申请上海交通大学博士学位论文

# 带多尺度和不确定性的输运与波动问题的计算方法


4em **论文作者** 马　征

4em **学号** 0120719009

4em **导师** 金石教授

4em **专业** 计算数学


4em **答辩日期** 2017 年 5 月 20 日

Submitted in total fulfillment of the requirements for the degree of Doctor in Computational Mathematics

# Numerical Methods for Transport Equations and Wave Propagations with Multiple Scales and Uncertainty

Zheng Ma

Advisor
Prof. Shi Jin

School of Mathematical Sciences
Shanghai Jiao Tong University
Shanghai, P.R.China

May 20, 2017

**附件四**

# 上海交通大学
# 学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：2017年5月17日

**附件五**

# 上海交通大学

# 学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

**保密**□，在___年解密后适用本授权书。

**不保密**☑。

（请在以上方框内打“√”）

学位论文作者签名：

日期：2017 年 5 月 17 日

指导教师签名：

日期：2017 年 5 月 17 日

# 带多尺度和不确定性的输运与波动问题的计算方法

## 摘 要

首先，我们提出了一种基于广义多项式混沌（gPC）的随机伽辽金方法（SG）用于计算具有随机和奇异系数的双曲方程。由于解的奇异性，标准 gPC-SG 方法收敛速度会很慢甚至不收敛。通过利用中心型有限差分或有限体积方法的离散解在空间和时间上较为光滑的特性，我们先离散原方程，然后再使用 gPC-SG 近似离散的系统。间断处的界面条件使用 [1-2] 中的方法处理，这样整个方法具有很快的收敛速度，对于固定的网格大小和时间步长其为谱收敛。我们使用带有不连续和随机系数的线性对流方程，以及带有不连续和随机的势能 Liouville 方程作为例子来说明我们的想法，提出了一阶和二阶的格式，并用数值例子验证了我们的想法。

其次，我们研究随机输入和扩散尺度下的线性传输方程的随机伽辽金方法。我们首先建立解在随机空间一致的（关于克努恩数）正则性结果，然后证明随机伽辽金方法的一致谱收敛（以及推出的随机渐近保持特性）。提出了对于该问题采用基于 micro-macro 分解的全离散格式，并证明其具有一致的稳定性。大量的数值实验用以证明该方法的稳定性和渐近性质。

第三，我们提出了一种用于具有向量势的半经典薛定谔方程的新的时间分裂傅立叶谱方法。与 [3] 的结果相比，我们的方法通过应用非均匀快速傅里叶变换（NUFFT）算法，使得对流步中傅立叶谱插值的使用变得可行。该算法在保持傅里叶方法的高空间精度的同时，效率上从 $O(N^2)$（直接计算）提高到 $O(N \log N)$，其中 $N$ 是网格点的总数。动能步骤和势能步骤通过具有伪谱近似的解析解来解决，因此，我们在整个方法中获得了空间的谱精度。我们证明该方法是无条件稳定的，并且我们对波函数和物理观察值进行了改进的误差估计，这与 [4] 中不带有向量势能的结果一致，并且优于 [3] 。我们进行了大量的一维和二维数值研究来验证所提出的方法的性质，并展示了 3D 问题的仿真，以展示其未来实际应用的潜力。

最后，我们研究带 Caputo 导数（分数阶）的标量守恒定律的数值近似，该方程的特点是引入了记忆效应。我们构造了这种方程的一阶和二阶显式迎风格式，这些格式被证明为有条件的 $\ell^1$ 递减以及 TVD 的。然而，Caputo 导数存在使得我们得到修正的 CFL 稳定性条件，即 $(\Delta t)^\alpha = O(\Delta x)$，其中 $\alpha \in (0, 1]$ 是分数。当 $\alpha$ 很小时，这样强的约束使得数值实现非常不切实际，然后我们提出了隐式迎风格式来克服这个问题，这被证明是无条件的 $\ell^1$ 递减和 TVD 的。我们进行了了各种数值实验来验证方法的性质，并提供了更多的数值证据来解释守恒律中的记忆效应。

**关键词**：双曲方程　随机系数　势垒　随机伽辽金方法　多项式混沌　线性输运方程　随机输入　扩散极限　不确定量化　渐近保持格式　半经典薛定谔方程　向量势　半拉格朗日时间分裂方法　非均匀快速傅立叶变换

# Numerical Methods for Transport Equations and Wave Propagations with Multiple Scales and Uncertainty

## ABSTRACT

First, we develop a generalized polynomial chaos (gPC) based stochastic Galerkin (SG) for hyperbolic equations with random and singular coefficients. Due to the singular nature of the solution, the standard gPC-SG methods may suffer from a poor or even non convergence. Taking advantage of the fact that the discrete solution, by the central type finite difference or finite volume approximations in space and time for example, is smoother, we first discretize the equation by a smooth finite difference or finite volume scheme, and then use the gPC-SG approximation to the discrete system. The jump condition at the interface is treated using the immersed upwind methods introduced in [1-2]. This yields a method that converges with the spectral accuracy for finite mesh size and time step. We use a linear hyperbolic equation with discontinuous and random coefficient, and the Liouville equation with discontinuous and random potential, to illustrate our idea, with both one and second order spatial discretizations. Spectral convergence is established for the first equation, and numerical examples for both equations show the desired accuracy of the method.

Secondly, we study the stochastic Galerkin approximation for the linear transport equation with random inputs and diffusive scaling. We first establish uniform (in the Knudsen number) stability results in the random space for the transport equation with uncertain scattering coefficients, and then prove the uniform spectral convergence (and consequently the sharp stochastic Asymptotic-Preserving property) of the stochastic Galerkin method. A micro-macro decomposition based fully discrete scheme is adopted for the problem and proved to have a uniform stability. Numerical experiments are conducted to demonstrate the stability and asymptotic properties of the method.

Thirdly, we propose a new time splitting Fourier spectral method for the semi-classical Schrödinger equation with vector potentials. Compared with the results in [3], our method achieves spectral accuracy in space by interpolating the Fourier series via the NonUniform Fast Fourier Transform (NUFFT) algorithm in the convection step. The NUFFT algorithm helps maintain high spatial accuracy of Fourier method, and at the same time improve the efficiency from $O(N^2)$ (of direct computation) to $O(N \log N)$ operations, where $N$ is the total number of grid points. The kinetic step and potential step are solved by analytical solution with pseudo-spectral approximation, and, therefore, we obtain spectral accuracy in space for the whole method. We prove that the method is unconditionally stable, and we show improved error estimates for both the wave function and physical observables, which agree with the results in [4] for vanishing potential cases and are superior to those in [3]. Extensive one and two dimensional numerical studies are presented to verify

the properties of the proposed method, and simulations of 3D problems are demonstrated to show its potential for future practical applications.

Finally, we investigate numerical approximations of the scalar conservation law with the Caputo derivative, which introduces the memory effect. We construct the first order and the second order explicit upwind schemes for such equations, which are shown to be conditionally $\ell^1$ contracting and TVD. However, the Caputo derivative leads to the modified CFL-type stability condition, $(\Delta t)^\alpha = O(\Delta x)$, where $\alpha \in (0, 1]$ is the fractional exponent in the derivative. When $\alpha$ is small, such strong constraint makes the numerical implementation extremely impractical. We have then proposed the implicit upwind scheme to overcome this issue, which is proved to be unconditionally $\ell^1$ contracting and TVD. Various numerical tests are presented to validate the properties of the methods and provide more numerical evidence in interpreting the memory effect in conservation laws.

**KEY WORDS:** hyperbolic equation, random coefficient, potential barrier, stochastic Galerkin method, polynomial chaos, linear transport equation, random inputs, diffusion limit, uncertainty quantification, asymptotic-preserving scheme, semi-classical Schrödinger equation, vector potential, semi-Lagrangian time splitting method, nonuniform FFT

# 目 录

# 插图索引

# 表格索引

# 算法索引

# 第一章　带多尺度和不确定性的输运与波动问题的计算方法：简介

本文中，我们主要关注于一大类常见的波动方程和 Kinetic 方程，包括对流方程和刘维尔方程、输运方程、薛定谔方程及分数阶的守恒律方程。我们主要的目的是相关的计算方法的设计与分析。这些方程通常来源于物理问题或工程问题，在实际应用中具有重要的意义，在数学尤其是应用数学的研究中又具有非常强的代表性。如对于对流方程、刘维尔方程及输运方程，相关的经典计算方法的设计和分析虽然相对成熟，但是在实际的物理问题或者工程问题中，由于各种误差和不确定性的存在使得这些经典方法的应用大大受限，这时，如何对这种误差或不确定性带来的影响进行数值估计和分析就变得非常重要了，这也正是我们在本文中讨论的重点之一（第二章和第三章）。而对于薛定谔方程，由于属于所谓的高频波问题，会使通常的数值格式的计算效率非常低下甚至不可用，这类问题目前还没有特别完美的数值格式可以在实际中大规模使用，针对这一点，本文试图通过改进一类特定的薛定谔方程的快速算法来做出一些自己的贡献（第四章）。最后，对于分数阶的守恒律方程，属于近年来新兴的一类问题，特别是在物理中的所谓多孔介质中的记忆效应的研究中非常重要，并且这类问题的已有研究相对较少，本文中通过构造相应的数值格式，期望增加我们对相关现象的理解（第五章）。

在本章中，我们将对相关问题作简要的介绍，主要包括研究的背景，研究的动机以及研究的内容和成果。

## 1.1　物理方程的不确定性量化

数值模拟的终极目标是预测物理事件或工程系统的行为。人们投入了大量的努力来开发精确的数值算法，目的是使得在数值误差得到很好的控制和理解的意义上，模拟预测是可靠的。这是数值分析的主要目标，并且仍然保持着很高的活跃度。然而，在传统意义的数值分析中，对于参数值、初始和边界条件等“数据”中错误或不确定性的影响的很少关心。不确定性量化（Uncertainty Quantification，简称 UQ）的目标是研究这些错误对数据的影响，并随后为实际问题提供更可靠的预测。这个领域在过去几年中受到越来越多的关注，特别是在复杂系统的背景下，数学模型只能作为真实物理学的简化和缩小的表示。虽然许多模型已经成功地揭示了预测和观测之间的量化关系，但是它们的使用受到我们在控制方程中为各种参数分配精确数值的能力的限制。由于我们对潜在的物理规律或不可避免的测量误差了解的并不完整，不确定性正代表了数据的这种变异性并且时刻存在。因此，为了充分了解模拟结果和随后的真实物理学，必须从模拟一开始就引入不确定性，而不是之后。

很多人在土木工程，水利，控制论等学科领域长期以来已经认识到了解不确定性的重要性。随之而来的是解决这个问题的方法。由于不确定性的“不确定”的性质，最主要的方法是将数据不确定性视为随机变量或随机过程，并将原始确定性系统重写为随机系统。我们指出，这种类型的随机系统与经典的“随机微分方程”（SDE）不同，其中随机输入是一些理想化的过程，如维纳过程，泊松过程等，并且诸如随机分析的工具已经被广泛发展，仍在积极研究之中，参见例如 [5-8]。

### 1.1.1 已有方法简介

#### 1.1.1.1 蒙特卡罗和基于抽样的方法

最常用的方法之一是蒙特卡罗取样（Monte Carlo Sampling，MCS）或其变体之一。在 MCS 中，根据预先给定的概率分布，生成随机输入的一些（独立）实例。对于每个实例，数据是固定的，问题变成确定。在解决问题的确定性实例之后，人们收集一组解决方案，即随机解的实例。从这个集合可以提取统计信息，例如均值，方差等。尽管 MCS 的应用是非常直接的，因为它只需要重复执行确定性的模拟，通常需要大量的重复，解的统计量的计算相对较慢。例如，平均值通常收敛为 $\frac{1}{\sqrt{K}}$，其中 K 是实例的数量（例如 [9]）。对于准确结果的大量实现的需要可能导致过多的计算负担，特别是对于在其确定性设置中已经具有计算密集度的系统。人们已经开发了用于暴力加速 MCS 的收敛的技术，例如拉丁超立方体采样（参见 [10-11]），准蒙特卡罗（参见 [12-14]）等等。其中值得一提的进年来发展迅速、具有很大潜力的多级蒙特卡罗方法（Multilevel Monte Carlo，MLMC），这种方法一种递归控制变量策略，使用一些随机输出量的廉价不准确的近似作为控制变量，以获得更准确但成本更高随机量的近似值，详细的介绍可以参考文献 [15]。然而，这些方法的设计同常有很多额外的限制，并且它们的适用性通常是有限的。

#### 1.1.1.2 扰动方法

最流行的非抽样方法是扰动方法，其中随机场通过泰勒级数在平均值附近展开，并在某些项截断。通常，最多使用二阶展开，因为高阶展开所得到的等式的结果变得非常复杂。这种方法已被广泛应用于各种工程领域 [16-17]。扰动方法的固有局限性在于，输入和输出的不确定性的大小不能太大（通常小于 10%），而且这些方法在其他方面表现不佳。

#### 1.1.1.3 矩方程方法

在这种方法中，人们试图直接计算随机解的矩。未知数是解的各阶矩，其方程是通过取原始随机控制方程的平均得出的。例如，平均场由控制方程的平均值确定。困难在于，矩方程的导出，除了在极少数情况下，总是需要更高阶矩的信息。这就引出了所谓的“矩封闭”问题，这通常是通过利用一些关于更高阶矩的特别性质来解决的。在水利学背景下，矩方程的更详细的介绍可以在 [18] 中找到。

#### 1.1.1.4 基于算子的方法

这些方法是基于控制方程中随机算子的运算。它们包括 Neumann 展开，其表示 Neumann 级数中的随机算子的逆 [19-20] 和加权积分法 [21-22]。类似于扰动方法，这些基于运算子的方法也被限制在很小的不确定性的情况下。它们的适用性通常强烈依赖于底层操作员，并且通常限于静态问题。

### 1.1.2 广义多项式混沌（gPC）

最近发展的，广义多项式混沌（generalized Polynomial Chaos, gPC）[23-24]，即经典多项式混沌的推广，已经成为最广泛使用的方法之一。使用 gPC，随机解被表示为输入随机参数的正交多项

式，并且可以选择不同类型的正交多项式以实现更好的收敛。它本质上是随机空间中的谱表示，并且当解光滑地依赖于随机参数时，其表现出快速收敛。基于 gPC 的方法将成为本文的重点。

### 1.1.3 gPC 方法的发展与应用

gPC 的发展开始于 R.Ghanem 及其同事们在 PC（多项式混沌）上的创新工作。受 Wiener-Hermite 均匀混沌理论的启发（Ghanem）[25]），Ghanem 采用 Hermite 多项式作为正交基来表示随机过程，并将该技术应用于许多工程问题的获得了成功。概述可以在 [24] 中找到。

Hermite 多项式的使用虽然在数学上是完备的，但在一些应用中，特别是在非高斯问题的收敛和概率近似的情况下，存在困难 [26-27]。随后，在 [23] 中提出了广义多项式混沌（gPC）来减轻困难。在 gPC 中，根据随机输入的概率分布，选择不同种类的正交多项式作为基。可以通过选择适当的基来实现最优收敛。在一系列论文中，gPC 的强大在各种 PDE 中得到了证明 [28-29]。

对于 gPC 的工作进一步推广是不需要作为基的多项式全局光滑。原则上任何一组完备的基都是可行的选择，就像有限元法一样，取决于给定的问题。这种推广包括分段多项式基 [30]，小波基 [31] 和多元 gPC[32]。

### 1.1.4 侵入性方法（Intrusive Method）：随机伽辽金方法

当应用于具有随机输入的微分方程时，要求解的量是 gPC 展开的系数。典型的方法是进行伽辽金投影以最小化有限阶 gPC 展开的误差，并且因为用于扩展系数的求解方程组是确定性的，可以通过常规数值方法来解决。这种随机伽辽金（stochastic Galerkin，SG）方法，已经在 PC 的早期工作中应用，并被证明是有效的，也是本文中主要使用的方法。

### 1.1.5 非侵入性的方法（Non-intrusive Method）：随机配点法

近年来，随着 [33] 的工作，高阶随机配点（stochastic collocation，SC）方法的兴趣激增。这在某种程度上是对旧技术“确定性抽样方法”的重新发现。随机配点方法的早期工作包括 [34-35]，并使用一维正交点的张量积为“抽样点”。虽然已经表明这种方法可以实现高阶逼近，但参见 [36]，其适用性仅限于较少数量的随机变量，因为采样点的数量以指数速度增长。[33] 的工作从多元插值分析中引入了“稀疏网格”技术，可以显着减少较高随机维数的采样点数。SC 算法的实现与 MCS 类似，即只需要通过选择适当的采样点集合，使用确定性求解器进行重复实现。原始高阶随机配点中，基函数是由节点定义的拉格朗日多项式，即稀疏网格 [33] 或张量网格 [36]。[37] 提出了一种更实用的“伪谱”方法，可以在 gPC 逼近的基础上与配点法进行结合。伪光谱 gPC 方法在实践中被证明比拉格朗日插值法更容易实现。

### 1.1.6 gPC 方法小结

由于 gPC-SG 方法在数值分析和实践中具有相当的简洁性，所以本文中我们将以 gPC-SG 方法为主，而使用 gPC-SC 方法作为参考对比。关于两者之间的优缺点及相关分析研究，读者可以参考 [38]。

## 1.2 本文的主要内容与创新点

### 1.2.1 带有随机、间断系数的双曲型方程方程：提出离散 gPC-SG 方法

这类问题出现在异质介质中的波传播中，通过不同介质之间的界面或潜在的障碍，使这些方程中的系数产生间断。随机或不确定性来自建模或实验误差。

为了处理间断系数，我们需要在间断点提供额外的物理条件，并在哈密顿保持格式的框架内 [1-2,39-40]，将这种条件构建到数值通量中。为了处理不确定性，我们使用 gPC-SG 方法。标准 gPC-SG 方法从原始方程的 gPC 近似开始，得到 gPC 系数的确定性方程组，然后通过空间和时间的标准格式离散化。这里的主要困难在于 gPC 近似仅当原始问题的解在随机空间中是光滑时才是准确的，而在我们的问题中解却具有高度奇异性。

我们提出了克服这个困难的新方法，主要想法是逆转上述 gPC-SG 过程。也就是说，我们首先用空间和时间离散原始方程，使用光滑的数值通量，然后将 gPC 近似应用于该离散方程。由于离散的解比连续的解更光滑，所以 gPC 近似被应用于更光滑滑的函数（对于固定时间步长和网格大小），因此可以得到很好的收敛速度。

### 1.2.2 带有不确定性的输运方程：gPC-SG 方法的一致（关于克努森数）收敛性分析与 micro-macro 格式的构建

在这个问题上，我们并没有遇到不光滑的解，但这里的问题是不确定性和多尺度并存。我们研究了在碰撞横截面，初始数据或边界数据中包含不确定性的线性输运方程。多尺度性质的特征由克努森数号 $\varepsilon$ 表示，其在所谓的光学薄区域 ($\varepsilon \ll 1$) 中，由于粒子的高散射率，导致线性传输方程趋近到扩散方程，称为扩散极限。近来，开发具有不确定性和扩散尺度的线性传输方程的渐近保持（AP）格式非常活跃（在随机 Galerkin 方法的框架下，成为 s-AP 方法）[41]。如果当 $\varepsilon \to 0$ 时，用于线性传输方程的随机 Galerkin 方法收敛到用于扩散极限方程的随机 Galerkin 方法，则方法是 s-AP 的。

这个研究的主要困难在于，当 $\varepsilon \ll 1$ 时，能量估计和收敛速度通常取决于 $\varepsilon$ 的倒数，这意味着需要使用的 gPC 近似多项式的阶数随着 $\varepsilon$ 减少而剧烈增加。虽然 AP 格式可以使用独立于 $\varepsilon$ 的数值参数，但是要严格证明这一点非常困难。

在本研究中，我们为随机碰撞横截面的线性传输方程提供了随机 Galerkin 方法的最佳误差估计，这是首次由人得到关于这类问题的一致收敛结果。对于数值部分，我们使用基于 micro-macro 分解的方法开发完全离散的 s-AP 方法。这种方法的优点在于它允许我们的到不依赖于 $\varepsilon$ 的稳定性条件。

### 1.2.3 使用非均匀快速傅立叶变换（NUFFT）改进具有向量势的半经典薛定谔方程的快速算法

为了设计解带有向量势的电磁场中的薛定谔方程的无条件稳定的格式，Jin 和 Zhou 在 [3] 中引入了一种半拉格朗日时间算子分裂方法，其中网格划分策略 $\Delta t = O(\varepsilon)$ 和 $\Delta x = O(\varepsilon)$ 就足以保证波函数的精确近似。这里的小参数 $\varepsilon$ 代表缩放的普朗克常数。类似地，可以使用与 $\varepsilon$ 独立的时间步长来捕获正确的物理可观察量。在对流骤中，多项式插值技术在 [3] 中进行了分析和实现，为了效率考虑牺牲了空间精度。实际上，如果使用谱插值就可以提高空间精度，不幸的是，它会将计算复杂

度从 $O(N)$（多项式插值）增加到 $O(N^2)$（直接傅里叶级数求和）其中 $N$ 是网格点的数量。这里的主要问题是标准逆 FFT 不再适用，因为取样的点不一定均匀分布。

由于非均匀快速傅里叶变换（NUFFT）（见 [42-43]）的提出，问题可以理想地解决。这是我们工作的主要动力。我们将 NUFFT 算法合并到时间分裂半拉格朗日方法中，计算复杂度为 $O(N \log N)$。我们证明了该方法是无条件稳定的。与多项式插值的情况不同，现在通过全局谱近似来进行插值。当需要还原时间和空间振荡时，即 $\Delta x = O(\varepsilon)$ 及 $\Delta t = O(\varepsilon)$，我们证明我们的方法在空间和时间上的额准确的。我们还在 Wigner 变换的框架中证明，使用与 $\varepsilon$ 独立的时间步长即可允许我们计算正确的物理观测值。

### 1.2.4　具有分数阶导数（Caputo）的守恒律方程的数值分析与计算

这项研究属于经典数值分析和计算，问题不具有不确定性。但是与经典守恒律不同，对于时间导数，我们使用 Caputo 导数 $(\partial_t^\alpha, \alpha \in (0, 1])$，其引入了多孔流体扩散与记忆这种非局部记忆效应，以及与非线性守恒定律相结合，我们对数值分析和计算两个方面的兴趣。

我们提出了一阶和二阶显示格式，并显示了使用修改的 CFL 条件，数值方案是 TVD（总变差减小）的。然而，修改的 CFL 条件越来越受限于 $\alpha \to 0$，这使得显式格式对于小的 $\alpha$ 不可行。受此影响，我们进一步设计了一种隐式的迎风格式，该方法被证明是无条件地稳定的，并且同样也是 TVD 的。在这些格式的帮助下，数值计算证明了方法的有效性，并提供数值证据来解释了守恒律中的记忆效应。据我们所知，这是对分数导数的非线性守恒定律的少数尝试之一。

# 第二章　gPC-SG 方法在带有间断及随机系数的双曲型方程中的应用

我们的目标是发展有效的数值方法来解决具有不光滑和不确定系数的线性双曲型方程。这样的问题通常出现在异质介质中的波传播中，由于不同介质之间具有界面或势垒，方程出现间断或甚至更很强的奇异性。而随机或不确定性的出现则是由于建模或实验中的误差，由于双曲方程中的通量通常是由*经验定律*、状态方程或矩封闭方法给出，导致了这种误差*非常特殊*，通常无法通过其他手段去掉或者减小。

当双曲方程包含奇异系数时，通常需要在奇异点提供额外的物理条件以使得初边界值问题是适定的，同时能够反应在界面或势垒处的波的正确物理行为 [1-2]。在势垒的情况下，一个自然的物理条件就是折射和反射条件，在所谓的哈密尔顿守恒格式（Hamiltonian-Preserving）的框架中 [2,40]，这些条件可以自然的构建到数值格式的通量中。这是我们将采用的解决奇异系数引起的困难的方法。

为了处理由随机不确定性带来的困难，我们将利用广义多项式混沌（generalized Polynomial Chaos (gPC)）展开为基础的随机伽辽金（stochastic Galerkin (SG)）方法 [23-24,44-49]。如果方程的解在随机空间具有足够的正则性，这种方法实现谱收敛，远快于经典的蒙特卡罗方法，因此对于这种随机不确定性的问题更有效率。不幸的是，对于双曲型问题，解经常不会有这样好的正则性，这导致方法收敛的速度非常缓慢甚至给出由于吉布斯现象导致的非收敛结果 [50-51]。本章中研究的对象就是由于在物理空间的界面、势垒导致的解的跳跃、间断，随之波动方程的演化而传播到随机空间，从而导致解在随机空间具有很差的光滑性的一类问题。

标准的 gPC-SG 方法过程如下：先对原始的微分方程在随机空间做 gPC 逼近（随机空间的正交多项式逼近），这样会得到一个关于 gPC 展开系数的*确定性的*方程组（而随机性蕴含在 gPC 正交多项式的基函数）。然后再用通常的方法将其空间和时间数值离散化（有限差分，有限体积，有限元或谱方法）。如果原始问题的解在随机空间是充分光滑的，那么这种 gPC 逼近方法是非常准确的。不幸的是，我们研究的问题恰恰不属于这种情况。

本章中我们的主要思想是*逆转*上述 gPC-SG 过程。也就是说，我们首先使用*光滑的*数值通量在空间和时间上对原始方程进行离散，然后再将 gPC 近似应用于该离散方程。由于离散后的数值解解比原连续方程的解具有更好的正则性，gPC 逼近则应用于该光滑的函数（对于固定的时间步长和网格大小），因此可以期望有更好的收敛速度。我们称这样的 gPC-SG 方法为*离散 gPC-SG 方法*。在 [50] 的随机配点法的框架下，提出并分析了一个类似的想法，他们也注意到，虽然解确实是不连续的，但是一些人们感兴趣的量（quantities of interests, QoI）通常具有更好的正则性，从而可以期待更好的收敛速度。

对于双曲型方程，光滑的数值通量通常由中心差分得到，因为中心差分格式不依赖于特征线的信息（例如 Lax-Friedrichs 格式，Lax-Wendroff 格式，等等）。而迎风类型的格式则通常会导致不光滑的数值通量，因为它们依赖于于特征线传播速度的绝对值。对于二阶（或高阶）格式，为了抑制数值粘性，通常使用斜率限制器（slope limiter 或者 flux limiter）或 ENO 或 WENO 型重构 [52-54]，

而其通常不是光滑函数。为了保持数值通量光滑，我们使用文章 [55] 中引入的光滑的 BAP 斜率限制器。

在本章中，我们将用这个新想法的来研究两个问题。首先是具有间断和随机系数的标量双曲型方程：

$$u_t(x,t,z) + \big[c(x,z)u(x,t,z)\big]_x = 0, \quad t > 0. \tag{2–1}$$

这里 $c(x,z)$ 是随机系数（随机波速），其中 $z$ 是在完备样本空间 $\Omega$ 中概率密度分布函数为 $\rho(z)$ 的随机变量。$c(x,z)$ 关于 $x$ 是简短的，这对应于不同介质之间的界面。第二个问题是是粒子密度分布 $u(x,v,t,z) > 0$ 的刘维尔（Liouville）方程：

$$u_t + vu_x - V_x u_v = 0, \quad t > 0, \quad x, v \in \mathbb{R}, \tag{2–2}$$

其中势函数 $V(x,z)$ 关于 $x$ 中是不连续的，这对应于一个势垒的情形。在这些问题中我们感兴趣的量包括 $u$ 的期望值，

$$\mathbb{E}[u] = \int u(z)\rho(z)\,\mathrm{d}z. \tag{2–3}$$

和方差

$$\mathbb{V}[u] := \mathbb{E}\big[(u - \mathbb{E}(u))^2\big] = \int u(z)^2\rho(z)\,\mathrm{d}z - (\mathbb{E}[u])^2 \tag{2–4}$$

对于方程 (2–1)，通过先使用 Lax-Friedrichs 格式再应用 gPC-SG 方法，我们将建立所提出的方法的正则性结果以及由此带来的在随机空间的谱收敛结果，而空间和时间的数值收敛是与在 [39] 中建立的确定性问题的分析类似。对于这些分析结果，我们都会给出相应的数值验证。

在这些问题中，不确定性也可能来自初始数据。这是个问题已经在文章 [51,56] 中给出了仔细的分析，我们的方法显然可以在这种情况下使用。

本章的结构如下。在2.1节中，我们将介绍离散 gPC-SG 方法应用于对流方程 (2–1)，并进行完整的正则性和全离散格式的数值收敛分析。在2.2节中，我们将展示如何将这个想法用于一阶和二阶空间离散的刘维尔方程。在2.3节中，我们将给出对应于这两个方程的数值示例，它们将证实该方法在随机空间的谱收敛。

## 2.1 带有间断、随机波速的对流方程的离散 gPC-SG 方法

我们考虑如下的标量对流方程模型

$$\begin{cases} u_t(x,t,z) + \big[c(x,z)u(x,t,z)\big]_x = 0, \quad t > 0, \\ u(x,0,z) = u_0(x,z). \end{cases} \tag{2–5}$$

这里 $c(x,z)$ 关于 $x$ 在某些地方是间断的，例如，

$$c(x,z) = \begin{cases} c^-(z) > 0, & \text{如果 } x < 0, \\ c^+(z) > 0, & \text{如果 } x > 0. \end{cases} \tag{2–6}$$

根据文章 [1]，我们需要在 $x = 0$ 处给一个界面条件使得整个问题是适定的：

$$u(0^-,t,z) = \alpha(z)u(0^+,t,z). \tag{2–7}$$

其中 $\alpha(z)=1$ 对应于质量守恒定律，而对于通量的守恒定义为 $\alpha(z)=c^-(z)/c^+(z)$，后者是我们在本章中要考虑的情形。注意，这里我们假设 $c(x,z)$ 关于随机变量 $z$ 足够光滑，并且只有一个不连续点位于 $x=0$。

由于界面条件 (2–7) 导致的 $u(x,t,z)$ 的间断将从物理空间传播到随机空间，带来的吉布斯现象（Gibb's phenomenon）使得 gPC-SG 方法收敛速度非常慢。这里我们提出一个与传统 gPC-SG 方法略有不同的方法：我们首先在空间和时间离散方程 (2–5)，如同 [39] 中引入的方法，这时随机变量 $z$ 视为固定的参数。[39] 中的关键想法是将界面条件 (2–7)“浸没”（immersed）到格式中。之后再对 gPC-SG 方法应用于离散的系统。

### 2.1.1　格式

设空间网格为 $x_i=i\Delta x$，其中 $i\in\mathbb{Z}$ 是所有整数的集合，$\Delta x$ 是网格大小。令 $t^n=n\Delta t$ 是时间离散，其中 $\Delta t$ 是时间步长。令 $U_i^n(z)=U(x_i,t^n,z)$ 是 $u(x_i,t^n,z)$ 的数值逼近。Jin 和 Qi 在 [39] 中为 (2–5) (2–7) 提出的浸没迎风格式（immersed upwind scheme）为

$$\begin{cases} U_i^{n+1}(z)=(1-\lambda^-(z))U_i^n(z)+\lambda^-(z)U_{i-1}^n(z), & \text{如果 } i\le 0,\\ U_i^{n+1}(z)=(1-\lambda^+(z))U_i^n(z)+\lambda^-(z)U_{i-1}^n(z), & \text{如果 } i=1,\\ U_i^{n+1}(z)=(1-\lambda^+(z))U_i^n(z)+\lambda^+(z)U_{i-1}^n(z), & \text{如果 } i\ge 2, \end{cases} \tag{2–8}$$

其中 $\lambda^\pm(z)=c^\pm(z)\Delta t/\Delta x$。

注意，从这个离散格式 (2–8) 看出，如果假设 $U_i^n(z)$ 对于每个固定的 $i$ 是关于 $z$ 的光滑函数，那么在一个时间步之后，$U_i^{n+1}(z)$ 仍然是 $z$ 的光滑函数。原因很简单：$\lambda^\pm(z)=c^\pm(z)\Delta t/\Delta x$ 是 $z$ 的光滑函数！由于我们假设初始数据相对于 $z$ 是光滑的，所以任何时刻 $t^n$ 的数值解也应当相对于 $z$ 是光滑的。那么，如果这时候将标准 gPC-SG 方法应用于该离散系统，则当物理网格 $\Delta x$ 和 $\Delta t$ 固定时，可以期望将 gPC-SG 方法会谱收敛到这个数值离散的解。

根据标准的 gPC-SG 方法，我们写出 $U_i^n(z)$ 关于 $z$ 的 gPC 展开。即用如下的有限 gPC 展开级数来做逼近，

$$U_{i,(K)}^n(z)=\sum_{k=0}^{K}\hat{U}_{i,(k)}^n P_k(z), \tag{2–9}$$

其中的 $P_k(z)$ 是以 $\rho(z)$ 为权重的正交多项式基，满足下述正交条件

$$\langle P_i,P_j\rangle=\int P_i(z)P_j(z)\rho(z)\,\mathrm{d}z=\delta_{ij}, \tag{2–10}$$

其中的内积定义为

$$\langle f,g\rangle=\int f(z)g(z)\rho(z)\,\mathrm{d}z, \tag{2–11}$$

$\delta_{ij}$ 是克罗内克 $\delta$ 函数。展开系数由如下式子确定

$$\hat{U}_{i,(k)}^n=\int U_{i,(K)}^n(z)P_k(z)\rho(z)\,\mathrm{d}z. \tag{2–12}$$

通过使用 (2–9) 展开和应用伽辽金（Galerkin）投影，可以得到系数 $\hat{U}^n_{i,(k)}$ 满足如下方程组

$$\begin{cases} \hat{\boldsymbol{U}}_i^{n+1} = (1-\boldsymbol{\lambda}^-)\hat{\boldsymbol{U}}_i^n + \boldsymbol{\lambda}^-\hat{\boldsymbol{U}}_{i-1}^n, & \text{如果 } i \le 0, \\ \hat{\boldsymbol{U}}_i^{n+1} = (1-\boldsymbol{\lambda}^+)\hat{\boldsymbol{U}}_i^n + \boldsymbol{\lambda}^-\hat{\boldsymbol{U}}_{i-1}^n, & \text{如果 } i = 1, \\ \hat{\boldsymbol{U}}_i^{n+1} = (1-\boldsymbol{\lambda}^+)\hat{\boldsymbol{U}}_i^n + \boldsymbol{\lambda}^+\hat{\boldsymbol{U}}_{i-1}^n, & \text{如果 } i \ge 2. \end{cases} \tag{2–13}$$

这里 $\hat{\boldsymbol{U}}_i^n = (\hat{U}^n_{i,(0)}, \dots, \hat{U}^n_{i,(K)})^T$ 是 $(K+1)$ 维的向量，$\boldsymbol{\lambda}^\pm$ 是 $(K+1)\times(K+1)$ 的矩阵，其元素为 $\{\lambda^\pm_{k,m}\}_{0\le k,m\le K}$，其中

$$\lambda^\pm_{k,m} = \int c^\pm(z)P_k(z)P_m(z)\rho(z)\,\mathrm{d}z\,. \tag{2–14}$$

### 2.1.2 误差估计和收敛性分析

我们首先介绍一些将在分析中使用的符号，空间和范数。我们假设 $u(x,t,z)$ 在区域 $D=[a,b]$ 中具有紧支集，其中 $a<0$ 和 $b>0$ 使得该区域包含界面 $x=0$。$-M\le i\le M$ 是空间离散指标，$\Delta x=(b-a)/(2M+1)$。时间步长指标是 $n=0,1,\dots$。

在随机空间 $\Omega$ 上定义一个带权的 $L^2$ 范数

$$\|f(\cdot)\|^2_{L^2(\Omega)} = \int f^2(z)\rho(z)\,\mathrm{d}z. \tag{2–15}$$

同时我们定义范数

$$\|u^n(\cdot)\|^2_H := \int \|u^n(z)\|^2_{\ell^1(D)}\rho(z)\,\mathrm{d}z, \tag{2–16}$$

其中

$$\|u^n(z)\|_{\ell^1(D)} = \sum_{i=-M}^{M}|u_i^n(z)|\Delta x. \tag{2–17}$$

#### 2.1.2.1 离散数值解在随机空间的正则性

为了获得误差估计，我们需要研究离散解 $U_i^n(z)$ 在随机空间的正则性。自然地，我们需要对给定数据的做适当的一些假设。更确切地说，我们做出以下假设（参见 [51,57]）。

**假设 2.1.**

$$\max_{z\in\Omega}|\partial_z^s\lambda^\pm(z)|\le\gamma_\ell,\quad \max_{D\otimes\Omega}|\partial_z^s u_0(x,z)|\le\eta_\ell,\quad \forall 0\le s\le\ell, \tag{2–18}$$

*其中 $0\le\lambda^\pm(z)=c^\pm(z)\Delta t/\Delta x\le 1$，$\ell=1,2,\dots$，$\gamma$，$\eta$ 是正的常数。不失一般性，我们假设存在一个有限的 $\tau=\max\{\gamma_\ell,\eta_\ell,1\}$。*

注意在假设 (2.1) 中常数 $\gamma_\ell$ 和 $\eta_\ell$ 不依赖于 $x$。现在我们可以陈述并证明如下正则性结果：

**定理 2.1.** *假设 (2.1)，那么离散的数值解 $U_i^n(z)$ 满足*

$$\max_{i\in\mathbb{N},z\in\Omega}|\partial_z^\ell U_i^n(z)|\le C_\ell(n)(2\tau)^n\tau, \tag{2–19}$$

*对于 $\forall\ell\in\mathbb{N}$，其中*

$$C_\ell(n)=\sum_{s=0}^{n}\binom{n}{s}(1+s)^\ell\le 2^{(\ell+1)n}. \tag{2–20}$$

**证明.** 将 (2–8) 对 $z$ 微分 $\ell$ 次，

$$\begin{cases} \partial_z^\ell U_i^{n+1}(z) = \partial_z^\ell U_i^n(z) - \sum_{s=0}^{\ell} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^-(z)\partial_z^s U_i^n(z) + \sum_{s=0}^{\ell} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^-(z)\partial_z^s U_{i-1}^n(z) & \text{如果 } i \le 0, \\ \partial_z^\ell U_i^{n+1}(z) = \partial_z^\ell U_i^n(z) - \sum_{s=0}^{l} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^+(z)\partial_z^s U_i^n(z) + \sum_{s=0}^{l} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^-(z)\partial_z^s U_{i-1}^n(z) & \text{如果 } i = 1, \\ \partial_z^\ell U_i^{n+1}(z) = \partial_z^\ell U_i^n(z) - \sum_{s=0}^{\ell} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^+(z)\partial_z^s U_i^n(z) + \sum_{s=0}^{l} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^+(z)\partial_z^s U_{i-1}^n(z) & \text{如果 } i \ge 2. \end{cases}$$

我们将对 $n$ 使用数学归纳法。当 $n = 1$ 时，初值经过一个时间步长

$$\begin{cases} \partial_z^\ell U_i^1(z) = \partial_z^\ell U_i^0(z) - \sum_{s=0}^{\ell} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^-(z)\partial_z^s U_i^0(z) + \sum_{s=0}^{\ell} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^-(z)\partial_z^s U_{i-1}^0(z) & \text{如果 } i \le 0, \\ \partial_z^\ell U_i^1(z) = \partial_z^\ell U_i^0(z) - \sum_{s=0}^{l} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^+(z)\partial_z^s U_i^0(z) \sum_{s=0}^{l} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^-(z)\partial_z^s U_{i-1}^0(z) & \text{如果 } i = 1, \\ \partial_z^\ell U_i^1(z) = \partial_z^\ell U_i^0(z) - \sum_{s=0}^{\ell} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^+(z)\partial_z^s U_i^0(z) + \sum_{s=0}^{l} \binom{\ell}{s} \partial_z^{\ell-s}\lambda^+(z)\partial_z^s U_{i-1}^0(z) & \text{如果 } i \ge 2. \end{cases}$$

根据假设2.1，

$$\max_{i\in\mathbb{N},z\in\Omega} |\partial_z^s U_i^0(z)| = \max_{i\in\mathbb{N},z\in\Omega} |\partial_z^s u_0(x_i,z)| \le \max_{D\otimes\Omega} |\partial_z^s u_0(x,z)| \le \tau, \tag{2–21}$$

以及

$$\max_{z\in\Omega} |\partial_z^{\ell-s}\lambda^\pm(z)| \le \tau. \tag{2–22}$$

我们有

$$\begin{aligned} \max_{i\in\mathbb{N},z\in\Omega} |\partial_z^\ell U_i^1(z)| \le & \max_{i\in\mathbb{N},z\in\Omega} |\partial_z^\ell U_i^0(z)| + \sum_{s=0}^{\ell} \binom{\ell}{s} \max_{z\in\Omega} |\partial_z^{\ell-s}\lambda^\pm(z)| \max_{i\in\mathbb{N},z\in\Omega} |\partial_z^s U_i^0(z)| \\ & + \sum_{s=0}^{\ell} \binom{\ell}{s} \max_{z\in\Omega} |\partial_z^{\ell-s}\lambda^\pm(z)| \max_{i\in\mathbb{N},z\in\Omega} |\partial_z^s U_{i-1}^0(z)| \\ \le & \tau + 2\tau^2 \sum_{s=0}^{l} \binom{\ell}{s} \le 2\tau(2^\ell + 1)\tau, \end{aligned} \tag{2–23}$$

满足 (2–19) 当 $n = 1$ 的情形，奠基成立。

接下来假设当 $n = p$ 时，(2–19) 成立，即：

$$\max_{i\in\mathbb{N},z\in\Omega} |\partial_z^\ell U_i^p(z)| \le C_\ell(p)(2\tau)^p\tau, \quad \forall \ell \in \mathbb{N}. \tag{2–24}$$

那么当 $n=p+1$ 时，和前面一样，

$$\begin{aligned}\max_{i\in\mathbb{N},z\in\Omega}|\partial_z^\ell U_i^{p+1}(z)| \leq & \max_{i\in\mathbb{N},z\in\Omega}|\partial_z^\ell U_i^p(z)| + \sum_{s=0}^{\ell}\binom{\ell}{s}\max_{z\in\Omega}|\partial_z^{\ell-s}\lambda^{\pm}(z)|\max_{i\in\mathbb{N},z\in\Omega}|\partial_z^s U_i^p(z)| \\ & + \sum_{s=0}^{\ell}\binom{\ell}{s}\max_{z\in\Omega}|\partial_z^{\ell-s}\lambda^{\pm}(z)|\max_{i\in\mathbb{N},z\in\Omega}|\partial_z^s U_{i-1}^p(z)| \\ \leq & C_\ell(p)(2\tau)^p\tau + 2\sum_{s=0}^{\ell}\binom{\ell}{s}\tau C_{\ell-s}(p)(2\tau)^p\tau \\ \leq & \left(C_\ell(p)+\sum_{s=0}^{\ell}\binom{\ell}{s}C_{\ell-s}(p)\right)(2\tau)^{p+1}\tau \\ := & C_\ell(p+1)(2\tau)^{p+1}\tau.\end{aligned} \tag{2–25}$$

从最后一个等式我们得到 $C_\ell(p)$ 的递归关系，

$$C_\ell(n+1) = C_\ell(n) + \sum_{s=0}^{\ell}\binom{\ell}{s}C_{\ell-s}(n), \tag{2–26}$$

通过归纳法可得

$$C_\ell(n) = \sum_{s=0}^{n}\binom{n}{s}(1+s)^\ell, \tag{2–27}$$

证实我们想要的结果，证毕。 □

**注 1.** 系数满足

$$C_\ell(n) = \sum_{s=0}^{n}\binom{n}{s}(1+s)^\ell \leq 2^n(1+n)^\ell \leq 2^{(\ell+1)n}. \tag{2–28}$$

对于给定时刻 $T=n\Delta t$，

$$C_\ell(n) \leq 2^{\frac{T}{\Delta t}}\left(1+\frac{T}{\Delta t}\right)^\ell \leq 2^{\frac{(\ell+1)T}{\Delta t}}. \tag{2–29}$$

#### 2.1.2.2 gPC-SG 方法的谱收敛

记 $U_i^n(z)$ 是线性对流方程的数值格式 (2–8) 的解。我们定义 $K$ 阶投影算子

$$\mathcal{P}_K U_i^n(z) = \sum_{k=0}^{K}\left\langle U_i^n(z), P_k(z)\right\rangle P_k(z). \tag{2–30}$$

gPC-SG 方法的误差可以分为两部分 $r_{i,(K)}^n(z)$ 和 $e_{i,(K)}^n(z)$，

$$\begin{aligned}U_i^n(z) - U_{i,(K)}^n(z) &= U_i^n(z) - \mathcal{P}_K U_i^n(z) + \mathcal{P}_K U_i^n(z) - U_{i,(K)}^n(z) \\ &:= r_{i,(K)}^n(z) + e_{i,(K)}^n(z),\end{aligned} \tag{2–31}$$

其中 $r_{i,(K)}^n(z) = U_i^n(z) - \mathcal{P}_K U_i^n(z)$ 是截断误差，$e_{i,(K)}^n(z) = \mathcal{P}_K U_i^n(z) - U_{i,(K)}^n(z)$ 是投影误差。

对于截断误差 $r_{i,(K)}^n(z)$，我们有如下引理：

**引理 2.2.** *在假设2.1下，对于给定时刻 $T = n\Delta t$ 和任意整数 $\ell \in \mathbb{N}$，*

$$\|r^n_{\cdot,(K)}(\cdot)\|_H \le \frac{(b-a)C_\rho(2^{\ell+2}\tau)^n\tau}{K^\ell}, \quad \forall \ell \in \mathbb{N}, \tag{2–32}$$

