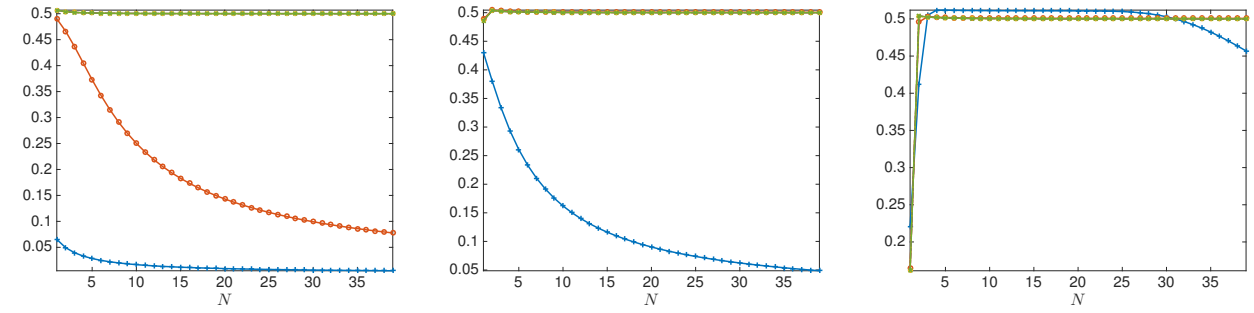


(d) $t = 0.1$, $g^{(2)}$ (e) $t = 1$, $g^{(2)}$ (f) $t = 10$, $g^{(2)}$

Figure 4: Results from Example 5.1. Top figures: $\|e_2^N\|$. Bottom figures: $\frac{\|e_2^{N+1}\|}{\|e_2^N\|\epsilon^2}$.

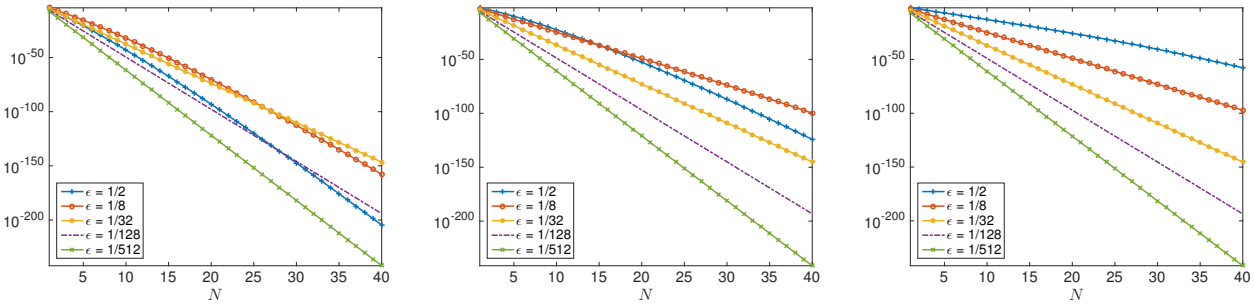


(a) $t = 0.1$, $g^{(3)}$ (b) $t = 1$, $g^{(3)}$ (c) $t = 10$, $g^{(3)}$

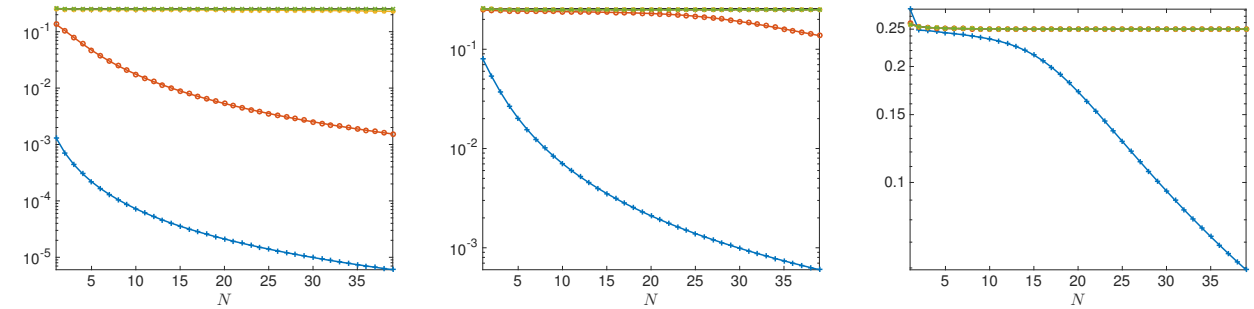


(d) $t = 0.1$, $g^{(3)}$ (e) $t = 1$, $g^{(3)}$ (f) $t = 10$, $g^{(3)}$

Figure 5: Results from Example 5.1. Top figures: $\|e^N\|$. Bottom figures: $\frac{\|e^{N+1}\|}{\|e^N\|\epsilon}$.