*其中 $C_\rho$ 是依赖于 $\{P_k(z)\}_{k\in\mathbb{N}}$ 的常数。*

**证明.** 根据 $r^n_{i,(K)}(z)$ 和范数 $\|\cdot\|_H$ 的定义，

$$\begin{aligned}
\|r^n_{\cdot,(K)}(\cdot)\| &= \|U^n_\cdot(\cdot) - \mathcal{P}_K U^n_\cdot(\cdot)\|_H \\
&= \left( \int \|U^n_\cdot(z) - \mathcal{P}_K U^n_\cdot(z)\|^2_{l^1(D)} \rho(z)\,\mathrm{d}z \right)^{1/2} \\
&= \left( \int \Big( \sum_{i=-M}^{M} |U^n_i(z) - \mathcal{P}_K U^n_i(z)|\Delta x \Big)^2 \rho(z)\,\mathrm{d}z \right)^{1/2} \\
&\le \sum_{i=-M}^{M} \left( \int |U^n_i(z) - \mathcal{P}_K U^n_i(z)|^2 \rho(z)\,\mathrm{d}z \right)^{1/2} \Delta x \\
&= \sum_{i=-M}^{M} \|U^n_i(\cdot) - \mathcal{P}_K U^n_i(\cdot)\|_{L^2(\Omega)} \Delta x,
\end{aligned} \tag{2–33}$$

这里我们已经应用了闵可夫斯基不等式（Minkowski inequality）。然后根据熟知的正交多项式逼近的结果 [58]，得到

$$\|U^n_i(\cdot) - \mathcal{P}_K U^n_i(\cdot)\|_{L^2(\Omega)} \le \frac{C_\rho \|\partial^\ell_z U^n_i(z)\|_{L^2(\Omega)}}{K^\ell}. \tag{2–34}$$

由**定理2.1**，得到

$$\|\partial^\ell_z U^n_i(\cdot)\|_{L^2(\Omega)} \le \max_{i\in\mathbb{N}, z\in\Omega} \big|\partial^\ell_z U^n_i(z)\big| \left( \int \rho(z)\,\mathrm{d}z \right)^2 \le C_l(n)(2\tau)^n\tau \le 2^{(\ell+1)n}(2\tau)^n\tau, \tag{2–35}$$

对于 $\forall l \in \mathbb{N}$，所以

$$\|U^n_i(\cdot) - \mathcal{P}_K U^n_i(\cdot)\|_{L^2(\Omega)} \le C_\rho(2^{\ell+2}\tau)^n\tau/K^\ell, \quad \forall \ell \in \mathbb{N}, \tag{2–36}$$

导致

$$\|U^n_\cdot(\cdot) - \mathcal{P}_K U^n_\cdot(\cdot)\|_H \le \sum_{i=-M}^{M} C_\rho(2^{\ell+2}\tau)^n\tau/K^\ell\Delta x = \frac{(b-a)C_\rho(2^{\ell+2}\tau)^n\tau}{K^\ell}. \tag{2–37}$$

证毕。 □

接下来要估计 $e^n_{i,(K)}(z)$。为此，首先注意到 $U^n_{i,(K)}(z)$ 满足

$$\begin{cases}
U^{n+1}_{i,(K)}(z) = U^n_{i,(K)}(z) - \mathcal{P}_K\big[\lambda^-(z)\big(U^n_{i,(K)}(z) - U^n_{i-1,(K)}(z)\big)\big] & \text{如果 } i \le 0, \\
U^{n+1}_{i,(K)}(z) = U^n_{i,(K)}(z) - \mathcal{P}_K\big[\big(\lambda^+(z)U^n_{i,(K)}(z) - \lambda^-(z)U^n_{i-1,(K)}(z)\big)\big] & \text{如果 } i = 1, \\
U^{n+1}_{i,(K)}(z) = U^n_{i,(K)}(z) - \mathcal{P}_K\big[\lambda^+(z)\big(U^n_{i,(K)}(z) - U^n_{i-1,(K)}(z)\big)\big] & \text{如果 } i \ge 2.
\end{cases} \tag{2–38}$$

另一方面，直接对格式 (2–8) 做 $K$ 阶投影

$$\begin{cases} \mathcal{P}_K U_i^{n+1}(z) = \mathcal{P}_K U_i^n(z) - \mathcal{P}_K\big[\lambda^-(z)\big(U_i^n(z) - U_{i-1}^n(z)\big)\big] & \text{如果} i \le 0, \\ \mathcal{P}_K U_i^{n+1}(z) = \mathcal{P}_K U_i^n(z) - \mathcal{P}_K\big[\big(\lambda^+(z)U_i^n(z) - \lambda^-(z)U_{i-1}^n(z)\big)\big] & \text{如果 } i = 1, \\ \mathcal{P}_K U_i^{n+1}(z) = \mathcal{P}_K U_i^n(z) - \mathcal{P}_K\big[\lambda^+(z)\big(U_i^n(z) - U_{i-1}^n(z)\big)\big] & \text{如果 } i \ge 2. \end{cases} \tag{2–39}$$

(2–39) 减去 (2–38) 得到

$$\begin{cases} e_{i,(K)}^{n+1} = e_{i,(K)}^n - \mathcal{P}_K\big[\lambda^-(z)(e_{i,(K)}^n - e_{i-1,(K)}^n)\big] \\ \qquad\qquad - \mathcal{P}_K\big[\lambda^-(z)(r_{i,(K)}^n - r_{i-1,(K)}^n)\big] & \text{如果 } i \le 0, \\ e_{i,(K)}^{n+1} = e_{i,(K)}^n - \mathcal{P}_K\big[(\lambda^+(z)e_{i,(K)}^n - \lambda^-(z)e_{i-1,(K)}^n)\big] \\ \qquad\qquad - \mathcal{P}_K\big[(\lambda^+(z)r_{i,(K)}^n - \lambda^-(z)r_{i-1,(K)}^n)\big] & \text{如果 } i = 1, \\ e_{i,(K)}^{n+1} = e_{i,(K)}^n - \mathcal{P}_K\big[\lambda^+(z)(e_{i,(K)}^n - e_{i-1,(K)}^n)\big] \\ \qquad\qquad - \mathcal{P}_K\big[\lambda^+(z)(r_{i,(K)}^n - r_{i-1,(K)}^n)\big] & \text{如果 } i \ge 2. \end{cases} \tag{2–40}$$

其中为了书写清楚我们省略了 $z$。

现在我们可以给出对于投影误差 $e_{i,(K)}^n(z)$ 的如下估计

**引理 2.3.** *在假设2.1下，对于给定时刻 $T = n\Delta t$ 和任意整数 $\ell \in \mathbb{N}$ 投影误差满足如下估计*

$$\|e_{\cdot,(K)}^n(\cdot)\|_H \le \frac{2\tau(b-a)C_\rho C'_\ell(n)}{K^\ell}, \quad \forall \ell \in \mathbb{N}, \tag{2–41}$$

*其中 $C'_\ell(n) = \dfrac{(2^{\ell+2}\tau)^n - 3^n}{2^{\ell+2}\tau - 3}$，$C_\rho$ 是只依赖于正交多项式 $\{P_k(z)\}_{k\in\mathbb{N}}$ 的常数。*

**证明.** 首先，根据 (2–40)，对于 $i \le 0$ 我们有如下估计，

$$\begin{aligned} \|e_{i,(K)}^{n+1}\|_{L^2(\Omega)} \le \|e_{i,(K)}^n\|_{L^2(\Omega)} &+ \|\mathcal{P}_K\|\big[\max_{z\in\Omega}(\lambda^-(z))(\|e_{i,(K)}^n\|_{L^2(\Omega)} + \|e_{i-1,(K)}^n\|_{L^2(\Omega)})\big] \\ &+ \|\mathcal{P}_K\|\big[\max_{z\in\Omega}(\lambda^-(z))(\|r_{i,(K)}^n\|_{L^2(\Omega)} + \|r_{i-1,(K)}^n\|_{L^2(\Omega)})\big]. \end{aligned} \tag{2–42}$$

注意到 $\|\mathcal{P}_K\| \le 1$（投影算子）和 $\max\limits_{z\in\Omega}(\lambda^\pm(z)) \le 1$，可以得出

$$\begin{aligned} \|e_{i,(K)}^{n+1}\|_{L^2(\Omega)} \le \|e_{i,(K)}^n\|_{L^2(\Omega)} &+ \|e_{i,(K)}^n\|_{L^2(\Omega)} + \|e_{i-1,(K)}^n\|_{L^2(\Omega)} \\ &+ \|r_{i,(K)}^n\|_{L^2(\Omega)} + \|r_{i-1,(K)}^n\|_{L^2(\Omega)}. \end{aligned} \tag{2–43}$$

根据 (2–36)，

$$\|r_{i,(K)}^n\|_{L^2(\Omega)} \le C_\rho(2^{\ell+2}\tau)^n\tau/K^\ell, \quad \forall i \in \mathbb{Z}, \forall \ell \in \mathbb{N}, \tag{2–44}$$

所以

$$\|e_{i,(K)}^{n+1}\|_{L^2(\Omega)} \le 2\|e_{i,(K)}^n\|_{L^2(\Omega)} + \|e_{i-1,(K)}^n\|_{L^2(\Omega)} + C_\rho 2\tau(2^{\ell+2}\tau)^n/K^\ell. \tag{2–45}$$

类似的，对于 $i = 1$ 和 $i \ge 2$，我们有和上面一样的估计，将这些估计加在一起并且乘以 $\Delta x$

$$\|e_{(K)}^{n+1}\|_H \le 3\|e_{(K)}^n\|_H + 2\tau(b-a)C_\rho(2^{\ell+2}\tau)^n/K^\ell. \tag{2–46}$$

使用这个递归关系以及注意到 $\|e^0_{(K)}\|_H = 0$，得到

$$\|e^n_{(K)}\|_H \le \frac{2\tau(b-a)C_\rho}{K^\ell}\frac{(2^{\ell+2}\tau)^n - 3^n}{2^{\ell+2}\tau - 3} := \frac{2\tau(b-a)C_\rho C'_\ell(n)}{K^\ell}. \tag{2–47}$$

证毕。 □

现在我们可以陈述 gPC-SG 方法对于离散系统的收敛性定理：

**定理 2.4.** *在假设2.1下，对于给定时刻 $T = n\Delta t$ 和任意给定整数 $\ell \in \mathbb{N}$，gPC-SG 方法应用于离散的格式所产生的误差*

$$\|U^n - U^n_{(K)}\|_H \le \frac{(b-a)C_\rho C(\ell,n)}{K^\ell}, \quad \forall \ell \in \mathbb{N}, \tag{2–48}$$

*其中 $C(\ell,n) = (2^{\ell+2}\tau)^n\tau + 2\tau C'_\ell(n)$。*

**证明.** 根据**引理2.2**和**引理2.3**，我们有

$$\|U^n - U^n_{(K)}\|_H \le \|r^n_{(K)}\|_H + \|e^n_{(K)}\|_H \le \frac{(b-a)C_\rho(2^{\ell+2}\tau)^n\tau}{K^\ell} + \frac{2\tau(b-a)C_\rho C'_\ell(n)}{K^\ell} := \frac{(b-a)C_\rho C(\ell,n)}{K^\ell},$$

证毕。 □

**注 2.** 常数 $C(\ell,n) = O\big(2^{(\ell+1)n}\big) = O\left(2^{\frac{(\ell+1)T}{\Delta t}}\right)$，这说明对于固定的网格参数，gPC-SG 方法是谱收敛。

#### 2.1.2.3 离散 gPC-SG 方法的误差估计

现在我们可以证明误差估计的主要结果。其中用到的部分误差估计引自 Jin 和 Qi 关于确定性问题的结果 [39]。

**引理 2.5.** *设 $u_0(x,z)$ 是 $z$ 的有界变差函数。那么格式 (2–8)，在 CFL 条件 $0 < \lambda^\pm(z) < 1$ 下，有如下的 $\ell^1$ 界：*

$$\|U^n(z) - u(\cdot,t^n,z)\|_{\ell^1(D)} \le C_1(z)\Gamma(c^-(z)) + C_2(z)\Gamma(c^+(z)), \quad \text{对每个固定的 } z, \tag{2–49}$$

*其中*

$$\Gamma(c^\pm(z)) = 2\sqrt{c^\pm(z)\Delta x(1 - c^\pm(z)\frac{\Delta t}{\Delta x})t_n} + \Delta x, \tag{2–50}$$

*$C_1(z)$，$C_2(z)$ 是 $z$ 的有界函数。*

**证明.** 对于每个固定的 $z$，这相当于一个确定性的问题，故可以直接应用文章 [39] 中的**定理 1**。注意这里我们假设 $c^\pm(z)$ 时恒正的并且有界的，所以可以得到界 $C_1(z)$ 和 $C_2(z)$。 □

接下来我们证明下述估计：

**定理 2.6.** *在假设2.1以及 $u_0(x,z)$ 是 $z$ 的有界变差函数，关于离散的 gPC-SG 方法的误差估计：*

$$\|U^n_{(K)} - u(\cdot,t^n,\cdot)\|_H \le C(T)(\sqrt{\Delta x + \Delta t} + \Delta x) + \frac{(b-a)C_\rho C(\ell,n)}{K^\ell}, \quad \forall l \in \mathbb{N}, \tag{2–51}$$

*$C(T)$ 只依赖于 $T$，$C(\ell,n)$ 依赖于 $\Delta t$ 和 $\ell$。*

**证明.** 首先我们将误差分成两部分：

$$\|U^n_{(K)} - u(\cdot, t^n, \cdot)\|_H \le \|U^n - u(x_i, t^n, z)\|_H + \|U^n_{(K)} - U^n\|_H. \tag{2–52}$$

第一部分是数值格式 (2–8) 的误差，根据**引理2.5**

$$\begin{aligned}\|U^n_{(K)} - u(\cdot, t^n, \cdot)\|_H &= \left( \int \|U^n(z) - u(\cdot, t^n, z)\|^2_{l^1(D)} \rho(z)\,\mathrm{d}z \right)^{1/2} \\ &\le \left( \int (C_1(z)\Gamma(c^-(z)) + C_2(z)\Gamma(c^+(z)))^2 \rho(z)\,\mathrm{d}z \right)^{1/2} \\ &\le \left( \int [(C_1(z)\Gamma(c^-(z))]^2 \rho(z)\,\mathrm{d}z \right)^{1/2} + \left( \int [(C_2(z)\Gamma(c^+(z))]^2 \rho(z)\,\mathrm{d}z \right)^{1/2}.\end{aligned} \tag{2–53}$$

最后一个不等式为闵可夫斯基不等式。注意到 $C_1(z)$ 有界和

$$\begin{aligned}\left( \int [\Gamma(c^\pm(z))]^2 \rho(z)\,\mathrm{d}z \right)^{1/2} &\le 2\left( \int c^\pm(z)\Delta x(1 - c^\pm(z)\frac{\Delta t}{\Delta x}) t_n \rho(z)\,dz \right)^{1/2} + \left( \int \Delta x^2 \rho(z)\,\mathrm{d}z \right)^{1/2} \\ &= 2\left( t_n \Delta x \int c^\pm(z)\rho(z)\,\mathrm{d}z - t_n \Delta t \int (c^\pm(z))^2 \rho(z)\,\mathrm{d}z \right)^{1/2} + \Delta x \\ &\le C(T)\sqrt{\Delta x + \Delta t} + \Delta x.\end{aligned} \tag{2–54}$$

因此得到

$$\|U^n - u(\cdot, t^n, \cdot)\|_H \le C(T)(\sqrt{\Delta x + \Delta t} + \Delta x). \tag{2–55}$$

对于第二部分，根据**定理2.4**我们有

$$\|U^n - U^n_{(K)}\|_H \le \frac{(b-a)C_\rho C(\ell, n)}{K^\ell}, \quad \forall \ell \in \mathbb{N}. \tag{2–56}$$

最后我们将这两部分相加即得结论，证毕。 □

## 2.2 用于带随机势函数的刘维尔方程的离散 gPC-SG 方法

在本节中，我们研究具有随机不确定性的经典力学中的 Liouville 方程：

$$u_t + v u_x - V_x u_v = 0, \quad t > 0, \quad x, v \in \mathbb{R}, \tag{2–57}$$

初值为

$$u(x, v, 0, z) = u_0(x, v, z), \tag{2–58}$$

其中 $u(x, v, t, z)$ 是在位置 $x$，速度为 $v$，时刻 $t$ 的经典粒子的密度分布函数。$V(x, z)$ 是依赖于随机变量 $z$ 的势函数。

刘维尔方程具有由牛顿第二定律定义的双特征线：

$$\frac{dx}{dt} = v, \quad \frac{dv}{dt} = -V_x(x, z), \tag{2–59}$$

是一个带有随机哈密顿量的哈密顿系统。

$$H = \frac{1}{2}v^2 + V(x,z). \tag{2–60}$$

如果 $V(x,z)$ 关于 $x$ 是间断的（对应于随机势垒的情况），则由 (2–59) 给出的 Liouville 方程的特征速度在间断处是无穷大，常规数值方法会遇到困难。另一方面，从经典力学可知，哈密尔顿算子在势垒上保持守恒。基于这个原理，Jin 和 Wen 提出了一个框架，被称为哈密顿守恒格式（Hamiltonian preserving scheme），其中他们根据粒子跨越势垒时的物理行为建立界面条件并加入到格式中，见 [2], [59]。

如同前面一节，我们首先用哈密顿守恒格式离散方程 (2–57)，这时我们把随机变量 $z$ 视为固定的参数。不失一般性，我们在 $x$ 和 $v$ 方向使用均匀网格，$x_{i+1/2}, i = 0, \ldots, N$ 和 $v_{j+1/2}, j = 0, \ldots, M$。网格中心在 $(x_i, v_j), i = 1, \ldots, N, j = 1, \ldots, M$，其中 $x_i = (x_{i+1/2}+x_{i-1/2})/2, v_j = (v_{j+1/2}+v_{j-1/2})/2$。网格大小记为 $\Delta x = x_{i+1/2} - x_{i-1/2}$ 和 $\Delta v = v_{i+1/2} - v_{i-1/2}$。此外，我们假设 $V$ 的间断点落在网格点上。令在 $x_{i+1/2}$ 处的 $V$ 的左和右极限分别为 $V^+_{i+1/2}$ and $V^-_{i+1/2}$。那么格式可以写成：

$$\partial_t u_{ij}(z) + v_j \frac{u^-_{i+1/2,j}(z) - u^+_{i-1/2,j}(z)}{\Delta x} - \mathrm{D}V_i(z) \frac{u_{i,j+1/2}(z) - u_{i,j-1/2}(z)}{\Delta v} = 0, \tag{2–61}$$

这里

$$\mathrm{D}V_i(z) := \frac{V^-_{i+1/2}(z) - V^+_{i-1/2}(z)}{\Delta x}. \tag{2–62}$$

我们还要确定网格边界上的数值通量 $u_{i,j+1/2}(z)$ 和 $u^\pm_{i+1/2,j}(z)$。

### 2.2.1　一阶空间离散格式

在这里，我们可以对通量 $u^\pm_{i+1/2,j}(z)$ 使用标准的一阶迎风格式，因为在这个方向上的波速度是 $v_j$，随机变量 $z$ 无关。因此，特性线实际上是确定性的。例如，我们考虑 $v_j > 0$，根据哈密顿守恒格式，

$$\begin{aligned} u^-_{i+1/2,j}(z) &= u_{ij}(z), \\ u^+_{i+1/2,j}(z) &= \begin{cases} c_1 u_{i+1,k}(z) + c_2 u_{i+1,k+1}(z), & \text{折射}, \\ u_{i+1,k}(z), & \text{反射}, \end{cases} \end{aligned} \tag{2–63}$$

$k$ 速度 $v$ 满足如下跨界面能量守恒条件 (见 [2]) 的指标，

$$\frac{1}{2}(v_j)^2 + V^-_{i+1/2} = \frac{1}{2}(v^+)^2 + V^+_{i+1/2}, \tag{2–64}$$

其中 $v^+$ 是粒子穿过势垒后的速度。如果 $(v_j)^2 + 2(V^-_{j+1/2} - V^+_{j+1/2}) > 0$，粒子会穿过势垒（折射），

$$v^+ = \sqrt{(v_j)^2 + 2(V^-_{j+1/2} - V^+_{j+1/2})}, \tag{2–65}$$

$k$ 满足

$$v_k \le v^+ < v_{k+1}, \tag{2–66}$$

$c_1$ 盒 $c_2$ 是线性插值的系数，

$$c_1 = \frac{v^+ - v_k}{\Delta v}, \quad c_2 = \frac{v_{k+1} - v^+}{\Delta v}, \quad c_1 + c_2 = 1. \tag{2–67}$$

如果 $(v_j)^2+2(V^-_{j+1/2}-V^+_{j+1/2})<0$，粒子会被势垒反射，这时 $k$ 满足

$$v_k=-v_j. \tag{2–68}$$

当 $v_j<0$ 时，同理我们可以由 (2–64) 确定 $c_1$,, $c_2$ 和 $k$。

对于通量 $u_{i,j+1/2}(z)$，在处理它时应该小心。与 $x$ 方向的通量不同，$v$ 方向的波速 $\mathrm{D}V_i(z)$ 依赖于随机变量 $z$，从而使得特性先是随机的。这将使得物理空间中的解的间断性传播到随机空间中，如果我们使用依赖特征线信息的格式（如迎风格式），这将导致最终的离散解在随机空间具有比较差的正则性。

这里我们用 Lax-Friedrichs 数值通量，它是不依赖于特征线信息的，

$$u_{i,j+1/2}(z)=\frac{1}{2}\left[\frac{\alpha}{DV_i(z)}\big(u_{i,j+1}(z)-u_{ij}(z)\big)-\big(u_{ij}(z)+u_{i,j+1}(z)\big)\right], \tag{2–69}$$

$\alpha$ 为常数满足 $\alpha\geq\max\limits_{i,z}|DV_i(z)|$。

从上面的讨论中，可以容易地看到 $x$ 方向和 $v$ 方向的通量都是关于 $z$ 的光滑函数。如2.1节所述，我们可以得出结论，该离散格式的解 $u_{ij}(z)$，对于每个固定的 $i,j$ 是 $z$ 的光滑函数。然后，我们将标准 gPC-SG 方法应用于该离散系统，与2.1节相同，当网格尺寸 $\Delta x$ 和时间步长 $\Delta t$ 固定时，可以期望该方法快速收敛到离散的解，其理由与2.1节中的相同。

**注 3.** 这里可以使用任何中心格式，例如局部 Lax-Friedrichs 格式等等。

### 2.2.2 二阶空间离散格式

在前两节中，我们提出了一个具有一阶空间离散的离散 gPC 格式。接下来，我们将给出空间上为二阶的离散格式。具体来说，空间数值通量根据哈密顿守恒格式 [2]，由下式给出（考虑 $v_j>0$ 时的情况）

$$\begin{aligned}
&u^-_{i+1/2,j}(z)=u_{ij}(z)+\frac{\Delta x}{2}s_{ij}(z),\\
&u^+_{i+1/2,j}(z)=\begin{cases}c_1\left(u_{i,k}(z)+\dfrac{\Delta x}{2}s_{i,k}(z)\right)+c_2\left(u_{i,k+1}(z)+\dfrac{\Delta x}{2}s_{i,k+1}(z)\right),\\ u_{i+1,k}(z)-\dfrac{\Delta x}{2}s_{i+1,k}(z),\end{cases}
\end{aligned} \tag{2–70}$$

$s_{ij}$ 为斜率限制器，$c_1,c_2,k$ 同样由 (2–63)–(2–68) 根据哈密顿守恒格式给出，正如一阶的情况。

由于解包含不连续性，因此二阶方案将必然引入数值振荡。为了抑制这些振荡，根据总变差减小的思想（total-variation-diminishing, TVD[52]，可以使用斜率限制器。然而，大多数用于捕捉激波斜率限制器都是非常不光滑的，而这在我们的格式中是不能容忍的。为此，我们使用了被称为 BAP 的光滑的斜率限制器，最早在文章 [55] 中被引入。分别定义在点 $(x_i,v_j)$ 向后和向前的查分为，

$$\begin{aligned}
s_l(z)&=(u_{ij}(z)-u_{i-1,j}(z))/\Delta x,\\
s_r(z)&=(u_{i+1,j}(z)-u_{ij}(z))/\Delta x,
\end{aligned} \tag{2–71}$$

那么 BAP 斜率限制器为

$$s_{ij}(z)=\mathcal{B}^{-1}\left(\frac{\mathcal{B}(s_l(z))+\mathcal{B}(s_r(z))}{2}\right). \tag{2–72}$$

一些光滑的 $\mathcal{B}(x)$ 的例子包括

$$\begin{aligned} &\mathcal{B}(x)=\arctan(x), && \mathcal{B}^{-1}(x)=\tan(x),\\ &\mathcal{B}(x)=\tanh(x), && \mathcal{B}^{-1}(x)=\tanh^{-1}(x),\\ &\mathcal{B}(x)=\frac{x}{\sqrt{1+x^2}}, && \mathcal{B}^{-1}(x)=\frac{x}{\sqrt{1-x^2}}. \end{aligned} \tag{2–73}$$

#### 2.2.2.1　全离散格式

接下来，我们需要在 $v$ 方向中定义数值通量。到这个阶段，为了得到一个光滑滑的离散解（关于 $z$），我们还需要选择一些不依赖于特征信息的通量格式，这里我们使用 Lax-Wendroff 格式：

$$u_{i,j+1/2}(z)=\frac{1}{2}(u_{i,j+1}(z)+u_{i,j}(z))+(\mathrm{D}V_i(z))\frac{\Delta t}{2\Delta v}(u_{i,j+1}(z)-u_{ij}(z)). \tag{2–74}$$

结合2–61)，(2–70) 和 (2–74)，我们得到在相空间二阶，并且对 $z$ 光滑的格式

$$\partial_t u_{ij}(z)=\mathrm{RHS}(z). \tag{2–75}$$

我们现在对这个离散系统使用 gPC-SG 方法，对于 gPC 展开中的第 $k$ 项 $u_{ij}^{n,k}$，我们有：

$$\partial_t u_{ij}^k(z)=\langle \mathrm{RHS}(z),P_k(z)\rangle, \tag{2–76}$$

$P_k(z)$ 为 $k$ 阶正交多项式，$\langle\cdot\rangle$ 为随机空间的内积。由于 $\mathrm{RHS}(z)$ 的非线性和其复杂的形式，我们将使用数值积分（即高斯积分）来计算 (2–76) 的右端。

$$\langle \mathrm{RHS}(z),P_k(z)\rangle=\sum_{m=0}^{M}\mathrm{RHS}(z_m)P_k(z_m)w_m, \tag{2–77}$$

$M$ 为高斯积分的点数，$z_i$, $w_i$ 为高斯积分点及对应的权重。这里我们将整个算法在每个时间步总结如下，

- 首先，使用 gPC 展开 $u_{ij}^n(z)=\sum_{k=0}^{K}u_{i,j}^{n,k}P_k(z)$ 来计算 $u_{ij}^n(z_m)$。注意，只需要计算 $P_k(z_m)$，而这是独立于时间的，因此可以预先计算好。
- 对于每个 $i,j,m$，用 $u_{ij}^n(z_m)$ 和 (2–70) 及 (2–74) 得到 $\mathrm{RHS}(z_m)$
- 最后对于每个 $i,j,k$，根据 (2–77)，使用向前欧拉或者龙格库塔向前迭代一个时间步长。

**注　4.** 对于对流方程 (2–5)，可以简单地在2–8) 中将 $U_i^n(z)$ 替换为 $U_i^n(z)+s_i(z)\Delta x/2$，并按照上述过程得到二阶格式。

## 2.3　数值例子

在本节中，我们将进行一些数值实验来展示我们提出的方法的性能并检查其数值精度。

### 2.3.1 例一：带有随机间断系数的标量对流方程

我们考虑初值问题

$$\begin{cases} u_t + \big[(c(x,z)u\big]_x = 0, & t>0, x\in\mathbb{R}, \\ u(x,0) = u_0(x), & x\in\mathbb{R}, \end{cases} \tag{2–78}$$

其中

$$c(x,z) = 0.3z + \begin{cases} c^- > 0, & \text{如果 } x<0, \\ c^+ > 0, & \text{如果 } x>0, \end{cases} \tag{2–79}$$

其中 $z$ 是 $[-1,1]$ 上的均匀分布（因此 gPC 展开的基应该是归一化后的勒让德多项式），我们将随机变量 $z$ 视为一个微小扰动，使得 $(c^{\pm}+0.3z)>0$ 在 $z\in[-1,1]$ 上成立。

在这个例子中，我们取初值为

$$u_0(x) = \cos(0.25\pi x), \quad \text{在 } [-1,3], \tag{2–80}$$

界面位于 $x=0$，界面条件为：

$$u(0^+,t,z) = \rho(z)u(0^-,t,z), \tag{2–81}$$

其中

$$\rho(z) = \frac{c^- + 0.3z}{c^+ + 0.3z}, \tag{2–82}$$

来使得通量守恒。

这个简单模型问题的显式解可以很容易地通过使用包含界面条件的特征方法来获得 [2]：

$$u(x,t,z) = \begin{cases} u_0\big(x-(c^+ + 0.3z)t\big), & x > (c^+ + 0.3z)t, \\ \rho(z)u_0\big(\rho(z)[x-(c^+ + 0.3z)t]\big), & 0 < x < (c^+ + 0.3z)t, \\ u_0\big(x-(c^- + 0.3z)t\big), & x<0. \end{cases} \tag{2–83}$$

在下面的例子中，我们设 $c^-=1$，$c^+=2$，最终时刻 $T=1$。可以由 (2–3) 和 (2–4) 得到显式解的期望与方差。

对于数值解，我们通过下式计算期望

$$\mathbb{E}_i(t^n) := \mathbb{E}[u(x_i,t^n,z)] = \int u(x_i,t^n,z)\rho(z)\,\mathrm{d}z = \hat{U}^n_{i,(0)},$$

和方差

$$\mathbb{V}_i(t^n) := \mathbb{E}\big[(u-\mathbb{E}(u))^2\big] = \sum_{k=1}^{K}\big(\hat{U}^n_{i,(k)}\big)^2.$$

显式解和数值解之间的误差用 $\ell^1$ 范数来描述。

图2–1显示当 $x=2$ 时，显式解 (2–83) 在 $z=0$ 处有间断。图2–2画出了显式解的期望值和方差。在这种情况下，可以预期标准 gPC-SG 方法的收敛很慢。

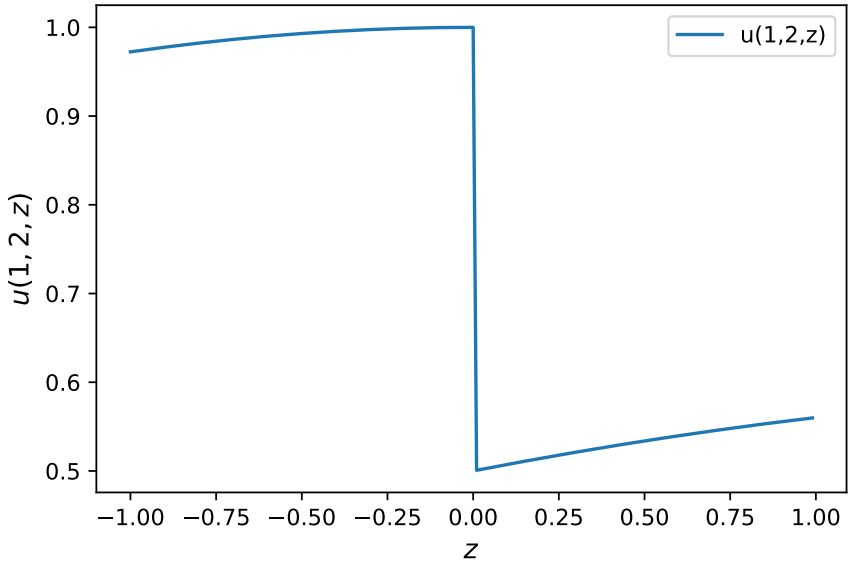


图 2–1 例一：显式解 (2–83) 在 $t = 1$ 和 $x = 2$ 处关于 $z$ 间断。

Fig 2–1 Example 1. The analytic solution (2–83) at $t = 1$ and $x = 2$ is a discontinuous function of $z$.

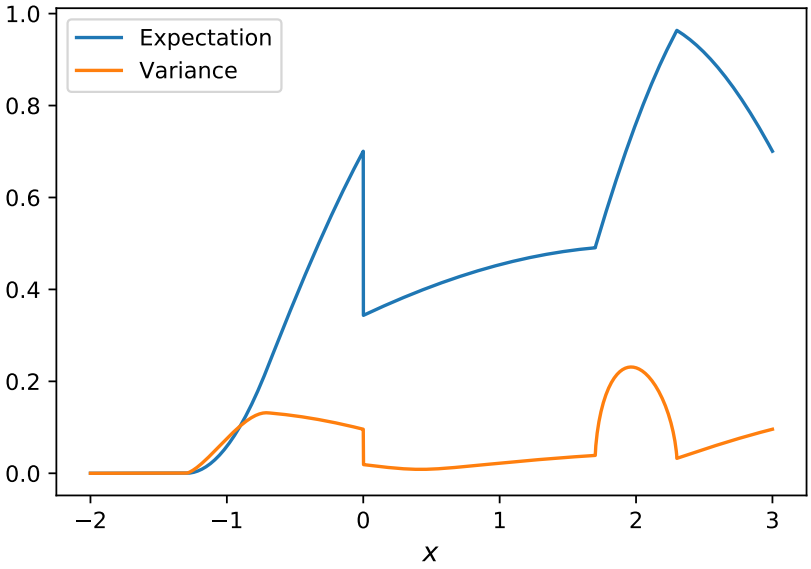


图 2–2 例一：显式解的期望与方差。

Fig 2–2 Example 1. The expectation and variance of the analytic solution.

#### 2.3.1.1 一阶差分逼近

在本小节中，我们将给出离散 gPC-SG 方法的数值结果。图2–3显示使用 $\Delta x = 0.001$，$\Delta t = \frac{1}{4}\Delta x$，gPC-SG 阶数 $K = 20$ 的数值解与显式解的期望与方差的比较。方差的误差是由一阶空间离散的低分辨率导致的。后面我们将使用二阶离散格式对其进行改进。

接下来，我们对 gPC-SG 进行收敛性测试。我们在所有计算中固定 $\Delta x = 0.005$ 和 $\Delta t = \frac{1}{5}\Delta x$。图2–4显示 $\ell^1$ 误差相对于 gPC 阶数 $K$ 衰减得非常快。当 $K = 4$ 时，基本衰减到有限差分法的误差。

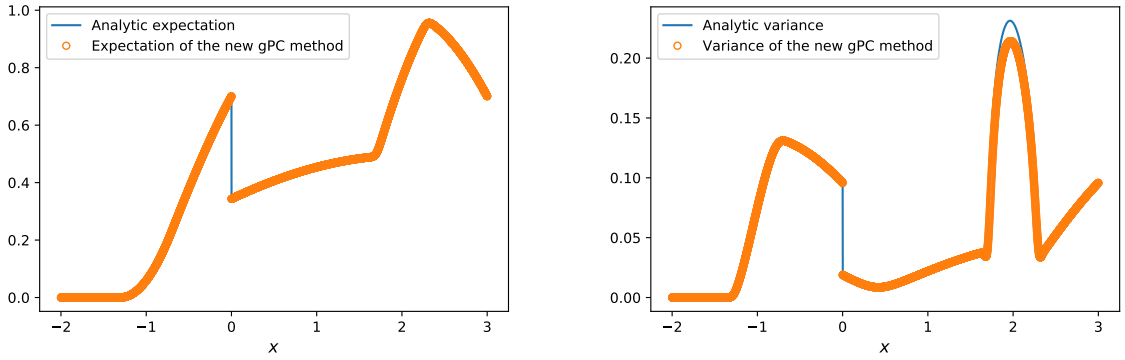


图 2–3 例一：一阶格式的数值解与显示解比较，$\Delta x = 0.001$，$\Delta t = \frac{1}{4}\Delta x$，gPC 阶数 $K = 20$。

Fig 2–3 Example 1. The analytic solution compared with the new gPC-SG method using first order finite difference approximation with $\Delta x = 0.001$, $\Delta t = \frac{1}{4}\Delta x$, gPC order $K = 20$.

然而，在图2–4中，由于有限差分的误差占主导地位（远大于 gPC-SG 方法的误差），这样很难验证 gPC 方法的收敛速度。为了验证 gPC-SG 的收敛性，我们固定 $\Delta x$ 和 $\Delta t$，用作为 $K = 30$ 时的解作为参考解，比较 $K$ 不同时与参考解的差别。我们测量每个 $K = 2, 3, \ldots, 20$ 和 $K = 30$ 之间的 $\ell^1$ 误差，结果如图2–5所示。通过对数图可以观察到 gPC-SG 方法是指数收敛的。

#### 2.3.1.2 二阶差分逼近

对于二阶格式，我们用与前面一阶格式相同的设置。图2–6画出了与显示解相比的期望值和方差，可以看到比一阶格式的结果要好，特别是方差在 $x = 2$ 附近的近似。

图2–7和图2–8显示了 gPC-SG 方法的收敛，从中可以观察到收敛是非常之快的。比较图2–7和图2–4，我们可以看到二阶格式有更好的总 $\ell^1$ 误差。但是图2–8中显示的 gPC 收敛的速率却不如一节格式。这点并不奇怪，因为我们的收敛速率取决于离散解的光滑性，而一阶格式给出的数值粘性较大，从而光滑性更好。二阶格式因为它更接近于精确解（其是非常不光滑的），因此与一阶格式相比离散解的光滑性相对比较差，而这正是影响收敛速率的主要因素，所以二阶格式的收敛速度相比一阶而言较慢。然而这并不意味着二阶格式不如一阶，因为必须考虑整体的误差，包括空间离散的误差。通过比较图2–6与图2–3，图2–7与图2–4，显然二阶格式要优于一阶格式。

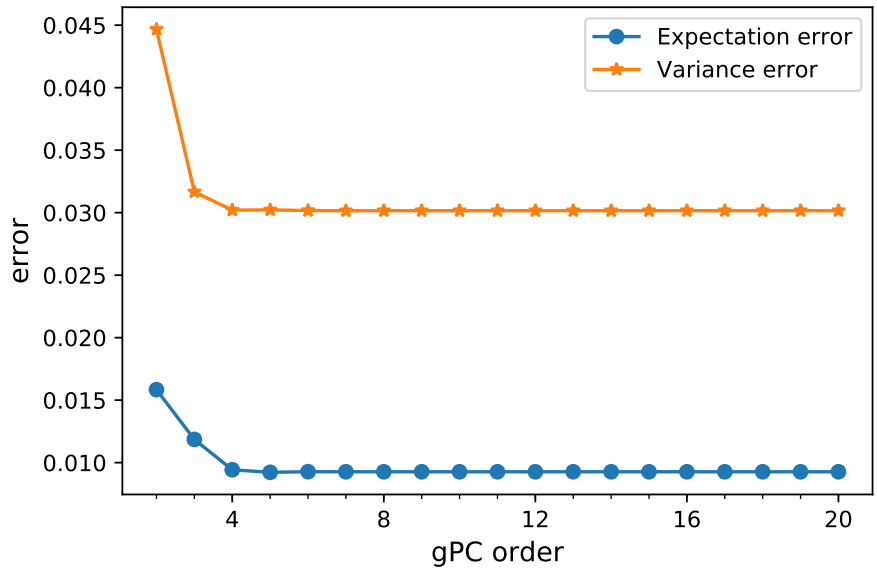


图 2–4 例一：一阶离散格式 $\ell^1$ 误差与 gPC 阶数的关系，$\Delta x = 0.005$, $\Delta t = \frac{1}{5}\Delta x$。

Fig 2–4 Example 1. The first order finite difference approximation with $\Delta x = 0.005$, $\Delta t = \frac{1}{5}\Delta x$: the $\ell^1$ error versus the gPC order.

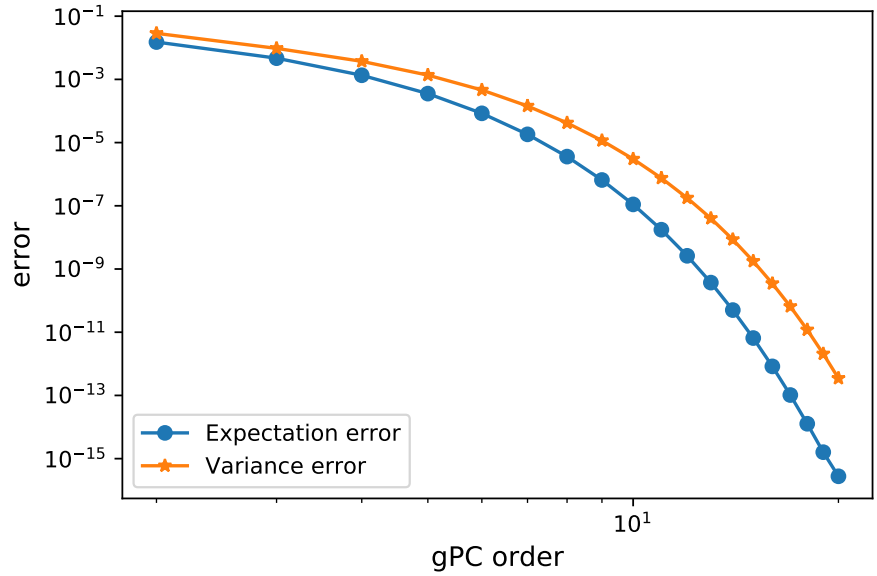


图 2–5 例一：一阶格式的 gPC-SG 方法误差与 gPC 阶数的对数关系，$\Delta x = 0.005$, $\Delta t = \frac{1}{5}\Delta x$。

Fig 2–5 Example 1. The first order finite difference approximation $\Delta x = 0.005$, $\Delta t = \frac{1}{5}\Delta x$: the gPC error versus the gPC order by a log-log plot (with other numerical parameters fixed).

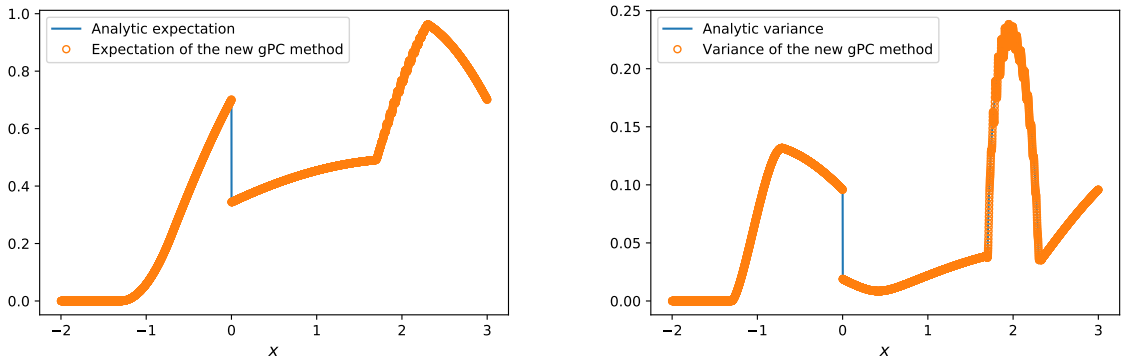


图 2–6 例一：显式解与二阶空间离散的 gPC-SG 方法比较，$\Delta x = 0.001$，$\Delta t = \frac{1}{4}\Delta x$，gPC-SG 的阶数为 $K = 20$。
Fig 2–6 Example 1. The analytic solution compared with the new gPC-SG method using the second order finite difference approximation with $\Delta x = 0.001$, $\Delta t = \frac{1}{4}\Delta x$, gPC order $K = 20$.

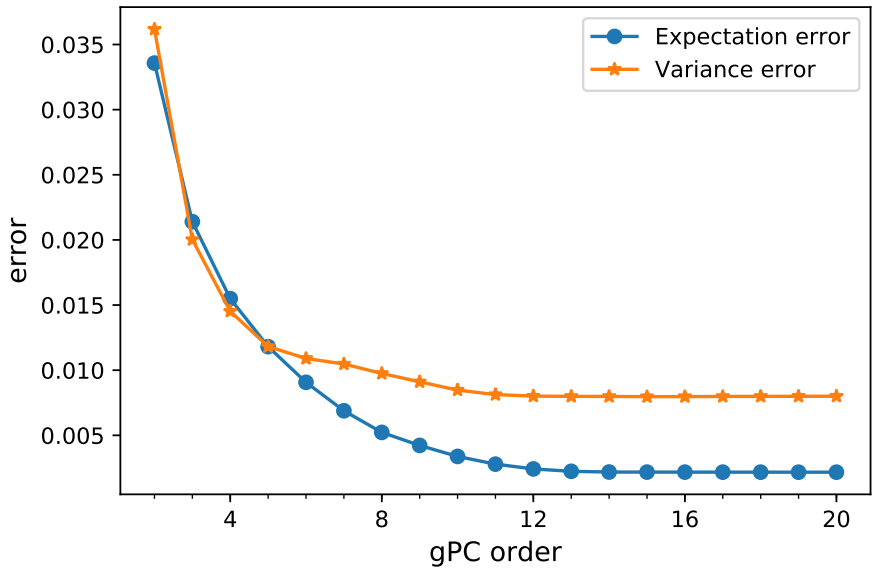


图 2–7 例一：二阶格式 $\ell^1$ 误差与 gPC 阶数的关系，$\Delta x = 0.005$，$\Delta t = \frac{1}{5}\Delta x$。
Fig 2–7 Example 1. The second order finite difference approximation $\Delta x = 0.005$, $\Delta t = \frac{1}{5}\Delta x$: the $\ell^1$ error versus the gPC order.