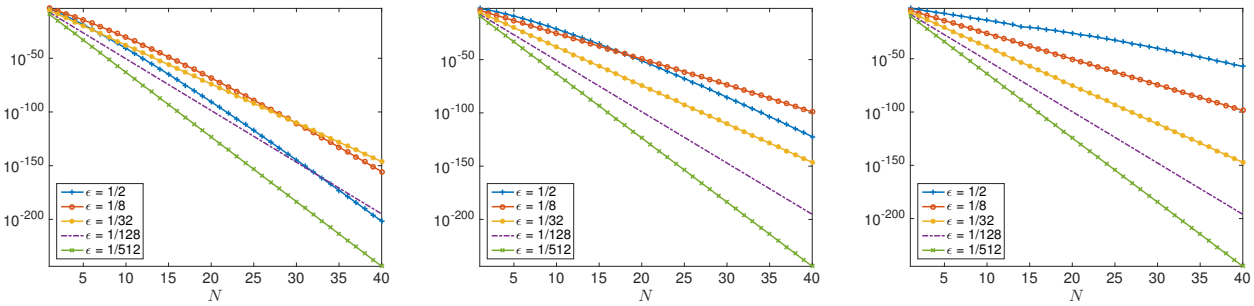


(a) $t = 0.1$, $g^{(3)}$ (b) $t = 1$, $g^{(3)}$ (c) $t = 10$, $g^{(3)}$

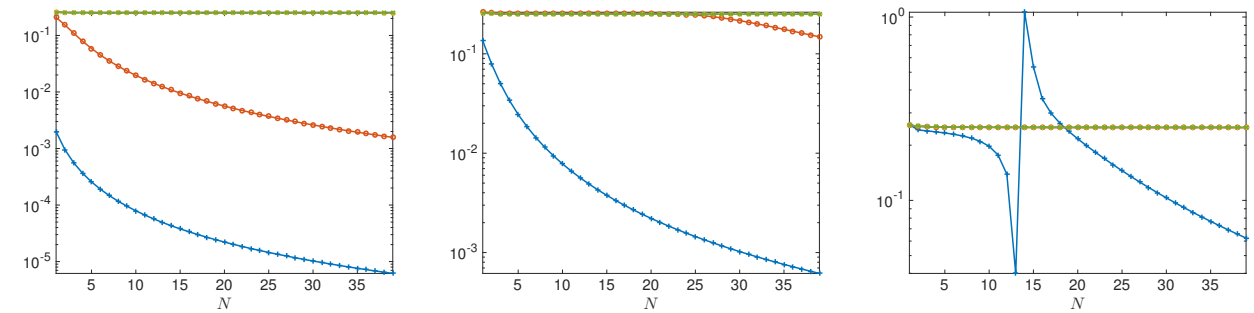


(d) $t = 0.1$, $g^{(3)}$ (e) $t = 1$, $g^{(3)}$ (f) $t = 10$, $g^{(3)}$

Figure 6: Results from Example 5.1. Top figures: $\|e_0^N\|$. Bottom figures: $\frac{\|e_0^{N+1}\|}{\|e_0^N\|\epsilon^2}$.

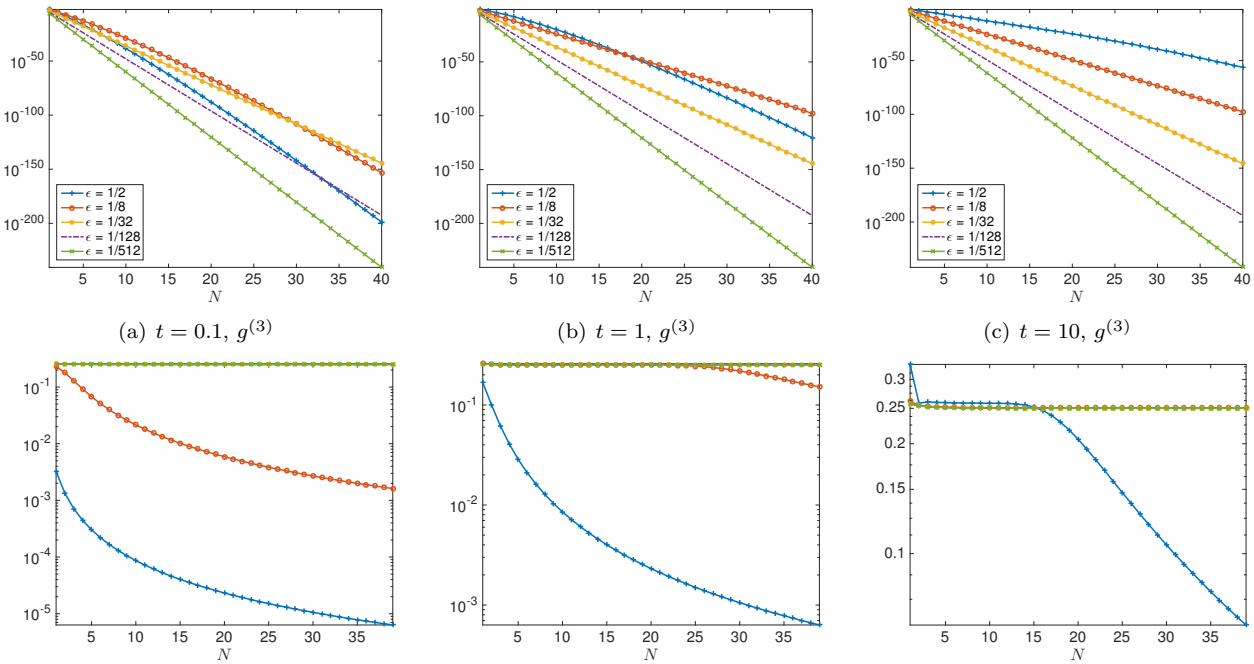


(a) $t = 0.1$, $g^{(3)}$ (b) $t = 1$, $g^{(3)}$ (c) $t = 10$, $g^{(3)}$



(d) $t = 0.1$, $g^{(3)}$ (e) $t = 1$, $g^{(3)}$ (f) $t = 10$, $g^{(3)}$

Figure 7: Results from Example 5.1. Top figures: $\|e_1^N\|$. Bottom figures: $\frac{\|e_1^{N+1}\|}{\|e_1^N\|\epsilon^2}$.



(a) $t = 0.1$, $g^{(3)}$ (b) $t = 1$, $g^{(3)}$ (c) $t = 10$, $g^{(3)}$

(d) $t = 0.1$, $g^{(3)}$ (e) $t = 1$, $g^{(3)}$ (f) $t = 10$, $g^{(3)}$

Figure 8: Results from Example 5.1. Top figures: $\|e_2^N\|$. Bottom figures: $\frac{\|e_2^{N+1}\|}{\|e_2^N\|\epsilon^2}$.

5.2 Quantifying coefficients in the error estimates

The manner in which the estimates in (1.9) and (1.10) depend on N and ℓ can ultimately be traced back to the coefficient $F(g,n,t)$, defined in (3.18). Indeed the results in (3.17), (3.28), (3.39), (3.73) and (3.74) all depend on $F(g,n,t)$ for some value of n: in (3.17), $n = \ell$; in (3.28), $n = N+1$; in (3.39), $n = N+2$; in (3.73), $n = 3N$, and in (3.74), $n = 3N + 4 - 2\ell$. The dependence of $F(g,n,t)$ on n arises via the term $a_n(2\lambda_2 t)$, where (recall that) $\lambda_2 = 4/45$ and

$$a_n(s) := \sum_{k>0} \left\{ (Ak)^{2n} e^{-k^2 s} \right\}. \tag{5.6}$$

For example, according to (1.9) and (3.79), after an initial layer,

$$\|e^N\|(t) \le \tilde{c}(t) \sqrt{a_{N+2}(2\lambda_2 t)}\, \epsilon^{N+1}, \quad \text{where} \quad \tilde{c}(t) = 2\left(\sqrt{2} + \frac{\sqrt{t}}{A}\right)\left(24 \max_{k>0} \mathcal{H}_k^0(g)\right)^{1/2}. \tag{5.7}$$

Similarly, it follows from (1.10), (3.75), and (3.18) that after an initial layer,

$$\|e_\ell^N\|(t) \le \tilde{d}\left(\frac{N - n_\ell + 2}{eA^2\lambda_2}\right)^{\frac{N-n_\ell+2}{2}} \sqrt{a_{3N+4-2n_\ell}(\lambda_2 t)}\, \epsilon^{2N+2-n_\ell}, \tag{5.8}$$

where

$$n_\ell = \begin{cases} 2, & \ell = 0 \\ \ell, & 1 \le \ell \le N, \end{cases} \qquad \text{and} \qquad \tilde{d} = 8\left(6\, e^{\lambda_2} \max_{k>0} \mathcal{H}_k^0(g)\right)^{1/2}. \tag{5.9}$$

By interpreting the right-hand side of (5.6) as a Riemann sum, we bound a_n as follows:

$$\begin{aligned} a_n(s) &\le A^{2n}\left(e^{-s} + \int_1^\infty (x+1)^{2n} e^{-sx^2} dx\right) \le A^{2n}\left(e^{-s} + \int_1^\infty (2x)^{2n} e^{-sx^2} dx\right) \\ &\le (2A)^{2n}\left(e^{-s} + \int_1^\infty x^{2n+1} e^{-sx^2} dx\right) = (2A)^{2n}\left(e^{-s} + \frac{1}{2} e^{-s} b_n(s)\right), \end{aligned} \tag{5.10}$$

where

$$b_n(s) = \sum_{k=0}^{n} \left(\frac{1}{s}\right)^{k+1} \frac{n!}{(n-k)!}, \qquad n \geq 0. \tag{5.11}$$

Setting (5.10) into (5.7) gives

$$\|e^N\|(t) \leq \tilde{c}(t)(2A)^{N+2} e^{-\lambda_2 t} \left(\frac{1}{2} b_{N+2}(2\lambda_2 t) + 1\right)^{1/2} \epsilon^{N+1} =: E^N(t), \tag{5.12}$$

and setting (5.10) into (5.8) gives

$$\|e_\ell^N\|(t) \leq \tilde{d} \left(\frac{N-n_\ell+2}{eA^2\lambda_2}\right)^{\frac{N-n_\ell+2}{2}} (2A)^{3N+4-2n_\ell} e^{-\lambda_2 t/2} \left(\frac{1}{2} b_{3N+4-2n_\ell}(\lambda_2 t) + 1\right)^{1/2} \epsilon^{2N+2-n_\ell} =: E_\ell^N(t). \tag{5.13}$$

Thus, the error-bound ratios

$$\frac{E^{N+1}(t)}{E^N(t)} = 2A \left(\frac{b_{N+3}(2\lambda_2 t)+2}{b_{N+2}(2\lambda_2 t)+2}\right)^{1/2} \epsilon \tag{5.14}$$

and

$$\begin{aligned} \frac{E_\ell^{N+1}(t)}{E_\ell^N(t)} &= \left(\frac{N-n_\ell+3}{eA^2\lambda_2^2}\right)^{1/2} \left(1 + \frac{1}{N-n_\ell+2}\right)^{\frac{N-n_\ell+2}{2}} (2A)^3 \left(\frac{b_{3N+7-2n_\ell}(\lambda_2 t)+2}{b_{3N+4-2n_\ell}(\lambda_2 t)+2}\right)^{1/2} \epsilon^2 \\ &\leq 8A^2 \left(\frac{N-n_\ell+3}{\lambda_2^2}\right)^{1/2} \left(\frac{b_{3N+7-2n_\ell}(\lambda_2 t)+2}{b_{3N+4-2n_\ell}(\lambda_2 t)+2}\right)^{1/2} \epsilon^2 \end{aligned} \tag{5.15}$$

can be used to quantity how much the estimates of $\|e^N\|$ and $\|e_\ell^N\|$ improve as N increases.