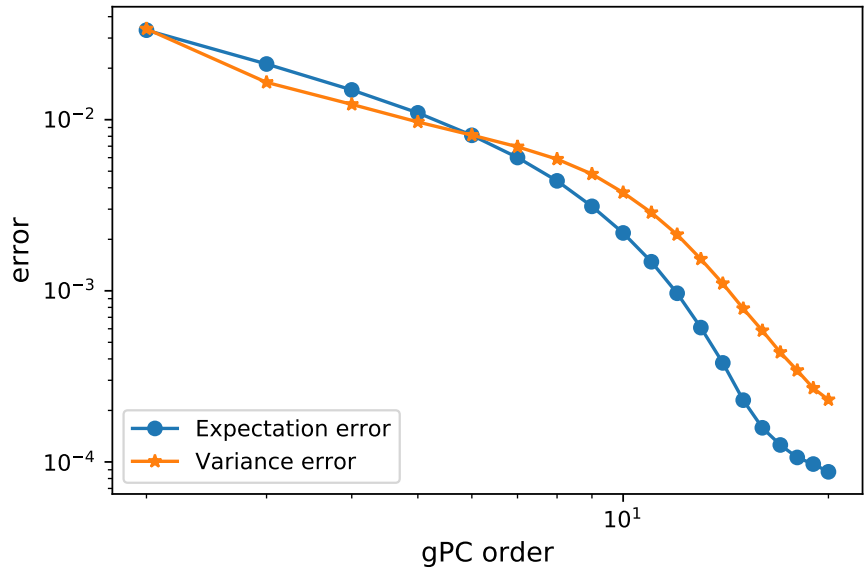


图 2–8 例一：二阶格式 $\ell^1$ 误差与 gPC 阶数的对数关系，$\Delta x = 0.005$，$\Delta t = \frac{1}{5}\Delta x$。

Fig 2–8 Example 1. The second order finite difference approximation $\Delta x = 0.005$, $\Delta t = \frac{1}{5}\Delta x$: the gPC error versus the gPC order by a log-log plot (with other numerical parameters fixed).

### 2.3.2　例二：带有随机间断势的刘维尔方程

再次写出刘维尔方程

$$u_t + vu_x - V_x u_v = 0, \quad t > 0, \quad x, v \in R, \tag{2–84}$$

其中随机势如下

$$V(x, y) = V_0(x) + 0.1xz, \tag{2–85}$$

$z$ 为 $(-1, 1)$ 上的均匀分布，

$$V_0(x) = \begin{cases} 0.2, & x < 0, \\ 0, & x > 0. \end{cases} \tag{2–86}$$

对于给定的初始数据，因为无法得到这个问题的显式精确解。所以，我们将使用**配点法**（stochastic collocation method）作为参考解。在配点法中，在对应的随机空间中取 $\{z_i\}_{1\le i\le M}$ 的离散点集合，称为抽样点，我们在只需在抽样点上解刘维尔方程 (2–2)。对于每个固定的 $z_i$，我们只需要使用用来解**确定性**的带间断势的刘维尔方程的哈密顿守恒格式。然后，可以通过 (2–3) 和 (2–4) 通过数值积分获得期望和方差。在以下示例中，我们选择 $\{z_i\}_{1\le i\le M}$ 作为 $M$ 阶勒让德多项式的根，并使用高斯积分来获得期望值和方差。

对于 gPC-SG 方法我们需要计算 $\int_{-1}^{1} V_0(z)P_j(z)P_k(z)\rho(z)\,\mathrm{d}z$ 也就是，

$$\int_{-1}^{1} V_0(z)P_j(z)P_k(z)\rho(z)\,\mathrm{d}z = \begin{cases} \dfrac{j+1}{\sqrt{(2j+1)(2j+3)}}, & k=j+1, \\ V_0'(x), & k=j, \\ \dfrac{j}{\sqrt{4j^2-1}}, & k=j-1. \end{cases} \tag{2–87}$$

这里我们得到一个对称的三对角矩阵。

为了说明由间断的势导致的解的奇异性，我们使用连续的初值：

$$u(x,v,0) = \begin{cases} \sin[2\pi(0.25-(x^2+v^2))], & x^2+v^2<0.25, \\ 0, & \text{其他}. \end{cases} \tag{2–88}$$

由配点法 ($M=20$ 个抽样点) 和我们的 $K=4$ 阶离散 gPC-SG 方法期望与方差如下图2–9所示。虽

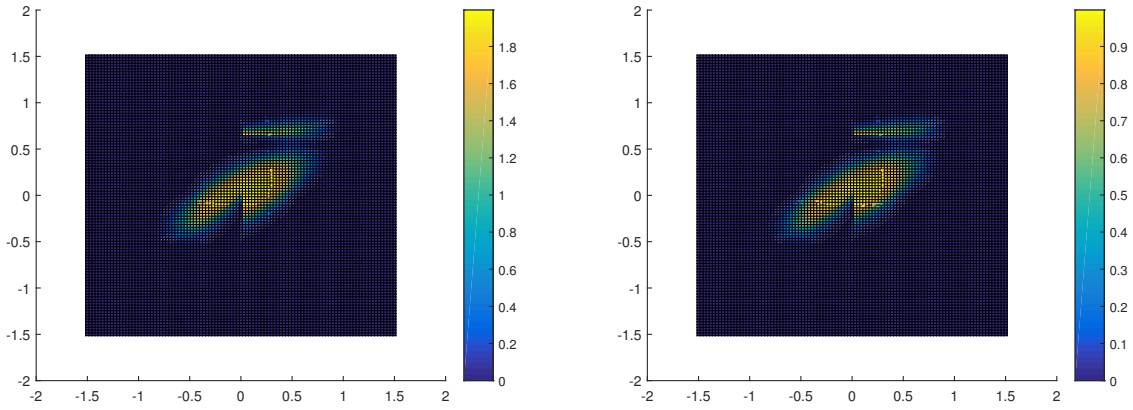


图 2–9 例二：初值为 (2–88)。由配点法（左）与离散 gPC-SG 方法（右）得到的期望。

Fig 2–9 Example 2 with initial data (2–88). Expectation of the solution. Left: the collocation method with 20 sample points. Right: the new gPC-SG method with gPC order $K=4$.

然初值是连续的，但由于界面条件的存在，解仍然是间断的。而这种奇异性会大大影响 gPC-SG 方法的收敛性。

#### 2.3.2.1 一阶差分逼近

在这个例子里，我们考虑初值

$$u(x,v,0) = \begin{cases} 1, & x\geq 0, v<0, x^2+v^2<1, \\ 1, & x\leq 0, v>0, x^2+v^2<1, \\ 0, & \text{其他}. \end{cases} \tag{2–89}$$

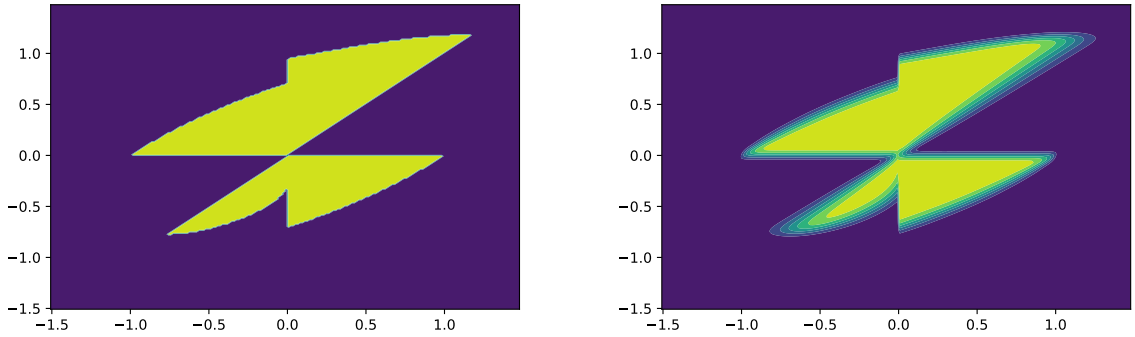


图 2–10 例二的确定性版本，初值 (2–89)。显式精确解（左），数值解（右），$\Delta x = \Delta v = 0.015$，$\Delta t = 0.001$。
Fig 2–10 The deterministic case of Example 2 with initial data (2–89). Left: analytic solution of the deterministic problem with $z = 0$ and $t = 1$. Right: numerical solution using the first order Hamiltonian preserving scheme with $\Delta x = \Delta v = 0.015$, $\Delta t = 0.001$.

注意，由于初始数据和势的不连续性从而导致解具有奇异性。该示例的确定性版本在 [2] 中被使用过，并且可以通过使用特征线的方法获得显式精确解。我们首先在画出对应于 [2] 中的确定性示例的图2–10，包括显示精确解和数值解（使用一阶通量），其中固定 $z = 0$。

然后我们与用配点法计算的解（$M = 20$ 样本点）（图2–11和2–12左）进行比较。图2–11和2–12右显示我们的新 gPC-SG 方法的解，其中 gPC 阶数 $K = 10$。这里的网格大小是 $\Delta x = \Delta v = 0.03$，时间步长是 $\Delta t = 0.002$。可以看到随机方程的解的期望和当 $z = 0$ 的确定性情况之间的差异，并且这种差异也可以容易地在方差图上看出。随机方程的解的期望更加光滑，因为它对 $z$ 变量进行了积分（相当于平均），从而得到了更好的正则性（参见 [60-61]）。关于计算的开销，我们的新的离散 gPC-SG 方法运行比配点法要快得多。配点法要花费大约 20 倍确定性版本的成本，因为我们使用了 20 个样本点；然而，我们的新的离散 gPC-SG，阶数 $K = 10$（即可得到优于配点法 20 个样本点的解），花费大约 10 倍确定性问题的成本。

在图2–13，我们画出离散 gPC-SG 方法的 $\ell^1$ 误差随着 gPC 阶数 $K$ 增加的变化关系，显示出了谱收敛结果。

#### 2.3.2.2　二阶差分格式

这里对于二阶差分格式，我们使用和前面一阶格式类似的参数设定。

首先，如同一阶的情形，我们在2–14中画出由二阶格式得到的 $z = 0$ 时确定性的方程的数值解。二阶格式明显给出了高分辨率的结果（相比一阶格式，见图2–10和图2–15）。但是由于在 $v$ 方向我们用的是 Lax-Wendroff 格式，所以这里有一些数值振荡。

接下来我们对比由配点法和离散 gPC-SG 方法计算的期望与方差，见图2–15和2–16： 可以看到二者的结果没有显著差异，在街的间断处都给出了比一阶格式更好的结果。这里对于计算的花费，我们的离散 gPC-SG 方法比配电法要稍快一些，因为在计算 $\langle \mathrm{RHS}(z) \rangle$ 时我们用了和配点法类似的

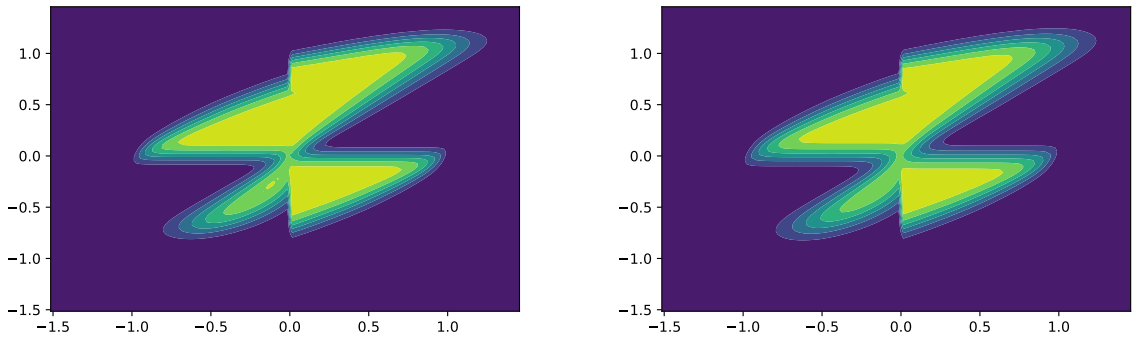


图 2–11 例二: 初值为 (2–89), 由一阶差分格式得到的期望, 配点法 (左) 与离散 gPC-SG 方法 (右), $\Delta x = \Delta v = 0.03$, $\Delta t = 0.002$。

Fig 2–11 Example 2 with initial data (2–89) by the first order finite difference approximation with $\Delta x = \Delta v = 0.03$ and $\Delta t = 0.002$. The expectation of the solution. Left: the collocation method with $M = 20$ samples points. Right: the new gPC-SG method using first order finite difference approximation.

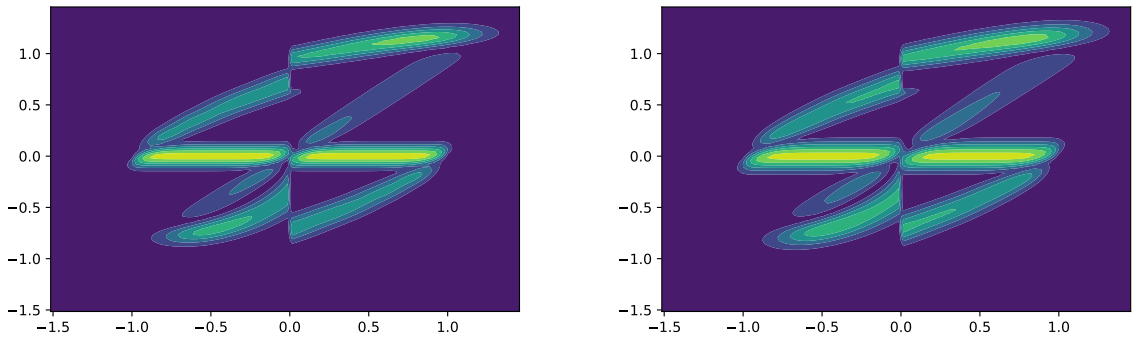


图 2–12 例二: 初值为 (2–89), 由一阶差分格式得到的方差, 配点法 (左) 与离散 gPC-SG 方法 (右), $\Delta x = \Delta v = 0.03$, $\Delta t = 0.002$。

Fig 2–12 Example 2 with initial data (2–89) by the first order finite difference approximation with $\Delta x = \Delta v = 0.03$ and $\Delta t = 0.002$. The variance of the solution. Left: the collocation method. Right: the new gPC-SG method.

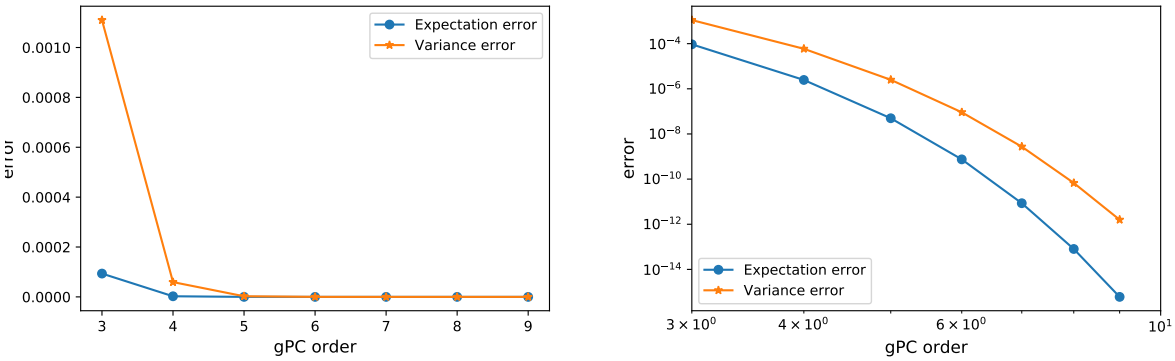


图 2–13 例二：初值 (2–89)，一阶离散 gPC-SG 方法的 $\ell^1$ 误差随着 gPC 阶数 $K$ 增加的变化关系（左）及对数图（右）。

Fig 2–13 Example 2 with initial data (2–89). Convergence of the new gPC-SG method using the first order finite difference approximation. Left: the $\ell^1$ error versus gPC order. Right: the gPC error versus the gPC order by a log-log plot (with other numerical parameters fixed).

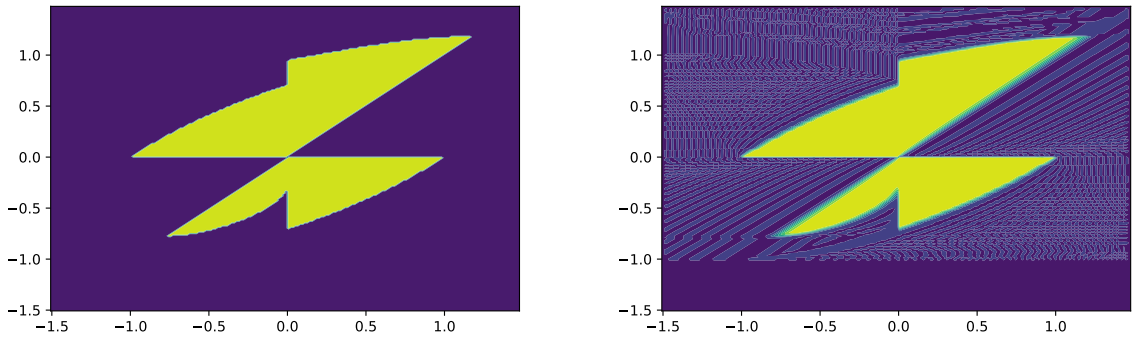


图 2–14 例二：初值为 (2–89) 的确定性版本。左图为显式精确解（$z = 0$，$t = 1$）；右图为二阶格式的数值解，$\Delta x = \Delta v = 0.015$，$\Delta t = 0.001$。

Fig 2–14 The deterministic case of Example 2 with initial data (2–89). Left: analytic solution of the deterministic problem with $z = 0$ and $t = 1$. Right: numerical solution using the second order Hamiltonian preserving scheme with $\Delta x = \Delta v = 0.015$, $\Delta t = 0.001$

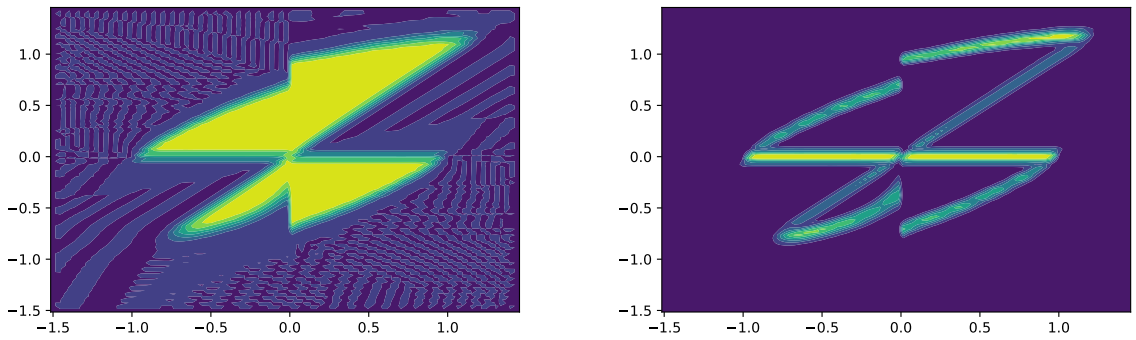


图 2–15 例二：初值为2–89)，二阶离散配点法的解，期望（左）和方差（右），$\Delta x = \Delta v = 0.03$，$\Delta t = 0.002$。
Fig 2–15 Example 2 with initial data (2–89) by the second order finite difference approximation with $\Delta x = \Delta v = 0.03$, $\Delta t = 0.002$. Reference solution by the collocation method with 20 sample points at $t = 1$. Left: expectation. Right: variance.

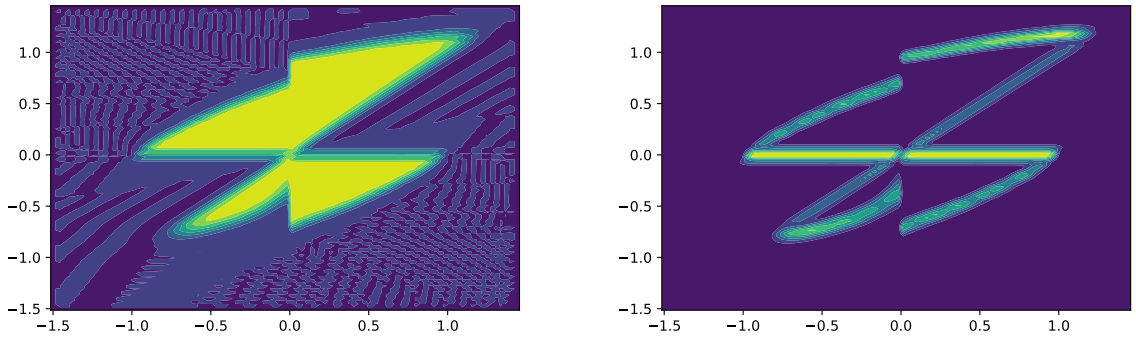


图 2–16 例二：初值为2–89)，二阶离散 gPC-SG 的结果，期望（左）和方差（右），$\Delta x = \Delta v = 0.03$，$\Delta t = 0.002$。
Fig 2–16 Example 2 with initial data (2–89) by the second order finite difference approximation with $\Delta x = \Delta v = 0.03$, $\Delta t = 0.002$. Solution at $t = 1$ computed by the new gPC-SG method. Left: expectation. Right: variance.

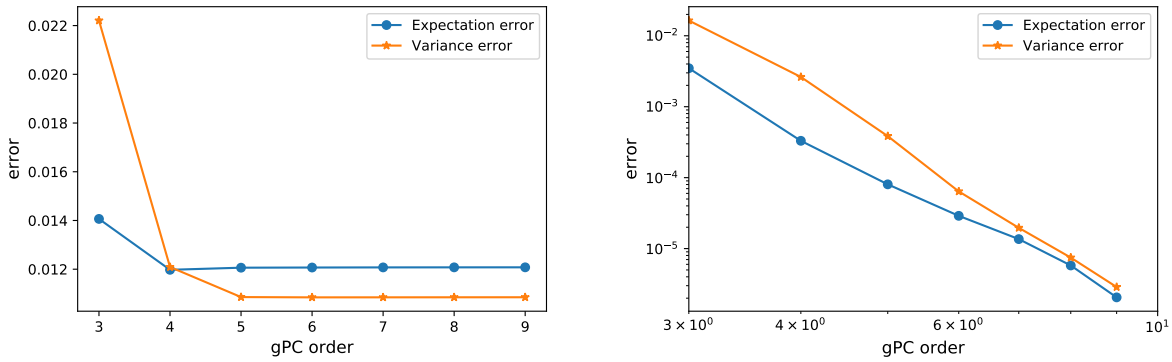


图 2–17 例二：初值 (2–89)，二阶离散 gPC-SG 方法的收敛性，$\ell^1$ 误差与 gPC 阶数的关系图（左）和相应的对数图（右）。

Fig 2–17 Example 2 with initial data (2–89). Convergence of the new gPC-SG method using second order finite difference approximation. Left: the $\ell^1$ error versus the gPC order. Right: the gPC error versus the gPC order by a log-log plot (with other numerical parameters fixed).

思想。

最后，我们来测试离散 gPC-SG 方法的收敛速率。首先固定网格参数：$\Delta x = \Delta v = 0.03$，$\Delta t = 0.02$，$t = 1$。用 20 个点的高斯积分来计算内积 (2–76) 并选择 gPC 阶数 $K = 10$ 为参考解来看 $K$ 从 3 增加到 10 时误差的变化。从图2–17可以看出谱收敛的结果。

## 2.4　本章总结与展望

在本章中，我们研究了带有间断与随机系数的双曲型方程的数值解法。为了克服由解的奇异性导致的 gPC-SG 方法收敛速度很慢的问题，我们提出了离散 gPC-SG 方法，利用离散的解具有较好的正则性，进而改进 gPC-SG 方法的收敛速度。对于对流方程我们进行了收敛性分析与误差估计。同时为了说明方法的有效性，我们对于对流方程与刘维尔方程进行了大量的数值实验。

对于不光滑的、带有不确定性的问题，gPC-SG 方法的应用仍然有很多问题没有解决。上述离散 gPC-SG 方法在高阶格式构造、非线性问题上仍然存在一定困难，同时如何将该方法向更广泛的问题上进行推广仍然需要大量的研究，会作为未来的研究方向之一。

# 第三章　gPC-SG 方法在带有随机输入及多尺度的线性输运方程中的应用

我们考虑一维平板中的线性输运方程：

$$\varepsilon\partial_t f + v\partial_x f = \frac{\sigma}{\varepsilon}\mathcal{L}f - \varepsilon\sigma^a f + \varepsilon S, \qquad \sigma(x,z) \geq \sigma_{\min} > 0, \tag{3–1}$$

$$\mathcal{L}f(t,x,v,z) = \frac{1}{2}\int_{-1}^{1} f(t,x,v',z)\,\mathrm{d}v' - f(t,x,v,z). \tag{3–2}$$

这种方程通常用来描述粒子（如中子）在背景介质中（如一些原子核）转移、传播的过程，例如中子输运方程、辐射输运方程等等。其中 $f(t,x,v,z)$ 表示粒子在 $t \geq 0$ 时刻，位置 $x \in (0,1)$，速度 $v = \Omega\cdot e_x = \cos\theta \in [-1,1]$ 的密度分布函数，这里 $\theta$ 是粒子前进方向与 $x$ 轴的夹角。$\sigma(x,z)$，$\sigma^a(x,z)$ 分别代表总的和吸收部分的碰撞横截面；$S(x,z)$ 代表源项；$\varepsilon$ 是无量纲化后的克努森数（Knudsen number），表示粒子平均自由程与系统的特征长度的比值（这里是问题的区间长度）。我们考虑如下的狄利克雷边界条件（只需要对于流入区域的粒子）：

$$\begin{aligned} f(t,0,v,z) &= f_{\mathrm{L}}(t,v,z), \qquad \text{for } v \geq 0\,, \\ f(t,1,v,z) &= f_{\mathrm{R}}(t,v,z), \qquad \text{for } v \leq 0\,, \end{aligned} \tag{3–3}$$

初值条件：

$$f|_{t=0}(x,v,z) = f_0(x,v,z). \tag{3–4}$$

这里我们特别感兴趣的问题是如果方程中的参量像碰撞横截面、源项，甚至是初值或者边界条件有一定的**不确定性**（*Uncertainty*），这种不确定性最终会随方程如何演化？其中的不确定性我们用概率密度函数为 $\omega(z)$ 的随机变量 $z \in \mathbb{R}^d$ 来刻画。所以在我们的问题中 $f$，$\sigma$、$\sigma^a$ 和 $S$ 都是依赖于 $z$ 的。

近几年来，对于带有不确定性的偏微分方程问题和一些相应的工程问题的研究逐渐引起了许多数学家和工程研究人员的重视，提出了很多新的数值算法。正如同我们在前面几章提到的，我们主要感兴趣的是研究基于所谓的多项式混沌（Polynomial Chaos，简称 PC，最早由 Wiener 引入 [25]）的随机伽辽金方法（stochastic Galerkin）。这种方法已经在许多应用中证明了其高效性，参照 [23,62-63]，并且已经被用于研究带随机系数的线性输运方程 [64]。显然这种问题同时包含了不确定性（随机性）和多尺度两种困难，这里的多尺度是由克努森数（Kundsen number）$\varepsilon$ 刻画的。在所谓的光学薄的区域（$\varepsilon \ll 1$），由于粒子具有很高的散射率，线性输运方程会趋于一个扩散方程，这就是大家熟知的扩散极限 [65-67]。在过去的几十年间，相当多的工作都在研究如何发展（确定性的）线性输运方程在扩散的尺度上渐进保持（asymptotic-preserving（AP））格式，例如 [68-75]。而直到最近，对于既含有不确定性同时由在扩散尺度下的线性输运方程的研究才被引入，见 [41]（在随机伽辽金（stochastic Galerkin）的框架下，被称作 s-AP 格式），更多的关于这方面的工作可以参考 [76-78]。我们称一个格式是 s-AP 如果随着 $\varepsilon \to 0$，线性输运方程的随机伽辽金格式变成对应的扩散极限方程的随机伽辽金格式，最早见于文章 [41]，其作者意识到对于确定性问题的 AP 框架可以稍加改动来

研究带有随机系数的线性输运方程。更进一步，在文献 [76-77] 中提到，kinetic 方程（包括线性和非线性）随着时间演化可以保持初值在随机空间的正则性，使得我们很自然的得到随机伽辽金方法的谱收敛结果。

但是在前面的工作中，当 $\varepsilon \ll 1$ 时，能量估计和由该估计得到的收敛速率依赖并且反比于 $\varepsilon$，这就意味着我们在随机伽辽金方法中需要的多项式阶数会随着 $\varepsilon$ 变小而急剧增长。事实上这是很多数值方法应用含有小参数或多尺度问题中常见的现象。AP 格式中，数值参数（如网格）不需要依赖于 $\varepsilon$（格式收敛关于 $\varepsilon$ 一致），而要严格的证明这点并不容易，仅仅在少数情况下可行，见 [70,79]。证明一致收敛的其中一种方法是通过相应的扩散极限，像文章 [70] 中对于确定性和稍后的文章 [80] 中对于带随机的输运方程的证明，也可以参考相关的综述文章 [75]。但是通常这种做法不能给出一个精确的收敛速率。

在本章中，对于含有随机变量的线性输运方程 (3–1) 的随机伽辽金方法，我们将给出一个精确的误差估计。这需要我们对于方程的解 $f$ 在随机空间的高阶导数给出精确的（不依赖于 $\varepsilon$，或者说关于 $\varepsilon$ 一致的）估计，同时还需要证明 $[f]-f$（其中 $[f]$ 表示 $f$ 的速度平均，见 (3–5)）在随机空间的高阶导数在 $\varepsilon \to 0$ 时一致有界。这样我们就可以不用借助于扩散极限得到关于 $\varepsilon$ 一致的谱收敛结果。

在文献 [41] 中的 s-AP 格式使用了基于对输运方程奇偶速度分解的 AP 框架 [72]。在本章中，我们将使用基于 micro-macro 分解的方法来构造一个全离散的 s-AP 格式（见 [74]）。这种方法的好处是使得我们可以证明一致的（关于 $\varepsilon$）稳定性，类似于相应的确定性问题 [81]。事实上，我们将指出关于 s-AP 格式的证明可以从 [81] 中的证明直接推广得到。

本章的组织结构如下，在3.1中我们对于线性输运方程的扩散极限做一个简单的总结。在3.2中，我们将给出基于广义多项式混沌（generalized polynomial chaos，gPC) 的随机伽辽金方法（gPC-SG）用于求解带有随机系数的线性输运方程，同时我们形式上的证明这个方法是 s-AP 的。在3.3中，我们将证明随机伽辽金方法会一致（关于 $\varepsilon$）的保持解在随机空间的正则性，进而证明该方法是关于 $\varepsilon$ 一致的谱收敛。接下来，在3.5中我们将给出一个基于 micro-macro 分解的全离散格式并在3.5中证明该格式具有一致的（关于 $\varepsilon$）稳定性。在3.6中我们会给出相应的数值实验结果。最后，在3.7中我们将对本章进行总结。

## 3.1 扩散极限

定义

$$[\phi] = \frac{1}{2}\int_{-1}^{1} \phi(v)\,\mathrm{d}v, \tag{3–5}$$

为一个速度依赖函数 $\phi$ 的平均。对于每个确定的随机变量 $z$，存在一个恒为正的函数 $\phi(v) > 0$，和所谓的绝对平衡态满足 $[\phi] = 1$，$[v\phi(v)] = 0$（根据 Perron-Frobenius 定理, cf.[67]）。

定义希尔伯特空间 $L^2\big((-1,1);\ \phi^{-1}\,\mathrm{d}v\big)$ 中的内积和范数

$$\langle f, g\rangle_\phi = \int_{-1}^{1} f(v)g(v)\phi^{-1}\,\mathrm{d}v, \qquad \|f\|_\phi^2 = \langle f, f\rangle_\phi. \tag{3–6}$$

其中线性算子 $\mathcal{L}$ 具有如下性质 [67]:

- $[\mathcal{L}f] = 0$，对于每个 $f \in L^2([-1,1])$;

- $\mathcal{L}$ 的零空间 $\mathcal{N}(\mathcal{L}) = \mathrm{Span}\,\{\,\phi \mid \phi = [\phi]\,\}$;
- $\mathcal{L}$ 的值域 $\mathcal{R}(\mathcal{L}) = \mathcal{N}(\mathcal{L})^{\perp} = \{\,f \mid [f] = 0\,\}$;
- $\mathcal{L}$ 是空间 $L^2((-1,1);\phi^{-1}\,\mathrm{d}v)$ 中的非负、自共轭算子，即存在一个正的常数 $s_m$ 使得

$$\langle f, \mathcal{L}f\rangle_{\phi} \le -2s_m \|f\|_{\phi}^2, \qquad \forall\, f \in \mathcal{N}(\mathcal{L})^{\perp}; \tag{3–7}$$

- $\mathcal{L}$ 存在一个伪逆（pseudo-inverse），记为 $\mathcal{L}^{-1}$，为从 $\mathcal{R}(\mathcal{L})$ 到 $\mathcal{R}(\mathcal{L})$ 的映射。

设 $\rho = [f]$。对于每个固定的 $z$，经典的线性输运方程的扩散极限理论 [65-67] 告诉我们，当 $\varepsilon \to 0$ 时，$\rho$ 收敛到如下的随机扩散方程：

$$\partial_t \rho = \partial_x(\kappa(z)\partial_x \rho) - \sigma^a(z)\rho + S, \tag{3–8}$$

其中扩散系数为

$$\kappa(z) = \frac{1}{3}\sigma(z)^{-1}\,. \tag{3–9}$$

Micro-macro 是研究玻尔兹曼方程及其流体力学极限 [82] 和构造相应的渐近保持格式的有力工具 [74,83-84]，它具有如下形式

$$f(t,x,v,z) = \rho(t,x,z) + \varepsilon g(t,x,v,z) \tag{3–10}$$

其中 $[g] = 0$。将 (3–10) 带入 (3–1)，我们得到相应的 micro-macro 形式：

$$\partial_t \rho + \partial_x\,[vg] = -\sigma^a \rho + S, \tag{3–11a}$$

$$\partial_t g + \frac{1}{\varepsilon}(I - [.])(v\partial_x g) = -\frac{\sigma(z)}{\varepsilon^2} g - \sigma^a g - \frac{1}{\varepsilon^2} v\partial_x \rho. \tag{3–11b}$$

现在很容易得到其扩散极限 (3–8)。当 $\varepsilon \to 0$，(3–11b) 给出

$$g = -\frac{v}{\sigma(z)}\partial_x \rho$$

带入到 (3–11a) 中，即得到扩散方程 (3–8)–(3–9)。

## 3.2　基于推广多项式混沌的随机伽辽金方法（gPC-SG）在输运方程中的应用

假设在希尔伯特空间 $H(\mathbb{R}^d;\omega(z)\,\mathrm{d}z)$ 中权为 $\omega(z)$ 的一组完备的正交多项式的基 $\{\phi_i(z), i = 0,1,\cdots,\}$，其中 $\phi_i(z)$ 是度为 $i$ 的多项式且满足

$$\langle \phi_i, \phi_j\rangle_{\omega} = \int \phi_i(z)\phi_j(z)\omega(z)\,\mathrm{d}z = \delta_{ij}.$$

这里 $\phi_0(z) = 1$，$\delta_{ij}$ 是克罗内克 $\delta$（Kronecker delta）函数。在这个空间中定义的内积和范数分别为，

$$\langle f, g\rangle_{\omega} = \int_{\mathbb{R}^d} f g\,\omega(z)\,\mathrm{d}z, \quad \|f\|_{\omega}^2 = \langle f, f\rangle_{\omega}\,. \tag{3–12}$$

因为 $f(t,\cdot,\cdot,\cdot)$ 定义在 $L^2\big((0,1)\times(-1,1)\times\mathbb{R}^d;\omega(z)\,\mathrm{d}x\,\mathrm{d}v\,\mathrm{d}z\big)$ 上，我们有广义多项式混沌展开（generalized Polynomial Chaos expansion，简称 gPC 展开）

$$f(t,x,v,z) = \sum_{i=0}^{\infty} f_i(t,x,v)\,\phi_i(z), \quad \hat{f} = \big(\,f_i\,\big)_{i=0}^{\infty} := \big(\bar{f}, \hat{f}_1\big).$$

$f$ 的均值和方差可以由展开的系数算出，as

$$\bar{f} = E(f) = \int_{\mathbb{R}} f\omega(z)\,\mathrm{d}z = f_0, \quad \text{var}\,(f) = |\hat{f}_1|^2\,.$$

随机伽辽金方法（stochastic Galerkin (SG) approximation）[23,62] 的思想是在有限项截断上述无穷级数

$$f_M = \sum_{i=0}^{M} f_i\,\phi_i, \quad \hat{f}^M = \big(f_i\big)_{i=0}^{M} := \big(\bar{f}, \hat{f}_1^M\big), \tag{3–13}$$

$f_M$ 的均值和方差同样可由展开系数算出

$$E(f_M) = \bar{f}, \quad \text{var}\,(f_M) = |\hat{f}_1^M|^2 \le \text{var}\,(f)\,.$$

更进一步，我们定义

$$\sigma_{ij} = \big\langle \phi_i,\, \sigma\phi_j \big\rangle_\omega, \quad \Sigma = \big(\sigma_{ij}\big)_{M+1,M+1},$$
$$\sigma_{ij}^a = \big\langle \phi_i,\, \sigma^a\phi_j \big\rangle_\omega, \quad \Sigma^a = \big(\sigma_{ij}^a\big)_{M+1,M+1},$$

其中 $0 \le i,j \le M$， Id 是 $(M+1)\times(M+1)$ 的单位矩阵，$\Sigma, \Sigma^a$ 为正定对称矩阵满足 [63]

$$\Sigma \ge \sigma_{\min}\,\text{Id}\,.$$

将截断的 gPC 展开 (3–13) 带入到输运方程 (3–1) 中，然后做伽辽金投影（Galerkin projection），得到 [41,64]：

$$\varepsilon\partial_t\hat{f} + v\partial_x\hat{f} = -\frac{1}{\varepsilon}(I - [\cdot])\Sigma\hat{f} - \varepsilon\Sigma^a\hat{f} - \hat{S} \tag{3–14}$$

其中 $\hat{S}$ 的定义和 (3–13) 类似。

接下来应用 micro-macro 分解

$$\hat{f}(t,x,v) = \hat{\rho}(t,x) + \varepsilon\hat{g}(t,x,v) \tag{3–15}$$

其中 $\hat{\rho} = [\hat{f}]$，$[g] = 0$，代回到 (3–14) 中得到

$$\partial_t\hat{\rho} + \partial_x\,[v\hat{g}] = -\Sigma^a\hat{\rho} + \hat{S}, \tag{3–16a}$$
$$\partial_t\hat{g} + \frac{1}{\varepsilon}(I - [.])(v\partial_x\hat{g}) = -\frac{1}{\varepsilon^2}\Sigma\hat{g} - \Sigma^a\hat{g} - \frac{1}{\varepsilon^2}v\partial_x\hat{\rho}, \tag{3–16b}$$

初值为

$$\hat{\rho}(0,x) = \hat{\rho}_0(x), \quad \hat{g}(0,x,v) = \hat{g}_0(x,v)\,,$$

满足

$$\frac{1}{2}\int_{-1}^{1}(\hat{\rho}(0,x) + \varepsilon\hat{g}(0,x,v))^2\,\mathrm{d}v = \hat{\rho}(0,x)^2 + \frac{\varepsilon^2}{2}\int_{-1}^{1}\hat{g}(0,x,v))^2\,\mathrm{d}v \le C\,.$$

显然上述系统 (3–16) 形式上当 $\varepsilon \to 0$ 扩散极限为：

$$\partial_t\hat{\rho} = \partial_x(K\partial_x\hat{\rho}) - \Sigma^a\hat{\rho} + \hat{S}\,, \tag{3–17}$$

其中

$$K = \frac{1}{3}\Sigma^{-1}\,.$$

这样 gPC-SG 方法是 s-AP 的，见 [41]。

我们可以很容易地得到系统 (3–16) 的能量估计

$$\int_0^1 \hat{\rho}(t,x)^2\,\mathrm{d}x + \frac{\varepsilon^2}{2}\int_0^1\int_{-1}^1 \hat{g}(t,x,v)^2\,\mathrm{d}v\,\mathrm{d}x$$
$$\leq \int_0^1 \hat{\rho}(0,x)^2\,\mathrm{d}x + \frac{\varepsilon^2}{2}\int_0^1\int_{-1}^1 \hat{g}(0,x,v)^2\,\mathrm{d}v\,\mathrm{d}x\,.$$

另一方面，对随机扩散方程 (3–8)–(3–9) 直接应用 gPC-SG 方法得到：

$$\partial_t\hat{\rho} = \partial_x(K_d\partial_x\hat{\rho}) - \Sigma^a\hat{\rho} + \hat{S}, \tag{3–18}$$

其中 $K_d = (\kappa_{ij})$，$\kappa_{i,j} = \langle\phi_i, \kappa\phi_j\rangle_\omega$。

## 3.3 gPC-SG 方法在随机空间的正则性和一致谱收敛分析

在这节中，为了方便起见假设 $\sigma^a = S = 0$ 和周期边界条件

$$f(t,0,v,z) = f(t,1,v,z). \tag{3–19}$$

我们将证明在给 $\sigma(z)$ 一定的假设条件下，带有随机参数的线性输运方程的解会一致的（关于 $\varepsilon$）保持初值在随机空间的正则性。然后基于该正则性的结果，给出 gPC-SG 方法谱收敛的分析并给出误差估计，并且该误差估计是不依赖于 $\varepsilon$ 的。

### 3.3.1 证明中要用到的记号

首先回忆前面一节引入的随机变量的希尔伯特空间3.2,

$$H(\mathbb{R}^d;\,\omega\,\mathrm{d}z) = \Big\{\, f \mid \mathbb{R}^d \to \mathbb{R},\, \int_{\mathbb{R}^d} f^2(z)\omega(z)\,\mathrm{d}z < +\infty \Big\}, \tag{3–20}$$

和相应的内积与范数 (3–12)。记关于 $z$ 的 $k$ 阶求导算子为

$$D^k f(t,x,v,z) := \partial_z^k f(t,x,v,z), \tag{3–21}$$

和 $H$ 中的索伯列夫范数

$$\|f(t,x,v,\cdot)\|_{H^k}^2 := \sum_{\alpha\leq k}\|D^\alpha f(t,x,v,\cdot)\|_\omega^2. \tag{3–22}$$

最后，我们引入如下的关于相空间的范数

$$\|f(t,\cdot,\cdot,\cdot)\|_\Gamma^2 := \int_Q \|f(t,x,v,\cdot)\|_\omega^2\,\mathrm{d}x\,\mathrm{d}v, \qquad t\geq 0, \tag{3–23}$$

$$\|f(t,\cdot,\cdot,\cdot)\|_{\Gamma^k}^2 := \int_Q \|f(t,x,v,\cdot)\|_{H^k}^2\,\mathrm{d}x\,\mathrm{d}v, \qquad t\geq 0, \tag{3–24}$$

其中 $Q = [0,1]\times[-1,1]$ 表示相空间。方便起见，我们将在证明中省略 $t$，用 $\|f\|_\Gamma, \|f\|_{\Gamma^k}$ 表示上述范数。

### 3.3.2 随机空间的正则性

为了研究 $f$ 关于随机变量 $z$ 的正则性，首先我们给出如下引理。简单起见，下述引理及其证明的叙述均是一维情况，对于高维情况证明是一样的，仅仅是部分系数略有不同。

**引理 3.1.** 假设 $\sigma(z) \geq \sigma_{\min} > 0$，对于任意正整数 $k$ 和 $\sigma \in W^{k,\infty}$，$g \in H^k$ 我们有

$$-\langle D^k(\sigma g), D^k g\rangle_\omega \leq -\frac{\sigma_{\min}}{2}\|D^k g\|_\omega^2 + \frac{4^k}{2\sigma_{\min}}\big(\max_{0\leq j\leq k}\|D^j\sigma\|_{L^\infty}^2\big)\|g\|_{H^{k-1}}^2 . \tag{3–25}$$

**证明.** 因为

$$D^k(\sigma g) = \sum_{j=0}^{k}\binom{k}{j}(D^{k-j}\sigma)(D^j g) = \sigma D^k g + \sum_{j=0}^{k-1}\binom{k}{j}(D^{k-j}\sigma)(D^j g)\,, \tag{3–26}$$

我们有

$$\begin{aligned} -\langle D^k\big(\sigma g\big), D^k g\rangle_\omega &= -\langle \sigma D^k g, D^k g\rangle_\omega - \Big\langle \sum_{j=0}^{k-1}\binom{k}{j}(D^{k-j}\sigma)(D^j g), D^k g\Big\rangle_\omega \\ &\leq -\sigma_{\min}\|D^k g\|_\omega^2 - \Big\langle \sum_{j=0}^{k-1}\binom{k}{j}(D^{k-j}\sigma)(D^j g), D^k g\Big\rangle_\omega . \end{aligned} \tag{3–27}$$

根据 Young 不等式（Young's inequality）

$$\begin{aligned} -\Big\langle \sum_{j=0}^{k-1}\binom{k}{j}(D^{k-j}\sigma)(D^j g), D^k g\Big\rangle_\omega \leq \frac{\sigma_{\min}}{2}\|D^k g\|_\omega^2 \\ + \frac{1}{2\sigma_{\min}}\Big\|\sum_{j=0}^{k-1}\binom{k}{j}(D^{k-j}\sigma)(D^j g)\Big\|_\omega^2 , \end{aligned} \tag{3–28}$$

和柯西不等式（Cauchy-Schwarz inequality）

$$\begin{aligned} \Big\|\sum_{j=0}^{k-1}\binom{k}{j}(D^{k-j}\sigma)(D^j g)\Big\|_\omega^2 &\leq \Big(\sum_{j=0}^{k-1}\binom{k}{j}^2\|D^{k-j}\sigma\|_{L^\infty}^2\Big)\Big(\sum_{j=0}^{k-1}\|D^j g\|_\omega^2\Big) \\ &\leq \Big\{\sum_{j=0}^{k}\binom{k}{j}^2\Big\}\max_{0\leq j\leq k}\|D^j\sigma\|_{L^\infty}^2\|g\|_{H^{k-1}}^2, \\ &\leq 4^k\big(\max_{0\leq j\leq k}\|D^j\sigma\|_{L^\infty}^2\big)\|g\|_{H^{k-1}}^2 . \end{aligned} \tag{3–29}$$

结合 (3–27)，(3–28) 和 (3–29)，得到

$$-\langle D^k\big(\sigma\cdot g\big), D^k g\rangle \leq -\frac{\sigma_{\min}}{2}\|D^k g\|_\omega^2 + \frac{4^k}{2\sigma_{\min}}\big(\max_{0\leq j\leq k}\|D^j\sigma\|_{L^\infty}^2\big)\|g\|_{H^{k-1}}^2 . \tag{3–30}$$

证毕。 □

现在我们可以给出如下正则性结果

**定理 3.2 (一致正则性).** 假设

$$\sigma(z) \geq \sigma_{\min} > 0\,.$$

如果对于整数 $m \geq 0$,

$$\|D^k\sigma(z)\|_{L^\infty} \leq C_\sigma, \qquad \|D^k f_0\|_\Gamma \leq C_0, \qquad k = 0, \ldots, m, \tag{3–31}$$

输运方程 *(3–1)–(3–2)* 的解 $f$，在 $\sigma^a = S = 0$ 和周期边界条件 *(3–19)* 下，满足

$$\|D^k f\|_\Gamma \leq C, \qquad k = 0, \cdots, m, \qquad \forall t > 0, \tag{3–32}$$

其中 $C_\sigma$, $C_0$ 和 $C$ 是不依赖 $\varepsilon$ 的常数。

**证明.** 因为 $\sigma^a = S = 0$，形式上方程 (3–1) 关于 $z$ 的 $k(0 \leq k \leq m)$ 阶导数为

$$\varepsilon^2\partial_t(D^k f) + \varepsilon v\partial_x(D^k f) = D^k\big(\sigma(z)([f] - f)\big), \tag{3–33}$$

其中 $[\cdot]$ 是速度平均算符，见 (3–5)。方程 (3–33) 两边同乘 $D^k f$ 并在 $Q = [0,1] \times [-1,1]$ 上积分，得到

$$\begin{aligned} \frac{\varepsilon^2}{2}\partial_t\|D^k f\|_\Gamma^2 + \varepsilon\int_Q v\langle D^k f, \partial_x(D^k f)\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v \\ = \int_Q \langle D^k\big(\sigma(z)([f] - f)\big), D^k f\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v\,. \end{aligned} \tag{3–34}$$

分部积分得到

$$\varepsilon\int_Q v\langle D^k f, \partial_x(D^k f)\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v = \frac{\varepsilon}{2}\int_{Q\times\mathbb{R}^d} v\partial_x(D^k f)^2\omega \,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}v = 0, \tag{3–35}$$

注意这里用到了周期边界条件 (3–19)。注意到

$$\int_Q \langle D^k\big(\sigma(z)([f] - f)\big), \big[D^k f\big]\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v = 0\,, \tag{3–36}$$

结合 (3–34) 可以推出

$$\frac{\varepsilon^2}{2}\partial_t\|D^k f\|_\Gamma^2 = -\int_Q \langle D^k\big(\sigma(z)([f] - f)\big), D^k([f] - f)\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v \tag{3–37}$$

**能量估计**：我们将对 $k$ 用数学归纳法建立如下能量估计，对于任意正整数 $k \geq 0$，存在 $k$ 个常数 $c_{kj} > 0$，$j = 0, \ldots, k-1$ 使得

$$\varepsilon^2\partial_t\Big(\|D^k f\|_\Gamma^2 + \sum_{j=0}^{k-1} c_{kj}\|D^j f\|_\Gamma^2\Big) \leq \begin{cases} -2\sigma_{\min}\big\|\,[f] - f\big\|_\Gamma^2, & k = 0, \\ -\sigma_{\min}\big\|D^k([f] - f)\big\|_\Gamma^2, & k \geq 1. \end{cases} \tag{3–38}$$

当 $k = 0$ 时，(3–37) 变成

$$\begin{aligned} \varepsilon^2\partial_t\|f\|_\Gamma^2 &= -2\int_Q \langle\sigma(z)([f] - f), ([f] - f)\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v \\ &\leq -2\sigma_{\min}\|[f] - f\|_\Gamma^2\,, \end{aligned} \tag{3–39}$$

满足 (3–38)，归纳奠基成立。

假设对于任意 $k \le p$，$p \in \mathbb{N}$，(3–38) 成立。将这些不等式加在一起得到

$$\varepsilon^2 \partial_t \Big( \frac{1}{2} \|f\|_\Gamma^2 + \sum_{i=1}^{p} \|D^i f\|_\Gamma^2 + \sum_{i=1}^{p} \sum_{j=0}^{i-1} c_{ij} \|D^j f\|_\Gamma^2 \Big) \le -\sigma_{\min} \big\| [f] - f \big\|_{\Gamma^p}^2 , \tag{3–40}$$

等价于

$$\varepsilon^2 \partial_t \Big( \sum_{j=0}^{p} c'_{p+1,j} \|D^j f\|_\Gamma^2 \Big) \le -\sigma_{\min} \big\| [f] - f \big\|_{\Gamma^p}^2 , \tag{3–41}$$

其中

$$c'_{p+1,j} = \begin{cases} \dfrac{1}{2} + \sum\limits_{i=1}^{p} c_{i0}, & j = 0, \\ 1 + \sum\limits_{i=1}^{p} c_{ij}, & 1 \le j \le p-1, \\ 1, & j = p. \end{cases} \tag{3–42}$$

当 $k = p+1$ 时，(3–37) 为

$$\varepsilon^2 \partial_t \|D^{p+1} f\|_\Gamma^2 = -2 \int_Q \langle D^{p+1} \big( \sigma(z)([f] - f) \big), D^{p+1}([f] - f) \rangle_\omega \, \mathrm{d}x \, \mathrm{d}v \, . \tag{3–43}$$

在**引理**3.1中令 $g = D^{p+1}([f] - f)$ 以及假设 $\|D^k \sigma(z)\|_{L^\infty} \le C_\sigma$，等式右端有估计

$$\begin{aligned} \mathrm{RHS} \le & -\sigma_{\min} \int_Q \big\| D^{p+1}([f] - f) \big\|_\omega^2 \, \mathrm{d}x \, \mathrm{d}v \\ & + \frac{4^{p+1}}{\sigma_{\min}} \big( \max_{0 \le j \le p+1} \|D^j \sigma\|_{L^\infty}^2 \big) \int_Q \big\| [f] - f \big\|_{H^p}^2 \, \mathrm{d}x \, \mathrm{d}v \\ \le & -\sigma_{\min} \big\| D^{p+1}([f] - f) \big\|_\Gamma^2 + \frac{C_\sigma^2 C'_{p+1}}{\sigma_{\min}} \big\| [f] - f \big\|_{\Gamma^p}^2 . \end{aligned} \tag{3–44}$$

其中 $C'_{p+1} = (p+1) 4^{p+1}$。现在可以得到估计

$$\varepsilon^2 \partial_t \|D^{p+1} f\|_\Gamma^2 \le -\sigma_{\min} \big\| D^{p+1}([f] - f) \big\|_\Gamma^2 + \frac{C_\sigma^2 C'_{p+1}}{\sigma_{\min}} \big\| [f] - f \big\|_{\Gamma^p}^2 . \tag{3–45}$$

将 (3–41) 乘以 $C_\sigma^2 C'_{p+1} / \sigma_{\min}^2$ 然后与 (3–45) 相加，

$$\varepsilon^2 \partial_t \Big( \|D^{p+1} f\|_\Gamma^2 + \sum_{j=0}^{p} c_{p+1,j} \|D^j f\|_\Gamma^2 \Big) \le -\sigma_{\min} \big\| D^{p+1}([f] - f) \big\|_\Gamma^2 , \tag{3–46}$$

其中

$$c_{p+1,j} = \frac{C_\sigma^2 C'_{p+1}}{\sigma_{\min}} c'_{p+1,j} \, . \tag{3–47}$$

这说明 (3–38) 对于 $k = p+1$ 仍然成立。根据数学归纳法，(3–38) 对所有整数 $k \in \mathbb{N}$ 成立。

最后，根据 (3–38)，我们有

$$\partial_t \Big( \|D^k f\|_\Gamma^2 + \sum_{j=0}^{k-1} c_{kj} \|D^j f\|_\Gamma^2 \Big) \le 0, \qquad c_{kj} > 0, \ k \in \mathbb{N}. \tag{3–48}$$

推出

$$\begin{aligned}\|D^k f\|_\Gamma^2 &\le \|D^k f\|_\Gamma^2 + \sum_{j=0}^{k-1} c_{kj}\|D^j f\|_\Gamma^2 \\ &\le \|D^k f_0\|_\Gamma^2 + \sum_{j=0}^{k-1} c_{kj}\|D^j f_0\|_\Gamma^2 \\ &\le C_0^2\Big(1+\sum_{j=0}^{k-1} c_{kj}\Big) := C^2,\end{aligned} \tag{3–49}$$

这里 $C$ 显然不依赖于 $\varepsilon$，证毕。 □

**定理3.2**说明解关于 $z$ 的导数可以被初值控制住。特别的，$\|D^k f\|_\Gamma$ 的界**不依赖于**$\varepsilon$! 这在证明格式是 s-AP 中非常关键。但是仅有这个估计是不能够保证整个 gPC-SG 方法是关于 $\varepsilon$ 一致的谱收敛（因为投影误差前面的系数是 $O(1/\varepsilon^2)$，所以我们需要一个关于 $[f]-f$ 的导数的 $O(\varepsilon^2)$ 的估计来抵消这个系数）。我们接下来给出一个引理。

**引理 3.3.** *假设对于某个整数 $m \ge 0$,*

$$\|D^k(\partial_x f_0)\|_\Gamma \le C_x, \qquad k = 0,\dots,m, \qquad t > 0. \tag{3–50}$$

*那么有如下结果:*

$$\int_Q \varepsilon\langle vD^k(\partial_x f), D^k([f]-f)\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v \le \frac{\sigma_{\min}}{4}\|D^k([f]-f)\|_\Gamma^2 + \frac{C_1\varepsilon^2}{\sigma_{\min}}. \tag{3–51}$$

**证明.** 首先注意到 $\partial_x f$ 和 $f$ 满足一样的方程，

$$\varepsilon^2\partial_t(\partial_x f) + \varepsilon v\partial_x(\partial_x f) = \sigma(z)([\partial_x f] - \partial_x f). \tag{3–52}$$

所以由**定理3.2**和假设 (3–50),

$$\|D^k(\partial_x f)\|_\Gamma \le C, \qquad t > 0, \tag{3–53}$$

$C$ 与 $\varepsilon$ 无关。根据 Young 不等式，

$$\begin{aligned}&\int_Q \varepsilon\langle vD^k(\partial_x f), D^k([f]-f)\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v \\ &\le \frac{\sigma_{\min}}{4}\|D^k([f]-f)\|_\Gamma^2 + \frac{\varepsilon^2}{\sigma_{\min}}\|vD^k(\partial_x f)\|_\Gamma^2 \\ &\le \frac{\sigma_{\min}}{4}\|D^k([f]-f)\|_\Gamma^2 + \frac{\varepsilon^2}{\sigma_{\min}}\|D^k(\partial_x f)\|_\Gamma^2 \\ &\le \frac{\sigma_{\min}}{4}\|D^k([f]-f)\|_\Gamma^2 + \frac{C_1\varepsilon^2}{\sigma_{\min}},\end{aligned} \tag{3–54}$$

$C_1 = C^2$ 为常数，证毕。 □

现在我们来证明如下定理。

**定理 3.4** ($\varepsilon^2$**-estimate on** $[f]-f$)**.** *在定理3.2和引理3.3中的所有假设下，对于给定的时间 $T>0$，有如下 $[f]-f$ 的正则性结果：*

$$\begin{aligned}&\|D^k([f]-f)\|_\Gamma^2\\&\le \mathrm{e}^{-\sigma_{\min}t/2\varepsilon^2}\|D^k([f_0]-f_0)\|_\Gamma^2+C'\varepsilon^2\\&\le C\varepsilon^2,\end{aligned}\tag{3–55}$$

*对任意 $t\in(0,T]$ 和 $0\le k\le m$，其中 $C'$ 与 $C$ 是与 $\varepsilon$ 无关的常数。*

**证明.** 首先注意到 $[f]$ 满足

$$\varepsilon^2\partial_t[f]+\varepsilon\partial_x[vf]=0,\tag{3–56}$$

所以 $[f]-f$ 满足如下方程

$$\varepsilon^2\partial_t([f]-f)+\varepsilon\partial_x([vf]-vf)=-\sigma(z)([f]-f).\tag{3–57}$$

如同**定理3.2**的证明，将该方程对 $z$ 求导 $k$ 次，然后乘以 $D^k([f]-f)$，在 $Q$ 上积分，得到

$$\begin{aligned}\varepsilon^2\partial_t\big\|D^k([f]-f)\big\|_\Gamma^2=&-2\int_Q\varepsilon\langle D^k(\partial_x[vf]-v\partial_x f),D^k([f]-f)\rangle_\omega\,\mathrm{d}x\,\mathrm{d}v\\&-2\int_Q\langle D^k\big(\sigma(z)([f]-f)\big),D^k([f]-f)\rangle_\omega\,\mathrm{d}x\,\mathrm{d}v\\:=&\,I+II.\end{aligned}\tag{3–58}$$

注意到

$$\int_Q\varepsilon\langle D^k(\partial_x[vf]),D^k([f]-f)\rangle_\omega\,\mathrm{d}x\,\mathrm{d}v=0,\tag{3–59}$$

和**引理3.3**，推出

$$I\le\frac{\sigma_{\min}}{2}\|D^k([f]-f)\|_\Gamma^2+\frac{2C_1\varepsilon^2}{\sigma_{\min}}.\tag{3–60}$$

对于第二部分根据**引理3.1**，

$$II\le-\sigma_{\min}\big\|D^k([f]-f)\big\|_\Gamma^2+\frac{C_\sigma^2 4^k}{\sigma_{\min}}\big\|[f]-f\big\|_{\Gamma^{k-1}}^2.\tag{3–61}$$

这样得到如下估计，

$$\begin{aligned}\varepsilon^2\partial_t\big\|D^k([f]-f)\big\|_\Gamma^2\le&-\frac{\sigma_{\min}}{2}\big\|D^k([f]-f)\big\|_\Gamma^2+\frac{2C_1\varepsilon^2}{\sigma_{\min}}\\&+\frac{C_\sigma^2 4^k}{\sigma_{\min}}\big\|[f]-f\big\|_{\Gamma^{k-1}}^2.\end{aligned}\tag{3–62}$$

接下来我们仍然使用数学归纳法，但 $k=0$ 时 (3–62)

$$\varepsilon^2\partial_t\big\|[f]-f\big\|_\Gamma^2\le-\frac{\sigma_{\min}}{2}\big\|[f]-f\big\|_\Gamma^2+\frac{2C_1\varepsilon^2}{\sigma_{\min}}.\tag{3–63}$$

由 Grönwall 不等式（Grönwall's inequality），

$$\begin{aligned}\big\|[f]-f\big\|_\Gamma^2&\le\mathrm{e}^{-\sigma_{\min}t/2\varepsilon^2}\big\|[f_0]-f_0\big\|_\Gamma^2+\frac{4C_1}{\sigma_{\min}^2}\varepsilon^2\\&\le C_0\varepsilon^2,\qquad\text{for } t>0,\end{aligned}\tag{3–64}$$

满足 (3–55)，奠基成立。

假设对于任意，$k \le p$ 其中 $p \in \mathbb{N}$，(3–55) 都成立。意味着

$$\big\|[f]-f\big\|^2_{\Gamma^p(t)} \le C_p\varepsilon^2. \tag{3–65}$$

所以当 $k=p+1$ 时由 (3–62)，

$$\begin{aligned}\varepsilon^2\partial_t\big\|D^{p+1}([f]-f)\big\|^2_\Gamma \le & -\frac{\sigma_{\min}}{2}\big\|D^{p+1}([f]-f)\big\|^2_\Gamma+\frac{2C_1\varepsilon^2}{\sigma_{\min}}\\ &+\frac{C_\sigma^2C'_{p+1}}{\sigma_{\min}}C_p\varepsilon^2,\end{aligned} \tag{3–66}$$

即

$$\partial_t\big\|D^{p+1}([f]-f)\big\|^2_\Gamma \le -\frac{\sigma_{\min}}{2\varepsilon^2}\big\|D^{p+1}([f]-f)\big\|^2_\Gamma+C''_{p+1}. \tag{3–67}$$

同样的，用 Grönwall 不等式得到

$$\begin{aligned}\big\|D^{p+1}([f]-f)\big\|^2_\Gamma &\le \mathrm{e}^{-\sigma_{\min}t/2\varepsilon^2}\big\|D^{p+1}([f_0]-f_0)\big\|^2_\Gamma+C''_{p+1}\varepsilon^2\\ &\le C_{p+1}\varepsilon^2, \qquad \text{for } t>0,\end{aligned} \tag{3–68}$$

$C_{p+1}$ 是不依赖 $\varepsilon$ 的常数。根据数学归纳法，证毕。 □

**注　5.** 以上的引理和定理都是在 $z\in\mathbb{R}$ 并且 $\sigma$ 只依赖与 $z$ 的条件下证明的。但是，我们的结论和技术并不局限于这些简单情况。对于 $z\in\mathbb{R}^d$，推广是直接的而对于 $\sigma(x,z)$ 也依赖于 $x$ 的情况，只需要用**定理3.2**的证明中的类似技术来修改**引理3.3**的证明即可。

### 3.3.3　一致的谱收敛

令 $f$ 是线性输运方程 (3–1)–(3–2) 的解，我们定义 $M$ 阶投影算子

$$\mathcal{P}_Mf=\sum_{i=0}^{M}\langle f,\phi_i\rangle_\omega\phi_i.$$

gPC-SG 方法产生的误差可以分为两部分 $r_N$ 和 $e_N$，

$$f-f_M=f-\mathcal{P}_Mf+\mathcal{P}_Mf-f_M:=r_M+e_M, \tag{3–69}$$

其中 $r_M=f-\mathcal{P}_Mf$ 是截断误差，$e_M=\mathcal{P}_Mf-f_M$ 是投影误差。

对于截断误差 $r_M$，我们有以下引理

**引理　3.5.** *在定理3.2和定理3.4的所有假设下，我们有对 $t\in(0,T]$ 和任意整数 $k=0,\dots,m$，*

$$\|r_M\|_\Gamma \le \frac{C_1}{M^k}. \tag{3–70}$$

*进一步的我们有，*

$$\big\|\,[r_M]-r_M\big\|_\Gamma \le \frac{C_2}{M^k}\varepsilon, \tag{3–71}$$

*其中 $C_1$ 和 $C_2$ 不依赖于 $\varepsilon$。*

**证明.** 根据正交多项式近似的标准误差估计和**定理3.2**，对于 $0 \le t \le T$,

$$\|r_M\|_\Gamma \le CM^{-k} \, |D^k f \, |_\Gamma \le \frac{C_1}{M^k} \tag{3–72}$$

其中 $C$ 独立于 $M$。

用相同的方法，根据**定理3.4**，

$$\begin{aligned} \big\| \, [r_M] - r_M \|_\Gamma &= \big\| ([f] - f) - ([\mathcal{P}_M f] - \mathcal{P}_M f) \big\|_\Gamma \\ &\le CM^{-k} \, |D^k([f] - f) \, |_\Gamma \\ &\le \frac{C_2}{M^k} \varepsilon \end{aligned} \tag{3–73}$$

证毕。 □

接下来我们要估计 $e_M$，为了这个目的，我们首先注意到 $f_M$ 满足

$$\varepsilon^2 \partial_t f_M + \varepsilon v \partial_x f_M = \mathcal{P}_M \big\{ \sigma(z)([f_M] - f_M) \big\}. \tag{3–74}$$

另一方面，通过直接对原始线性输运方程做 $M$ 阶投影，我们得到

$$\varepsilon^2 \partial_t (\mathcal{P}_M f) + \varepsilon v \partial_x (\mathcal{P}_M f) = \mathcal{P}_M \big\{ \sigma(z)([f] - f) \big\}. \tag{3–75}$$

(3–75) 减去 (3–74) 得到

$$\begin{aligned} \varepsilon^2 \partial_t e_M + \varepsilon v \partial_x e_M &= \mathcal{P}_M \Big\{ \sigma(z) \big\{ \, [f] - f - ([f_M] - f_M) \big\} \Big\} \\ &= \mathcal{P}_M \Big\{ \sigma(z) \big\{ \, [f] - f - ([\mathcal{P}_M f] - \mathcal{P}_M f) \\ &\qquad + ([\mathcal{P}_M f] - \mathcal{P}_M f) - ([f_M] - f_M) \big\} \Big\} \\ &= \mathcal{P}_M \Big\{ \sigma(z) \big( \, [r_M] - r_M \big) \Big\} + \mathcal{P}_M \Big\{ \sigma(z) \big( \, [e_M] - e_M \big) \Big\}. \end{aligned} \tag{3–76}$$

现在我们可以给出投影误差 $e_M$ 的如下估计，

**引理 3.6.** *在定理3.2和定理3.4的所有假设下，我们有对 $t \in (0, T]$ 和任意整数 $k = 0, \ldots, m$,*

$$\|e_M\|_\Gamma \le \frac{C(T)}{M^k}, \tag{3–77}$$

*$C(T)$ 是与 $\varepsilon$ 无关的常数。*

**证明.** 我们使用与之前基本相同的能量估计：用 $e_M$ 乘 (3–76) 并在 $Q$ 上积分，注意到

$$\int_Q \langle \mathcal{P}_M \big\{ \sigma(z) \big( \, [r_M] - r_M \big) \big\}, [e_M] \rangle_\omega \, \mathrm{d}x \, \mathrm{d}v = 0, \tag{3–78}$$

$$\int_Q \langle \mathcal{P}_M \big\{ \sigma(z) \big( \, [e_M] - e_M \big) \big\}, [e_M] \rangle_\omega \, \mathrm{d}x \, \mathrm{d}v = 0, \tag{3–79}$$

然后得到

$$\begin{aligned} \varepsilon^2 \partial_t \|e_M\|_\Gamma^2 = \; &- \int_Q \langle \mathcal{P}_M \big\{ \sigma(z) \big( \, [e_M] - e_M \big) \big\}, [e_M] - e_M \rangle_\omega \, \mathrm{d}x \, \mathrm{d}v \\ &- \int_Q \langle \mathcal{P}_M \big\{ \sigma(z) \big( \, [r_M] - r_M \big) \big\}, [e_M] - e_M \rangle_\omega \, \mathrm{d}x \, \mathrm{d}v. \end{aligned} \tag{3–80}$$

注意到投影算符 $\mathcal{P}_M$ 是一个自共轭算符

$$\langle \mathcal{P}_M f, g\rangle_\omega = \langle f, \mathcal{P}_M g\rangle_\omega,$$

以及

$$\mathcal{P}_M e_M = e_M,$$

所以

$$\begin{aligned}
\varepsilon^2 \partial_t \|e_M\|_\Gamma^2 = & -\int_Q \langle \sigma(z)\big([e_M]-e_M\big), [e_M]-e_M\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v \\
& -\int_Q \langle \sigma(z)\big([r_M]-r_M\big), [e_M]-e_M\rangle_\omega \,\mathrm{d}x\,\mathrm{d}v \\
\le & -\sigma_{\min}\big\|[e_M]-e_M\big\|_\Gamma^2 + \frac{\sigma_{\min}}{2}\big\|[e_M]-e_M\big\|_\Gamma^2 \\
& + \frac{C_\sigma}{2\sigma_{\min}}\big\|[r_M]-r_M\big\|_\Gamma^2 \\
\le & -\frac{\sigma_{\min}}{2}\big\|[e_M]-e_M\big\|_\Gamma^2 + \frac{C_\sigma}{2\sigma_{\min}}\Big(\frac{C'}{M^k}\Big)^2\varepsilon^2 \\
\le & \Big(\frac{C}{M^k}\Big)^2\varepsilon^2,
\end{aligned} \tag{3–81}$$

其中对于最后两个不等式，我们使用了 Young 不等式和**引理3.5**。然后对 $t$ 积分得到

$$\|e_M\|_\Gamma^2 \le \|e_M^0\|_\Gamma^2 + \Big(\frac{C(T)}{M^k}\Big)^2, \tag{3–82}$$

因为 $e_M^0 = \mathcal{P}_M f_0 - f_M^0 = 0$，证毕。 □

最后，我们陈述主要的收敛定理:

**定理 3.7 (关于$\varepsilon$ 的一致收敛).** *假设*

$$\sigma(z) \ge \sigma_{\min} > 0\,.$$

*如果对于某个整数 $m \ge 0$,*

$$\|\sigma(z)\|_{H^k} \le C_\sigma, \quad \|D^k f_0\|_\Gamma \le C_0, \quad \|D^k(\partial_x f_0)\|_\Gamma \le C_x, \quad k = 0,\dots,m, \tag{3–83}$$

*那么整个 gPC-SG 方法的误差*

$$\|f - f_M\|_\Gamma \le \frac{C(T)}{M^k}, \tag{3–84}$$

*$C(T)$ 是与 $\varepsilon$ 无关的常数。*

**证明.** 由**引理3.5**和**引理3.6**,

$$\|f - f_M\|_\Gamma \le \|r_M\|_\Gamma + \|e_M\|_\Gamma \le \frac{C(T)}{M^k},$$

证毕。 □

**注 6. 定理3.7**给出了一个关于 $\varepsilon$ 一致的谱收敛速率，因此可以选择 $M$ 独立于 $\varepsilon$，这是一个非常强的 s-AP 性质。在各向异性散射的情况，即 $\sigma$ 依赖于 $v$，则通常获得需要 $M \gg \varepsilon$ 的收敛速率（参见例如 [77]）。在这种情况下，s-AP 的证明要困难得多，并且通常需要借助于扩散极限方程。在确定性情况的情况下参见 [70]，在确定性情况的情况下参见 [80]。

## 3.4 全离散格式

如同在 [41] 中所指出的，通过使用 gPC-SG 方法，我们会得到关于原始的确定性输运方程的向量化版本。这使得能够应用原本对于确定性问题构造的 AP 格式。在本节中，我们将对 gPC-SG 系统 (3–16) 应用在文章 [74] 中提出的 AP 格式。

线性输运方程中非常重要而又具有挑战性的问题之一是边界条件的处理，这里我们并不考虑这个问题。相关工作可以参考参考 Jin 和 Levermore 的早期工作 [85] 以及 Lemou 和 Méhats 的最近的工作 [86]，他们对 AP 性质和物理边界条件的数值处理做了相当深入的研究。

我们采用均匀网格 $x_i = ih, i = 0, 1, \cdots N$，其中 $h = 1/N$ 是网格大小，时间方向 $t^n = n\Delta t$，$\rho_i^n$ 是 $\rho$ 在网格点 $(x_i, t^n)$ 的近似值，$g_{i+\frac{1}{2}}^{n+1}$ 则定义在交错格点 $x_{i+1/2} = (i + 1/2)h$, $i = 0, \cdots N - 1$ 上。

gPC-SG 系统 (3–11) 的全离散格式为

$$\frac{\hat{\rho}_i^{n+1} - \hat{\rho}_i^n}{\Delta t} + \left[ v \frac{\hat{g}_{i+\frac{1}{2}}^{n+1} - \hat{g}_{i-\frac{1}{2}}^{n+1}}{\Delta x} \right] = -\Sigma_i^a \hat{\rho}_i^{n+1} + \hat{S}_i, \tag{3–85a}$$

$$\begin{aligned} &\frac{\hat{g}_{i+\frac{1}{2}}^{n+1} - \hat{g}_{i+\frac{1}{2}}^n}{\Delta t} + \frac{1}{\varepsilon \Delta x}(I - [.]) \left( v^+ (\hat{g}_{i+\frac{1}{2}}^n - \hat{g}_{i-\frac{1}{2}}^n) + v^- (\hat{g}_{i+\frac{3}{2}}^n - \hat{g}_{i+\frac{1}{2}}^n) \right) \\ &\quad = -\frac{1}{\varepsilon^2} \Sigma_i \hat{g}_{i+\frac{1}{2}}^{n+1} - \Sigma^a \hat{g}_{i+\frac{1}{2}}^{n+1} - \frac{1}{\varepsilon^2} v \frac{\hat{\rho}_{i+1}^n - \hat{\rho}_i^n}{\Delta x}. \end{aligned} \tag{3–85b}$$

当 $\varepsilon \to 0$ 时可以很容易得出其形式上的扩散极限，

$$\frac{\hat{\rho}_i^{n+1} - \hat{\rho}_i^n}{\Delta t} - K \frac{\hat{\rho}_{i+1}^n - 2\hat{\rho}_i^n + \hat{\rho}_{i-1}^n}{\Delta x^2} = -\Sigma_i^a \hat{\rho}_i^{n+1} + \hat{S}_i, \tag{3–86}$$

其中 $K = \dfrac{1}{3}\Sigma^{-1}$，而这正是 (3–17) 的全离散格式。因此，按照 [41] 中的定义，该格式是 s-AP 的。

我们还注意到 $\left[ \hat{g}_{i+\frac{1}{2}}^n \right] = 0$ 对于每个 $n$，后面我们将会反复用到这个性质。

## 3.5 一致稳定性

AP 格式的一个重要性质是具有不依赖于 $\varepsilon$ 的稳定性条件。因此当 $\varepsilon$ 很小时，可以取 $\Delta t \gg O(\varepsilon)$。在本节中，我们将证明上述结果，证明基本上按照文章 [81] 中证明确定性问题的思路进行。

为了说明清楚，在本节中，我们假设 $\sigma^a = S = 0$。关于稳定性的主要结果是以下定理：

**定理 3.8.** *记*

$$\sigma_{ij} = \langle \phi_i, \sigma \phi_j \rangle_\omega, \quad \Sigma = (\sigma_{ij}), \quad \Sigma \geq \sigma_{\min}\ Id\ .$$

*如果 $\Delta t$ 满足如下的 CFL 条件*

$$\Delta t \leq \frac{\sigma_{\min}}{3} \Delta x^2 + \frac{2\varepsilon}{3} \Delta x, \tag{3–87}$$

*那么格式 (3–85) 中的 $\hat{\rho}^n$ 和 $\hat{g}^n$ 满足能量估计*

$$\Delta x \sum_{i=0}^{N-1} \left( (\hat{\rho}_i^n)^2 + \frac{\varepsilon^2}{2} \int_{-1}^{1} \left( \hat{g}_{i+\frac{1}{2}}^n \right)^2 \mathrm{d}v \right) \leq \Delta x \sum_{i=0}^{N-1} \left( (\hat{\rho}_i^0)^2 + \frac{\varepsilon^2}{2} \int_{-1}^{1} \left( \hat{g}_{i+\frac{1}{2}}^0 \right)^2 \mathrm{d}v \right)$$

*对每个 $n$ 成立，故格式 (3–85) 是稳定的。*

**注 7.** 因为式 (3–87) 的右端当 $\varepsilon \to 0$ 时有下界 (且这个下界恰好是离散后的扩散方程 (3–86) 的稳定性条件)，所以格式是渐近稳定的并且 $\Delta t$ 即使在 $\varepsilon \to 0$ 时仍然可以是有限的。

### 3.5.1　一些记号和相关引理

我们给出一些将在分析中用到的记号。对于每个在格点上有定义的函数 $\mu=(\mu_i)_{i=0}^{N-1}$:

$$\|\mu\|^2=\Delta x\sum_{i=0}^{N-1}\mu_i^2\,. \tag{3–88}$$

对于每个依赖速度的格点函数 $v\in[-1,1]\mapsto\phi(v)=(\phi_{i+\frac12}(v))_{i=0}^{N-1}$，定义:

$$|||\phi|||=\Delta x\sum_{i=0}^{N-1}\left[\phi_{i+\frac12}^2\right]. \tag{3–89}$$

如果 $\phi$ 和 $\psi$ 是两个依赖于速度的函数，他们的内积定义如下:

$$\langle\phi\,,\,\psi\rangle=\Delta x\sum_{i=0}^{N-1}\left[\phi_{i+\frac12}\psi_{i+\frac12}\right]. \tag{3–90}$$

现在我们定义在格式 (3–85) 中用到的有限差分算子。对于每个格点函数 $\phi=(\phi_{i+\frac12})_{i\in\mathbb{Z}}$, 定义单侧差分算子:

$$D^-\phi_{i+\frac12}=\frac{\phi_{i+\frac12}-\phi_{i-\frac12}}{\Delta x}\quad\text{ and }\quad D^+\phi_{i+\frac12}=\frac{\phi_{i+\frac32}-\phi_{i+\frac12}}{\Delta x} \tag{3–91}$$

以及中心差分算子:

$$D^c\phi_{i+\frac12}=\frac{\phi_{i+\frac32}-\phi_{i-\frac12}}{2\Delta x}\quad\text{ and }\quad D^0\phi_i=\frac{\phi_{i+\frac12}-\phi_{i-\frac12}}{\Delta x}(=D^-\phi_{i+\frac12}). \tag{3–92}$$

最后对于每个格点函数 $\mu=(\mu_i)_{i\in\mathbb{Z}}$，定义如下的中心算子:

$$\delta^0\mu_{i+\frac12}=\frac{\mu_{i+1}-\mu_i}{\Delta x}. \tag{3–93}$$

我们先指出一些基本事实: 对于 $\phi=(\phi_{i+\frac12})_{i=0}^{N-1}$, $\psi=(\psi_{i+\frac12})_{i=0}^{N-1}$, 和 $\mu=(\mu_i)_{i=0}^{N-1}$, 有 (见 [81])

$$\left(v^+D^-+v^-D^+\right)\phi_{i+\frac12}=vD^c\phi_{i+\frac12}-\frac{\Delta x}{2}|v|D^-D^+\phi_{i+\frac12}; \tag{3–94}$$

$$\Delta x\sum_{i\in\mathbb{Z}}\left(D^+\phi_{i+\frac12}\right)^2\le\frac{4}{\Delta x^2}\Delta x\sum_i\phi_{i+\frac12}^2; \tag{3–95}$$

$$\left|\left\langle\left(v^+D^++v^-D^-\right)\psi\,,\,\phi\right\rangle\right|\le\alpha|||\phi|||^2+\frac{1}{4\alpha}|||v|D^+\psi|||^2,\ \forall\alpha>0; \tag{3–96}$$

$$\Delta x\sum_{i\in\mathbb{Z}}\mu_iD^0\phi_i=-\Delta x\sum_{i\in\mathbb{Z}}\left(\delta^0\mu_{i+\frac12}\right)\phi_{i+\frac12}; \tag{3–97}$$

$$\Delta x\sum_{i\in\mathbb{Z}}\psi_{i+\frac12}D^-\phi_{i+\frac12}\Delta x=-\Delta x\sum_{i\in\mathbb{Z}}\left(D^+\psi_{i+\frac12}\right)\phi_{i+\frac12}; \tag{3–98}$$

$$\Delta x\sum_{i\in\mathbb{Z}}\phi_{i+\frac12}D^c\phi_{i+\frac12}=0; \tag{3–99}$$

$$\text{如果}\quad g\in L^2([-1,1]),\quad\text{那么}\quad[vg]^2\le\frac12\left[|v|g^2\right]. \tag{3–100}$$

### 3.5.2 能量估计

现在我们给出能量估计的细节，证明基本和确定性问题的证明 [81] 类似。

首先，用 $\hat{\rho}^{n+1}$ and $\varepsilon^2\hat{g}^{n+1}$ 分别乘 (3–85a) and(3–85b)。根据假设 $\sigma_i^a = 0$，$\hat{S}_i = 0$，以及 $\Sigma \geq \sigma_{\min} Id$，我们有

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\hat{\rho}^{n+1}\|^2 - \|\hat{\rho}^n\|^2 + \|\hat{\rho}^{n+1} - \hat{\rho}^n\|^2\right) + \Delta x \sum_{i=0}^{N-1} \hat{\rho}_i^{n+1} D^0\left[v\hat{g}_i^{n+1}\right] \\
&\quad + \frac{\varepsilon^2}{2\Delta t}\left(|||\hat{g}^{n+1}|||^2 - |||\hat{g}^n|||^2 + |||\hat{g}^{n+1} - \hat{g}^n|||^2\right) \\
&\quad + \varepsilon\left\langle \hat{g}^{n+1}, \left(v^+D^- + v^-D^+\right)\hat{g}^n\right\rangle \\
&\leq -\sigma_{\min}|||\hat{g}^{n+1}|||^2 + \Delta x \sum_{i=0}^{N-1}\left[vD^0\hat{g}_i^{n+1}\right]\hat{\rho}_i^n.
\end{aligned}$$

结合左边的第二项和右边的最后一项，得到

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\hat{\rho}^{n+1}\|^2 - \|\hat{\rho}^n\|^2 + \|\hat{\rho}^{n+1} - \hat{\rho}^n\|^2\right) \\
&\quad + \frac{\varepsilon^2}{2\Delta t}\left(|||\hat{g}^{n+1}|||^2 - |||\hat{g}^n|||^2 + |||\hat{g}^{n+1} - \hat{g}^n|||^2\right) \\
&\quad + \varepsilon\left\langle \hat{g}^{n+1}, \left(v^+D^- + v^-D^+\right)\hat{g}^n\right\rangle \\
&\leq -\sigma_{\min}|||\hat{g}^{n+1}|||^2 + \Delta x \sum_{i=0}^{N-1}\left[vD^0\hat{g}_i^{n+1}\right](\hat{\rho}_i^n - \hat{\rho}_i^{n+1}).
\end{aligned}$$

使用 Young 不等式，

$$\Delta x \sum_{i=0}^{N-1}\left[vD^0\hat{g}_i^{n+1}\right](\hat{\rho}_i^n - \hat{\rho}_i^{n+1}) \leq \frac{1}{2\Delta t}\|\hat{\rho}^{n+1} - \hat{\rho}^n\|^2 + \frac{\Delta t}{2}\Delta x \sum_{i=0}^{N-1}\left[vD^0\hat{g}_i^{n+1}\right]^2.$$

给出

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\hat{\rho}^{n+1}\|^2 - \|\hat{\rho}^n\|^2\right) + \frac{\varepsilon^2}{2\Delta t}\left(|||\hat{g}^{n+1}|||^2 - |||\hat{g}^n|||^2 + |||\hat{g}^{n+1} - \hat{g}^n|||^2\right) \\
&\quad + \varepsilon\left\langle \hat{g}^{n+1}, \left(v^+D^- + v^-D^+\right)\hat{g}^n\right\rangle \\
&\leq -\sigma_{\min}|||\hat{g}^{n+1}|||^2 + \frac{\Delta t}{2}\Delta x \sum_{i=0}^{N-1}\left[vD^0\hat{g}_{i+\frac{1}{2}}^{n+1}\right]^2.
\end{aligned}$$

做如下分解

$$\begin{aligned}
&\left\langle \hat{g}^{n+1}, \left(v^+D^- + v^-D^+\right)\hat{g}^n\right\rangle = \left\langle \hat{g}^{n+1}, \left(v^+D^- + v^-D^+\right)\hat{g}^{n+1}\right\rangle \\
&+ \left\langle \hat{g}^{n+1}, \left(v^+D^- + v^-D^+\right)(\hat{g}^n - \hat{g}^{n+1})\right\rangle =: A + B,
\end{aligned}$$

其中

$$A = \frac{\Delta x}{2}\Delta x \sum_{i=0}^{N-1}\left[|v|\left(D^+\hat{g}_{i+\frac{1}{2}}^{n+1}\right)^2\right].$$

$$B = -\left\langle \left(v^+D^+ + v^-D^-\right)\hat{g}^{n+1},\, \hat{g}^n - \hat{g}^{n+1}\right\rangle.$$

由 Young 不等式，

$$|B| \le \frac{\varepsilon}{2\Delta t}|||\hat{g}^{n+1} - \hat{g}^n|||^2 + \frac{\Delta t}{2\varepsilon}||||v|D^+\hat{g}^{n+1}|||^2.$$

推出

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\hat{\rho}^{n+1}\|^2 - \|\hat{\rho}^n\|^2\right) + \frac{\varepsilon^2}{2\Delta t}\left(|||\hat{g}^{n+1}|||^2 - |||\hat{g}^n|||^2\right)\\
&\quad + \varepsilon\frac{\Delta x}{2}\sum_{i=0}^{N-1}\left[|v|\left(D^+\hat{g}^{n+1}_{i+\frac{1}{2}}\right)^2\right]\Delta x - \frac{\Delta t}{2}||||v|D^+\hat{g}^{n+1}|||^2\\
&\le -\sigma_{\min}|||\hat{g}^{n+1}|||^2 + \frac{\Delta t}{2}\Delta x\sum_{i=0}^{N-1}\left[vD^0\hat{g}^{n+1}_{i+\frac{1}{2}}\right]^2.
\end{aligned}$$

因为 $|v| \le 1$，

$$\frac{\Delta t}{2}||||v|D^+\hat{g}^{n+1}|||^2 \le \frac{\Delta t}{2}\Delta x\sum_{i=0}^{N-1}\left[|v|\left(D^+\hat{g}^{n+1}_{i+\frac{1}{2}}\right)^2\right],$$

$$\frac{\Delta t}{2}\Delta x\sum_{i\in\mathbb{Z}}\left[vD^0\hat{g}^{n+1}_{i+\frac{1}{2}}\right]^2 \le \frac{\Delta t}{4}\Delta x\sum_{i=0}^{N-1}\left[|v|\left(D^+\hat{g}^{n+1}_{i+\frac{1}{2}}\right)^2\right].$$

这意味着

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\hat{\rho}^{n+1}\|^2 - \|\hat{\rho}^n\|^2\right) + \frac{\varepsilon^2}{2\Delta t}\left(|||\hat{g}^{n+1}|||^2 - |||\hat{g}^n|||^2\right)\\
&\le -\sigma_{\min}|||\hat{g}^{n+1}|||^2 + \left(\frac{3\Delta t}{4} - \varepsilon\frac{\Delta x}{2}\right)\Delta x\sum_{i=0}^{N-1}\left[|v|\left(D^+\hat{g}^{n+1}_{i+\frac{1}{2}}\right)^2\right].
\end{aligned}$$

注意

$$\begin{aligned}
&\left(\frac{3\Delta t}{4} - \varepsilon\frac{\Delta x}{2}\right)\Delta x\sum_{i=0}^{N-1}\left[|v|\left(D^+\hat{g}^{n+1}_{i+\frac{1}{2}}\right)^2\right] \le \left(\frac{3\Delta t}{4} - \varepsilon\frac{\Delta x}{2}\right)_+\Delta x\sum_{i=0}^{N-1}\left[\left(D^+\hat{g}^{n+1}_{i+\frac{1}{2}}\right)^2\right]\\
&\le \left(\frac{3\Delta t}{4} - \varepsilon\frac{\Delta x}{2}\right)_+\frac{4}{\Delta x^2}|||\hat{g}^{n+1}|||^2,
\end{aligned}$$

其中 $(a)_+ = \max(0, a)$ 表示 $a$ 中正的部分. 带入到 (3.5.2)，得到

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\hat{\rho}^{n+1}\|^2 - \|\hat{\rho}^n\|^2\right) + \frac{\varepsilon^2}{2\Delta t}\left(|||\hat{g}^{n+1}|||^2 - |||\hat{g}^n|||^2\right)\\
&\le \left(\left(\frac{3\Delta t}{4} - \varepsilon\frac{\Delta x}{2}\right)_+\frac{4}{\Delta x^2} - \sigma_{\min}\right)|||\hat{g}^{n+1}|||^2.
\end{aligned}$$

最终得到估计

$$\|\hat{\rho}^{n+1}\|^2 + \varepsilon^2|||\hat{g}^{n+1}|||^2 \le \|\hat{\rho}^n\|^2 + \varepsilon^2|||\hat{g}^n|||^2$$

如果 $\Delta t$ 满足

$$\left(\frac{3\Delta t}{4} - \varepsilon\frac{\Delta x}{2}\right)_+\frac{4}{\Delta x^2} \le \sigma_{\min}.$$

由于 $\sigma_{\min} > 0$，等价于 $(\frac{3\Delta t}{4} - \varepsilon\frac{\Delta x}{2})\frac{4}{\Delta x^2} \le \sigma_{\min}$，这给出了一个充分条件。

$$\Delta t \le \frac{\Delta x^2 \sigma_{\min}}{3} + \frac{2}{3}\varepsilon\Delta x\,.$$