It is easy to verify that b_n satisfies the following recurrence formula:

$$b_0(s) = \frac{1}{s} \qquad \text{and} \qquad b_{n+1}(s) = \frac{n+1}{s} b_n(s) + \frac{1}{s}, \quad \text{for } n \geq 0. \tag{5.16}$$

Hence, for any $n \geq 1$,

$$\left(\frac{b_{n+1}(s)+2}{b_n(s)+2}\right) = \left(\frac{\frac{n+1}{s} b_n(s) + \frac{1}{s} + 2}{b_n(s)+2}\right) = \frac{n+1}{s} \left(\frac{b_n(s) + \frac{1}{n+1}}{b_n(s)+2}\right) + \left(\frac{2}{b_n(s)+2}\right) \leq \frac{n+1}{s} + 1. \tag{5.17}$$

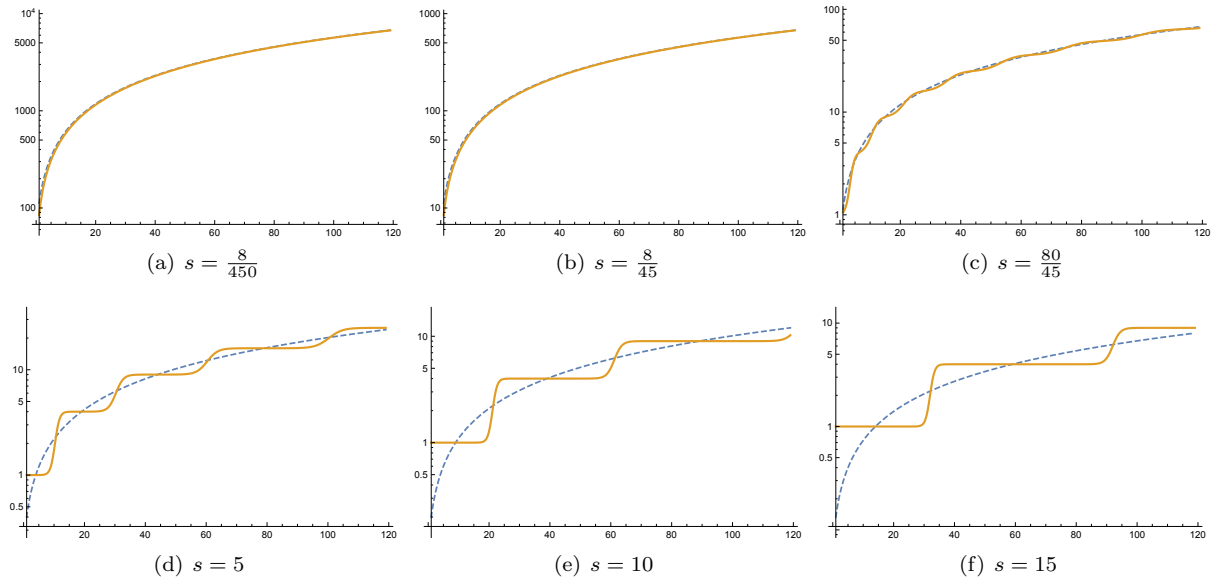
When applied to (5.14), (5.17) with $n = N+2$ implies that

$$\frac{E^{N+1}(t)}{E^N(t)} \leq 2A \left(\frac{N+3}{2\lambda_2 t} + 1\right)^{1/2} \epsilon. \tag{5.18}$$

This dependence on t suggests that the normalized true-error ratio $\|e^{N+1}\|/(\epsilon\|e^N\|)$ decreases as t increases, as observed in Example 5.1. Similarly, using (5.17) in (5.15) with $n = 3N+4-2n_\ell, \ldots, 3N+6-2n_\ell$ gives

$$\begin{aligned} \frac{E_\ell^{N+1}(t)}{E_\ell^N(t)} &\leq 8A^2 \left(\frac{N-n_\ell+3}{\lambda_2^2}\right)^{1/2} \left(\frac{3N+5-2n_\ell}{\lambda_2 t} + 1\right)^{1/2} \left(\frac{3N+6-2n_\ell}{\lambda_2 t} + 1\right)^{1/2} \left(\frac{3N+7-2n_\ell}{\lambda_2 t} + 1\right)^{1/2} \epsilon^2 \\ &\leq 8A^2 \left(\frac{N-n_\ell+3}{\lambda_2^2}\right)^{1/2} \left(\frac{3N+7-2n_\ell}{\lambda_2 t} + 1\right)^{3/2} \epsilon^2, \end{aligned} \tag{5.19}$$

This also suggests that the normalized true-error ratio $\|e_\ell^{N+1}\|/(\epsilon^2\|e_\ell^N\|)$ decreases as t increases, as observed for the first three moments in Example 5.1. However, in both cases, t needs to be sufficiently large in order for these ratios to be small. In particular, any increase in the coefficient $\lambda_2 = 4/45$ will yield better bounds for E^{N+1}/E^N and E_ℓ^{N+1}/E_ℓ^N. The numerical results in the following and final example suggest that this value of λ_2, which is established in Lemma 3.1, is probably not optimal.



(a) $s = \frac{8}{450}$ (b) $s = \frac{8}{45}$ (c) $s = \frac{80}{45}$

(d) $s = 5$ (e) $s = 10$ (f) $s = 15$

Figure 9: Results from Example 5.3 for different values of s. The orange curves are a_{n+1}/a_n vs. n. The blue dashed curves are of $(n+1)/s$.

Example 5.3. We investigate the ratio a_{n+1}/a_n numerically using the finite sum

$$a_n^K(s) = \sum_{0<k\leq K} \left\{ (Ak)^{2n} e^{-k^2 s} \right\}. \tag{5.20}$$

Numerical test suggest that a_n^K converges as $K \to \infty$ and that $K = 1000$ is sufficient to capture the behavior of the infinite sum in (5.6), and therefore use $a_n \approx a_n^{1000}$ in the remainder of the computation. We compute $a_n(s)$ for $n = 1, \ldots, 120$ and different values s. We then plot the ratios a_{n+1}/a_n in Figure 9 and make the following observations:

1) It appears from the plots in Figure 9 that

$$\frac{a_{n+1}(s)}{a_n(s)} \sim \frac{n+1}{s}. \tag{5.21}$$

This approximation is consistent with the theoretical bound in (5.17) for large n, and the profiles of the two ratios match quite well.

2) Recall again from Lemma 3.1 that $\lambda_2 = 4/45$. Thus if we set $s = s(t) = 2\lambda_2 t$, the first values of $s = 8/450, 8/45, 80/45$ in Figure 9(a)-(c) correspond to the values $t = 0.1, 1, 10$ that are used in Examples 5.1 and 5.2. As s increases (9(d)-(f)), we begin to see plateaus connected by sharp transitions. This behavior is most notable in Figures 9(d)–9(f), and it is reminiscent of the profiles of the normalized error ratios from Example 5.1 (cf. plots (e) and (f) of Figures 1–4), albeit at smaller values of t. Currently, we do not have any explanation for these jumps or their locations. However, the fact that this behavior emerges for larger values of s suggests that it may be possible to prove Lemma 3.1 with a larger value of λ_2.

6 Conclusion

In this paper, we give error estimates, in terms of a multiscale parameter ϵ, for the spectral approximation in the velocity variable of an idealized kinetic model. This approximation yields a linear, symmetric hyperbolic system of partial differential equations for the expansion coefficients, which are functions of x and t. Under the

assumption that the initial data g is isotropic, with $g \in L^2(d\mu dx)$ and $\partial_x g \in L^2(d\mu dx)$, we prove that the error in the spectral approximation with N modes is $\mathcal{O}(\epsilon^{N+1})$. In additional, we prove super-convergent results for the expansion coefficients. We also provide numerical results that support the theoretical estimates. These results exhibit the predicted order of convergence even when $\partial_x g \notin L^2(d\mu dx)$. Thus it remains open whether this condition is necessary for our result.

The coefficients of the error estimates are independent of ϵ but not N. Thus, in an effort to demonstrate the practical benefit when increasing N, we investigate these coefficients both theoretically and numerically. In particular, we find that the ratio of successive error bounds in N is itself bounded above by the product $\epsilon\alpha_N(t)$, where

$$\alpha_N(t) \leq 2A \left(\frac{N+3}{2\lambda_2 t} + 1 \right)^{1/2}, \tag{6.1}$$

with $\lambda_2 = 4/45$ and $A = A(\lambda_2) \simeq 1.2$. Meanwhile, the ratio in the error estimate for the moments is bounded above by the product $\epsilon^2 \beta_{N,\ell}(t)$, where

$$\beta_{N,\ell}(t) = 8A^2 \left(\frac{N - n_\ell + 3}{\lambda_2^2} \right)^{1/2} \left(\frac{3N + 7 - 2n_\ell}{\lambda_2 t} + 1 \right)^{3/2}. \tag{6.2}$$

Thus for reasonable (but not too large) values of N and t sufficiently large, our estimate of the spectral error improves significantly as N is increased. In our analysis, we are able to prove our theoretical results with $\lambda_2 = 4/45$. However, numerical results suggest that a larger value of λ_2 is possible and demonstrate that the theoretical benefit of having a larger value is significant.

In the future we intend to establish the theoretical results of this paper with a larger value of λ_2. In addition, we will explore the ϵ-dependent behavior of the error under more general initial conditions such as anisotropic initial conditions, real boundary conditions, non-zero absorption and sources, spatially dependent scattering, and higher-dimensional problems. We also hope to investigate alternative angular discretizations and nonlinear systems.

A Spectral Error Estimate

The purpose of the section is to show that, with sufficient regularity on the initital condition g, the standard estimate (1.7) holds with a constant C that is independent of $\epsilon \in [0, 1]$.

Definition A.1. *Let r, q, s, and S be non-negative integers. For any $u \in L^2(d\mu dx)$, define the shorthand $u^{(r,q)} = \partial_x^r \partial_\mu^q u$ and the semi-norm $|u|_{r,q} = \|u^{(r,q)}\|_{L^2(d\mu dx)}$. Then define the space*

$$V^s(d\mu dx) = \left\{ u \in L^2(d\mu dx) \colon \sum_{q=0}^{s} |u|_{s-q,q} < \infty \right\} \tag{A.1}$$

with the associated semi-norm $|\cdot|_{V^s(d\mu dx)} = \sum_{q=0}^{s} |\cdot|_{s-q,q}$. Finally, let

$$H^S(d\mu dx) = \left\{ u \in L^2(d\mu dx) \colon \sum_{s=0}^{S} |\cdot|_{V^s(d\mu dx)} < \infty \right\} \tag{A.2}$$

be the usual Sobolev space with norm $\|\cdot\|_{H^S} = \sum_{s=0}^{S} |\cdot|_{V^s(d\mu dx)}$.