至此**定理 6.1** 证明结束。

## 3.6 数值例子

在这部分，我们将给出一些数值结果来证明方法的有效性。我们考虑带有随机系数 $\sigma(z)$ 的线性输运方程：

$$\varepsilon\partial_t f + v\partial_x f = \frac{\sigma(z)}{\varepsilon}([f] - f), \quad 0 < x < 1\,, \tag{3–101}$$

初值为：

$$f(0, x, v, z) = 0\,,$$

边界条件为：

$$f(t, 0, v, z) = 1, \quad v \ge 0; \qquad f(t, 1, v, z) = 0, \quad v \le 0.$$

### 3.6.1 例一：收敛性测试

首先考虑随机系数依赖于一维的随机变量：

$$\sigma(z) = 2 + z, \quad z \text{ 均匀分布于}(-1, 1).$$

该方程 (3–101) 的随机扩散极限方程为

$$\partial_t \rho = \frac{1}{3\sigma(z)}\partial_{xx}\rho\,, \tag{3–102}$$

初边值条件为：

$$\rho(t, 0, z) = 1, \quad \rho(t, 1, z) = 0, \quad \rho(0, x, z) = 0.$$

在该条件下极限方程 (3–102) 有解析解

$$\rho(t, x, z) = 1 - \ \mathrm{erf}\left(\frac{x}{\sqrt{\frac{4}{3\sigma(z)}t}}\right). \tag{3–103}$$

当 $\varepsilon$ 很小时，我们使用它作为参考解，因为它于精确解的误差至多为 $O(\varepsilon^2)$。这里我们设 $\varepsilon = 10^{-8}$。对于 $\varepsilon$ 较大时，由于方程没有显示的解，我们将使用所谓的配点法（collocation method，参见 [63]），并对 micro-macro 系统 (3–11) 用相同的时间、空间离散来作为比较。关于配点法的更多细节，尤其是 s-AP 性质，请参考 Jin 和 Liu 的工作中的讨论 [77]。此外，在以下示例中，在速度空间我们用 30 个高斯积分点来计算 $\rho$。

为了测试方法的准确性，我们使用均值和标准差在 $x$ 方向的 $\ell^2$ 范数：

$$e_{mean}(t) = \left\|\mathbb{E}[u^h] - \mathbb{E}[u]\right\|_{\ell^2}\,,$$

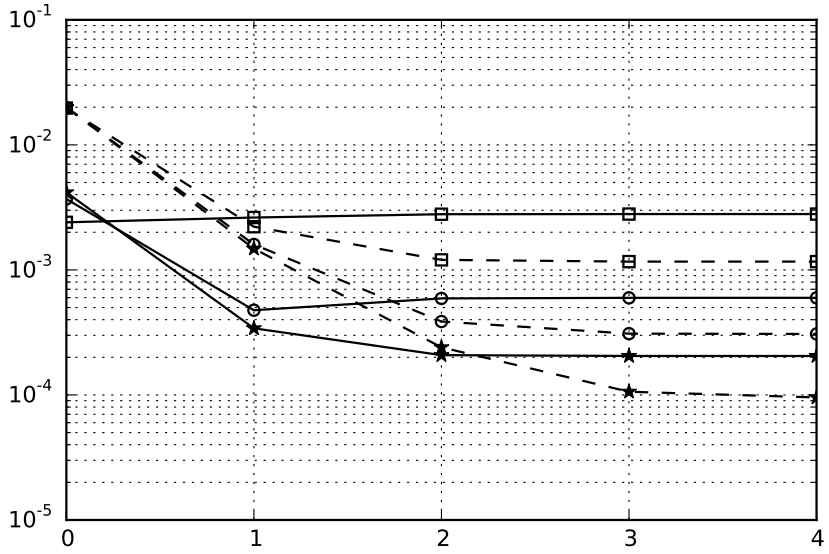


图 3–1 例一：$\rho$ 均值的误差（实线）和标准差的误差（虚线）与 gPC 阶数的关系。这里 $\varepsilon = 10^{-8}$: $\Delta x = 0.04$（方形），$\Delta x = 0.02$（圆圈），$\Delta x = 0.01$（星）。

Fig 3–1 Example 1. Errors of the mean (solid line) and standard deviation (dash line) of $\rho$ with respect to the gPC order at $\varepsilon = 10^{-8}$: $\Delta x = 0.04$ (squares), $\Delta x = 0.02$ (circles), $\Delta x = 0.01$ (stars).

$$e_{std}(t) = \left\| \sigma[u^h] - \sigma[u] \right\|_{\ell^2},$$

其中 $u^h, u$ 分别是数值解和参考解。

在图3–1中，我们绘制了 gPC 数值解的均值和标准偏差的误差在 $t = 0.01$ 与 gPC 阶数关系。三组结果包括：$\Delta x = 0.04$（正方形），$\Delta x = 0.02$（圆圈），$\Delta x = 0.01$（星）。这里 $\Delta t = 0.0002/3$。可以看到，误差随着 $N$ 增大快速衰减，然后在空间离散误差占优势变缓。很明显，即使 $\varepsilon = 10^{-8}$，由 gPC 展开导致的误差在 $M = 4$ 时就可以被忽略。解的均值和标准偏差的曲线分别显示在图3–2的左边和右边。

图3–3中我们还绘制了的通量 $vf$ 的平均值和标准偏差的曲线。在这里我们观察到 gPC-SG 方法和配点法以及参考解 (3–103) 之间良好的一致性。

在图3–4中，我们检查由四阶 gPC-SG 方法获得的解在 $t = 0.01$ 时 $\Delta x = 0.01$，$\Delta t = \Delta x^2/12$ 与极限方程解析解 (3–103) 的误差。正如预期，在数值误差占主导之前当 $\varepsilon^2$ 变小的时候误差会随之变小。

### 3.6.2　例二：混合尺度

在这个测试中，我们仍然令 $\sigma = 2 + z$。考虑如果 $\varepsilon > 0$ 也在一个较大的混合尺度上依赖于空间变量：

$$\varepsilon(x) = 10^{-3} + \frac{1}{2}[\tanh(6.5 - 11x) + \tanh(11x - 4.5)] \tag{3–104}$$

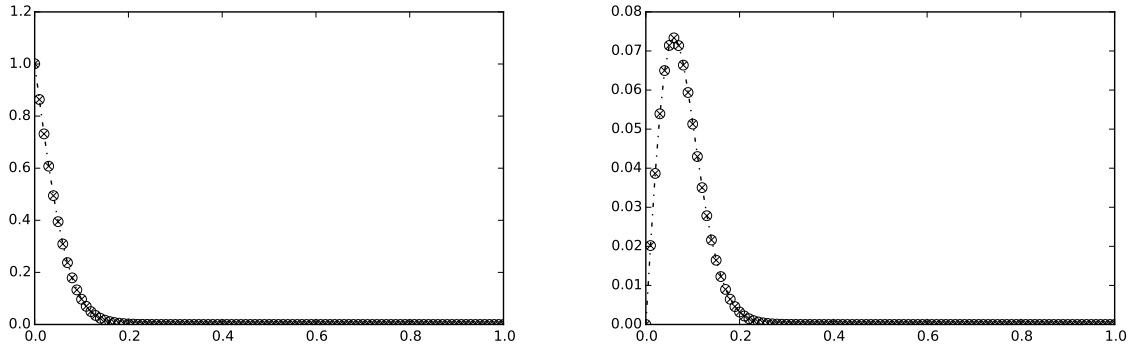


图 3–2 例一：$\rho$ 的均值（左边）和标准差（右边）。$\varepsilon = 10^{-8}$，gPC-SG 方法 $M = 4$（圆圈），配点法（叉）和极限解析解 (3–103)。

Fig 3–2 Example 1. The mean (left) and standard deviation (right) of $\rho$ at $\varepsilon = 10^{-8}$, obtained by the gPC Galerkin at order $M = 4$ (circles), the stochastic collocation method (crosses), and the limiting analytical solution (3–103).

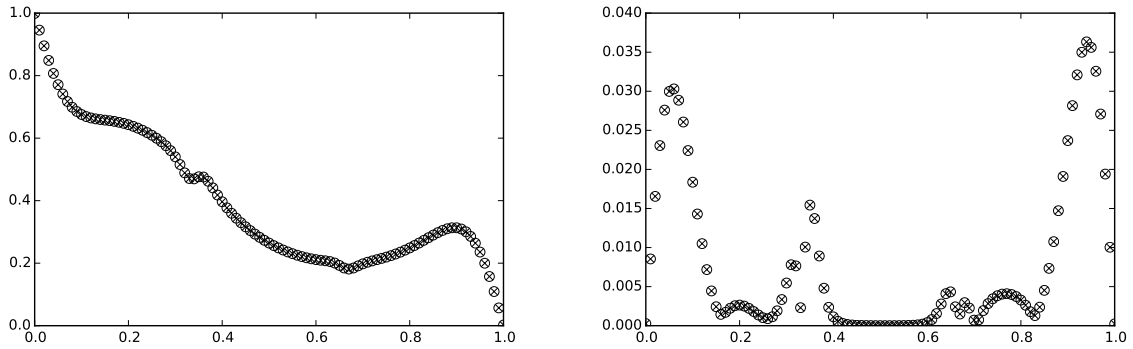


图 3–3 例一：均值（左）和标准差（右），gPC-SG 方法（圆圈）和配点法（叉），$t = 0.01$。

Fig 3–3 Example 1. The mean (left) and standard deviation (right) obtained by gPC-Galerkin (circle) and collocation method (cross) at time $t = 0.01$

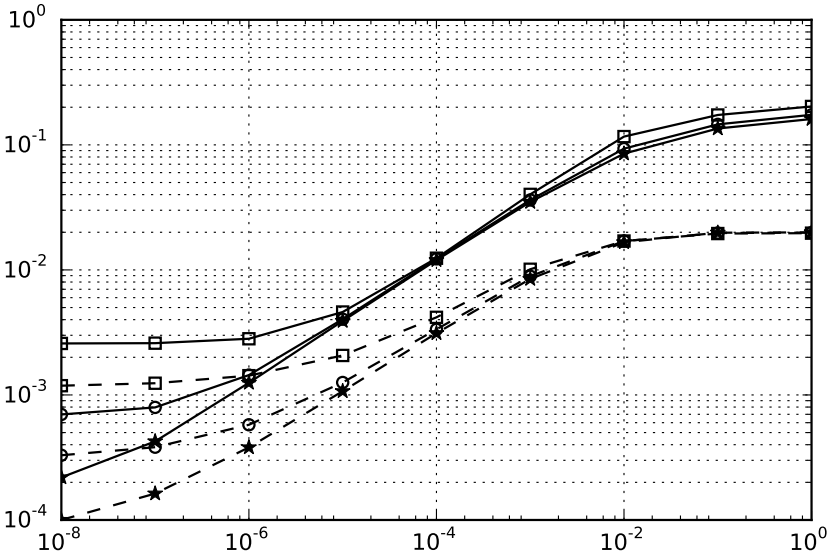


图 3–4 例一：两种解：解析解 (3–103) 和四阶 gPC-SG 方法，$\rho$ 的均值误差（实线）和标准差误差（虚线）与 $\varepsilon^2$ 的关系。$\Delta x = 0.04$ (方形), $\Delta x = 0.02$ (圆圈) and $\Delta x = 0.01$ (星)。

Fig 3–4 Example 1. Differences in the mean (solid line) and standard deviation (dash line) of $\rho$ with respect to $\varepsilon^2$, between the limiting analytical solution(3–103) and the 4th-order gPC solution with $\Delta x = 0.04$ (squares), $\Delta x = 0.02$ (circles) and $\Delta x = 0.01$ (stars).

并光滑地从 $O(10^{-3})$ 变到 $O(1)$，如图3–5。这种情况下会验证我们格式在混合多尺度下的适应能力，尤其是关于 $\varepsilon$ 的一致收敛性。

为了使质量仍然守恒，正确的线性输运方程为如下的形式，

$$\partial_t f + v\partial_x\Big(\frac{1}{\varepsilon(x)}f\Big) = \frac{\sigma}{\varepsilon^2(x)}\mathcal{L}f - \sigma^a f + S, \qquad \sigma(x,z) \geq \sigma_{\min} > 0, \tag{3–105}$$

那么 micro-macro 分解 (3–11) 修改为

$$\partial_t \rho + \partial_x [vg] = -\sigma^a \rho + S, \tag{3–106a}$$

$$\partial_t g + \frac{1}{\varepsilon(x)}(I - [.])(v\partial_x g) = -\frac{\sigma(z)}{\varepsilon^2(x)}g - \sigma^a g - \frac{1}{\varepsilon(x)}v\partial_x\Big(\frac{1}{\varepsilon(x)}\rho\Big). \tag{3–106b}$$

可以发现只有最后一项发生了变化。对于极限方程 (3–8)，也需要修改为

$$\partial_t \rho = \partial_x(\kappa(z)\partial_x\rho) - \partial_x(\kappa(z)a(x)\rho) - \sigma^a(z)\rho + S, \tag{3–107}$$

其中我们假设

$$a(x) = \lim_{\varepsilon\to 0}\frac{\varepsilon'(x)}{\varepsilon(x)}, \tag{3–108}$$

存在。对于相应的数值格式，我们也只需要把 (3–85b) 中的最后一项

$$-\frac{1}{\varepsilon^2}v\frac{\hat{\rho}^n_{i+1} - \hat{\rho}^n_i}{\Delta x} \tag{3–109}$$

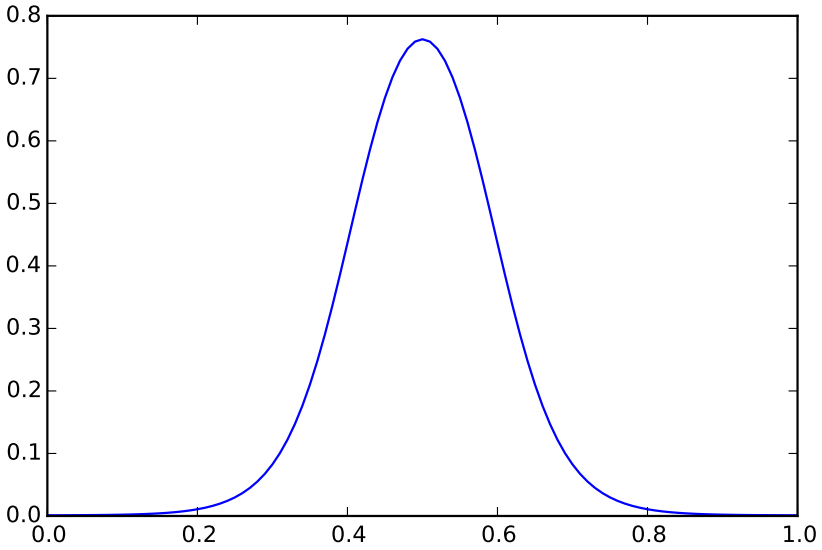


图 3–5 $\varepsilon(x)$
Fig 3–5 $\varepsilon(x)$

替换为

$$-\frac{1}{\varepsilon(x_{i+1/2})}v\left(\frac{\hat{\rho}_{i+1}^n}{\varepsilon(x_{i+1})}-\frac{\hat{\rho}_i^n}{\varepsilon(x_i)}\right)\frac{1}{\Delta x}. \tag{3–110}$$

初值为

$$f_{\text{in}}(x,v,z)=\frac{\rho_0}{2}\left[\exp\left(-(\frac{v-0.75}{T_0})^2\right)+\exp\left(-(\frac{v+0.75}{T_0})^2\right)\right] \tag{3–111}$$

其中

$$\rho_0(x)=\frac{2+\sin(2\pi x)}{2},\quad T_0(x)=\frac{5+2\cos(2\pi x)}{20}. \tag{3–112}$$

参考解是由配点法（30 样本）得到。相关参数如下：网格空间 $\Delta x=0.01$，时间 $\Delta t=\Delta x^2/3$。我们用五阶 gPC-SG 方法分别演化到时刻 $t=0.005$，$t=0.01$，$t=0.05$，$t=0.1$。对于 $v$ 方向的积分，我们使用 30 个点的高斯积分。

图3–6显示了均值和标准差的 $\ell^2$ 误差与 gPC 阶数的关系，可以看出收敛的非常快（谱收敛）。

### 3.6.3 例三：随机初值

接下来我们在初值上加入随机性（$\sigma=2+z$ 仍然随机）。

$$f(0,x,v,z)=f(0,x,v)+0.2z \tag{3–113}$$

其中 $f(x,v,0)$ 和 (3–111) 中一样。这次 $\Delta x=0.01$，$\Delta t=\Delta x^2/12$ 以及最终时间 $T=0.01$。首先我们测试流体极限 $\varepsilon=10^{-8}$，如图3–7。接下来我们测试 $\varepsilon=1$，如图3–8。可以看到两种方法吻合度很高。

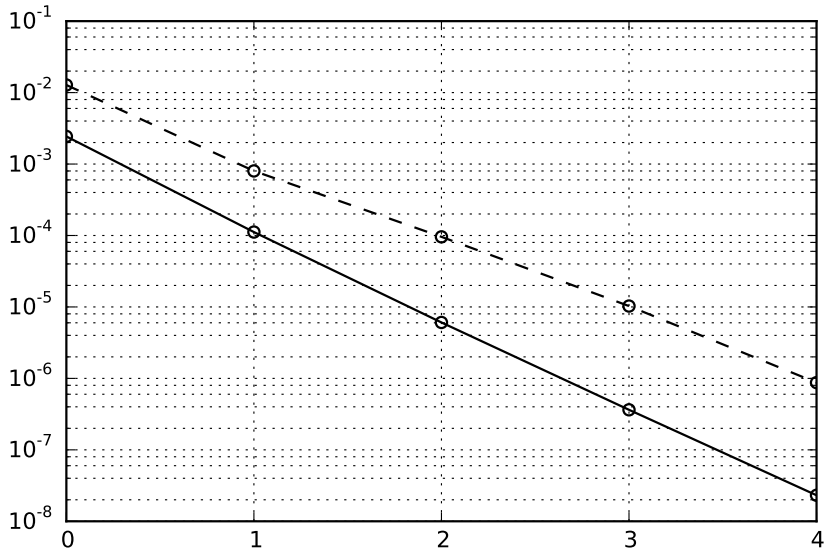


图 3–6 例二：均值（实线）和标准差（虚线）的 $\ell^2$ 误差与 gPC 阶数的关系

Fig 3–6 Example 2 with initial data(3–111)–(3–112). The $\ell^2$ error of mean and standard deviation (dash line) with respect to gPC order.

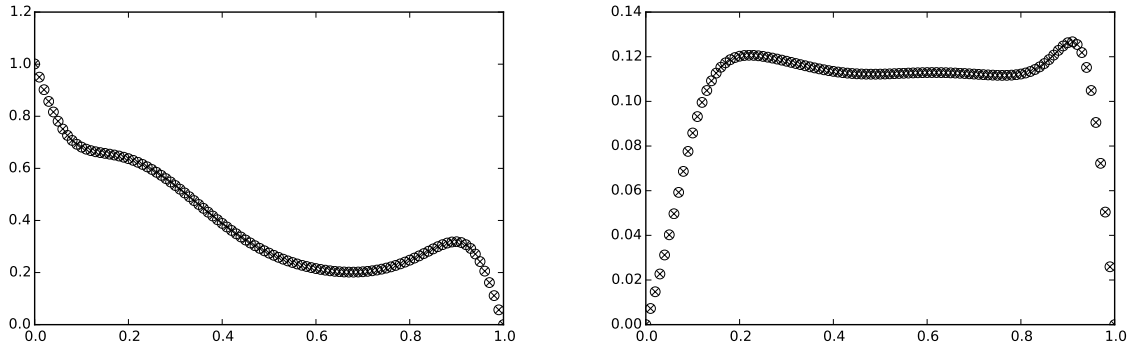


图 3–7 例三：均值（左）和标准差（右），gPC-SG 方法（圆圈）和配点法（叉），$t = 0.1$，$\varepsilon = 10^{-8}$。

Fig 3–7 Example 3. The mean (left) and standard deviation (right) obtained by gPC-Galerkin (circle) and collocation method (cross) at time $t = 0.1$, $\varepsilon = 10^{-8}$.

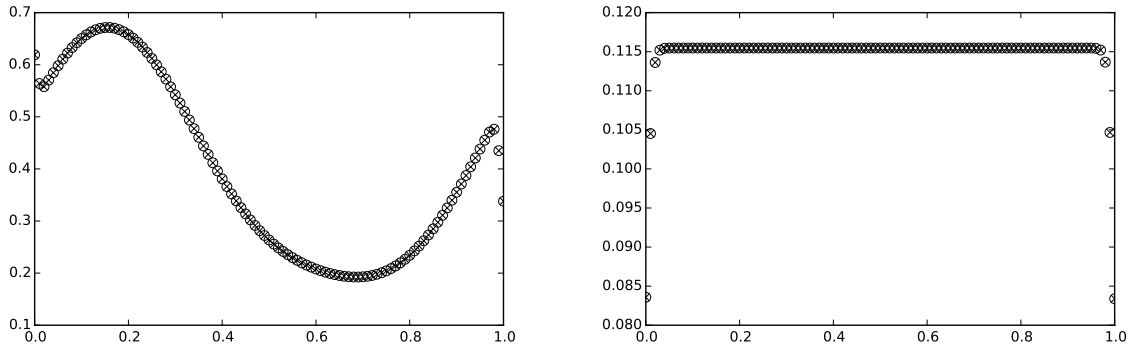


图 3–8 例三：均值（左）和标准差（右），gPC-SG 方法（圆圈）和配点法（叉），$t = 0.1, \varepsilon = 1$。
Fig 3–8 Example 3. The mean (left) and standard deviation (right) obtained by gPC-Galerkin (circle) and collocation method (cross) at time $t = 0.1$, $\varepsilon = 1$

### 3.6.4 例四：随机边界条件

这个例子中，我们在边界条件中加入随机性：

$$f_L(t, v, z) = 2 + z, \quad f_R(t, v, z) = 1 + z. \tag{3–114}$$

我们也测试了当 $\varepsilon = 10^{-8}$ 和 $\varepsilon = 10$ 两种情况，如图3–9和图3–10，同样的 gPC-SG 方法和参考解高度吻合。

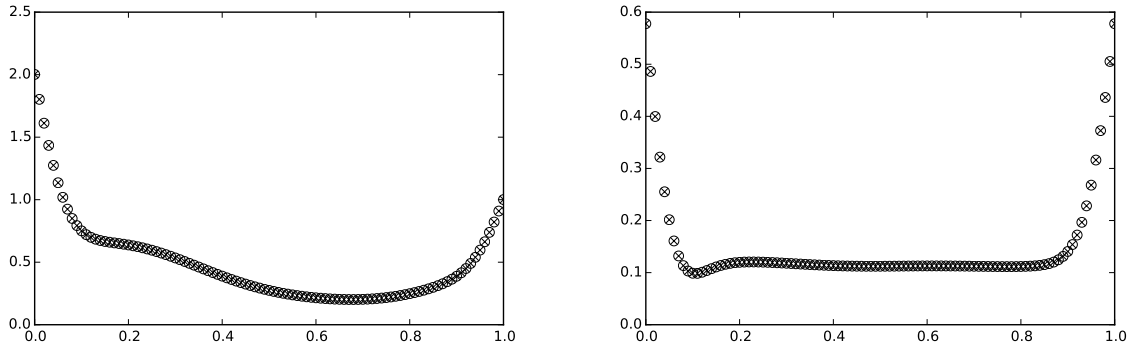


图 3–9 例四：均值（左）和标准差（右），gPC-SG 方法（圆圈）和配点法（叉），$t = 0.1, \varepsilon = 10^{-8}$。
Fig 3–9 Example 4. The mean (left) and standard deviation (right) obtained by gPC-Galerkin (circle) and collocation method (cross) at time $t = 0.1$, $\varepsilon = 10^{-8}$

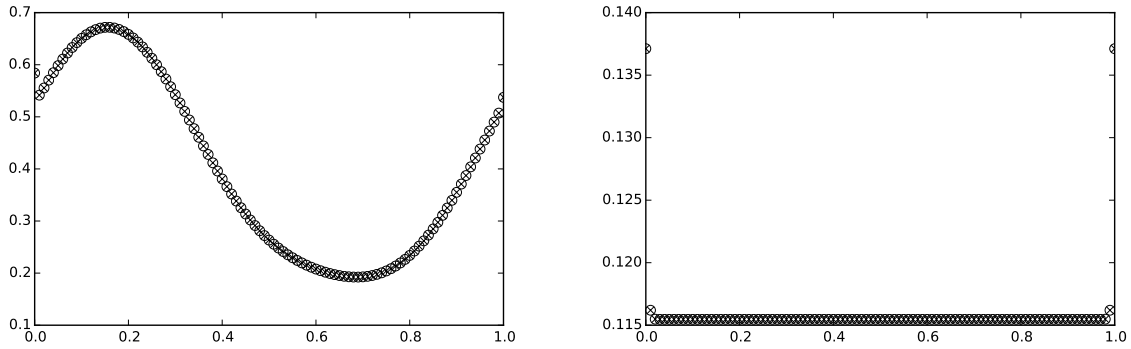


图 3–10 例四：均值（左）和标准差（右），gPC-SG 方法（圆圈）和配点法（叉），$t = 0.1, \varepsilon = 10$。
Fig 3–10 Example 4. The mean (left) and standard deviation (right) obtained by gPC-Galerkin (circle) and collocation method (cross) at time $t = 0.1$, $\varepsilon = 10$

### 3.6.5 例五：二维随机空间

最后，我们来考察具有如下形式的二维随机场：

$$\sigma(x, z_1, z_2) = 1 - \frac{\sigma z_1}{\pi^2}\cos(2\pi x) - \frac{\sigma z_2}{4\pi^2}\cos(4\pi x) \tag{3–115}$$

其中我们设置 $\sigma = 4$ 和 $z_1, z_2$ 为均匀分布在 $(-1, 1)$ 上的随机变量。由 $\Delta x = 0.025$，$\Delta t = 0.0002/3$ 的五阶 gPC-SG 方法得到在 $t = 0.01$ 的解 $\rho$ 的均值和标准偏差见图3–11。然后我们使用高阶配点法（40×40 高斯 - 勒让德积分点）来计算解的参考解的均值和标准差。在图3–12中，我们画出 $\rho$ 的均值（实线）和标准差（虚线）的误差与 gPC 阶数的关系。可以清楚地看到误差的快速谱收敛。

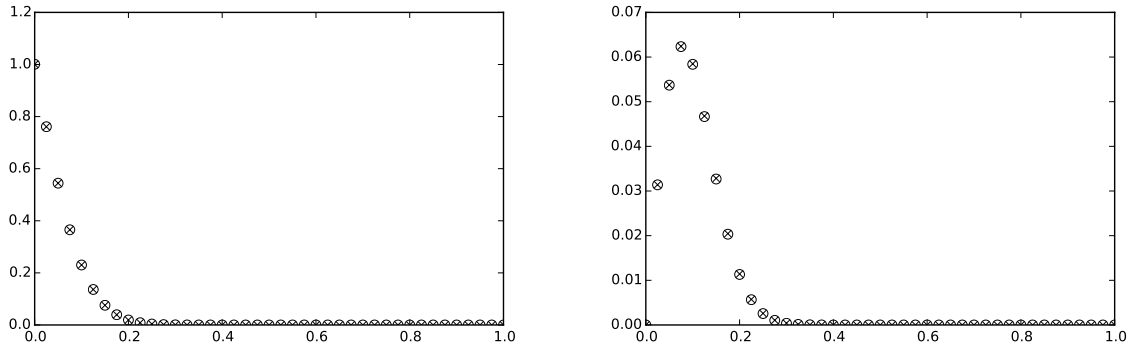


图 3–11 均值（左）和标准差（右），五阶 gPC-SG 方法（圆圈）和配点法（叉），二维随机变量。
Fig 3–11 The mean (left) and standard deviation (right) of $\rho$ at $\varepsilon = 10^{-8}$, obtained by 5th-order gPC Galerkin (circles) and the stochastic collocation method (crosses). The random input has dimension $d = 2$.

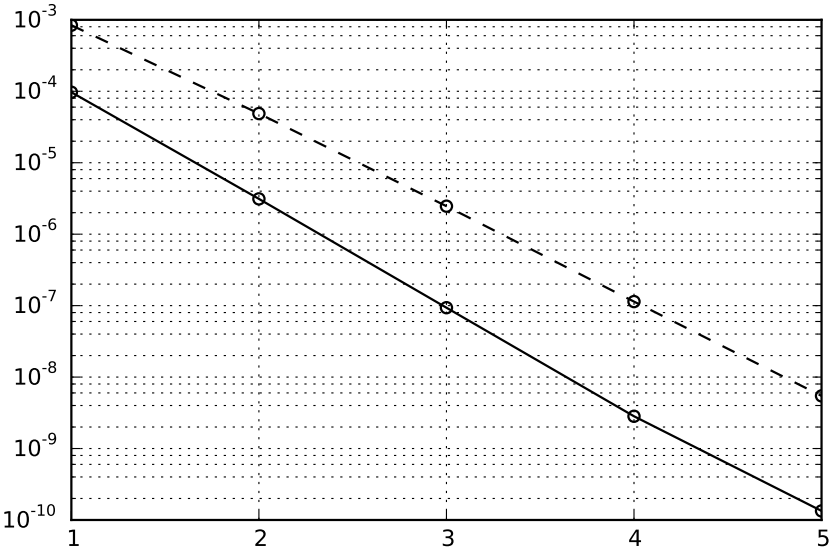


图 3–12 $\rho$ 的均值误差（实线）和标准差误差（虚线）与 gPC-SG 阶数的关系，二维随机变量。
Fig 3–12 Errors of the mean (solid line) and standard deviation (dash line) of $\rho$ with respect to gPC order, with the $d = 2$ dimensional random input.

## 3.7 本章总结与展望

在本章中，我们建立了随机伽辽金方法对带有随机散射系数的线性输运方程关于克努森数一致的谱精度分析，从而允许我们证明该方法具有随机渐近保持性质（s-AP）。对于基于 micro-macro 分解的全离散格式，我们证明了一致的稳定性结果。这是首次有人证明了关于这类问题的一致性结果。

关于一致性的分析在带有多尺度与不确定性的问题中非常重要。我们的分析对于线性问题具有一般的指导意义，进而可以推广到更多的 kinetic 方程或其他方程上去。而对于非线性的问题，仍然非常困难，将作为未来的研究的内容。

# 第四章　使用 NUFFT 的半拉格朗日时间算子分裂法在具有向量势的薛定谔方程的应用

量子效应在许多科学和工程领域中发挥重要作用，例如理论化学，固态力学和量子光学。而对于薛定谔方程的数学分析和数值模拟具有根本的重要性。这种类型的方程形成了一类的色散偏微分方程，即不同波长的波在不同相速度下传播的方程。而当考虑磁场时，我们需要将向量势函数引入到薛定谔方程中。

在本章中，我们考虑具有向量势函数的半经典薛定谔方程，其具有形式

$$i\varepsilon\partial_t u^\varepsilon = \frac{1}{2}\left(-i\varepsilon\nabla_x - \mathbf{A}(x)\right)^2 u^\varepsilon + V(x)u^\varepsilon, \quad t\in\mathbb{R}^+, \quad x\in\mathbb{R}^3, \tag{4–1}$$

$$u^\varepsilon(x,0) = u_0(x), \quad x\in\mathbb{R}^3, \tag{4–2}$$

其中 $u^\varepsilon(x,t)$ 是复值波函数，$V(x)\in\mathbb{R}$ 是标量势函数，$\mathbf{A}(x)\in\mathbb{R}^3$ 是向量势函数。数学上我们用标量势和向量势来描述电磁场，即电场 $\mathbf{E}(x)\in\mathbb{R}^3$ 和磁场 $\mathbf{B}(x)\in\mathbb{R}^3$ 如下

$$\mathbf{E} = -\nabla V(x), \qquad \mathbf{B} = \nabla\times\mathbf{A}\,(x)\,. \tag{4–3}$$

薛定谔方程(4–1)可以由不带向量势的方程通过局部规范变换（参见 [87]）导出。外部电磁场的存在的量子力学演化会导致许多深远的结果，例如朗道能级、塞曼效应和超导性。在分析方面，哈密顿算子在光谱和散射性质上具有不同的特征（见 [88]）。数值上，它也带来了新的挑战，特别是在半经典格式中。向量势函数的存在使得薛定谔方程中引入对流项并且同时标量势函数也产生了一定影响（参见 [3]）。

事实上，可以通过额外添加一个条件来简化势函数，即给出一个指定的规范。电场 $\mathbf{E}(x)\in\mathbb{R}^3$ 和磁场 $\mathbf{B}(x)\in\mathbb{R}^3$ 具有所谓的规范不变性。一个自然的选择是 $\nabla_x\cdot\mathbf{A}=0$，这就是所谓的库仑规范。在这个规范中，矢量势和正则动量互为对易，$[\mathbf{A}, -i\varepsilon\nabla_x]=0$ 使得修改后的薛定谔方程(4–1)的动能（“kinetic”）部分可以简化为如下

$$\frac{1}{2}(-i\varepsilon\nabla_x - \mathbf{A})^2 u^\varepsilon = -\frac{\varepsilon^2}{2}\Delta_x u^\varepsilon + i\varepsilon\mathbf{A}\cdot\nabla_x u^\varepsilon + \frac{1}{2}|\mathbf{A}|^2 u^\varepsilon. \tag{4–4}$$

在薛定谔方程中，波函数作为辅助量用于计算宏观物理量（物理可观测量），例如位置密度

$$n(x,t) = |u^\varepsilon(x,t)|^2, \tag{4–5}$$

和修正的电流密度

$$\mathbf{J}(x,t) = \frac{1}{2}\left(\overline{u^\varepsilon}\left(-i\varepsilon\nabla_x - \mathbf{A}\right)u^\varepsilon - u^\varepsilon\left(-i\varepsilon\nabla_x - \mathbf{A}\right)\overline{u^\varepsilon}\right), \tag{4–6}$$

其中 $\bar{f}$ 表示 $f$ 的复共轭。实际上，我们有以下质量守恒方程

$$\frac{\partial}{\partial t}n + \nabla_x\cdot\mathbf{J} = 0. \tag{4–7}$$

我们指出 $n$ 和 $\mathbf{J}$ 都是规范不变量。另外两个重要的物理量是质量

$$m(t) := \|u^\varepsilon(x,t)\|_{L^2}^2 = \int_{\mathbb{R}^3} n(t,x)dx, \tag{4–8}$$

和能量

$$\mathcal{E}(t) := \frac{1}{2}\|(-i\varepsilon\nabla - \mathbf{A})u^\varepsilon\|_{L^2}^2 + \langle u^\varepsilon, Vu^\varepsilon\rangle, \tag{4–9}$$

其中 $\langle f,g\rangle \equiv \int_{\mathbb{R}^d} f(x)\overline{g(x)}\,dx$ 是标准的内积。对于 $u^\varepsilon \in C(\mathbb{R}_t; L^2(\mathbb{R}^d)\cap \mathcal{S}(\mathbb{R}^d))$，这些量在动力学演化中是守恒的。我们把详细的证明放在本章的附录A中。

在半经典格式下，即 $\varepsilon \ll 1$ 时，波函数 $u^\varepsilon$ 在空间和时间上都是高度振荡的，振荡的尺度为 $O(\varepsilon)$。因此当 $\varepsilon \to 0$ 时，它在强意义下并不收敛。当 $\varepsilon \ll 1$ 时，相比于直接求解薛定谔方程，更多的是一些近似方法，如水平集方法和基于 WKB 分析和 Wigner 变换的矩封闭方法，参见 [89-92]。高斯波束法（或高斯波包法）是另一类重要的方法，它允许精确计算焦散和捕获相位信息（参见例如 [93-96]），这种模型误差为 $O(\varepsilon^{1/2})$。为了提高近似精度，人们引入了高阶高斯波束方法，其具有误差 $C_k(T)\varepsilon^{k/2}$（见 [97-98]）。然而，在 [99-100] 中已经证明对于固定的 $\varepsilon$，更高阶的高斯波束方法可能不是减少误差的实用方法。然而，Hagedorn 在 [101] 中提出，在 [100-102] 中分析和实现的 Hagedorn 波包可以有效地减少所有 $\varepsilon \in (0,1]$ 中的误差。在 [100] 中，Zhou 已经将该方法扩展到向量势函数，并为伽辽金近似的高阶收敛提供了严格的证明。最近，Russo 和 Smereka 在文献 [103-104] 中提出了一种基于所谓的高斯波包变换的新方法，这是另一个值得选择的方法。

在数值上，如果想直接模拟薛定谔方程(4–1)，波函数的振荡性质会使计算开销十分巨大。对于物理可观测量的计算，如 $n(x,t)$ 和 $\mathbf{J}(x,t)$，也是如此。据我们所知，最好的方法之一是时间分裂谱方法，由 Bao，Jin 和 Markowich 在 [4,92,105] 中提出，其中网格策略 $\Delta t = O(\varepsilon)$ 和 $\Delta x = O(\varepsilon)$ 就足以保证波函数的准确近似。而为了计算正确的物理可观察量，时间步长可以放松到 $O(1)$。

由于向量势函数的存在，与经典情况相比，在(4–4)中存在两个主要变化：修正的标量势和新的对流项。为了设计一个无条件稳定的格式，Jin 和 Zhou 在 [3] 中引入了半拉格朗日时间分裂方法（semi-Lagrangian time splitting method），其中网格划分策略 $\Delta t = O(\varepsilon)$ 和 $\Delta x = O(\varepsilon)$ 足以保证准确的近似的波函数。类似地，可以使用 $\varepsilon$ 时间步长来捕获正确的物理可观察量。在对流步骤中，在 [3] 中分析和实现了多项式插值技术，其中为了效率考虑牺牲了空间精度。事实上，可以代替地应用谱插值以改善空间精度，不幸的是，它将使计算复杂度从 $O(N)$（多项式插值）增加到 $O(N^2)$（直接傅里叶级数求和）其中 $N$ 是网格点的数量。这里主要问题是取样点不一定均匀分布，标准逆 FFT 不再适用，因此，半拉格朗日方法需要效率和精度之间的平衡。

由于非均匀快速傅立叶变换（NUFFT）（参见例如 [42-43]），可以是该问题得到理想的解决。这是我们工作的主要出发点。非均匀傅里叶变换在各种应用领域中产生，从医学成像到射电天文学到偏微分方程的数值解。当采样是均匀的并且在等间隔频率处需要傅立叶变换时，经典快速傅里叶变换（FFT）在计算中起到重要作用，其仅需要 $O(N\log N)$ 复杂度来计算 $N$ 个傅里叶系数而不是 $O(N^2)$ 的复杂度。然而，当数据不是在“物理”或“频率”域中的均匀分割的网格上进行采样时，不幸的是，FFT 并不适用。在过去几年中，已经开发了许多算法来克服了这种限制，通常被称为非均匀 FFT（NUFFT）。

在本章中，我们将 NUFFT 算法结合到时间分裂半拉格朗日方法中，计算复杂度为 $O(N\log N)$。可以证明新方法是无条件稳定的。与多项式插值的情况不同（插值模板需要特别处理），现在通过全

局谱近似来进行插值。当需要还原时间和空间振荡时，即 $\Delta x = O(\varepsilon)$ 和 $\Delta t = O(\varepsilon)$，我们证明我们的方法在空间具有谱精度和一阶时间精度。我们还在 Wigner 变换的框架中证明，允许不依赖于 $\varepsilon$ 时间步长来计算正确的物理可观察量。

我们进行了大量的数值实验来验证我们的方法。通过高阶分裂方案可以容易地提高时间方向的精度。在数值中，我们使用 Strang 算子分裂。在一维情况下，我们已经验证了该方法在时间上二阶收敛，在空间上谱收敛。我们还表明，当计算物理可观测量时，没有必要还原时间振荡。我们同时给出了二维和三维数值实验。

本章的其余部分按以下方式组织。在 4.1 节中，我们提出了数值方法的详细结构以及对 NUFFT 算法的简要回顾。在 4.2 节中提供了波函数的严格稳定性分析和误差估计，其中我们还分析了计算物理可观测量时网格划分策略。在 4.3 节，我们提出了各种数值测试来验证我们的方法的性质。我们在最后一节总结一些评论和未来的方向。

## 4.1 数值方法

### 4.1.1 时间算子分裂与谱逼近

在本节中，我们将应用在 [3-4] 中引入的时间分裂谱方法。为了简单起见，我们考虑周期边界条件的一维问题。对多维情况的可以通过张量积直接推广。这里，我们只描述一阶时间分裂格式，在 [3] 中描述了对高阶格式的推广。

考虑在计算区域 $[a,b]$ 上的均匀网格 $x_j = a + j\Delta x$, $j = 0, \ldots, N-1$, 其中 $\Delta x = (b-a)/N$，$N$ 正的偶数。时间步长 $\Delta t$，定义 $t_n = n\Delta t$，$U^n = (U_0^n, \ldots, U_{N-1}^n)^T$ 的分量 $U_j^n$ 为 $u^\varepsilon(x_j, t_n)$ 的数值近似，$V_j$ 是 $V(x_j)$ 的数值近似。

我们考虑一维薛定谔方程，即库仑规范下的 (4–1)，

$$i\varepsilon\partial_t u^\varepsilon = -\frac{\varepsilon^2}{2}\Delta u^\varepsilon + i\varepsilon\mathbf{A}\cdot\nabla u^\varepsilon + \frac{1}{2}|\mathbf{A}|^2 u^\varepsilon + Vu^\varepsilon, \quad a < x < b, \quad t > 0, \tag{4–10}$$

和周期边界条件

$$u^\varepsilon(a,t) = u^\varepsilon(b,t), \quad u_x^\varepsilon(a,t) = u_x^\varepsilon(b,t), \tag{4–11}$$

及初值

$$u^\varepsilon(x,0) = u_0^\varepsilon(x). \tag{4–12}$$

注意，在一维情况下，$\mathbf{A}$ 是一个标量函数，这时势函数的规范计没有明确定义，但方程(4–10)数值方法可以推广到多维情况。在时间算子分裂法的框架中，要将(4–10)从 $t_n$ 演变为 $t_{n+1}$，我们可以先街薛定谔方程的动能算子部分，

$$i\varepsilon\partial_t u^\varepsilon = -\frac{\varepsilon^2}{2}\Delta u^\varepsilon, \quad t \in [t_n, t_{n+1}], \tag{4–13}$$

接着解势函数的部分

$$i\varepsilon\partial_t u^\varepsilon = \frac{1}{2}|\mathbf{A}|^2 u^\varepsilon + Vu^\varepsilon, \quad t \in [t_n, t_{n+1}], \tag{4–14}$$

最后解对流的部分

$$\partial_t u^\varepsilon = \mathbf{A}\cdot\nabla u^\varepsilon, \quad t \in [t_n, t_{n+1}]. \tag{4–15}$$

为了数值解上面的方程，我们首先引入如下函数空间 $S_N$

$$S_N = \text{span}\{e^{i\mu_k(x-a)},\ \mu_k = (2\pi k)/(b-a) \quad k = -N/2, \ldots, N/2-1\}. \tag{4–16}$$

设 $\text{II}_\text{N} : S_\text{p} := \{u(x)|u \in C^1([a,b]), u(a) = u(b), u'(a) = u'(b)\} \to S_N$ 为标准的投影算子 [106], 即

$$(\text{II}_\text{N}\, u)(x) = \sum_{k=-N/2}^{N/2-1} \widetilde{u}_k\, e^{i\mu_k\,(x-a)}, \quad x \in [a,b], \qquad \forall\, u(x) \in S_\text{p}, \tag{4–17}$$

以及

$$\widetilde{u}_k = \frac{1}{b-a}\int_a^b u(x)\, e^{-i\mu_k\,(x-a)}dx, \qquad k = -N/2, \ldots, N/2-1. \tag{4–18}$$

为了计算傅立叶系数 $\widetilde{u}_k$，我们用数值积分（梯形法）在均匀网格点上来近似(4–18)中的积分，并且所得的求和是由 FFT 实现。等价地，这个数值近似允许我们在网格点上定义 $u(x)$ 的插值，如下

$$u_I(x) = \sum_{k=-N/2}^{N/2-1} \hat{u}_k\, e^{i\mu_k(x-a)}, \quad x \in [a,b], \tag{4–19}$$

其中

$$\hat{u}_k = \frac{1}{N}\sum_{j=0}^{N-1} U_j\, e^{-i\mu_k(x_j-a)} = \frac{1}{N}\sum_{j=0}^{N-1} U_j e^{-i\frac{2\pi jk}{N}}, \quad k = -N/2, \ldots, N/2-1. \tag{4–20}$$

数值上，我们可以在傅立叶空间精确的得到(4–13)的解

$$U_j^* = \sum_{k=-N/2}^{N/2-1} e^{-i\frac{\Delta t}{2}\varepsilon\mu_k^2}\, \hat{u}_k^n\, e^{i\mu_k(x_j-a)}, \tag{4–21}$$

势能的方程由可以在物理空间得到精确解

$$U_j^{**} = e^{-i(\frac{1}{2}|\mathbf{A}|^2 u_j^* + V_j)\Delta t/\varepsilon}\, U_j^*. \tag{4–22}$$

一般来说，对于任意向量势函数 $\mathbf{A}(x)$，不可能得到对流方程(4–15)的精确解。虽然许多数值方法可用于解该方程，但是大多数主要的方法具有 CFL 条件，使得时间步长不能选的很大。为了用较大的时间步长来得到物理观测量，有必要应用无条件稳定的方法。

为了用改进的稳定性条件数值地解对流方程，在 [3] 中提出了具有多项式插值的半拉格朗日方法。半拉格朗日方法包括沿着特性线回溯和插值。准确地说，我们用周期边界条件求解对流方程

$$\partial_t u^\varepsilon - \mathbf{A}\cdot\nabla u^\varepsilon = 0, \quad t \in [t_n, t_{n+1}]. \tag{4–23}$$

对应的特征线方程为

$$\frac{dx(t)}{dt} = -\mathbf{A}(x(t)), \quad x(t_{n+1}) = x_j. \tag{4–24}$$

我们令常微分方程(4–24)的数值解 $x(t_n) = x_j^0$。沿着特性线，我们有 $u^\varepsilon(x_j, t_{n+1}) = u^\varepsilon(x_j^0, t_n)$。然而，由于经移位的目标点不一定是网格点，因此需要插值以近似估计 $u^\varepsilon(x_j^0, t_n)$。我们可以采用全局谱插值，例如傅里叶伪谱差值，或局部多项式插值，例如 $M$ 阶拉格朗日多项式插值 [3]。

当应用局部多项式插值时，对于每个移位的目标点 $x_j^0$，需要从其相邻网格点构建多项式插值。以 $M$ 阶的拉格朗日多项式插值为例，对于每个 $x_j^0$，选择 $M$ 个相邻的网格点来构造具有 $O(\Delta x^M)$ 误差的拉格朗日多项式插值。局部多项式插值的总计算量每个时间步为 $O(N)$。

而在全局谱插值中，我们首先在网格点上构建谱插值，然后算出在移位的目标点处的插值函数。对于傅立叶谱插值，所有傅里叶系数在用 FFT 的计算量为 $O(N\log N)$。值得指出的是，FFT 不能用于计算在移位的目标点处的差值函数，因为 $\{x_j^0\}$ 不一定是均匀分布的网格点。对每个目标点的有限傅立叶级数的直接估计 $x_j^0$ 需要 $O(N)$ 的计算量。因此，总评估过程成本是 $O(N^2)$，并且它是相当耗时的，特别是在高维时。

与局部多项式插值相比，傅里叶插值在空间上是谱精度的，但是在效率方面是瓶颈。使用 NUFFT 算法 [43]，我们可以把效率从 $O(N^2)$ 提高到到 $O(N\log N)$，而不牺牲其谱精度。注意，NUFFT 仅仅是用于计算这里遇到的离散傅里叶加法的快速算法。我们将介绍使用 NUFFT 的半拉格朗日方法给出一个简要的介绍。

### 4.1.2　使用 NUFFT 的半拉格朗日方法解对流方程

按照前面一节的分析，为了使用半拉格朗日方法解对流方程 (4–15)，我们需要计算

$$U_j^{n+1}=u^\varepsilon(x_j^0,t_n)\approx\sum_{k=-N/2}^{N/2-1}\hat{u}_k^n e^{i\mu_k(x_j^0-a)}=\sum_{k=-N/2}^{N/2-1}\hat{u}_k^n e^{ik\frac{2\pi(x_j^0-a)}{b-a}}. \tag{4–25}$$

一般来说，移位的目标点 $x_j^0$ 不一定均匀分布，因此由于变换矩阵的代数结构的破坏，FFT 不适用于傅里叶级数求和(4–25)。直接求和，需要 $O(N^2)$ 的运算量，将大大影响效率，特别是对于二维和三维问题。使用 NUFFT 算法，计算(4–25)可以在 $O(N\log N)$ 的复杂度内完成。与 $O(N^2)$ 复杂度相比，效率提高到 $O(N\log N)$ 是相当可观的。然后我们将 NUFFT 算法结合到半拉格朗日方法中，并将改进方法的细节给出如下

**算法 4–1** 使用 NUFFT 的半拉格朗日方法

1: 回溯解出错位的目标点 $x_j^0$.
2: 使用 FFT，由网格点 $x_j$ 计算出 $u^\varepsilon(x,t_n)$ 的傅立叶谱插值。
3: 使用 NUFFT 计算(4–25)中的 $U_j^{n+1}=u^\varepsilon(x_j^0,t_{n+1})$。

算法4–1的总计算复杂度由三部分组成。步骤 1 可以通过 $O(N)$ 的 ODE 求解器求解。步骤 2 中的傅立叶系数可以由 $O(N\log N)$ 复杂度的向前 FFT（forward FFT）来计算。步骤 3 中的计算可以在 NUFFT 的 $O(N\log N+N)$ 操作内完成。总之，总复杂度是 $O(N+N\log N)$，空间精度从多项式精度提高到谱精度。我们将给出出数值验证。对多维情况的推广是简单和直接的。

### 4.1.3　NUFFT 算法简介

在本节中，我们将简要介绍一下 NUFFT 算法。该算法旨在加速傅里叶级数的计算，其涉及物理空间和频率空间的不均匀点，复杂度最多为 $O(N\log N)$。该算法有有很多版本，我们在这里按照 [43] 中描述的使用高斯内核进行插值的简单且快速的实现。

我们在一维中定义了类型 1 和 2 的不均匀离散傅里叶变换，如下

$$\text{类型 1:} \qquad F(k) = \sum_{j=0}^{N-1} f_j e^{-ikx_j}, \quad k = -M/2, \ldots, M/2-1, \tag{4–26}$$

$$\text{类型 2:} \qquad f(x_j) = \sum_{k=-M/2}^{M/2-1} F(k) e^{ikx_j}, \quad j = 0, 1, \ldots, N, \tag{4–27}$$

其中 $x_j \in [0, 2\pi]$ 是不均匀的格点，$f_j$ 为复数。

简单起见，为了说明基本的基本思想，我们考虑一维类型 1 的情况。在我们的半拉格朗日方法中使用的类型 2 求和可以看作是类型 1 的逆，可以参见 [43]。请注意，方程 (4–26) 描述了函数的精确傅里叶系数

$$f(x) = \sum_{j=0}^{N-1} f_j \delta(x - x_j), \tag{4–28}$$

可以看作 $[0, 2\pi]$ 上的周期函数。这里 $\delta(x)$ 是狄拉克 $\delta$ 函数。通过与一维热核（heat kernal）在 $[0, 2\pi]$ 上进行卷积，即 $g_\tau(x) = \sum_{l=-\infty}^{\infty} e^{-(x-2l\pi)^2/4\tau}$，我们可以构建一个周期为 $2\pi$，$C^\infty$ 的函数 $f_\tau$ 如下

$$f_\tau(x) = f * g_\tau(x) = \int_0^{2\pi} f(y) g_\tau(x-y)\, dy. \tag{4–29}$$

实际上，$f_\tau$ 是 $f$ 的一个很好的逼近并且可以在 $x$ 的均匀网格上被计算出来。傅立叶系数 $F_\tau(k) = \frac{1}{2\pi}\int_0^{2\pi} f_\tau(x) e^{-ikx}\, dx$ 可由标准的 FFT 在一个过抽样（oversampled）的网格上准确的逼近，

$$F_\tau(k) \approx \frac{1}{M_r} \sum_{m=0}^{M_r-1} f_\tau(2\pi m/M_r) e^{-ik2\pi m/M_r}, \tag{4–30}$$

这里

$$f_\tau(2\pi m/M_r) = \sum_{j=0}^{N-1} f_j g_\tau(2\pi m/M_r - y_j). \tag{4–31}$$

一旦我们知道了 $F_\tau(k)$，根据卷积理论我们有，

$$F(k) = \sqrt{\frac{\pi}{\tau}} e^{k^2\tau} F_\tau(k). \tag{4–32}$$

相关参数的最佳选择涉及一些深入的分析，我们在此省略。根据在 [43] 中的讨论，我们选择 $M_r = 2M$ 和 $\tau = 12/M^2$，并使用高斯函数将每个源扩展到最近的 24 个点，那么它产生大约 12 位数的精度。对于 6 位数精度，我们选择 $\tau = 6/M^2$ 并将每个源分散到最接近的 12 点。在数值实验中，如果没有另外说明，我们选择 12 位精度。

## 4.2 数值分析

在这节中，我们将研究波函数与物理观测量的稳定性、收敛性。

### 4.2.1　稳定性分析

对任意函数 $u(x) \in S_{\mathrm{p}}$，令 $\mathbf{U} = (u(x_0), \cdots, u(x_{N-1}))^T$ 为 $u$ 的网格向量。定义 $\|\cdot\|_{l^2}$ 为离散的 $l^2$ 范数，$\|\cdot\|_{L^2}$ 为函数空间 $S_{\mathrm{p}}$ 的 $L^2$ 范数

$$\|\mathbf{U}\|_{l^2} = \Big(\Delta x \sum_{j=0}^{N-1} |U_j|^2\Big)^{1/2}, \quad \|u\|_{L^2} = \Big(\int_a^b |u(x)|^2\, dx\Big)^{1/2}. \tag{4–33}$$

经过简单的计算，我们有 $\|u_I(x)\|_{L^2} = \|U\|_{l^2}$ 对于任意的 $u \in S_{\mathrm{p}}$ 成立。

**引理　4.1.** *对于每个时间步 $t \in [t_n, t_{n+1}]$，在解动能方程(4–13)和势能方程(4–14)后，我们有*

$$\|\mathbf{U}^{**}\|_{l^2} = \|\mathbf{U}^n\|_{l^2}. \tag{4–34}$$

**证明.** 证明和在 [4] 中的非常类似，简洁起见我们将其省略。 □

为了用半拉格朗日方法求解对流方程,首先沿着特征回溯。特征线方程在任意给定的初值点 $x_0 \in [a, b]$,

$$\frac{dx(t)}{dt} = -\mathbf{A}(x(t)), \quad x(t_0) = x_0, \quad x_0 \in [a, b]. \tag{4–35}$$

我们定义 $S_{\mathrm{p}}$ 上的映射 $E(t, t_0)$ 如下

$$(E(t, t_0)v)(x_0) := v(x(t)), \quad v \in S_{\mathrm{p}}. \tag{4–36}$$

由于 (4–35) 是一个自治系统，所以 $E(t, t_0)$ 仅是 $t - t_0$ 的函数。因此，我们将用 $E(t - t_0)$ 表示 $E(t, t_0)$。如果来自回溯和谱插值的误差是可以忽略的，即 $u_I^{n+1}(x_j)$ 是 “精确的”，使用 NUFFT 的半拉格朗日方法可以描述为

$$u_I^{n+1}(x) = \mathrm{II}_N E(\Delta t) u_I^{**}(x), \tag{4–37}$$

其中 $u_I^{**}(x)$ 是 $u^{**}(x)$ 的谱插值。

**引理　4.2.** *假设 $\mathbf{A} \in C^1([a, b])$ 并且是无旋的，即 $\nabla \cdot \mathbf{A} = 0$, 则半拉格朗日格式 (4–37) 是无条件稳定的并且我们有*

$$\|u_I^{n+1}\|_{L^2} \le \|u_I^{**}\|_{L^2}. \tag{4–38}$$

**证明.** 由 (4–37),

$$\|u_I^{n+1}\|_{L^2} = \|\mathrm{II}_N E(\Delta t) u_I^{**}\|_{L^2} \le \|E(\Delta t) u_I^{**}\|_{L^2}, \tag{4–39}$$

因为 $\mathbf{A}$ 是无旋的，我们有 $\|E(\Delta t)\|_{(L^2)^*} \le 1$ 及 $\|u_I^{n+1}\|_{L^2} \le \|u_I^{**}\|_{L^2}$. □

**注　8.** 该引理很容易推广到更一般的 $\mathbf{A}$，可以参见 [107] 中更多的讨论。

结合引理4.1和4.2，我们得到如下的稳定性结果，

**定理　4.3.** *使用 NUFFT 的半拉格朗日时间分裂谱方法，(4–21)、(4–22)及(4–25)，是无条件稳定的。事实上，对于任何网格大小和时间步*

$$\|\mathbf{U}^{n+1}\|_{l^2} \le \|\mathbf{U}^n\|_{l^2}, \quad n = 1, 2, \cdots \tag{4–40}$$

**证明.** 根据引理4.1，$\|\mathbf{U}^{**}\|_{l^2} = \|\mathbf{U}^n\|_{l^2}$。由引理4.2，我们有

$$\|\mathbf{U}^{n+1}\|_{l^2} = \|u_I^{n+1}\|_{L^2} \leq \|u_I^{**}\|_{L^2} = \|\mathbf{U}^{**}\|_{l^2} = \|\mathbf{U}^n\|_{l^2}.$$

□

### 4.2.2 波函数的误差估计

在这节中，我们研究波函数数值逼近的误差及网格策略。我们假设波函数载空间和时间上都是 $\varepsilon$ 振荡的。更具体地说，存在独立于 $t$，$x$ 和 $\varepsilon$ 的正的常数 $B_m$，$C_m$，$D_m$ 使得

$$\left\| \frac{\partial^{m_1+m_2}}{\partial x^{m_1}\partial t^{m_2}} u(x,t) \right\|_{C([0,T];L^2)} \leq \frac{1}{\varepsilon^{m_1+m_2}} C_{m_1+m_2}, \quad m = m_1 + m_2,\ m_1, m_2 \in \mathbb{N}^+, \tag{4–41}$$

$$\left\| \frac{\partial^m}{\partial x^m} \mathbf{A}(x) \right\|_{L^2} \leq D_m, \qquad \left\| \frac{\partial^m}{\partial x^m} V(x) \right\|_{L^2} \leq B_m. \tag{4–42}$$

注意，在4–41中，差分运算符对于一般光滑函数是无界的，但它在光滑 $L^2$ 函数的子空间中是有界的，最多为 $\varepsilon$ 振荡的。假设(4–42)意味着这些势能函数是光滑的，并且上解独立于 $\varepsilon$。我们使用 $f_I$ 来表示基于前面部分提到的离散数据 $f(x_j)$ 的谱近似。现在我们可以证明使用 NUFFT 的半拉格朗日一阶时间分裂谱方法法（缩写为 SL-TS）的以下误差估计。

证明基本上遵循 [3] 中的定理 4 和 [4] 中的定理 4.1，但与以前的版本不同，因为这是在向量势函数存在的情况下显示了空间中的谱精度。

**定理 4.4.** *设 $u^\varepsilon(x,t)$ 是方程(4–10)的精确解，$u^{\varepsilon,n}$ 是用一阶 SL-TS 方法的离散逼近。我们假设能够以可以忽略的误差数值解特征线方程(4–24)，而且在 SL-TS 方法中对流步的 NUFFT 的误差可以忽略。在假设(4–41)-(4–42)下，我们进一步假设 $\Delta x = O(\varepsilon)$ 和 $\Delta t = O(\varepsilon)$，那么对于任何时间 $t \in [0,T]$，我们有*

$$\|u^\varepsilon(t_n) - u_I^{\varepsilon,n}\|_{L^2} \leq G_m \frac{T}{\Delta t} \left( \frac{\Delta x}{\varepsilon} \right)^m + \frac{CT\Delta t}{\varepsilon}, \tag{4–43}$$

*其中 $m \in \mathbb{N}^+$ 是 $u^\varepsilon(x,t)$ 的正则指标，$C$ 为独立于 $\Delta t$，$\Delta x$，$\varepsilon$，$m$ 的正常数，$G_m$ 是不依赖于 $\Delta t$, $\Delta x$ 和 $\varepsilon$ 的正常数。*

**证明.** 在证明中，如果没有明确说明，所涉及的常数被假定为独立于 $\varepsilon$。为了清楚起见，我们重写方程(4–10)

$$\partial_t u^\varepsilon = (\mathcal{A} + \mathcal{B} + \mathcal{C}) u^\varepsilon, \tag{4–44}$$

其中

$$\mathcal{A} = \frac{i\varepsilon}{2}\Delta, \quad \mathcal{B} = -\frac{i}{\varepsilon}\left( \frac{1}{2}|\mathbf{A}|^2 + V \right), \quad \mathcal{C} = \mathbf{A} \cdot \nabla.$$

设 $u^\varepsilon(t_n)$ 是在 $t = t_n$ 时的精确解，那么

$$u^\varepsilon(t_{n+1}) = e^{(\mathcal{A}+\mathcal{B}+\mathcal{C})\Delta t} u^\varepsilon(t_n). \tag{4–45}$$

定义由（一阶）时间算子分裂获得的解（无空间离散）为

$$w^{n+1} = e^{\mathcal{C}\Delta t} e^{\mathcal{B}\Delta t} e^{\mathcal{A}\Delta t} u^\varepsilon(t_n). \tag{4–46}$$

请注意，由于算子分裂误差，$w^{n+1}$ 与 $u^\varepsilon(t_{n+1})$ 不同。如 [3] 所示，局部分裂误差是

$$\|u^\varepsilon(t_{n+1}) - w^{n+1}\|_{L^2} = O\left(\frac{\Delta t^2}{\varepsilon}\right). \tag{4–47}$$

由三角不等式，得到

$$\|u^\varepsilon(t_{n+1}) - u_I^{\varepsilon,n+1}\|_{L^2} \le \|u^\varepsilon(t_{n+1}) - w^{n+1}\|_{L^2} + \|w^{n+1} - w_I^{n+1}\|_{L^2} + \|w_I^{n+1} - u_I^{\varepsilon,n+1}\|_{L^2}, \tag{4–48}$$

其中 $w_I^{n+1}$ 是 $w^{n+1}$ 的谱插值逼近。(4–48)的第一项是分裂误差(4–47)，第二项给出了谱逼近的误差，其上界为 $C_m(\frac{\Delta x}{\varepsilon})^m$。到目前为止，分析与以前的结果一致。但是，我们需要分析最后一项，这是通过数值近似引入的误差项。特别的，由于在对流步骤中利用了谱逼近，因此产生的误差与 [3] 不同，因此需要仔细研究。

在 SL-TS 方法中，由算子 $\mathcal{B}$ 控制的势能步是通过解析求解的，而算子 $\mathcal{A}$ 和 $\mathcal{C}$ 的动能步骤和对流步是通过数值近似演化的，分别由 $\mathcal{A}_{SP}$ 和 $\mathcal{C}_{SL}$ 表示。由三角不等式：

$$\begin{aligned}
\|w_I^{n+1} - u_I^{\varepsilon,n+1}\|_{L^2} &= \|w^{n+1} - u^{\varepsilon,n+1}\|_{l^2} \\
&= \|e^{\mathcal{C}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}\Delta t}u^\varepsilon(t_n) - e^{\mathcal{C}_{SL}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^{\varepsilon,n}\|_{l^2} \\
&\le \|e^{\mathcal{C}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}\Delta t}u^\varepsilon(t_n) - e^{\mathcal{C}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^\varepsilon(t_n)\|_{l^2} \\
&+ \|e^{\mathcal{C}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^\varepsilon(t_n) - e^{\mathcal{C}_{SL}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^\varepsilon(t_n)\|_{l^2} \\
&+ \|e^{\mathcal{C}_{SL}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^\varepsilon(t_n) - e^{\mathcal{C}_{SL}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^{\varepsilon,n}\|_{l^2}.
\end{aligned} \tag{4–49}$$

(4–49)右端的第一项表示 $u^\varepsilon(t_n)$ 在动能步的谱逼近，所以如同在 [3-4] 中的分析，这项的误差为是 $O((\frac{\Delta x}{\varepsilon})^m)$（对于任意的正整数 $m$）。(4–49)右端的第二项表示 $e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^\varepsilon(t_n)$ 在对流步的数值逼近。根据对动能步和势能步的稳定性分析，

$$\|e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^\varepsilon(t_n)\|_{l^2} = \|u^\varepsilon(t_n)\|_{l^2}.$$

然后，第 2 节中的局部误差分析意味着，当计算偏移的网格点和 NUFFT 的误差很小时，第二项是 $O((\frac{\Delta x}{\varepsilon})^m)$，$m$ 为任意正整数，这里谱插值的误差占主导。

(4–49)右端的最后一项连接两个连续时间步长之间的数值解的数值误差，其中数值稳定性至关重要。也可以很容易地看出，运算符 $e^{\mathcal{A}\Delta t}$，$e^{\mathcal{B}\Delta t}$ 和 $e^{\mathcal{C}\Delta t}$（在库仑规范下）是 $L^2$ 范数下中的周期光滑函数类的酉运算符，这意味着 $\|e^{\mathcal{A}\Delta t}\|_{L^2} = \|e^{\mathcal{B}\Delta t}\|_{L^2} = \|e^{\mathcal{C}\Delta t}\|_{L^2} = 1$。通过上一节的稳定性分析，我们已经证明，

$$\|e^{\mathcal{A}_{SP}\Delta t}\|_{L^2} = 1 \quad \|e^{\mathcal{C}_{SL}\Delta t}\|_{L^2} \le 1.$$

因此，我们得出以下估估计，即(4–49)右侧的最后一项

$$\begin{aligned}
&\|e^{\mathcal{C}_{SL}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^\varepsilon(t_n) - e^{\mathcal{C}_{SL}\Delta t}e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^{\varepsilon,n}\|_{l^2} \\
&\le \|e^{\mathcal{C}_{SL}\Delta t}\|_{L^2}\|e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^\varepsilon(t_n) - e^{\mathcal{B}\Delta t}e^{\mathcal{A}_{SP}\Delta t}u^{\varepsilon,n}\|_{l^2} \\
&\le \|e^{\mathcal{C}_{SL}\Delta t}\|_{L^2}\|u^\varepsilon(t_n) - u_I^{\varepsilon,n}\|_{L^2} \\
&\le \|u^\varepsilon(t_n) - u_I^{\varepsilon,n}\|_{L^2}.
\end{aligned} \tag{4–50}$$

推出

$$\|w_I^{n+1} - u_I^{\varepsilon,n}\|_{L^2} \le \|u^\varepsilon(t_n) - u_I^{\varepsilon,n}\|_{L^2} + C'_m \left(\frac{\Delta x}{\varepsilon}\right)^m, \tag{4–51}$$

$C'_m$ 是独立于 $t$，$x$，$\varepsilon$ 的常数。现在我们有递推关系

$$\|u^\varepsilon(t_{n+1}) - u_I^{\varepsilon,n+1}\|_{L^2} \le \|u^\varepsilon(t_n) - u_I^{\varepsilon,n}\|_{L^2} + C_1(\frac{\Delta x}{\varepsilon})^m + C_2(\frac{\Delta t^2}{\varepsilon}), \tag{4–52}$$

$C_1$，$C_2$ 是独立于 $t$，$x$，$\varepsilon$ 的常数。

由 $\|u^\varepsilon(t_n) - u_I^{\varepsilon,n}\|_{L^2}$ 的递推关系和归纳法，我们有

$$\|u^\varepsilon(t_n) - u_I^{\varepsilon,n}\|_{L^2} \le G_m \frac{T}{\Delta t}(\frac{\Delta x}{\varepsilon})^m + \frac{CT\Delta t}{\varepsilon}. \tag{4–53}$$

证毕。 □

这个定理意味着如果 $\delta > 0$ 是 $L^2$ 范数中的误差界，那么 $\|u^\varepsilon(t_n) - u_I^{\varepsilon,n}\|_{L^2} < \delta$，相应的网格划分策略是

$$\frac{\Delta t}{\varepsilon} = O(\delta), \quad \frac{\Delta x}{\varepsilon} = O(\delta^{1/m}\Delta t^{1/m}), \tag{4–54}$$

$m \ge 1$ 为任意整数。对于高阶算子分裂技术，可以进行类似的分析，这在本文中被省略。

我们指出，网格划分策略(4–54)与 [4] 中的结果（没有向量势函数）一致，并且明显优于 [3]，在这篇参考文献中多项式插值应用于半拉格朗日方法，因此整个数值方案在空间上不具有谱精度。

### 4.2.3 物理观测量的误差估计

一般来说，如果只关心物理观测值，则网格划分策略中的条件较弱可能就足够了（参见 [3-4]），其中可以使用 Wigner 变换来说明这一点。对于 $f, g \in L^2(\mathbb{R}^d)$，Wigner 变换被定义为相空间函数

$$w^\varepsilon(f,g)(t,x,\xi) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{iy\cdot\xi}\bar{f}(x - \frac{\varepsilon}{2}y)g(x + \frac{\varepsilon}{2}y)\,dy. \tag{4–55}$$

定义 $w^\varepsilon = w^\varepsilon(u^\varepsilon, u^\varepsilon)$，当 $\varepsilon \to 0$ 时，Wigner 变换收敛到 Wigner 测度 $w^0 = \lim\limits_{\varepsilon\to 0} w^\varepsilon(u^\varepsilon, u^\varepsilon)$，其中的收敛是在弱意义下的。

设 $a(x,\xi)$ 是一个光滑的实值相空间函数，在无限远处有足够的衰减，称为半经典符号。那么，自共轭的拟微分算子 $A^\varepsilon := a(x, \varepsilon D)^W$ 被称为可观察值，这里 $D = i\nabla_x$ 和 $W$ 表示 Weyl 量化。那么在这种状态下可观察到的平均值被定义为

$$E_a^\varepsilon(t) = \int_{\mathbb{R}^d} \bar{u}^\varepsilon(t,x)(a(x,\varepsilon D)^W u^\varepsilon(t,x))\,dx. \tag{4–56}$$

一个重要的性质是对偶等式

$$\int_{\mathbb{R}^d} \bar{u}^\varepsilon(t,x)(a(x,\varepsilon D)^W u^\varepsilon(t,x))\,dx = \int_{\mathbb{R}^d\times\mathbb{R}^d} w^\varepsilon(t,x,\xi)a(x,\xi)\,dxd\xi. \tag{4–57}$$

$E_a^\varepsilon(t)$ 可以取半经典极限

$$\lim_{\varepsilon\to 0} E_a^\varepsilon(t) = \int_{\mathbb{R}^d\times\mathbb{R}^d} w^0(t,x,\xi)a(x,\xi)\,dxd\xi. \tag{4–58}$$

$\tilde{w}^\varepsilon$ 是数值近似解的 Wigner 变换。可以很容易地证明以下的不等式

$$|E_a^\varepsilon - \tilde{E}_a^\varepsilon| \le \|a\|_{\mathcal{E}} \cdot \|w^\varepsilon - \tilde{w}^\varepsilon\|_{\mathcal{E}^*} \le C\|a\|_{\mathcal{E}} \cdot \|u^\varepsilon - \tilde{u}^\varepsilon\|_{L^2(a,b)}, \tag{4–59}$$

对于 $a \in \mathcal{E}$，是如下的巴拿赫空间

$$\mathcal{E} = \left\{\phi \in C_0(\mathbb{R}_x^d \times \mathbb{R}_\xi^d) : (\mathcal{F}_{\xi\to v}\phi) \in L^1(\mathbb{R}_v^d; C_0(\mathbb{R}_x^d))\right\}.$$

$\mathcal{F}$ 表示傅里叶变换，$\mathcal{E}^*$ 是 $\mathcal{E}$ 的对偶空间。我们指出，当波函数在无穷远下衰减得足够快时，可以对 Banach 空间进行延拓。

在算子分裂后的每个时间步长 $t \in [t_n, t_{n+1}]$ 中，波函数由于谱逼近和 NUFFT（很小）导致。根据定理4.4及上面的不等式(4–59)，可以估计相应 Wigner 变换中的误差。

估计(4–59)意味着，空间网格划分策略 $\Delta x/\varepsilon = O(\delta^{1/m}\Delta t^{1/m})$ 足以保证在时间间隔 $[0,T]$ 上由谱插值引起的所有物理观察值中的 $O(\delta)$ 误差。

如 [3] 中讨论的计算物理观测值的分裂误差为 $O(\varepsilon)$，因为经典极限方程独立于 $\varepsilon$，并且薛定谔方程的时间分裂对应于 Wigner 方程的分裂。在动能步骤和势能步骤中，时间的演化是精确的，在对流部分中，向后特征回溯在预处理步骤中完成（足够精细的、独立于 $\varepsilon$ 的时间步长）因此，在时间离散化中，根本没有依赖于 $\varepsilon$ 的误差。

在所有这些考虑之后，我们得出结论，使用 NUFFT 的 SL-TS 可以用于捕获正确的物理观测值。这意味着用时间步长 $\Delta t = O(\delta)$ 和空间网格划分策略 $\Delta x = O(\varepsilon)$，数值解在 Wigner 变换中得到 $O(\delta)$ 误差（当 $\varepsilon \to 0$ 时），因此所有物理观察值的误差为 $O(\delta)$。

## 4.3　数值例子

在本节中，我们将通过大量的一维数值例子来确认所提方法的准确性和有效性，并在二维和三维情况下提供仿真实例。参考解是通过具有精细网格大小和时间步长的时间显式谱方法（TESP）获得的，其中通过标准四阶 Runge-Kutta 方法来求解对流方程，空间导数近似由傅立叶谱方法 [3] 得到。波函数的误差以 $l^2$ 范数计算，而物理观测值的误差则被测量通过它们的累积函数的 $l^2$ 范数。

**例　4.1.** *依赖时间的向量势函数*

在这个例子中，我们使用与 [3] 中相同的一维示例，计算域为 $C = [0, 2\pi]$，最终时间为 $T = 0.4$。标量势为 $V(x) = 1$，向量势为 $\mathbf{A}(x,t) = \sin(x - 2t)/10$。初始值为 $u_0(x) = e^{-10(x-\pi)^2} e^{i\cos(x)/\varepsilon}$。

波函数，位置密度和电流密度的误差定义如下：

$$E_u = \|u_{\Delta t}^N - u^{\text{ref}}\|_{l^2}, \quad E_n = \|\tilde{n}_{\Delta t}^N - \tilde{n}^{\text{ref}}\|_{l^1}, \quad E_I = \|\tilde{I}_{\Delta t}^N - \tilde{I}^{\text{ref}}\|_{l^1}, \tag{4–60}$$

其中 $u_{\Delta t}^N$ 为数值解（$\Delta x = \frac{2\pi}{N}$，$\Delta t$），$u^{\text{ref}}$ 是有非常精细的网格用 TESP 方法得到的。$\tilde{n}, \tilde{I}$ 是相应的累积函数

$$\tilde{n}(x) = \int_0^x n(s)ds, \qquad \tilde{I}(x) = \int_0^x I(s)ds, \quad 0 < x < 2\pi. \tag{4–61}$$

电流密度定义如下，$I(t,x) = \varepsilon \text{Im}(\overline{u}^\varepsilon(t,x)\nabla_x u^\varepsilon(t,x)) = \frac{\varepsilon}{2i}(\overline{u}^\varepsilon \nabla_x u^\varepsilon - u^\varepsilon \nabla_x \overline{u^\varepsilon})$。

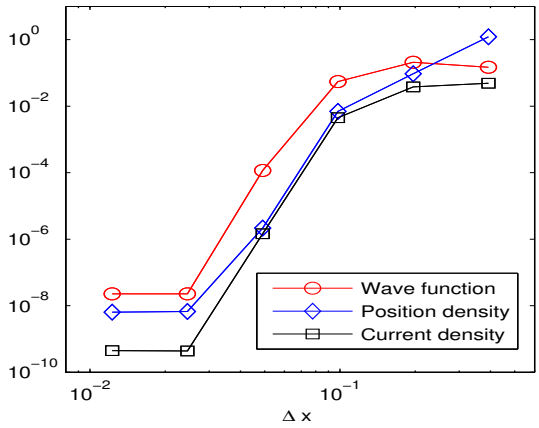


图 4–1 例4.1中波函数的 $l^2$ 误差，位置密度的 $l^1$ 误差和电流密度的 $l^1$ 误差与 $\Delta x$ 的对数曲线，$\varepsilon = 1/32$。
Fig 4–1 Log-log plot of errors of the wave function ($l^2$ norm) , the position densities ($l^1$ norm), and current densities ($l^1$ norm) versus mesh sizes $\Delta x$ for $\varepsilon = 1/32$ in Example 4.1.

我们在表4–1和图4–1中看出，当 $\Delta x = O(\varepsilon)$ 时，使用足够的精细时间步长 $\Delta t = 10^{-6}\varepsilon$，随着空间网格点的增加，波函数和物理观测值中的误差呈指数级降低，直到达到最小，这是由 NUFFT 算法中的误差所主导的。因此，如果 NUFFT 的误差可以忽略不计，我们已经确认了所提出的方法可以实现空间的谱精度。

对于不同的 $\varepsilon$，使用足够精细的空间网格大小，$\Delta x = \frac{2\pi}{32}\varepsilon$，在表4–2中列出关于时间步长的收敛关系，并将数字误差绘制在图4–2中。显然曲线表明，对于波函数和物理观察值，我们都有 $\Delta t$ 的二阶收敛。这与我们在时间分裂中使用的 Strang 分裂一致。另外，从表4–2我们看到，即使 $\Delta t \gg \Delta x$ 及 $\Delta t \gg \varepsilon$，数值方法仍然是稳定的。这验证了全时间分裂谱方法的无条件稳定性。

另一个重要观察是通过检查表4–2的每一列，我们看到，对于固定的 $\Delta t$，当 $\varepsilon$ 减少时，波函数中的误差成比例地增加，而物理观察中的误差几乎保持不变。特别地，我们观察到（例如，在表4–2的第一列中），当 $\Delta t \gg \Delta x$ 和 $\Delta t \gg \varepsilon$ 时，物理观测的准确性非常高。它证明我们可以采取 $\varepsilon$ 独立的时间步骤来捕获正确的物理观测值。如果我们只计算物理观察值，时间步长只需 $O(1/\varepsilon)$，如果需要模拟波函数，步长就要 $O(1/\varepsilon^2)$。

表 4–1 例4.1的空间误差。$\Delta x = \frac{2\pi}{N}$，$\Delta t = 10^{-6}\varepsilon$，$\varepsilon = 1/32$。参考解由 TESP 得出，$\Delta x = \frac{2\pi}{4096}$ 和 $\Delta t = 10^{-6}\varepsilon$。
Table 4–1 Spatial errors computed with $\Delta x = \frac{2\pi}{N}$ and very fine time step $\Delta t = 10^{-6}\varepsilon$ for $\varepsilon = 1/32$ in Example 4.1. Reference solution is obtained by TESP with $\Delta x = \frac{2\pi}{4096}$ and $\Delta t = 10^{-6}\varepsilon$.

| $N$ | 8 | 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|---|
| $E_u$ | 1.4844E-01 | 2.0897E-01 | 5.4851E-02 | 1.1685E-04 | 2.2790E-08 | 2.2602E-08 |
| $E_n$ | 1.2187 | 9.3894E-02 | 6.9707E-03 | 2.1646E-06 | 6.6955E-09 | 6.3930E-09 |
| $E_I$ | 4.8682E-02 | 3.8011E-02 | 4.5217E-03 | 1.4634E-06 | 4.3334E-10 | 4.4653E-10 |

表 4–2 时间方向误差，$\Delta x = \frac{2\pi}{32}\varepsilon$，$\Delta t_j = \frac{1}{10\times 2^j}, j = 1,\ldots,6$。参考解由 TESP 得到。

Table 4–2 Temporal errors computed with $\Delta x = \frac{2\pi}{32}\varepsilon$ and different time steps $\Delta t_j = \frac{1}{10\times 2^j}, j = 1,\ldots,6$ in Example 4.1. Reference solution is obtained by TESP with $\Delta x = \frac{2\pi}{32}\varepsilon$ and $\Delta t = 10^{-6}\varepsilon$.

| $E_u$ | $\Delta t_1$ | $\Delta t_2$ | $\Delta t_3$ | $\Delta t_4$ | $\Delta t_5$ | $\Delta t_6$ |
|---|---|---|---|---|---|---|
| $\varepsilon = \frac{1}{16}$ | 1.1461E-05 | 2.8641E-06 | 7.1595E-07 | 1.7898E-07 | 4.4740E-08 | 1.1181E-08 |
| $\varepsilon = \frac{1}{32}$ | 1.2923E-05 | 3.2295E-06 | 8.0728E-07 | 2.0181E-07 | 5.0446E-08 | 1.2606E-08 |
| $\varepsilon = \frac{1}{64}$ | 2.0866E-05 | 5.2144E-06 | 1.3034E-06 | 3.2585E-07 | 8.1458E-08 | 2.0361E-08 |
| $\varepsilon = \frac{1}{128}$ | 3.9232E-05 | 9.8038E-06 | 2.4507E-06 | 6.1266E-07 | 1.5316E-07 | 3.8293E-08 |
| $\varepsilon = \frac{1}{256}$ | 7.7212E-05 | 1.9295E-05 | 4.8231E-06 | 1.2057E-06 | 3.0140E-07 | 7.5325E-08 |
| $E_n$ | $\Delta t_1$ | $\Delta t_2$ | $\Delta t_3$ | $\Delta t_4$ | $\Delta t_5$ | $\Delta t_6$ |
| $\varepsilon = \frac{1}{16}$ | 1.2604E-06 | 3.1537E-07 | 7.8930E-08 | 1.9808E-08 | 5.0278E-09 | 1.3343E-09 |
| $\varepsilon = \frac{1}{32}$ | 1.0910E-06 | 2.7306E-07 | 6.8388E-08 | 1.7209E-08 | 4.4144E-09 | 1.2172E-09 |
| $\varepsilon = \frac{1}{64}$ | 1.0466E-06 | 2.6205E-07 | 6.5737E-08 | 1.6647E-08 | 4.3745E-09 | 1.3086E-09 |
| $\varepsilon = \frac{1}{128}$ | 1.0356E-06 | 2.5952E-07 | 6.5318E-08 | 1.6756E-08 | 4.6157E-09 | 1.5836E-09 |
| $\varepsilon = \frac{1}{256}$ | 1.0338E-06 | 2.5990E-07 | 6.6230E-08 | 1.7801E-08 | 5.6972E-09 | 2.6858E-09 |
| $E_I$ | $\Delta t_1$ | $\Delta t_2$ | $\Delta t_3$ | $\Delta t_4$ | $\Delta t_5$ | $\Delta t_6$ |
| $\varepsilon = \frac{1}{16}$ | 7.1615E-07 | 1.7901E-07 | 4.4738E-08 | 1.1171E-08 | 2.7788E-09 | 6.8085E-10 |
| $\varepsilon = \frac{1}{32}$ | 5.1362E-07 | 1.2839E-07 | 3.2084E-08 | 8.0066E-09 | 1.9871E-09 | 4.8226E-10 |
| $\varepsilon = \frac{1}{64}$ | 4.6330E-07 | 1.1581E-07 | 2.8932E-08 | 7.2115E-09 | 1.7815E-09 | 4.2398E-10 |
| $\varepsilon = \frac{1}{128}$ | 4.5076E-07 | 1.1267E-07 | 2.8138E-08 | 7.0051E-09 | 1.7219E-09 | 4.0109E-10 |
| $\varepsilon = \frac{1}{256}$ | 4.4759E-07 | 1.1183E-07 | 2.7889E-08 | 6.9018E-09 | 1.6551E-09 | 3.4344E-10 |

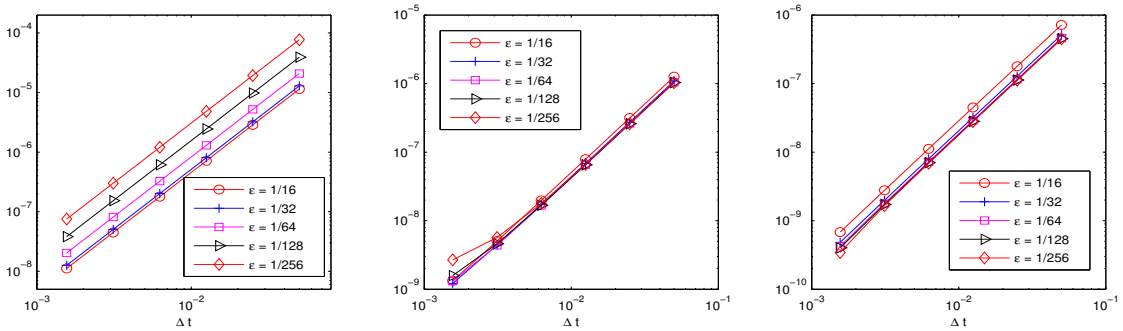


图 4–2 例4.1的波函数（左）、位置密度（中）及电流密度（右）的误差在不同的 $\varepsilon$ 下与时间步长 $\Delta t$ 的对数关系

Fig 4–2 Log-log plot of the errors of the wave function (left), the position densities (middle), and current densities (right) versus time steps $\Delta t$ for different $\varepsilon$ in Example 4.1.

**例 4.2.** 二维系统的模拟

在这个例子中，我们将新方法应用于在 [3] 中详细描述的基本 2D 模型。向量势为 $\mathbf{A}$ =

$\frac{1}{2}(-\cos(y),\sin(x))^T$，标量势 $V=0$。初始的波位于 $(x_0,y_0)=(0.1,-0.02)$，振荡为 $O(\varepsilon)$

$$u_0(x,y)=e^{-20(x-x_0)^2-20(y-y_0)^2}e^{i\sin(x)\sin(y)/\varepsilon}. \tag{4–62}$$

计算区域为 $[-\pi,\pi]\times[-\pi,\pi]$。这里我们对于不同的 $\varepsilon=1/16,1/32,1/64,1/128$ 数值计算其密度，网格 $h_x=h_y=\frac{2\pi}{16}\varepsilon$，时间步长 $\Delta t=1/50$。参考解是通过我们的方法以相同的网格大小获得的，但使用非常精细的时间步长，即 $\Delta t=1/100\ \varepsilon$。表4–3显示了不同的 $\varepsilon$ 的波函数的空间误差和时间 $T=0.4$ 的位置密度。这里的位置密度误差不是针对累积函数计算的，而是位置密度本身。图4–3显示了对于不同的 $\varepsilon=1/32,1/64,1/128$（从上到下的行，每行对应于相同的 $\varepsilon$）在不同的时间 $t=0.4,0.8$，第二列和第四列是以非常精细 $\Delta t$ 计算的参考密度。表4–3和图4–3中显示的结果确认我们的方法可以捕获具有较大时间步长的正确观察值，空间误差与我们的分析相符。

表 4–3 例4.2：对于不同的 $\varepsilon$，空间方向的误差。$\Delta x=\frac{2\pi}{16}\varepsilon$，$\Delta t=1/50$。参考解的由非常小的时间步长 $\Delta t=\varepsilon/100$ 算出。

Table 4–3 Spatial errors computed with $\Delta x=\frac{2\pi}{16}\varepsilon$ and a fixed time step $\Delta t=1/50$ for different $\varepsilon$ in Example 4.2. Reference solution is obtained with the same mesh size $\Delta x=\frac{2\pi}{16}\varepsilon$ and a fine time step $\Delta t=\varepsilon/100$.

| $\varepsilon$ | 1/16 | 1/32 | 1/64 | 1/128 |
|---|---|---|---|---|
| $E_u$ | 4.7093E-06 | 7.3482E-06 | 1.3777E-05 | 2.7109E-05 |
| $E_n$ | 1.0472E-06 | 1.0528E-06 | 1.2366E-06 | 1.3789E-06 |

**例 4.3.** 三维系统的模拟

我们要研究的最后一个例子是一个 3D 模型，这是最具物理意义的，因为量子系统的大多数磁效应发生在 3D 空间中，通常不能将其减小到较小维度的子空间。

值得指出的是，我们研究的模型(4–1)与 Pauli 方程高度相关，Pauli 方程描述了外部电磁场中自旋半粒子的量子演化。在由向量势 $\mathbf{A}$ 和标量势 $V(x)$ 描述的电磁场中，Pauli 方程为

$$i\varepsilon\partial_t\boldsymbol{u}^\varepsilon=\big[\frac{1}{2}(-i\varepsilon\nabla-\boldsymbol{A})^2+V(x)\big]\boldsymbol{I}\boldsymbol{u}^\varepsilon-\frac{1}{2}(\boldsymbol{\sigma}\cdot\mathbf{B})\boldsymbol{u}^\varepsilon, \tag{4–63}$$

其中 $\boldsymbol{\sigma}=(\sigma_x,\sigma_y,\sigma_z)$ 是 Pauli 矩阵，$\boldsymbol{u}^\varepsilon=(u_+^\varepsilon,u_-^\varepsilon)^T$ 是双分量旋转波函数。这里，$\boldsymbol{I}$ 是作为运算符的 $2\times2$ 单位矩阵，$\boldsymbol{B}=\nabla\times\boldsymbol{A}$ 是磁场。

$$i\varepsilon\partial_t\boldsymbol{u}^\varepsilon=\big[\frac{1}{2}(-i\varepsilon\nabla-\boldsymbol{A})^2+V(\boldsymbol{x})\big]\boldsymbol{I}\boldsymbol{u}^\varepsilon. \tag{4–64}$$

然后，$\boldsymbol{u}^\varepsilon$ 可以解耦，每个分量的方程与 (4–1) 的形式相同。

在这个测试中，我们研究了具有恒定磁场的三维系统 $\boldsymbol{B}=B\,(0,0,1)^T$。向量势选择为 $\mathbf{A}=\frac{B}{2}(-y,x,0)^T$。初始波函数为具有 $O(\varepsilon)$ 振荡的双阱函数，即，

$$u_0(x,y,z)=(e^{-20(x-x_0)^2-20y^2-20z^2}+e^{-20(x+x_0)^2-20y^2-20z^2})e^{i\sin(y)\sin(z)/\varepsilon}, \tag{4–65}$$

其中 $x_0=0.5$。计算域为 $[-4,4]^3$。在这里，我们以网格大小 $h_x=h_y=h_z=\frac{1}{32}$ 的数值计算 $\varepsilon=1/16$ 的密度演化，时间步长 $\Delta t=1/40$。

图4–4展示了在不同的时间呈现密度的等值面，$n(\mathbf{x})=10^{-4}$。

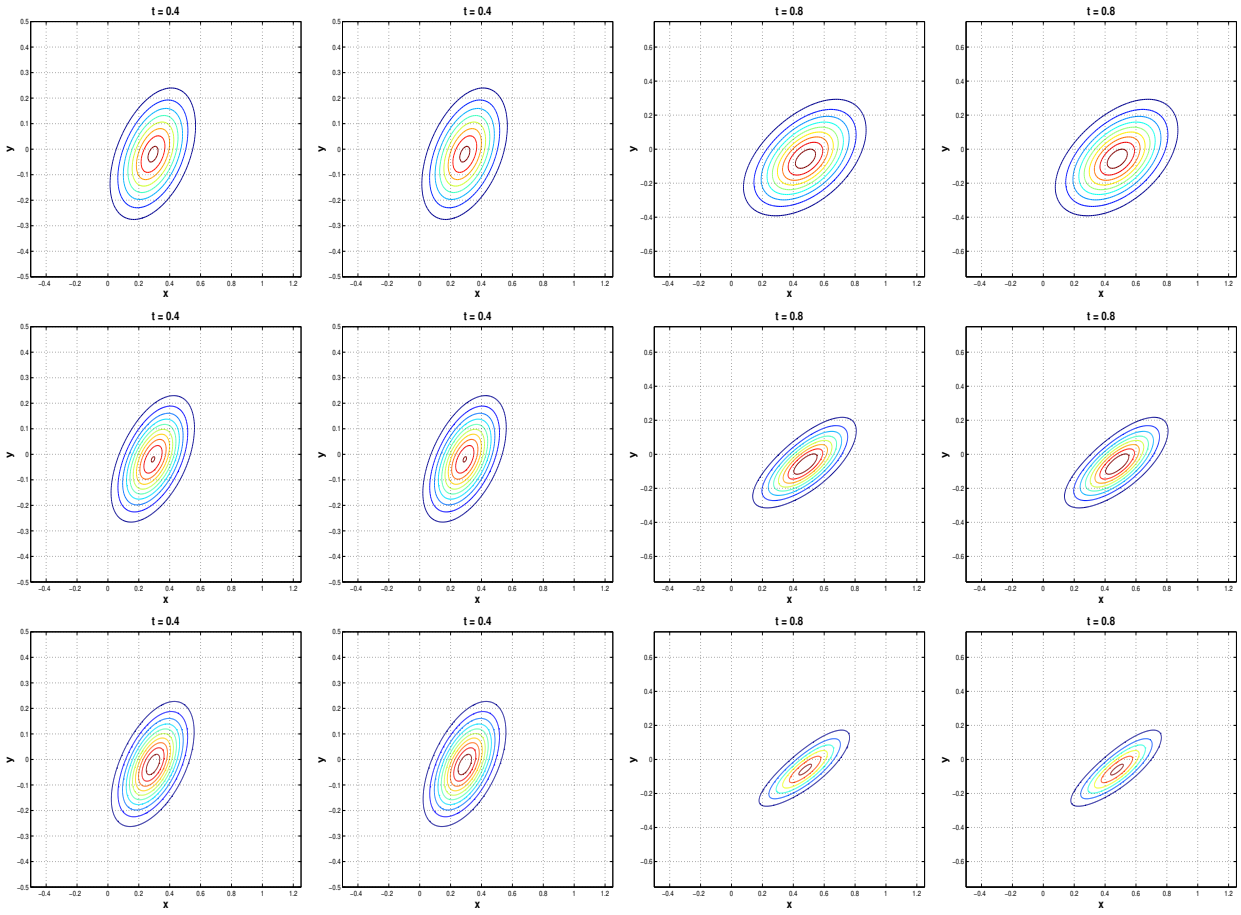


图 4–3 例4.2：不同时刻 $t = 0.4, 0.8$、不同的 $\varepsilon = 1/32, 1/64, 1/128$（从上到下）密度的等值线图，其中第二、四列为参考解。

Fig 4–3 Contour plot of the density computed with $\Delta t = 1/50$ at different times $t = 0.4, 0.8$ for $\varepsilon = 1/32, 1/64, 1/128$ (rows from top to bottom) in Example 4.2, where the second and the forth columns are reference solutions.