Lemma A.2. *Let f solve (1.1) with initial condition $g \in V^s(d\mu dx)$ for some positive integer s. Then $f \in C([0,\infty); V^s(d\mu dx))$ with*

$$|f|_{s-q,q}(t) \leq (q+1)!\, |g|_{V^s(d\mu dx)} \tag{A.3}$$

for all integers $q \in [0, s]$ and $t \geq 0$.

Proof. Given $h \in C([0,\infty); L^2(d\mu dx))$ and $v \in L^2(d\mu dx)$, the equation

$$\begin{cases} \partial_t u(x,\mu,t) + \dfrac{1}{\epsilon}\mu\partial_x u(x,\mu,t) + \dfrac{1}{\epsilon^2}u(x,\mu,t) = h(x,\mu,t), & (x,\mu,t) \in [-\pi,\pi)\times[-1,1]\times(0,\infty), \quad \text{(A.4a)} \\ u(x,\mu,0) = v(x,\mu), & (x,\mu) \in [-\pi,\pi)\times[-1,1], \quad \text{(A.4b)} \end{cases}$$

has a mild solution (see, for example, [29, p.402]) $u \in C([0,\infty); L^2(d\mu dx))$, given by

$$u(x,\mu,t) = e^{-\frac{t}{\epsilon^2}} v(x - \frac{1}{\epsilon}\mu t, \mu) + \int_0^t e^{-\frac{t-\tau}{\epsilon^2}} h(x - \frac{1}{\epsilon}\mu(t-\tau), \mu, \tau)\, d\tau. \tag{A.5}$$

where the argument $x - \epsilon^{-1}\mu t$ is understood with respect to the periodicity of the spatial domain. Applying the triangle equality to (A.5) gives, for each $t \geq 0$,

$$\begin{aligned} ||u||_{L^2(d\mu dx)}(t) &\leq e^{-\frac{t}{\epsilon^2}} ||v||_{L^2(d\mu dx)} + \int_0^t e^{-\frac{t-\tau}{\epsilon^2}} ||h||_{L^2(d\mu dx)}(\tau)\, d\tau \\ &\leq e^{-\frac{t}{\epsilon^2}} ||v||_{L^2(d\mu dx)} + \epsilon^2(1 - e^{-\frac{t}{\epsilon^2}}) \max_{\tau\in[0,t]} ||h||_{L^2(d\mu dx)}(\tau). \end{aligned} \tag{A.6}$$

We now proceed by induction on q. If $q = 0$, then differentiation of (1.1a) in x gives

$$\epsilon\partial_t f^{(r,0)} + \mu\partial_x f^{(r,0)} + \frac{1}{\epsilon} f^{(r,0)} = \frac{1}{\epsilon}\overline{f^{(r,0)}} \tag{A.7}$$

for any integer $r \geq 0$. Hence $u = f^{(s,0)}$ satisfies (A.4) with source $h = \frac{1}{\epsilon^2}\overline{f^{(s,0)}} \in C([0,\infty); L^2(d\mu dx))$ and initial condition $v = g^{(s,0)} \in L^2(d\mu dx)$. Thus (A.6) gives

$$\begin{aligned} |f|_{s,0}(t) &\leq e^{-\frac{t}{\epsilon^2}} |g|_{s,0} + (1 - e^{-\frac{t}{\epsilon^2}}) \max_{\tau\in[0,t]} ||\overline{f^{(s,0)}}||_{L^2(d\mu dx)}(\tau) \\ &\leq e^{-\frac{t}{\epsilon^2}} |g|_{V^s(d\mu dx)} + (1 - e^{-\frac{t}{\epsilon^2}}) \max_{\tau\in[0,t]} |f|_{s,0}(\tau), \end{aligned} \tag{A.8}$$

Let $t_* \in [0,t]$ be such that $|f|_{s,0}(t_*) = \max_{\tau\in[0,t]} |f|_{s,0}(\tau)$. Then $|f|_{s,0}(t_*) = \max_{\tau\in[0,t_*]} |f|_{s,0}(\tau)$ so that, according to (A.8),

$$|f|_{s,0}(t_*) \leq e^{-\frac{t_*}{\epsilon^2}} |g|_{V^s(d\mu dx)} + (1 - e^{-\frac{t_*}{\epsilon^2}}) |f|_{s,0}(t_*). \tag{A.9}$$

Therefore

$$|f|_{s,0}(t) \leq |f|_{s,0}(t_*) \leq |g|_{V^s(d\mu dx)}, \tag{A.10}$$

which verifies (A.3).