## 4.4 本章总结与展望

在本章中，我们提出并分析了具有向量势的半经典薛定谔方程的新的时间分裂谱方法，其中在对流部分的半拉格朗日方法的插值步骤中应用 NUFFT 技术。我们分析了近似波函数和计算物理观测值的方法的稳定性和准确性。并且通过大量一维和三维情况下的各种数值测试来验证分析结果。

电磁场的半经典薛定谔方程在理论和应用上都具有重要的意义，这个模型的研究为更复杂和重要的量子力学模型提供了好的基础（例如，通过并入 Stern-Gerlach 项）。另外，如果要通过波函数模拟带电粒子的量子态，那么可以通过将电流方程与泊松方程与麦克斯韦方程联立进行计算，将作为未来的研究方向。

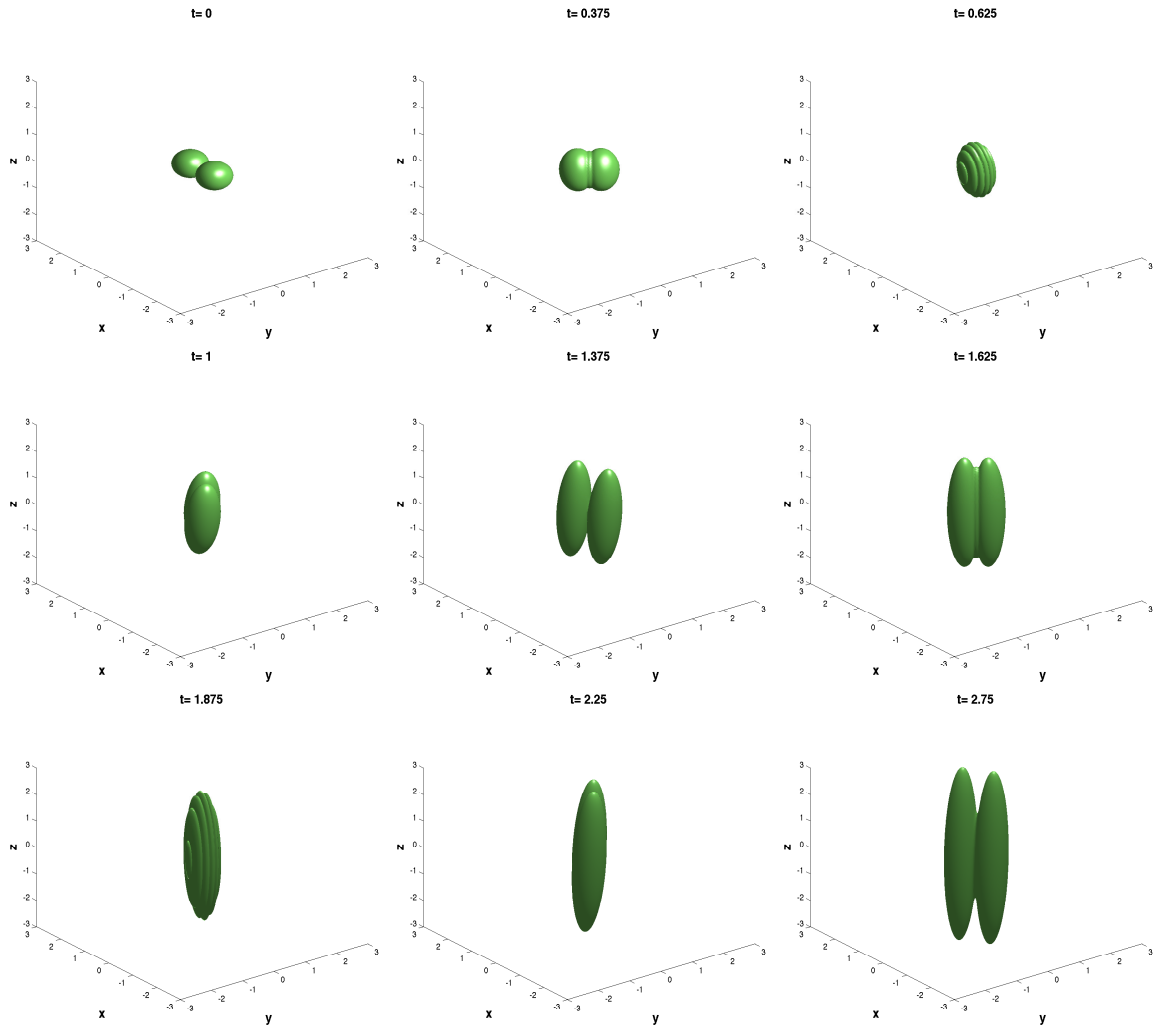


图 4–4 例4.3：不同时刻的密度的等值面。$n(x,y,z)=10^{-4}$，$\varepsilon=1/16$。

Fig 4–4 Isosurface of the density, $n(x,y,z)=10^{-4}$, at different times for $\varepsilon=1/16$ in Example 4.3.

# 第五章　带 Caputo 导数的分数阶守恒律方程的显示与隐式 TVD 算法

在本章中，我们将研究时间方向为 Caputo 导数的标量守恒定律的数值近似。控制方程是以下分数阶时间导数的守恒律：

$$\partial_t^\alpha u(x,t) + f(u(x,t))_x = 0, \quad x \in \mathbb{R},\ t > 0, \tag{5–1}$$

$u(x,t)$ 是密度或浓度函数，$f(u)$ 是通量。方程的初值为：

$$u(x,0) = u_0(x), \quad x \in \mathbb{R}. \tag{5–2}$$

这里，当 $\alpha \in (0,1)$ 时，$\partial_t^\alpha$ 为所谓的 Caputo 导数（分数阶导数）：

$$\partial_t^\alpha u(t) = C_\alpha \int_0^t (t-s)^{-\alpha} \partial_s u(s) ds,$$

其中 $C_\alpha = 1/\Gamma(1-\alpha)$。当 $\alpha = 1$ 时，$\partial_t^\alpha u(t)$ 即为通常的导数 $\partial_t u(t)$。

Caputo 导数最早在文献 [108] 中被引入，用于描述在具有记忆效应的多孔介质中的扩散。一些之前的工作已经表明，Caputo 导数对等离子体传输的建模是有效的（参见 [109-110]），并且已经利用分数函数来构建空间和时间中长距离或非局部相互作用的物理模型，参见例如 [111-112]。最近在 [113] 中，Allen，Caffarelli 和 Vasseur 证明了当有分数势能压力和分数时间导数时，该方程的弱解是存在的并且是 Hölder 连续的。带有 Caputo 导数的标量守恒定律方程(5–1)描述了具有一般通量函数和记忆效应的分布进行的时间演化，虽然相关的物理解释很清楚，但是对于这种方程并没有太多的数学研究。

在数值近似方面，已经有很多数值方法被提出，对带有 Caputo 导数的 ODE 或扩散方程也进行了大量分析，参见例如 [114-118]。其中大多数可以推广到分数阶时空平流扩散方程（参见 [119]），对空间（分数阶）导数中进行了特殊处理。此外，对于分数时空对流–扩散方程的数值近似，已经有一些值得研究的参考文献，参见 [120]。在不同的方程中的研究已经表明，Caputo 导数的相容的离散将引入时间耗散，这有助于稳定数值格式。因此，至少对于线性方程，Caputo 导数不会在数值近似中引起额外的挑战。在 [116] 中，Zhao，Sun 和 Karniadakis 提出了一种逼近粘性 Burgers 方程的数值方法，但是由于他们将傅里叶配点法应用于空间离散化，所以得到的解在间断处出现了吉布斯现象。据我们所知，这仍然是分数阶非线性守恒定律目前唯一的尝试。

在本章中，我们的目的是对带有 Caputo 导数的标量守恒律方程(5–1)构建显式和隐式迎风格式并进行分析。我们提出了方程(5–1)的一阶和二阶格式，并且显示了在修改的 CFL 条件下，数值格式是全变差下降的（total variation diminishing，TVD）。然而，由于修改的 CFL 条件当 $\alpha \to 0$ 时过于严格，这使得显式格式对于小的 $\alpha$ 不可行。在此基础上，我们进一步设计了一种隐式的迎风格式，并且证明了该格式的 $\ell^1$ 范数递减，因此是 TVD 的。特别地，对于线性对流的情况，我们还表明隐式格式也是能量稳定的，并满足熵条件。

本章的其余部分概述如下。我们在本节的第二部分总结了带 Caputo 导数的标量守恒定律的一些初步知识。在第 2 节中，我们简要总结了 Caputo 导数的近似值的现有结果。我们首先提出并展示了第 3 节第一阶和第二阶明显逆风方案的稳定性分析，然后引入了一种隐性的逆风方案来避免 CFL 条件。我们在第 4 节进行各种数值测试，不仅要验证拟议方案的性质，还要通过进行一些设计的测试来调查 Caputo 导数的解释。

## 5.1 基本知识和定义

在这部分中，我们首先介绍时间为分数阶导数的对流方程的 Mittag-Leffler 函数，并用 Caputo 导数讨论了非线性守恒律的弱解。我们考虑以含分数阶时间导数的方程

$$\partial_t^\alpha u = -Au, \quad u(x,0) = g(x). \tag{5–3}$$

$A$ 是一个某个给定的算子。根据拉普拉斯变换，

$$\mathcal{L}u(s) := \hat{u}(s) = \int_0^\infty e^{-st}u(t)dt,$$

方程(5–3)变成

$$(s^\alpha + A)\hat{u} = s^{\alpha-1}g.$$

由拉普拉斯逆变换

$$u(t) = E_\alpha(-t^\alpha A)g, \tag{5–4}$$

$E_\alpha(z)$ 是 Mittag-Leffler 函数

$$E_\alpha(z) = \sum_{n=0}^\infty \frac{z^n}{\Gamma(\alpha n+1)}. \tag{5–5}$$

如果令 $A = -a\partial_x$，其中 $a$ 为常数，则相应的对流方程为，

$$\partial_t^\alpha u + a\partial_x u = 0, \tag{5–6}$$

它的解可以写为

$$u(x,t) = \sum_{n=0}^\infty \frac{\partial_x^n g}{\Gamma(\alpha n+1)} a^n t^{\alpha n}. \tag{5–7}$$

注意，当 $\alpha = 1$ 时，该方程即变成通常的标量守恒律方程以及它的解

$$u(x,t) = g(x+at) = \sum_{n=0}^\infty \frac{\partial_x^n g}{n!} a^n t^n. \tag{5–8}$$

然而，对于一般守恒定律(5–1)，由于非线性的通量使得解不能够显式表示，使用弱解更方便。我们给出以下定义：

**定义 5.1.** $u(x,t)$ 是方程(5–1)的*弱解*，如果 $\partial_t^\alpha u \in L^1_{\text{loc}}(\mathbb{R})$，$f(u) \in L^1_{\text{loc}}(\mathbb{R})$ 及对于任何测试函数 $\phi \in \mathcal{D}(\mathbb{R})$，有

$$\int_{\mathbb{R}} \left(\partial_t^\alpha u\phi - f(u)\partial_x\phi\right) dx = 0. \tag{5–9}$$

人们可以很容易地验证弱解的概念是到经典解的推广：(5–1) 的每个经典解都是弱解。

然而，像标准守恒律一样，(5–1)的弱解可能不唯一。为了说明这点，我们考虑如下分数阶的 Burgers' 方程

$$\partial_t^\alpha u + \partial_x \left(\frac{1}{2}u^2\right) = 0, \tag{5–10}$$

及初值

$$u(x,0) = \begin{cases} 1, & x > 0, \\ -1, & x < 0. \end{cases} \tag{5–11}$$

首先初值本身就是一个弱解，这意味着我们有一个静止的间断解（见图5–1左）

$$u(x,t) = \begin{cases} 1, & x > 0, \\ -1, & x < 0. \end{cases} \tag{5–12}$$

我们来验证这个解满足弱解的定义。首先，我们有

$$\partial_t^\alpha u \equiv 0 \quad \text{for } \forall x \neq 0,$$

同时还有

$$f(u) = \frac{1}{2}u^2 \equiv \frac{1}{2} \quad \text{for } \forall x \neq 0.$$

根据**定义**5.1，对于每个 "测试" 函数 $\phi \in \mathcal{D}(\mathbb{R})$

$$\int_{\mathbb{R}} \left(\partial_t^\alpha u \phi - f(u)\partial_x \phi\right) dx = 0 - \frac{1}{2}\int_{\mathbb{R}} \phi_x \, dx = 0.$$

所以这是方程 (5–10) 的一个弱解。但是显然，这不是一个熵解，因为特征线从间断处向外移动。

为了构建另一个解，我们使用我们将在本章稍后介绍的数值方法，数值结果绘制在图5–1右侧，$t = 0.02$ 和 $\alpha = 0.8$。我们看到，直观地，该解处在静态不连续解和标准稀疏解之间，这表现出了记忆效应。

为了定义熵解，我们考虑了具有人工粘性的守恒律

$$\partial_t^\alpha u + \partial_x f(u) = \varepsilon \partial_{xx} u. \tag{5–13}$$

这里，$\varepsilon > 0$ 是扩散系数，$\varepsilon \ll 1$。可以证明，(5–13) 的 Cauchy 问题具有满足最大原理的唯一一个经典解 $u^\varepsilon$ 。如果当 $\varepsilon \to 0$ 时，序列 $\{u^\varepsilon\}$ 收敛到函数 $u$，我们称 $u$ 是方程 (5–1) 的熵解。

我们指出，通过构建一对函数，即凸熵函数和熵通量来定义熵解是完全有可能的，其粘性极限的定义是等价的，但这将涉及到大量分析，并不是本章的重点。

## 5.2 Caputo 导数的数值逼近

虽然人们对标准的常微分方程的数值近似有了深入的了解，但是由于分数阶时间导数的常微分方程（FODE）数值方法的严格分析会遇到额外的困难，因此对分数阶时间导数的 ODE（FODE）数值方法的研究非常有限。但是最近人们对于这个研究领域的兴趣逐渐增加。

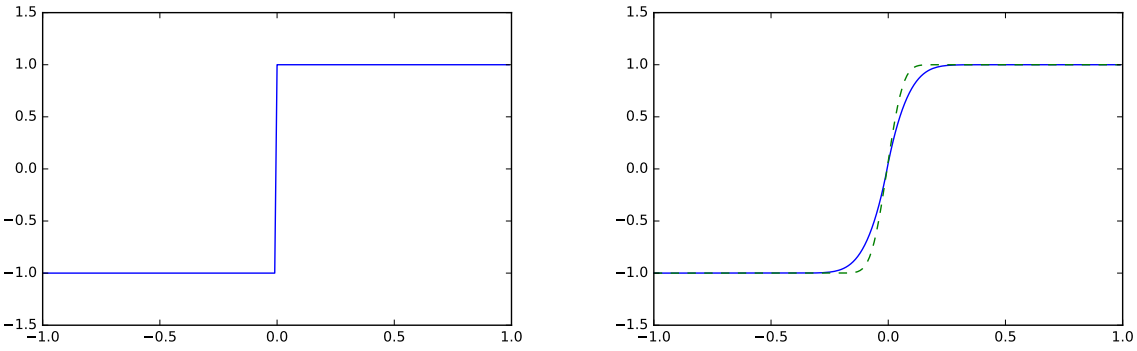


图 5–1 弱解不唯一性。左：静态的间断解；右：实线为带记忆效应的稀疏解，虚线普通为 Burgers' 方程的稀疏解。
Fig 5–1 Non-uniqueness of the solution. Left: the static discontinuous solution. Right: solid line is the rarefaction solution with memory effect; dashed line is the rarefaction solution for the Burgers' equation with the standard time derivative.

Langlands 和 Henry[121] 考虑了具有分数阶时间导数的扩散方程，并为此方程引入了 $L_1$ 稳定的数值格式。Sun 和 Wu[122] 为分数阶时间导数构造了一个 $L_1$ 近似的有限差分格式。Lin 和 Xu[115] 分析了分数阶扩散方程的时间离散化的有限差分格式，并证明了时间的收敛是 $2-\alpha$ 阶。Lv 和 Xu[123] 通过给出更准确的系数改进了误差估计。Zhao，Sun 和 Karniadakis[116] 导出了涉及异常扩散和波传播的分数阶时间导数的两个二阶近似公式。Lin 和 Liu[124] 分析了线性多步法，证明了该方法的稳定性和收敛性。Kumar 和 Agrawal[118] 提出了 FODE 的另一类数值方法，可以将其化减为 Volterra 型积分方程。基于这种方法，Cao 和 Xu[117] 提出了一种构建 FODE 高阶数值方案的一般技术。

在本章中，我们使用如下 Caputo 导数的数值逼近（见 [115]）。假设对于均匀的时间步长 $\tau$，记 $t^n = n\tau$, $n = 0, 1, 2, \cdots$，$u(t^n)$ 的数值逼近记为 $U^n$。为了构建一阶格式，假设

$$0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = t.$$

时间导数用向前差分逼近，即

$$\begin{aligned}
\partial_t^\alpha u(t^{n+1}) &\approx \frac{1}{\Gamma(1-\alpha)} \sum_{k=0}^{n} \int_{t_k}^{t_{k+1}} \frac{U^{k+1} - U^k}{\tau (t_{n+1} - s)^\alpha} \, ds \\
&= \frac{1}{\Gamma(1-\alpha)(1-\alpha)} \sum_{k=0}^{n} \frac{(n+1-k)^{1-\alpha} - (n-k)^{1-\alpha}}{\tau^\alpha} (U^{k+1} - U^k) \\
&= \frac{1}{\Gamma(2-\alpha)\tau^\alpha} \left( U^{n+1} - \sum_{n=0}^{k} c_k^{n+1} U^k \right) := D_t^\alpha U^{n+1},
\end{aligned} \tag{5–14}$$

其中

$$c_k^{n+1} = 2(n+1-k)^{1-\alpha} - (n+2-k)^{1-\alpha} - (n-k)^{1-\alpha}, \quad k = 1, \cdots, n.$$

$$c_0^{n+1} = (n+1)^{1-\alpha} - n^{1-\alpha}.$$

直接计算并注意到 $y = x^{1-\alpha}$ 是单调递增的凹函数，有

$$\sum_{k=0}^{n} c_k^{n+1} = 1, \quad c_k^{n+1} > 0, \quad k = 0, \cdots, n. \tag{5–15}$$

因此，正如标准时间导数可以给出瞬时变化率一样，从其数值近似，及(5–14)得出，Caputo 导数可以被解释为从其历史值的凸组合中的量的变化率，并且系数 $c_k^{n+1}$ 表示出由于记忆效应引起的影响的强度。历史的影响随着记忆效应的变弱而逐渐减小。

值得指出，对于固定的 $\alpha$，

$$c_n^{n+1} = 2 - 2^{1-\alpha} - 0^{1-\alpha} = 2 - 2^{1-\alpha},$$

独立于 $n$。因此，我们记 $c_n^{n+1}$ 为 $\tilde{c}$，它将是显式迎风格式的 CFL 中一个重要的量，将在后面予以说明。

相容性误差由 Lv 和 Xu 在 [123] 中证明，如下述定理，

**引理 5.1.** *对于任意 $\alpha \in (0,1)$，这个格式的截断误差为，*

$$\partial_t^\alpha u(t^n) = D_t^\alpha u(t^n) + r_\tau^n, \tag{5–16}$$

*满足如下误差估计*

$$|r_\tau^k| \leqslant CM(u)\tau^{2-\alpha}, \quad \forall k = 0, 1, \cdots, n, \tag{5–17}$$

*$C$ 独立于 $u$ 和 $\tau$，$M(u) = \max\limits_{t\in(0,t^n]} |\partial_t^2 u(t)|$.*

值得指出的是，本章的重点是用 Caputo 导数构建和分析守恒律的数值方法，所以我们选择使用了一个常见的 Caputo 导数的数值近似。实际上，由于 Caputo 导数有许多较高阶的近似（参见，例如 [116-117]），以下结果和提出的数值方法的时间精度可以很容易地改善。

## 5.3 数值方法和稳定性分析

在本节中，我们利用上一节介绍的 Caputo 导数的数值近似来设计 FODE 和守恒定律的数值格式。我们专注于每个格式的稳定性条件，特别是它们与标准时间导数的模型不同之处。

### 5.3.1 FODE 模型的向后欧拉格式

考虑如下 ODE 模型

$$\partial_t^\alpha u(t) = \lambda u(t). \tag{5–18}$$

$\lambda$ 为复数 $\mathbf{Re}(\lambda) \leqslant 0$，表示离散算子的特征值。

Caputo 导数的 ODE 模型的几个数值方案的稳定性分析已经在以前的一些文献中有所提及，见 [114-115,123,125-126]。然而，在这项工作中，由于我们的目标是设计非线性守恒律的全隐式格式，所

以我们需要利用 ODE 模型的数值方法，其中时间离散化由后向微分公式（BDF）给出，而这个工作尚未有人研究。因此，我们首先考虑这种 ODE 模型的向后欧拉方法。

$$D_t^\alpha U^{n+1} = \lambda U^{n+1}. \tag{5–19}$$

两边同乘 $\tau^\alpha \Gamma(2-\alpha)$ 得到

$$(1 - \lambda\tau^\alpha\Gamma(2-\alpha))\, U^{n+1} = \sum_{k=0}^{n} c_k^{n+1} U^k.$$

如果记 $z = \lambda\tau^\alpha\Gamma(2-\alpha)$，格式对应的稳定多项式 $\pi(\xi; z)$

$$\pi(\xi; z) = (1-z)\,\xi^{n+1} - \sum_{k=0}^{n} c_k^{n+1}\xi^k.$$

下面考虑两种情况，即当 $\lambda \neq 0$ 和当 $\lambda = 0$ 时，

当 $\lambda \neq 0$ 时，$\mathbf{Re}(z) \leqslant 0$，$z \neq 0$，可以得到 $|1-z| > 1$. 假设 $\xi_0$ 及 $|\xi_0| \geqslant 1$ 是 $\pi(\xi; z)$ 的根，对于 $k \leqslant n$ 有，

$$|\xi_0^k| \leqslant |\xi_0|^k \leqslant |\xi_0^{n+1}|.$$

可以得到

$$\begin{aligned} |(1-z)\xi_0^{n+1}| = |1-z||\xi_0|^{n+1} &= |\sum_{k=0}^{n} c_k^{n+1}\xi_0^k| \\ &\leqslant \sum_{k=0}^{n} c_k^{n+1}|\xi_0|^k \leqslant \left(\sum_{k=0}^{n} c_k^{n+1}\right)|\xi_0|^{n+1} = |\xi_0|^{n+1}, \end{aligned}$$

矛盾。这说明，稳定多项式只有模小于 1 的根，即格式是绝对稳定的。

当 $\lambda = 0$ 时，那么 $z = 0$，并且稳定性分析简化到时间离散化的零稳定性。如果稳定性多项式的根的模严格地大于 1，那么上面的分析仍然成立，并且可以证明没有这样的根存在。

如果多项式的根的模恰好是 1，可以假设根为 $\xi_0 = e^{i\theta}$。如果 $\theta = 0$，那么 $\xi_0 = 1$，我们有

$$\pi(1; 0) = 1^{n+1} - \sum_{k=0}^{n} c_k^{n+1} = 0.$$

计算

$$\frac{d\pi(\xi; 0)}{d\xi} = (n+1)\xi^n - \sum_{k=0}^{n} c_k^{n+1} k\xi^{k-1}.$$

注意，系数 $\{c_k^{n+1}\}_{k=0}^n$ 满足条件(5–15)，因此

$$\left|\sum_{k=0}^{n} c_k^{n+1} k\right| < \left|\sum_{k=0}^{n} c_k^{n+1} n\right| = n.$$

由此推出

$$\frac{d\pi(1; 0)}{d\xi} = (n+1) - \sum_{k=0}^{n} c_k k > n + 1 - n = 1 \neq 0.$$

所以 1 不是稳定性多项式的重根。

如果 $\theta \neq 0$，我们得到如下方程，

$$e^{i(n+1)\theta} = \sum_{k=0}^{n} c_k^{n+1} e^{ik\theta}.$$

两边同时除以 $e^{i(n+1)\theta}$，

$$1 = \sum_{k=0}^{n} c_k^{n+1} e^{i(k-1-n)\theta}.$$

由于 $\theta \neq 0$，至少有一个 $e^{i(k-1-n)\theta}$ 不是实值。因此，上述方程的右边是 $n+1$ 单位复数的凸组合。所以，我们得出结论，右边的和不可能超过 1。因此，$e^{i\theta}$ with $\theta \neq 0$ 不是稳定多项式的根。

最后我们得出，当 $\mathbf{Re}(\lambda) \leqslant 0$，ODE 模型的向后欧拉格式是无条件稳定的。换句话说，我们已经证明

**定理　5.2.** *对于 ODE 模型(5–18)的向后欧拉方法(5–19)是 A-稳定的。*

这个结果与对于 ODE 和抛物线方程的 [115] 的稳定性结果是一致的，这并不是令人惊讶，因为分数阶时间导数在相当于在时间上增加了耗散量 [115-116]。然而，分数阶时间导数的双曲线问题的稳定性分析还是公开的问题，在下一节中，我们主要关注标量守恒律。

为了提供一些对后向欧拉方法(5–19)的稳定性区域的直观了解，我们使用线性多步法的边界轨迹法来数值绘制不同的 $\alpha$ 和 $n$ 的 $z$ 变量的复平面边界点。如图5–2所示，稳定区域是封闭曲线的外部。我们观察到，稳定区域不仅取决于 $\alpha$，而且还取决于与记忆效应长度成比例的 $n$。作为 $n \to \infty$，我们看到边界渐近地接近某个极限曲线。

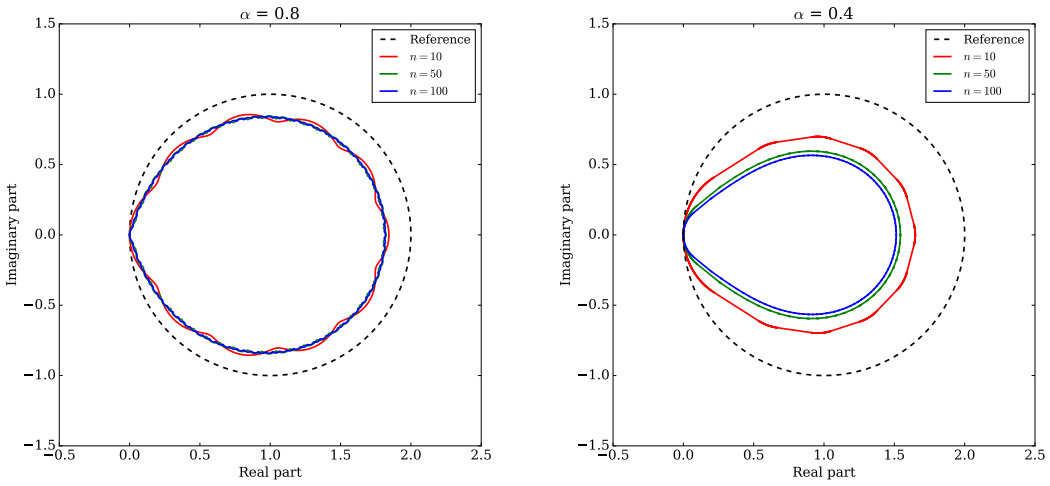


图 5–2 *左*：$\alpha = 0.8$，$n = 10, 50, 100$ 的绝对稳定区域。*右*：$\alpha = 0.4$，$n = 10, 50, 100$ 的绝对稳定区域。*参考*：普通导数的向后欧拉方法的稳定区域。

Fig 5–2 Left: Absolute stability zone for $\alpha = 0.8$, $n = 10, 50, 100$. Right: Absolute stability zone for $\alpha = 0.4$, $n = 10, 50, 100$. Reference: backward Euler for the standard derivative.

请注意，我们的稳定区域包括虚轴 $\mathbf{Re}\,\lambda = 0$，这与使用标准时间导数的向后欧拉方法是一样的，因此它可以应用双曲型方程。虽然，对于标准时间导数的双曲型问题，通常优先选择显式方法，我们将看到 Caputo 时间导数在稳定性中引起额外的约束，这促使我们为(5–1)设计隐式格式。

### 5.3.2 标量守恒律方程的显示迎风格式

5.3.2.1 一阶格式

考虑一维的守恒律方程

$$\partial_t^\alpha u + (f(u))_x = 0, \tag{5–20}$$

通量可以分解为

$$f = f^+ + f^-, \quad (f^+)' \geqslant 0, \quad (f^-)' \leqslant 0. \tag{5–21}$$

我们做这个假设来简化我们的分析。注意，在设计这种数值方案时，通量分解不是必不可少的。读者可以参考 [127-128] 关于这个问题的一般性讨论。不失一般性，通量函数 $f(u)$ 可能是非线性的。但是显然，当 $f = au$ 时，它也包括线性对流的情况。

再次，我们假设均匀的时间步长 $\tau$，并且记 $t^n = n\tau$，$n = 0, 1, 2\ \cdots$。此外，在计算域 $[a\ b]$，我们假设均匀的空间网格 $x_j = a + jh$，对于 $j = 0, 1\ \cdots\ M$，其中空间网格大小 $h = \frac{b-a}{M}$。我们用 $U_j^k$ 表示 $u\left(x = x_j, t = t^k\right)$ 的数值近似值。然后，非线性守恒律的一阶迎风法格式为，

$$D_t^\alpha U_j^{n+1} + \frac{1}{h}\left(f^+(U_j^n) - f^+(U_{j-1}^n)\right) + \frac{1}{h}\left(f^-(U_{j+1}^n) - f^-(U_j^n)\right) = 0. \tag{5–22}$$

如果记

$$\lambda_j^{+,n} = \frac{a_j^{+,n}\tau^\alpha\Gamma(2-\alpha)}{h}, \quad \lambda_j^{-,n} = \frac{a_j^{-,n}\tau^\alpha\Gamma(2-\alpha)}{h},$$

对于在 $U_{j-1}^n$ 和 $U_j^n$ 中的 $\xi_j^n$ 以在 $U_j^n$ and $U_{j+1}^n$ 中的 $\eta_j^n$，有

$$a_j^{+,n} = \frac{f^+(U_j^n) - f^+(U_{j-1}^n)}{U_j^n - U_{j-1}^n} = (f^+)'(\xi_j^n) \geqslant 0,$$

$$a_j^{-,n} = \frac{f^-(U_{j+1}^n) - f^-(U_j^n)}{U_{j+1}^n - U_j^n} = (f^-)'(\eta_j^n) \leqslant 0,$$

数值格式可以写为

$$U_j^{n+1} = (\tilde{c} - \lambda_j^{+,n} + \lambda_j^{-,n})U_j^n + \lambda_j^{+,n}U_{j-1}^n - \lambda_j^{-,n}U_{j+1}^n + \sum_{k=0}^{n-1} c_k^{n+1}U_j^k. \tag{5–23}$$

据此，我们给出如下的 CFL 条件

$$\frac{\tau^\alpha\Gamma(2-\alpha)}{h}\left(\max|(f^+)'| + \max|(f^-)'|\right) \leqslant \tilde{c}. \tag{5–24}$$

我们观察到，CFL 条件基本上与使用标准时间导数的守恒定律一致，除了时间步长 $\tau$ 获得由 Caputo 导数产生的指数 $\alpha$。

由 CFL 条件，有

$$\tilde{c} - \lambda_j^{+,n} + \lambda_j^{-,n} \geqslant 0.$$

容易得出最大值原理，$\forall n \in \mathbb{N}^+$

$$\max_j |U_j^n| \leqslant \max_j |U_j^0|.$$

也可以证明在 CFL 条件(5–24)下方法是 TVD 的。事实上，可以将(5–23)写作

$$U_j^{n+1} = \tilde{c}U_j^n - \delta f^+(U_j^n) + \delta f^+(U_j^n) + \delta f^+(U_{j-1}^n) - \delta f^+(U_{j+1}^n) + \sum_{k=0}^{n-1} c_k^{n+1} U_j^k, \tag{5–25}$$

其中 $\delta = \frac{\tau^\alpha \Gamma(2-\alpha)}{h}$。考虑满足同一个差分方程的另外一个解 $V_j^n$，

$$V_j^{n+1} = \tilde{c}V_j^n - \delta f^+(V_j^n) + \delta f^+(V_j^n) + \delta f^+(V_{j-1}^n) - \delta f^+(V_{j+1}^n) + \sum_{k=0}^{n-1} c_k^{n+1} V_j^k. \tag{5–26}$$

两式相减

$$\begin{aligned} U_j^{n+1} - V_j^{n+1} = {} & \tilde{c}\left(U_j^n - V_j^n\right) - \delta\left(f^+(U_j^n) - f^+(V_j^n)\right) + \delta\left(f^-(U_j^n) - f^-(V_j^n)\right) \\ & + \delta\left(f^+(U_{j-1}^n) - f^+(V_{j-1}^n)\right) - \delta\left(f^-(U_{j+1}^n) - f^-(V_{j+1}^n)\right) + \sum_{k=0}^{n-1} c_k^{n+1}\left(U_j^k - V_j^k\right). \end{aligned}$$

根据均值理论

$$\begin{aligned} U_j^{n+1} - V_j^{n+1} = {} & \left(\tilde{c} - \delta(f^+)'(\xi_j^+) + \delta(f^-)'(\xi_j^-)\right)\left(U_j^n - V_j^n\right) + \delta(f^+)'(\xi_{j-1}^+)\left(U_{j-1}^n - V_{j-1}^n\right) \\ & - \delta(f^-)'(\xi_{j+1}^-)\left(U_{j+1}^n - V_{j+1}^n\right) + \sum_{k=0}^{n-1} c_k^{n+1}\left(U_j^k - V_j^k\right), \end{aligned}$$

其中 $\xi_j^+$ 和 $\xi_j^-$ 为 $U_j^n$ 和 $V_j^n$ 之间的数。注意当满足 CFL 条件(5–24)时

$$\tilde{c} - \delta(f^+)'(\xi_j^+) + \delta(f^-)'(\xi_j^-) \geqslant 0,$$

根据三角不等式

$$\begin{aligned} \left|U_j^{n+1} - V_j^{n+1}\right| = {} & \left(\tilde{c} - \delta(f^+)'(\xi_j^+) + \delta(f^-)'(\xi_j^-)\right)\left|U_j^n - V_j^n\right| + \delta(f^+)'(\xi_{j-1}^+)\left|U_{j-1}^n - V_{j-1}^n\right| \\ & - \delta(f^-)'(\xi_{j+1}^-)\left|U_{j+1}^n - V_{j+1}^n\right| + \sum_{k=0}^{n-1} c_k^{n+1}\left|U_j^k - V_j^k\right|. \end{aligned}$$

对 $j$ 求和

$$\sum_j \left|U_j^{n+1} - V_j^{n+1}\right| \leqslant \sum_{k=0}^{n} c_k^{n+1} \sum_j \left|U_j^k - V_j^k\right|.$$

这里，通量项都消掉了，由归纳法，

$$\|U^n - V^n\|_{\ell^1} \leqslant \|U^0 - V^0\|_{\ell^1},$$

即，我们证明了如下定理

**定理 5.3.** *当满足 CFL 条件(5–24)时，方程(5–1)的一阶迎风格式(5–22)是 $\ell^1$ 递减的。*

作为一个直接的推论，如果取 $V_j^n$ 为 $U_{j+1}^n$，得到对于 $n \in \mathbb{N}^+$,

$$\mathrm{TV}\left[U^n\right] \leqslant \mathrm{TV}\left[U^0\right],$$

其中 $\mathrm{TV}\left[U^n\right] = \sum_j |U_{j+1}^n - U_j^n|$。也就是

**推论 5.4.** *当满足 CFL 条件(5–24)时，方程(5–1)的一阶迎风格式(5–22)是 TVD 的。*

5.3.2.2 MUSCL 格式

为了构造一个空间方向二阶的格式，正负通量可用分段线性函数逼近

$$f^{\pm,n}(x) = f_j^{\pm,n} + s_j^{\pm,n}(x - x_j), \quad x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}},$$

其中记 $f_j^{\pm,n} = f^{\pm,n}(U_j^n)$。斜率函数由通量限制器定义

$$s_j^{\pm,n} = \frac{f_j^{\pm,n} - f_{j-1}^{\pm,n}}{h}\phi^0\left(\frac{f_{j+1}^{\pm,n} - f_j^{\pm,n}}{f_j^{\pm,n} - f_{j-1}^{\pm,n}}\right).$$

本章中我们仅考虑 minmod 限制器

$$\phi^0(\theta) = \max(0, \min(1, \theta)),$$

或 Van Leer 限制器

$$\phi^0(\theta) = \frac{|\theta| + \theta}{1 + \theta}.$$

注意，这两种限制器都是对称的，即 $a\phi^0(b/a) = b\phi^0(a/b)$。二阶通量分裂格式为

$$D_t^\alpha U_j^{n+1} + \frac{1}{h}\left(f^{+,n}(x_{j+\frac{1}{2}}-) - f^{+,n}(x_{j-\frac{1}{2}}-)\right) + \frac{1}{h}\left(f^{-,n}(x_{j+\frac{1}{2}}+) - f^{-,n}(x_{j-\frac{1}{2}}+)\right) = 0.$$

或等价的

$$D_t^\alpha U_j^{n+1} + \psi_j^{+,n}\frac{f_j^{+,n} - f_{j-1}^{+,n}}{h} + \psi_j^{-,n}\frac{f_{j+1}^{-,n} - f_j^{-,n}}{h} = 0.$$

系数

$$\psi_j^{+,n} = 1 + \frac{1}{2}\phi^0\left(\frac{f_{j+1}^{+,n} - f_j^{+,n}}{f_j^{+,n} - f_{j-1}^{+,n}}\right) - \frac{1}{2}\phi^0\left(\frac{f_{j-1}^{+,n} - f_{j-2}^{+,n}}{f_j^{+,n} - f_{j-1}^{+,n}}\right),$$

$$\psi_j^{-,n} = 1 + \frac{1}{2}\phi^0\left(\frac{f_{j+2}^{-,n} - f_{j+1}^{-,n}}{f_{j+1}^{-,n} - f_j^{-,n}}\right) - \frac{1}{2}\phi^0\left(\frac{f_j^{-,n} - f_{j-1}^{-,n}}{f_{j+1}^{-,n} - f_j^{-,n}}\right).$$

同样的，可以对于 $f^+$ 和 $f^-$ 应用均值理论。非线性守恒律的二阶通量分裂格式可以改写为(5–23)的形式，其中

$$a_j^{+,n} = \psi_j^{+,n}(f^+)'(\xi_j^n), \quad a_j^{-,n} = \psi_j^{-,n}(f^-)'(\eta_j^n).$$

因为限制器满足

$$0 \leqslant \frac{\phi^0(\theta)}{\theta} \leqslant 2, \quad 0 \leqslant \phi^0(\theta) \leqslant 2,$$

所以系数 $\psi_j^{\pm,n} \in [0,2]$，这表明 $\pm a_j^{\pm,n}$ 是非负的。同理，可以提出如下 CFL 条件

$$2\frac{\tau^\alpha \Gamma(2-\alpha)}{h}\left(\max|(f^+)'| + \max|(f^-)'|\right) \leqslant \tilde{c}. \tag{5–27}$$

显然，在上述条件下，(5–23)右侧的系数都是非负的。因此，我们可以通过类似于一阶情况的论证来得出 TVD 和稳定性。

我们观察到，对于显式的格式，尽管由于 Caputo 导数，稳定性约束只改变为 $\tau^\alpha = O(h)$，但在实践中，这使得这些方法不可行。例如，当 $\alpha = \frac{1}{2}$ 时，该约束已经像抛物型方程的显式方法一样受到限制。而当 $\alpha \to 0$ 时，时间步长的选择由 CFL 条件可以知是非常非常小的。

### 5.3.3　标量守恒律方程的隐式格式

#### 5.3.3.1　稳定性分析

正如我们在上一节中所看到的，尽管我们能够得到标量守恒定律的修正 CFL 条件，但稳定性约束意味着时间步长 $\Delta t$ 是空间网格大小 $\Delta x$ 的较高阶量级。因此，我们需要分析方程(5–20)的隐式迎风方案，其中通量函数 $f$ 满足条件(5–21)，由

$$D_t^\alpha U_j^{n+1} + \frac{1}{h}\left(f^+(U_{j-1}^{n+1}) - f^+(U_{j-2}^{n+1})\right) + \frac{1}{h}\left(f^-(U_{j+1}^{n+1}) - f^-(U_j^{n+1})\right) = 0. \tag{5–28}$$

我们引入类似的记号来重写格式

$$\lambda_j^{+,n} = \frac{a_j^{+,n}\tau^\alpha\Gamma(2-\alpha)}{h}, \quad \lambda_j^{-,n} = \frac{a_j^{-,n}\tau^\alpha\Gamma(2-\alpha)}{h},$$

对于某些 $U_{j-1}^{n+1}$ 和 $U_j^{n+1}$ 之间的 $\xi_j^{n+1}$，$U_j^{n+1}$ 和 $U_{j+1}^{n+1}$ 之间的某些 $\eta_j^{n+1}$，有

$$a_j^{+,n+1} = \frac{f^+(U_j^{n+1}) - f^+(U_{j-1}^{n+1})}{U_j^{n+1} - U_{j-1}^{n+1}} = (f^+)'(\xi_j^{n+1}) \geqslant 0,$$

$$a_j^{-,n+1} = \frac{f^-(U_{j+1}^{n+1}) - f^-(U_j^{n+1})}{U_{j+1}^{n+1} - U_j^{n+1}} = (f^-)'(\eta_j^{n+1}) \leqslant 0,$$

格式重写为

$$U_j^{n+1} - \sum_{k=0}^{n} c_k^{n+1} U_j^k = -\lambda_j^{+,n+1}\left(U_j^{n+1} - U_{j-1}^{n+1}\right) - \lambda_j^{-,n+1}\left(U_{j+1}^{n+1} - U_j^{n+1}\right). \tag{5–29}$$