Next assume that (A.3) holds for $q = q_0$, with $0 \leq q_0 < s$. Differentiation of (1.1a) in x and μ gives

$$\epsilon\partial_t f^{(r,q_0+1)} + \mu\partial_x f^{(r,q_0+1)} + \frac{1}{\epsilon} f^{(r,q_0+1)} = -(q_0+1) f^{(r+1,q_0)} \tag{A.11}$$

for any $r \geq 0$. Therefore $u = f^{(s-(q_0+1),q_0+1)}$ satisfies (A.4) with the source $h = -\frac{q_0+1}{\epsilon} f^{(s-q_0,q_0)} \in C([0,\infty); L^2(d\mu dx))$ and initial condition $v = g^{(s-(q_0+1),q_0+1)} \in L^2(d\mu dx)$. Thus (A.6) gives

$$\begin{aligned} |f|_{s-(q_0+1),q_0+1}(t) &\leq e^{-\frac{t}{\epsilon^2}} |g|_{s-(q_0+1),q_0+1} + \epsilon(1 - e^{-\frac{t}{\epsilon^2}})(q_0+1) \max_{\tau\geq 0} |f|_{s-q_0,q_0}(\tau) \\ &\leq |g|_{V^s(d\mu dx)} + \epsilon(q_0+1)(q_0+1)!\, |g|_{V^s(d\mu dx)} \\ &\leq (q_0+2)!\, |g|_{V^s(d\mu dx)}. \end{aligned} \tag{A.12}$$

□

Remark A.3. *For sufficiently small ϵ, the bound*

$$|f|_{s-q,q}(t) \leq \left(\prod_{i=0}^{q}(1 + i\epsilon)\right) |g|_{V^s(d\mu dx)} \tag{A.13}$$

provides a sharper estimate than (A.3)*. The proof of this alternative bound uses the same arguments.*

Theorem A.4. *Suppose that $g \in H^{1+q}(d\mu dx)$ for some integer $q > 0$. Then there exists a constant $C = C(g, q)$, such that*

$$\|f - f^N\|_{L^2(d\mu dx)}(t) \leq C(1 + t^{1/2})N^{-q}, \quad \forall t \geq 0. \tag{A.14}$$

Proof. We begin by estimating $\xi = Pf - f^N$ in terms of η. A direct calculation using (1.1) and (1.5) shows that

$$\partial_t \xi + \frac{1}{\epsilon}\mathcal{P}\left(\mu \partial_x \xi\right) + \frac{1}{\epsilon^2}(\xi - \bar{\xi}) = -\frac{1}{\epsilon}\mathcal{P}\left(\mu \partial_x \eta\right), \tag{A.15}$$

which is equivalent to the system (2.9). Integrating (A.15) against ξ on the left gives

$$\begin{aligned}
\frac{1}{2}\partial_t \|\xi\|^2_{L^2(d\mu dx)} + \frac{1}{\epsilon^2}\|\tilde{\xi}\|^2_{L^2(d\mu dx)} &= -\frac{1}{\epsilon}\iint \xi\, \mathcal{P}\left(\mu \partial_x \eta\right)\, d\mu dx \\
&= -\frac{1}{\epsilon}\iint \xi\, \mu \partial_x \eta \, d\mu dx \\
&= -\frac{1}{\epsilon}\iint \bar{\xi}\, \mu \partial_x \eta \, d\mu dx - \frac{1}{\epsilon}\iint \tilde{\xi}\, \mu \partial_x \eta \, d\mu dx,
\end{aligned} \tag{A.16}$$

where $\tilde{\xi} = \xi - \bar{\xi}$. For $N \geq 1$, μ and η are orthogonal; hence the first term in the last line of (A.16) is zero. Meanwhile Young's inequality yields a bound on the second term:

$$-\frac{1}{\epsilon}\iint \tilde{\xi}\, \mu \partial_x \eta \, d\mu dx \leq \frac{1}{2\epsilon^2}\|\tilde{\xi}\|^2_{L^2(d\mu dx)} + \frac{1}{2}\|\mu \partial_x \eta\|^2_{L^2(d\mu dx)}. \tag{A.17}$$

Hence, (A.16) reduces to

$$\partial_t \|\xi\|^2_{L^2(d\mu dx)} + \frac{1}{\epsilon^2}\|\tilde{\xi}\|^2_{L^2(d\mu dx)} \leq \|\mu \partial_x \eta\|^2_{L^2(d\mu dx)}, \tag{A.18}$$

and therefore,

$$\partial_t \|\xi\|^2_{L^2(d\mu dx)} \leq \|\mu \partial_x \eta\|^2_{L^2(d\mu dx)}. \tag{A.19}$$

Since $\xi|_{t=0} = 0$, integrating (A.19) in time gives

$$\|\xi\|^2_{L^2(d\mu dx)}(t) \leq t \sup_{\tau \geq 0}\left\{\|\mu \partial_x \eta\|^2_{L^2(d\mu dx)}(\tau)\right\} \leq t \sup_{\tau \geq 0}\left\{\|\partial_x \eta\|^2_{L^2(d\mu dx)}(\tau)\right\} \tag{A.20}$$

Thus it remains only to bound $\|\partial_x \eta\|_{L^2(d\mu dx)}$.

We now turn to polynomial approximation theory: given a function $\psi \in H^q(d\mu)$, where $H^q(d\mu)$ is the Sobolev space of functions with q weak derivatives in $L^2(d\mu)$, there exists a constant $K_1 > 0$ such that [3, Lemma 2.2]

$$\|\psi - \mathcal{P}\psi\|_{L^2(d\mu)} \leq K_1 \|\psi\|_{H^q(d\mu)} N^{-q}. \tag{A.21}$$

We apply this result to $\partial_x \eta = \partial_x f - \mathcal{P}\partial_x f$, using also Lemma A.2, to find that

$$\begin{aligned}
\|\partial_x \eta\|_{L^2(d\mu dx)}(\tau) &\leq K_1 \|\partial_x f\|_{L^2(dx; H^q(d\mu))}(\tau) N^{-q} \\
&= K_1 \left(\sum_{0 \leq r \leq q} |f|_{1,r}(\tau)\right) N^{-q} \\
&\leq K_1 \sum_{0 \leq r \leq q} (r+1)!\, |g|_{V^{1+r}(d\mu dx)}\, N^{-q} \\
&\leq K_1 (q+1)!\, \|g\|_{H^{1+q}(d\mu dx)}\, N^{-q}.
\end{aligned} \tag{A.22}$$

This bound is independent of t. Thus combining (A.20) and (A.22) gives

$$\|\xi\|_{L^2(d\mu dx)}(t) \leq K_1 (q+1)!\; t^{1/2} \|g\|_{H^{1+q}(d\mu dx)} N^{-q}. \tag{A.23}$$

To complete the proof, we estimate $\eta = f - \mathcal{P}f$ using (A.21) and Lemma A.2,

$$\|\eta\|_{L^2(d\mu dx)}(t) \leq K_1 \|f\|_{L^2(dx; H^q(d\mu))}(t) N^{-q} \leq K_1 (q+1)!\, \|g\|_{H^q(d\mu dx)} N^{-q}, \quad \forall t \geq 0. \tag{A.24}$$

Combining (A.23) and (A.24) recovers (A.14) with $C = K_1(q+1)!\, \|g\|_{H^{1+q}(d\mu dx)}$.

□

References

[1] Multiprecision Computing Toolbox for MATLAB 4.3.3.12177, Advanpix LLC., Yokohama, Japan.

[2] C. Bardos, R. Santos, and R. Sentis. Diffusion approximation and computation of the critical size. *Transactions of the american mathematical society*, 284(2):617–649, 1984.

[3] G. Ben-Yu. *Spectral methods and their applications*. World Scientific, 1998.

[4] A. Bensoussan, J. L. Lions, and G. C. Papanicolaou. Boundary layers and homogenization of transport processes. *Publications of the Research Institute for Mathematical Sciences*, 15(1):53–157, 1979.

[5] T. J. M. Boyd and J. J. Sanderson. *The physics of plasmas*. Cambridge University Press, 2003.

[6] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[7] C. G. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods: Fundamentals in single domains*. Springer, 2010.

[8] K. M. Case and P. F. Zweifel. *Linear transport theory*. Addison-Wesley, 1967.

[9] C. Cercignani. *The Boltzmann Equation and its Applications*, volume 67 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1988.

[10] C. Cercignani, R. Illner, and M. Pulvirenti. *The Mathematical Theory of Dilute Gases*, volume 106 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1994.

[11] S. Chapman and T. G. Cowling. *The mathematical theory of non-uniform gases: an account of the kinetic theory of viscosity, thermal conduction and diffusion in gases*. Cambridge university press, 1970.

[12] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 1 Physical Origins and Classical Methods*. Springer Science & Business Media, 2012.

[13] B. Davison and J. B. Sykes. Neutron transport theory. 1957.

[14] J. Dolbeault, C. Mouhot, and C. Schmeiser. Hypocoercivity for linear kinetic equations conserving mass. *Transactions of the American Mathematical Society*, 367(6):3807–3828, 2015.

[15] L. Evans. *Partial differential equations*. American Mathematical Society, 1998.

[16] M. Frank, C. Hauck, and K. Kuepper. Convergence of filtered spherical harmonic equations for radiation transport. *Commun. Math. Sci*, 14(5):1443–1465, 2016.

[17] G. J. Habetler and B. J. Matkowsky. Uniform asymptotic expansions in transport theory with small mean free paths, and the diffusion approximation. *Journal of Mathematical Physics*, 16:846–854, Apr. 1975.

[18] C. D. Hauck and R. B. Lowrie. Temporal regularization of the p_n equations. *Multiscale Modeling & Simulation*, 7(4):1497–1524, 2009.

[19] R. D. Hazeltine and F. L. Waelbroeck. *The framework of plasma physics*. Westview, 2004.

[20] J. S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral methods for time-dependent problems*, volume 21. Cambridge University Press, 2007.

[21] E. W. Larsen and J. B. Keller. Asymptotic solution of neutron transport problems for small mean free paths. *Journal of Mathematical Physics*, 15:75–81, Jan. 1974.

[22] E. W. Larsen, J. E. Morel, and J. M. McGhee. Asymptotic derivation of the multigroup p 1 and simplified pn equations with anisotropic scattering. *Nuclear science and engineering*, 123(3):328–342, 1996.

[23] E. E. Lewis and W. F. Miller. *Computational methods of neutron transport*. John Wiley and Sons, Inc., New York, NY, 1984.

[24] E. E. Lewis and W. F. Miller. *Computational Methods of Neutron Transport*. John Wiley and Sons, 1984.

[25] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser. *Semiconductor Equations*. Springer-Verlag, New York, 1990.

[26] A. Mezzacappa and O. Messer. Neutrino transport in core collapse supernovae. *Journal of Computational and Applied Mathematics*, 109(1):281–319, 1999.

[27] D. Mihalas and B. Weibel-Mihalas. *Foundations of radiation hydrodynamics*. Courier Corporation, 1999.

[28] G. C. Pomraning. *Radiation Hydrodynamics*. Pergamon Press, New York, 1973.

[29] M. Renardy and R. C. Rogers. *An introduction to partial differential equations*, volume 13. Springer Science & Business Media, 2006.

[30] S. Selberherr. *Analysis and simulation of semiconductor devices*. Springer Science & Business Media, 2012.