现在我们来证明下述定理

**定理　5.5.** *守恒律方程(5–1) 的隐式迎风格式(5–28)为无条件 $\ell^1$ 递减的。*

**证明.** 重写(5–29)

$$U_j^{n+1} - \sum_{k=0}^{n} c_k^{n+1} U_j^k = -\delta\left(f^+(U_{j-1}^{n+1}) - f^+(U_{j-2}^{n+1})\right) - \delta\left(f^-(U_{j+1}^{n+1}) - f^-(U_j^{n+1})\right). \tag{5–30}$$

注意 $\delta = \frac{\tau^\alpha}{h}$。设 $V_j^n$ 是同一个方程相同数值格式的解，

$$V_j^{n+1} - \sum_{k=0}^{n} c_k^{n+1} V_j^k = -\delta\left(f^+(V_{j-1}^{n+1}) - f^+(V_{j-2}^{n+1})\right) - \delta\left(f^-(V_{j+1}^{n+1}) - f^-(V_j^{n+1})\right). \tag{5–31}$$

两式相减，应用均值理论

$$U_j^{n+1}-V_j^{n+1}+\delta(f^+)'(\xi_j^+)\left(U_j^{n+1}-V_j^{n+1}\right)-\delta(f^-)'(\xi_j^-)\left(U_j^{n+1}-V_j^{n+1}\right)=\sum_{k=0}^{n}c_k^{n+1}\left(U_j^k-V_j^k\right)-\delta(f^-)'(\xi_{j-1}^-)\left(U_{j+1}^{n+1}-V_{j+1}^{n+1}\right)-\delta(f^+)'(\xi_{j-1}^+)\left(U_{j-1}^{n+1}-V_{j-1}^{n+1}\right),$$

$\xi_j^+$ 和 $\xi_j^-$ 是 $U_j^{n+1}$ 和 $V_j^{n+1}$ 之间的数。

两边同乘 $\mathrm{Sgn}\left(U_j^{n+1}-V_j^{n+1}\right)$，对 $j$ 求和，得到

$$\begin{aligned}\text{L.H.S.}&=\sum_j\left|U_j^{n+1}-V_j^{n+1}\right|+\sum_j\delta(f^+)'(\xi_j^+)\left|U_j^{n+1}-V_j^{n+1}\right|-\sum_j\delta(f^-)'(\xi_j^-)\left(U_j^{n+1}-V_j^{n+1}\right)\\&=\sum_j\left|U_j^{n+1}-V_j^{n+1}\right|+\delta\sum_j\left(\left|\delta(f^+)'(\xi_j^+)\left(U_j^{n+1}-V_j^{n+1}\right)\right|+\left|(f^-)'(\xi_j^-)\left(U_j^{n+1}-V_j^{n+1}\right)\right|\right)\\&=\sum_j\left|U_j^{n+1}-V_j^{n+1}\right|+\delta\sum_j\left[\left|f^+\left(U_j^{n+1}\right)-f^+\left(V_j^{n+1}\right)\right|+\left|f^-\left(U_j^{n+1}\right)-f^+\left(V_j^{n+1}\right)\right|\right].\end{aligned}$$

右端由三角不等式，

$$\begin{aligned}\text{R.H.S.}&\leqslant\sum_j\sum_{k=0}^{n}c_k^{n+1}\left|U_j^k-V_j^k\right|+\delta\sum_j\left(\left|(f^+)'(\xi_{j-1}^+)\left(U_{j-1}^{n+1}-V_{j-1}^{n+1}\right)\right|+\left|(f^-)'(\xi_{j+1}^-)\left(U_{j+1}^{n+1}-V_{j+1}^{n+1}\right)\right|\right)\\&=\sum_j\sum_{k=0}^{n}c_k^{n+1}\left|U_j^k-V_j^k\right|+\delta\sum_j\left[\left|f^+\left(U_j^{n+1}\right)-f^+\left(V_j^{n+1}\right)\right|+\left|f^-\left(U_j^{n+1}\right)-f^+\left(V_j^{n+1}\right)\right|\right].\end{aligned}$$

得到

$$\sum_j\left|U_j^{n+1}-V_j^{n+1}\right|\leqslant\sum_j\sum_{k=0}^{n}c_k^{n+1}\left|U_j^k-V_j^k\right|.$$

根据归纳法，

$$\|U^n-V^n\|_{\ell^1}\leqslant\|U^0-V^0\|_{\ell^1},$$

即隐式迎风格式是 $\ell^1$ 递减的。 □

类似于之前的情况，我们有

**推论 5.6.** *守恒律方程(5–1) 的隐式迎风格式(5–28) 是无条件 VD 的。*

#### 5.3.3.2 能量估计和熵解

为了进一步研究近似 Caputo 导数时引入的数值耗散，我们应用以下能量法，并且证明隐式迎风方法对于线性对流方程，即 $f=au$ 无条件地为 $l^2$ 稳定。为了简单起见，我们取 $a>0$，那么右边的(5–29)化为

$$\text{R.H.S.}=-\lambda\left(U_j^{n+1}-U_{j-1}^{n+1}\right),$$

其中 $\lambda=a\tau^\alpha\Gamma(2-\alpha)/h$。

(5–29)两边乘 $U_j^{n+1}$，对 $j$ 求和，得到

$$\begin{aligned}
\text{L.H.S.} &= \sum_j U_j^{n+1}\left(U_j^{n+1}-\sum_{k=0}^{n} c_k^{n+1}U_j^k\right)\\
&= \sum_j\sum_{k=0}^{n} c_k^{n+1}\left[\left(U_j^{n+1}\right)^2-U_j^{n+1}U_j^k\right]\\
&= \frac{1}{2}\sum_j\sum_{k=0}^{n} c_k^{n+1}\left[\left(U_j^{n+1}\right)^2+\left(U_j^{n+1}-U_j^k\right)^2-\left(U_j^k\right)^2\right]\\
&= \frac{1}{2}\sum_j\left(U_j^{n+1}\right)^2+\frac{1}{2}\sum_j\sum_{k=0}^{n} c_k^{n+1}\left(U_j^{n+1}-U_j^k\right)^2-\frac{1}{2}\sum_j\sum_{k=0}^{n} c_k^{n+1}\left(U_j^k\right)^2\\
&= \frac{1}{2}\|U^{n+1}\|_{l^2}^2-\frac{1}{2}\sum_{k=0}^{n} c_k^{n+1}\|U^k\|_{l^2}^2+\frac{1}{2}\sum_j\sum_{k=0}^{n} c_k^{n+1}\left(U_j^{n+1}-U_j^k\right)^2.
\end{aligned}$$

和

$$\begin{aligned}
\text{R.H.S.} &= \sum_j -\lambda U_j^{n+1}\left(U_j^{n+1}-U_{j-1}^{n+1}\right)\\
&= -\frac{\lambda}{2}\sum_j\left[\left(U_j^{n+1}\right)^2-2U_j^{n+1}U_{j-1}^{n+1}+\left(U_{j-1}^{n+1}\right)^2\right]\\
&= -\frac{\lambda}{2}\sum_j\left(U_j^{n+1}-U_{j-1}^{n+1}\right)^2.
\end{aligned}$$

因此得到

$$\|U^{n+1}\|_{l^2}^2+\sum_j\sum_{k=0}^{n} c_k^{n+1}\left(U_j^{n+1}-U_j^k\right)^2+\lambda\sum_j\left(U_j^{n+1}-U_{j-1}^{n+1}\right)^2=\sum_{k=0}^{n} c_k^{n+1}\|U^k\|_{l^2}^2.$$

根据归纳法，可以容易的证明对于 $n\in\mathbb{N}^+$, $\|U^n\|_{l^2}^2\leqslant\|U^0\|_{l^2}^2$，所以我们有如下估计，

$$\|U^n\|_{l^2}^2+\sum_j\sum_{k=0}^{n-1} c_k^{n}\left(U_j^{n}-U_j^k\right)^2+\lambda\sum_j\left(U_j^{n}-U_{j-1}^{n}\right)^2\leqslant\|U^0\|_{l^2}^2.$$

左边的第二项对应于分数时间导数的阻尼效应，左侧的第三项对应于逆风方法的数值耗散。

最后，我们要验证，隐式迎风方法也满足线性对流方程的熵条件。假设 $\eta(u)$ 是一个凸熵函数，$\psi(u)$ 是其熵通量函数。对于线性对流方程，我们有 $\psi(u)=a\eta(u)$。在不失一般性的情况下，我们取 $a>0$，并以下列方式重写隐式迎风方格式：

$$U_j^{n+1}+\lambda U_j^{n+1}=\sum_{k=0}^{n} c_k^{n+1}U_j^k+\lambda U_{j-1}^{n+1},$$

其中 $\lambda=a\tau^\alpha/(hC_\alpha)$。两边同除 $1+\lambda$，

$$U_j^{n+1}=\sum_{k=0}^{n}\frac{c_k^{n+1}}{1+\lambda}U_j^k+\frac{\lambda}{1+\lambda}U_{j-1}^{n+1}.$$

显然，右端为一个凸组合

$$\sum_{k=0}^{n} c_k^{n+1} + \lambda = 1 + \lambda, \quad c_k^{n+1} > 0.$$

熵函数的凸性表明

$$\eta\left(U_j^{n+1}\right) = \eta\left(\sum_{k=0}^{n} \frac{c_k^{n+1}}{1+\lambda} U_j^k + \frac{\lambda}{1+\lambda} U_{j-1}^{n+1}\right) \leqslant \sum_{k=0}^{n} \frac{c_k^{n+1}}{1+\lambda} \eta\left(U_j^k\right) + \frac{\lambda}{1+\lambda} \eta\left(U_{j-1}^{n+1}\right).$$

对 $j$ 求和并记 $\eta(U^n) = \sum_j \eta(U_j^N)$，得到

$$\eta\left(U^{n+1}\right) \leqslant \sum_{k=0}^{n} c_k^{n+1} \eta\left(U^k\right).$$

根据归纳法

$$\eta\left(U^{n+1}\right) \leqslant \eta\left(U^0\right),$$

说明离散的熵是不减的。

对于具有 Caputo 导数的一般的守恒律，可以证明类似的结果。然而，Caputo 导数的熵解与守恒定律的分析方面尚未完全了解。因此，我们将把相关的数值分析作为未来可能的方向之一。但在这项工作中，我们将进行各种数值测试。

## 5.4 数值例子

### 5.4.1 显式格式的例子

首先我们给出标量对流方程的一阶、二阶显式格式的例子，考虑如下方程：

$$\partial_t^\alpha u + \partial_x u = 0, \tag{5–32}$$

及不连续的初值

$$u(x,0) = \begin{cases} 2, & \text{if } x < 0, \\ 1, & \text{if } x \geq 0. \end{cases} \tag{5–33}$$

#### 5.4.1.1 一阶格式的收敛性和稳定性测试

收敛性测试中，我们固定 $\Delta t = 0.0001$ 并计算在 $T = 0.2$ 时的解。参考解是用非常细的网格 $\Delta x = 0.001$ 盒 $\Delta t = 0.0001$ 得到的。误差使用 $\ell^1$ 来衡量：

$$\text{error} = \|u(x_j, T) - u_{\text{ref}}\|_{\ell^1}. \tag{5–34}$$

作为比较，分别令 $\alpha = 0.8$ 及 $\alpha = 0.9$。结果显示在图5–3中：从 log-log 图中可以看出这是一个一阶格式。

对于稳定性测试，我们将固定 $\Delta x = 0.01$，并且让 $\Delta t$ 从小到大增加，我们显示出使得我们的格式发散的 $\Delta t$ 的临界值，在图5–4中。该结果与满足 CFL 条件(5–24)的最大 $\Delta t$ 计算结果进行比较。我们观察到，我们得出的稳定性条件基本上是很准确的。

我们观察到，当使用显式方法时，对 $\Delta t$ 的约束是非常严格的，这使得计算非常耗时。特别地，稳定性条件非常受限于 $\alpha \to 0$，在应用中非常受限。

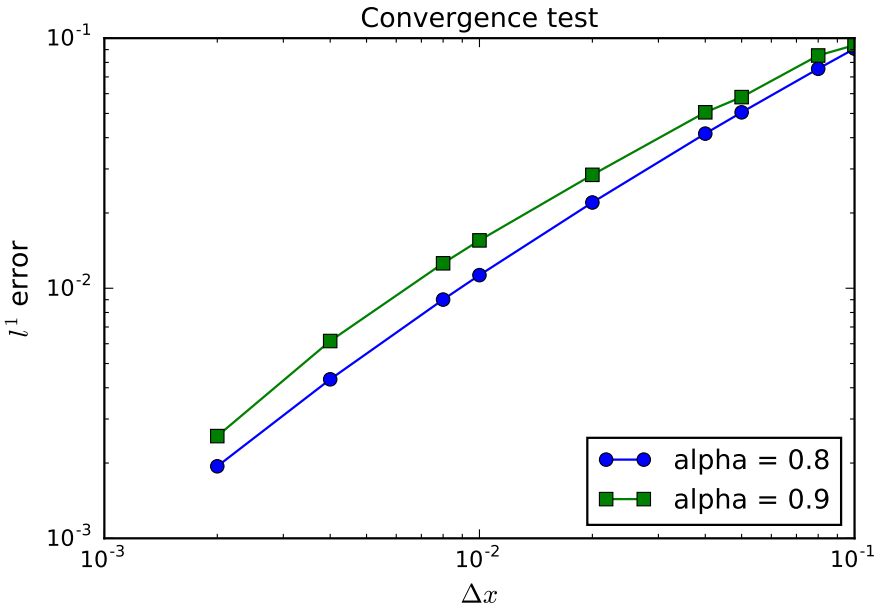


图 5–3 收敛性测试表明为 $\Delta x$ 的一阶收敛。

Fig 5–3 Convergence test shows it is a first order scheme in $\Delta x$.

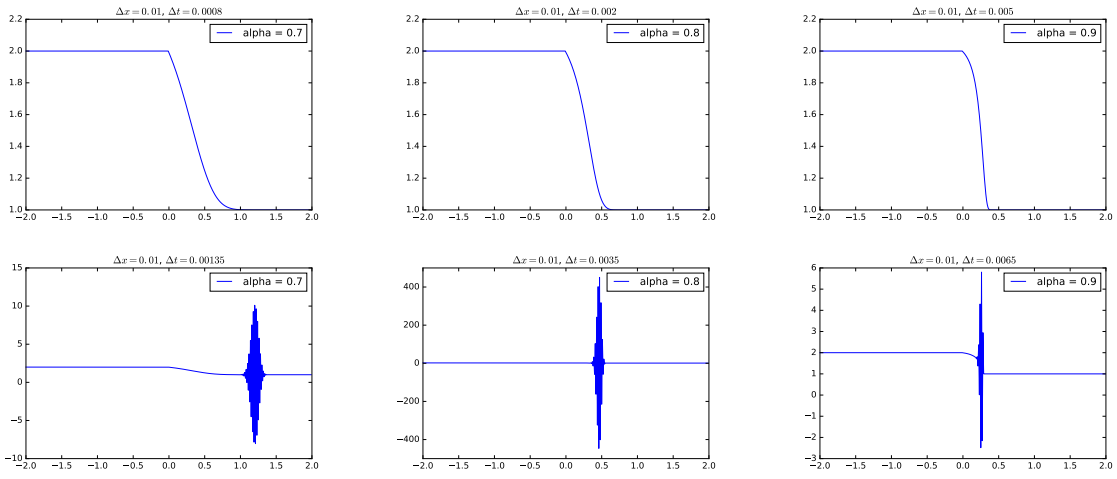


图 5–4 稳定性条件的测试。(a) $\alpha = 0.7$. 上：当 $\Delta t = 0.0008$ 时格式收敛。下：当 $\Delta t = 0.00135$ 时格式发散。(b) $\alpha = 0.8$。上：当 $\Delta t = 0.002$ 时格式收敛。下：当 $\Delta t = 0.0035$ 时格式发散。(c) $\alpha = 0.9$。上：当 $\Delta t = 0.005$ 时格式收敛。下：当 $\Delta t = 0.0065$ 时格式发散。

Fig 5–4 The stability condition test. (a) $\alpha = 0.7$. Up: scheme converges when $\Delta t = 0.0008$. Below: scheme diverges when $\Delta t = 0.00135$. (b) $\alpha = 0.8$. Up: scheme converges when $\Delta t = 0.002$. Below: scheme diverges when $\Delta t = 0.0035$. (c) $\alpha = 0.9$. Up: scheme converges when $\Delta t = 0.005$. Below: scheme diverges when $\Delta t = 0.0065$.

#### 5.4.1.2 二阶格式的收敛性和稳定性测试

在这一部分我们基本上重复了我们在一阶格式中进行的数值测试。对于收敛测试，我们仍然固定 $\Delta t = 0.0001$，并在时间 $T = 0.2$ 计算解。由于限制器的使用，MUSCL 方案的准确性可能会降到局部一阶。为了简单起见，我们选择使用连续的初始条件进行精度测试，

$$u(x, 0) = e^{-10x^2} + 1 \tag{5–35}$$

从 log-log 图5–5中可以看出这是一个二阶格式。

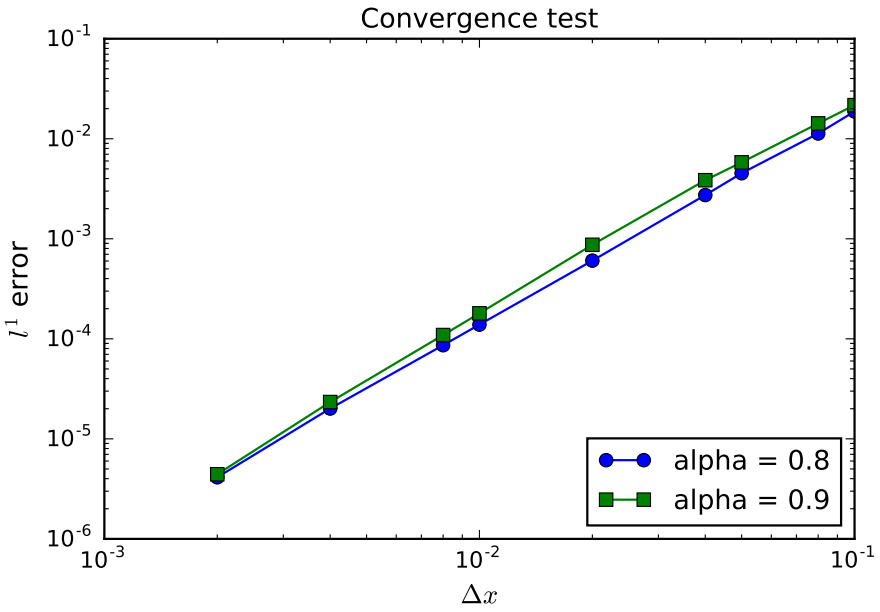


图 5–5 收敛性测试表明为 $\Delta x$ 的二阶收敛。
Fig 5–5 Convergence test shows it is a second order scheme in $\Delta x$.

稳定性测试如同前一节一样，结果显示在图5–6中。同样的可以看到对 $\Delta t$ 的强约束是的计算非常低效。

总结本节，我们指出，我们对 Burgers' 方程进行了相同的测试，并得到了类似的结果，将在本章中略去。

### 5.4.2 隐式格式的例子

在上一节中，我们说明了对于小的 $\alpha$ 使用显式方案几乎是不可行的。由于 CFL 条件的限制，一个显式格式是非常低效的，特别是当 $\alpha \to 0$。这促使我们改用隐式格式。通过使用隐式格式，我们可以进行更多的数值测试，以对 Caputo 导数的守恒律作更多地了解。

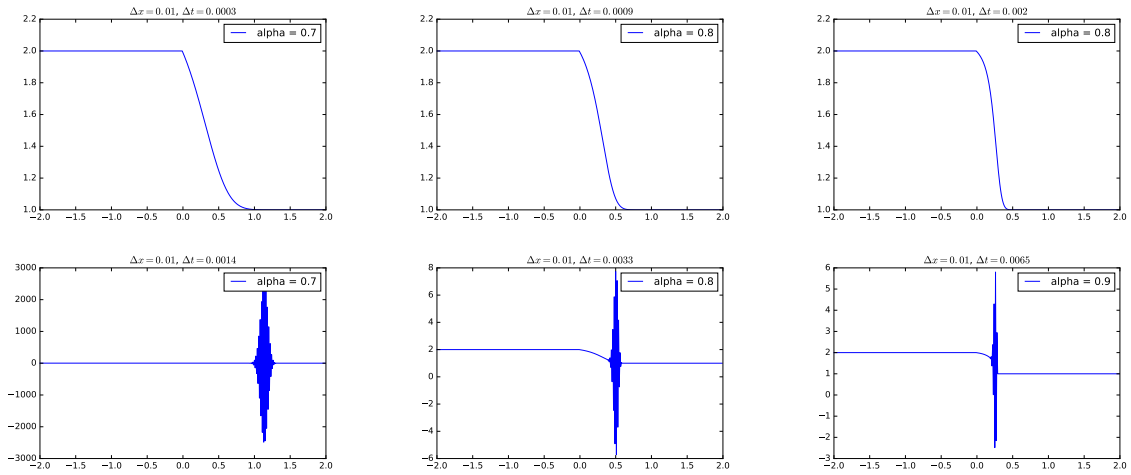


图 5–6 稳定性条件的测试。(a) $\alpha = 0.7$. 上：当 $\Delta t = 0.0003$ 时格式收敛。下：当 $\Delta t = 0.0014$ 时格式发散。(b) $\alpha = 0.8$。上：当 $\Delta t = 0.0009$ 时格式收敛。下：当 $\Delta t = 0.0033$ 时格式发散。(c) $\alpha = 0.9$。上：当 $\Delta t = 0.002$ 时格式收敛。下：当 $\Delta t = 0.0065$ 时格式发散。

Fig 5–6 The stability condition test. (a) $\alpha = 0.7$. Up: scheme converges when $\Delta t = 0.0003$. Below: scheme diverges when $\Delta t = 0.0014$. (b) $\alpha = 0.8$. Up: scheme converges when $\Delta t = 0.0009$. Below: scheme diverges when $\Delta t = 0.0033$. (c) $\alpha = 0.9$. Up: scheme converges when $\Delta t = 0.002$. Below: scheme diverges when $\Delta t = 0.0065$.

#### 5.4.2.1　隐式格式的收敛性

在本小节中，我们将测试隐性迎风格式的收敛性。标量对流方程的所有参数与上一节相同。我们仍然固定 $\Delta x = 0.01$，由于我们预计这个格式是无条件稳定的，所以我们选择 $\alpha = 0.2$，这是我们在显式的情况下做不到的。对于稳定性测试，我们选择 $\Delta t = 0.01, 0.02, 0.04, 0.06, 0.08$，这是 $O(\Delta x)$，如图5–7所示。对于收敛测试，我们现在可以固定 $\Delta t = 0.01$，这得益于无条件稳定的特征，并且观察到空间中的一阶收敛（见图5–7 右）。

无条件稳定的特性也使我们解非线性的守恒律方程，这里我们测试带 Caputo 导数的 Burges' 方程：

$$\begin{cases} \partial_t^\alpha u + u\partial_x u = 0, \\ u(x,0) = -\sin(\pi x). \end{cases} \tag{5–36}$$

对于固定的 $\alpha = 0.2, 0.5, 0.8$。和前面一样，稳定性测试我们固定 $\Delta x = 0.01$，让 $\Delta t$ 增加；收敛性测试我们固定 $\Delta t = 0.01$，让 $\Delta x$ 增加。结果显示在图5–8。在稳定性测试中，数值误差正比于 $\Delta t$ 的减少知道空间误差占主要地位，这解释了对于小的 $\Delta t$ 较平的误差曲线。

#### 5.4.2.2　数值实验：理解记忆效应

我们可以看出，由于隐式格式对于使用 Caputo 导数的守恒定律是有效和稳定的，我们将使用该格式来研究这些方程。将在本节中给出几个测试。

首先，我们分别用对流方程和 Burgers' 方程显示在不同的 $\alpha$ 下的解。在图5–9（左）中，我们

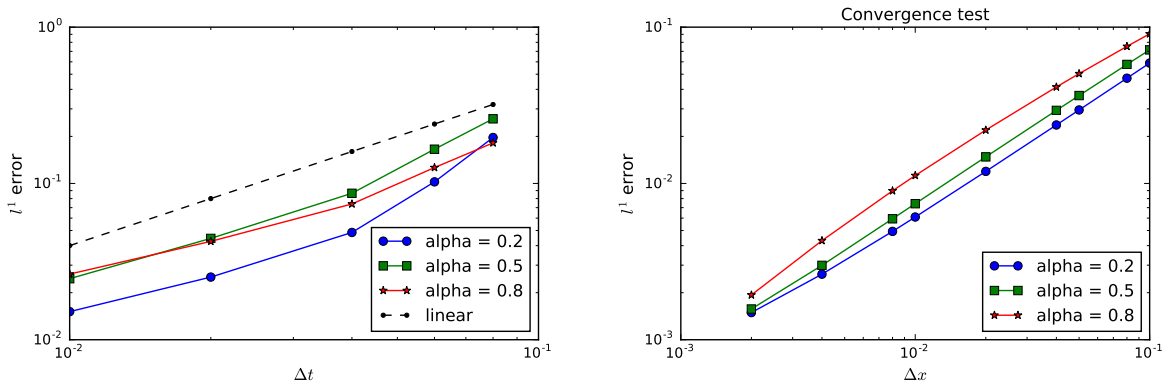


图 5–7 对于不同的 $\alpha$ 的线性对流方程的隐式迎风格式。左：稳定性测试；右：关于 $\Delta x$ 的收敛性测试。

Fig 5–7 Implicit upwind scheme for the linear advection equation with different $\alpha$. Left: stability test. Right: convergence test in $\Delta x$.

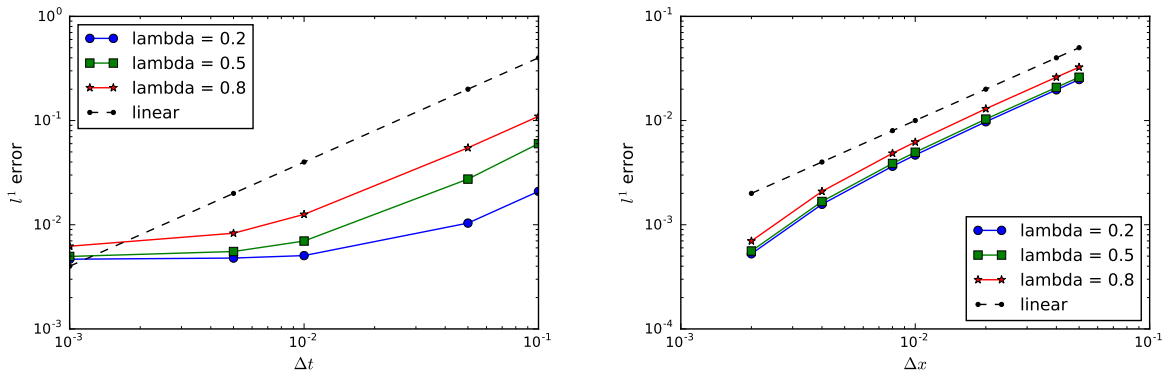


图 5–8 Burgers' 方程的隐式迎风格式。左：稳定性测试；右：关于 $\Delta x$ 的收敛性测试。

Fig 5–8 Implicit upwind scheme for the Burgers' equation. Left: stability test. Right: convergence test in $\Delta x$.

观察到间断的解表现的收敛行为，当 $\alpha \to 1$ 时，最终收敛于当 $\alpha = 1$ 时标准对流方程的解，对于 Burgers' 方程，同样的行为如图5–9（右）所示。

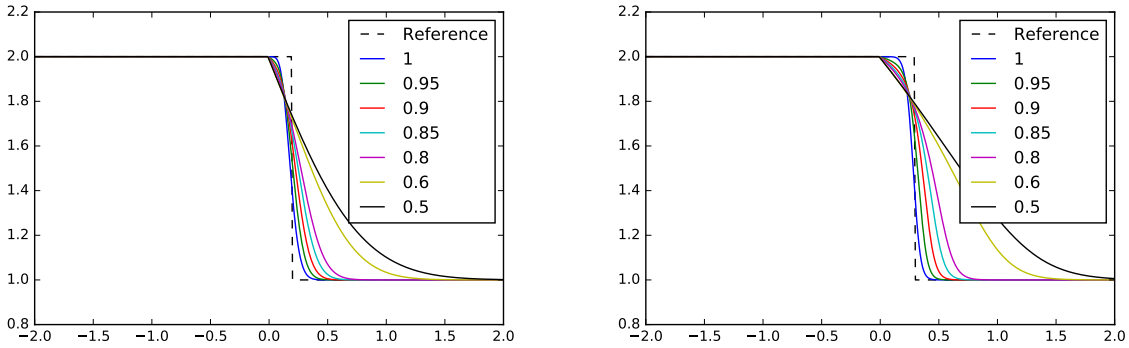


图 5–9 由隐式迎风格式计算的不同的 $\alpha$ 的解，$T = 0.2$，$a = 1$，$\Delta t = \Delta x = 0.01$。左：线性对流方程；右：Burgers 方程。

Fig 5–9 Solutions at $T = 0.2$ with different $\alpha$, $a = 1$, $\Delta t = \Delta x = 0.01$, by the implicit upwind method. Left: linear advection equation. Right: Burgers' equation

接下来考虑不均匀的记忆效应，即 $\alpha$ 依赖于 $x$ 和 $t$。在线性对流的情况，考虑

$$\alpha(x, \lambda) = 1 - \lambda \exp(-30x^2 - 7000 \times (0.5)^{12}), \tag{5–37}$$

初值为

$$u(x, 0) = \begin{cases} 0.5\cos(\pi(2x + 4)) + 0.5, & x \in [-1.5, -0.5] \\ 0, & \text{其他} \end{cases} \tag{5–38}$$

在图5–10，图5–11，图5–12中，展示了不同 $\alpha(x, t)$ 下的解。观察到由记忆效应导致的垂直方向的压缩和水平方向的扩展。

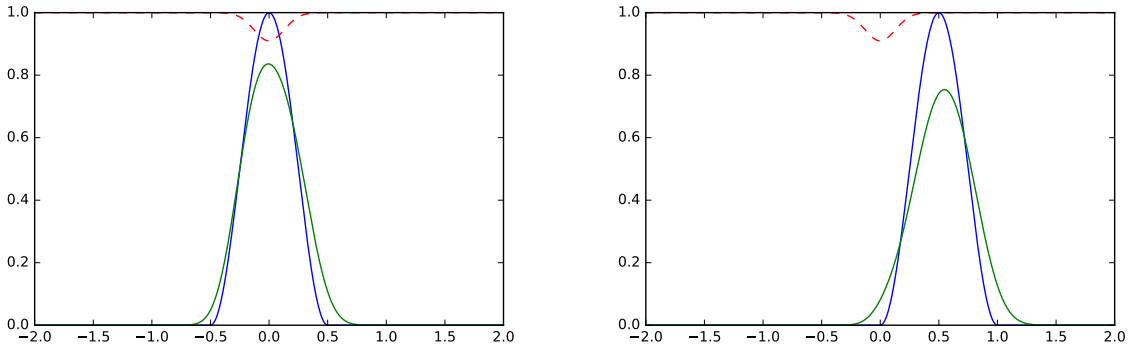


图 5–10 左：数值解（绿线）在 $T = 1$ 及 $\lambda = 0.5$（虚线）和精确解在 $T = 1$，$\alpha = 1$（蓝线）。右：数值解（绿线）在 $T = 1.5$ 及 $\lambda = 0.5$（虚线）和精确解在 $T = 1.5$，$\alpha = 1$（蓝线）。虚线为对于相应的 $\lambda$ 下的 $\alpha(x, \lambda)$。

Fig 5–10 Left: solution (green line) at $T = 1$ with $\lambda = 0.5$ (dash line) and exact solution with $T = 1$, $\alpha = 1$ (blue line). Right: solution (green line) at $T = 1.5$ with $\lambda = 0.5$ (dash line) and exact solution with $T = 1.5$, $\alpha = 1$ (blue line). Dash line is $\alpha(x, \lambda)$ with corresponding $\lambda$s.

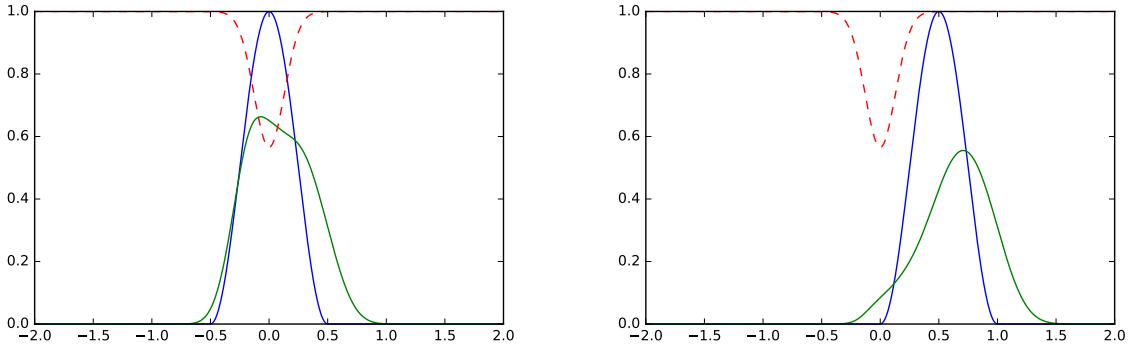


图 5–11 左：数值解（绿线）在 $T = 1$ 及 $\lambda = 2.4$（虚线）和精确解在 $T = 1$，$\alpha = 1$（蓝线）。右：数值解（绿线）在 $T = 1.5$ 及 $\lambda = 2.4$（虚线）和精确解在 $T = 1.5$，$\alpha = 1$（蓝线）。虚线为对于相应的 $\lambda$ 下的 $\alpha(x, \lambda)$。

Fig 5–11 Left: solution (green line) at $T = 1$ with $\lambda = 2.4$ (dash line) and exact solution with $T = 1$, $\alpha = 1$ (blue line). Right: solution (green line) at $T = 1.5$ with $\lambda = 2.4$ (dash line) and exact solution with $T = 1.5$, $\alpha = 1$ (blue line). Dash line is $\alpha(x, \lambda)$ with corresponding $\lambda$s.

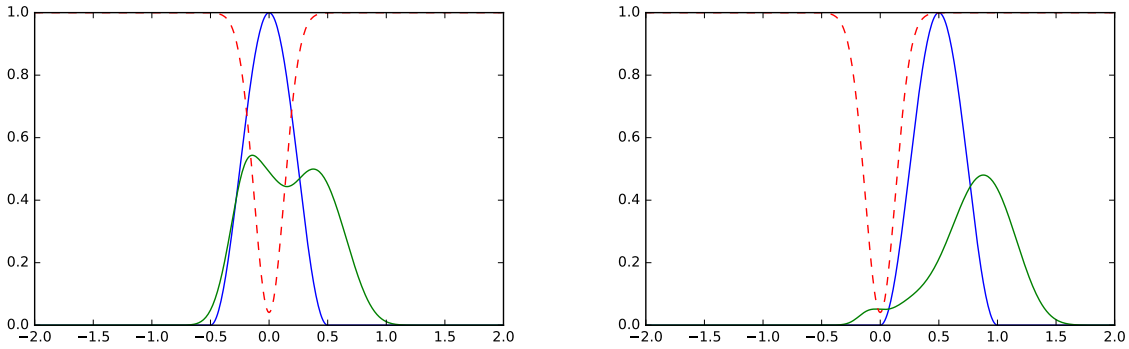


图 5–12 左：数值解（绿线）在 $T=1$ 及 $\lambda=5.3$（虚线）和精确解在 $T=1$，$\alpha=1$（蓝线）。右：数值解（绿线）在 $T=1.5$ 及 $\lambda=5.3$（虚线）和精确解在 $T=1.5$，$\alpha=1$（蓝线）。虚线为对于相应的 $\lambda$ 下的 $\alpha(x,\lambda)$。
Fig 5–12 Left: solution (green line) at $T=1$ with $\lambda=5.3$ (dash line) and exact solution with $T=1$, $\alpha=1$ (blue line). Right: solution (green line) at $T=1.5$ with $\lambda=5.3$ (dash line) and exact solution with $T=1.5$, $\alpha=1$ (blue line). Dash line is $\alpha(x,\lambda)$ with corresponding $\lambda$s.

最后我们考虑 Burgers' 方程 (5–36)，其中 $\alpha_1(x,t)=1-0.9\exp(-8|x|-7000(t-0.8)^{12})$ 及 $\alpha_2(x,t)=1$，这个例子和 Karniadakis 的文章 [129] 中的一样。然而，它们文章中得到的解受到 Gibbs 现象的破坏，这是由于它们在间断处使用了伪谱方法。这里，我们的结果是没有数值振荡的，同时也观察到了压缩效应，见图5–13和图5–14。

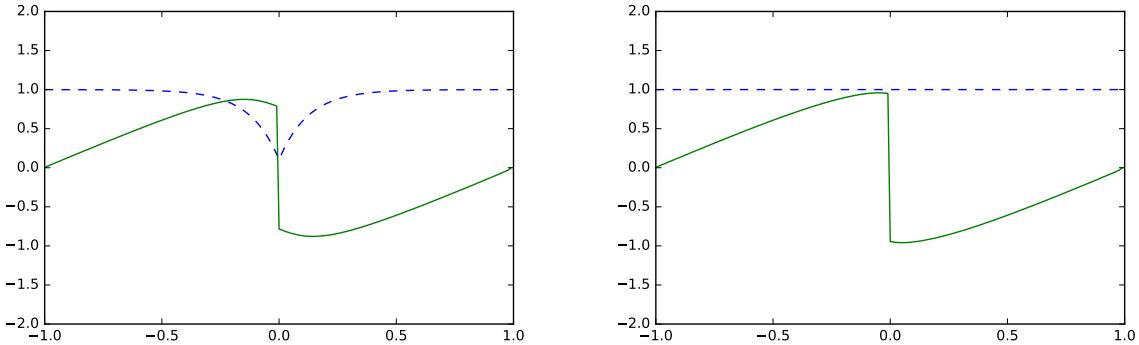


图 5–13 左：数值解（实线）在 $T=0.5$ 及 $\alpha=\alpha_1(x,t)$ (虚线)，$\Delta t=\Delta x=0.01$。右：数值解（实线）在 $T=0.5$ 及 $\alpha=1$，$\Delta t=\Delta x=0.01$。由隐式迎风格式得到。
Fig 5–13 Left: solution (solid line) at $T=0.5$ with $\alpha=\alpha_1(x,t)$ (dash line), $\Delta t=\Delta x=0.01$. Right: solution (solid line) at $T=0.5$ with $\alpha=1$, $\Delta t=\Delta x=0.01$ by using the implicit upwind method with fast sweeping method [130].

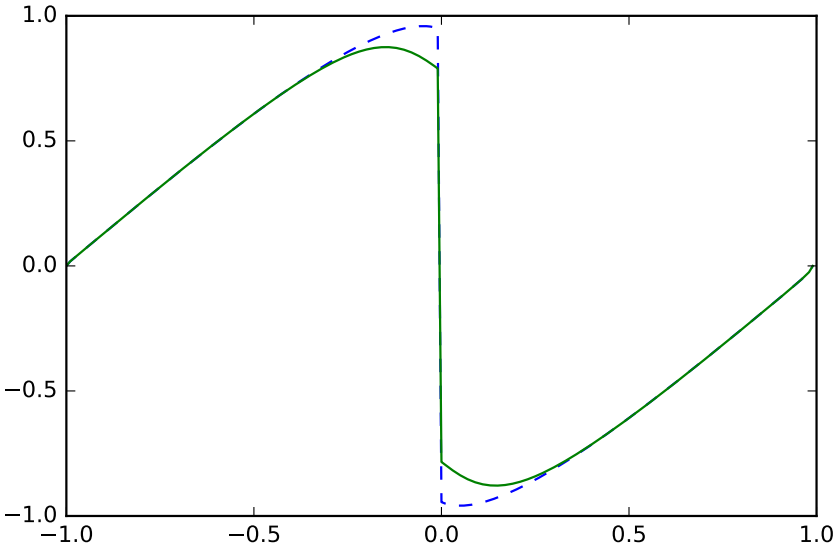


图 5–14 左：数值解（实线）在 $T = 0.5$ 及 $\alpha = \alpha_1(x,t)$ (虚线)，$\Delta t = \Delta x = 0.01$。右：数值解（实线）在 $T = 0.5$ 及 $\alpha = 1$，$\Delta t = \Delta x = 0.01$。由隐式迎风格式得到。

Fig 5–14 Solution (green solid line) at $T = 0.5$ with $\alpha = \alpha_1(x,t)$ and solution (blue dash line) at $T = 0.5$ with $\alpha = 1$, $\Delta t = \Delta x = 0.01$ by using the implicit upwind method with fast sweeping method [130].

## 5.5 本章总结与展望

本章中，我们考虑了带有 Caputo（分数阶）时间导数的守恒律方程，提出了相应的一阶、二阶显式、隐式迎风格式，对于稳定性和 TVD 特性进行了分析。我们通过大量的数值实验，包括 Burgers 方程等来说明我们的格式的可行性。同时，利用构造的格式来研究分数阶导数带来的“记忆效应”的现象，加深了对这类方程的理解。

对于带有分数阶的 PDE，是近年来非常热门的研究领域，无论在分析还是在计算方法的研究上，都非常有价值，尤其是在多空介质中的物理问题的记忆效应的研究中。然而，相关的数值分析与理论分析结果都非常有限，在未来将会对此继续进行深入的研究。

# 全文总结与展望

在本文中，我们讨论了对于一些物理中的波动方程与输运方程相关问题的数值算法与分析。主要包括带有不确定性的（随机性）的和经典的两大类问题。

对于不确定量化，首先，我们研究了带有间断与随机系数的双曲型方程的数值解法。为了克服由解的不光滑导致的 gPC-SG 方法收敛速度很慢的问题，我们提出了离散 gPC-SG 方法，利用离散的解具有较好的光滑性，进而改进 gPC-SG 方法的收敛速度。对于对流方程我们进行了收敛性的分析与误差估计。同时为了说明方法的有效性，我们进行了大量的数值实验，包括对流方程与刘维尔方程；对于线性输运方程，由于多尺度与不确定性的同时存在，我们建立了 gPC-SG 方法对带有随机散射系数的线性输运方程关于克努森数一致的谱精度分析，从而允许我们证明该方法随机渐近保持性质（s-AP）。对于基于 micro-macro 分解的全离散格式，我们证明了一致的稳定性结果。这是首次有人证明了关于这类问题的一致性结果。

对于传统的算法的改进，我们在第四章中，提出并分析了具有向量势的半经典薛定谔方程的新的时间分裂谱方法，其中在对流部分的半拉格朗日方法的插值步骤中应用 NUFFT 技术。分析了近似波函数和计算物理观测值的方法的稳定性和准确性。在最后，我们对近年来刚刚兴起的分数阶导数的波动方程进行了数值研究和分析，提出了相应的一阶、二阶显式、隐式迎风格式，对于稳定性和 TVD 特性进行了分析。我们通过大量的数值实验，包括 Burgers 方程等来说明我们的格式的可行性。借助于这类格式，我们得以通过数值实验来帮助我们理解所谓的记忆效应。

所有以上的问题中我们都进行了严格的分析论证，同时进行了大量的数值实验来说明我们方法的优越性。但是，还有很多没有解决的问题，例如离散 gPC-SG 方法在高阶格式、非线性问题的构造上仍然存在一定困难，同时如何将该方法向更广泛的问题上进行推广仍然需要大量的研究；关于一致性的分析在带有多尺度与不确定性的问题非常重要，我们的分析对于线性问题具有非常一般的指导作用，但对于非线性的问题，仍然非常困难；不确定量化中 gPC-SG 方法如何保持原方程的双曲性、一致谱收敛的结果能否向更多类型的方程推广；对于带有分数阶的 PDE 无论在分析还是在计算方法的研究上，都非常有价值，然而相关的数值分析与理论分析结果都非常有限，如何更好的理解记忆效应及相关的数学理论等等；这些将作为我以后的主要研究方向。

# 附录 A 质量和能量守恒的证明

在本附录中，我们简要介绍了第四章中薛定谔方程的质量和能量守恒的证明。我们假设波函数在 Schwartz 空间 $S(\mathbb{R}^d)$ 中。

## A.1 质量守恒

**证明.** 对于 $u^\varepsilon \in C(\mathbb{R}_t; L^2(\mathbb{R}^d))$，首先注意到 $(-i\varepsilon\nabla - \mathbf{A})$ 是自共轭算子，即，

$$\langle(-i\varepsilon\nabla - \mathbf{A})f, g\rangle = -i\varepsilon\langle\nabla f, g\rangle - \langle\mathbf{A}f, g\rangle = \langle f, (-i\varepsilon\nabla - \mathbf{A})g\rangle. \tag{A–1}$$

直接计算得出

$$\frac{d}{dt}\|u^\varepsilon\|_{L^2} = \frac{d}{dt}\langle u^\varepsilon, u^\varepsilon\rangle = \langle\partial_t u^\varepsilon, u^\varepsilon\rangle + \langle u^\varepsilon, \partial_t u^\varepsilon\rangle = 0.$$

□

## A.2 能量守恒

**证明.** 对 $\mathcal{E}(t)$ 求导，我们有

$$\frac{d}{dt}\mathcal{E}(t) = (I) + (II),$$

其中

$$\begin{aligned}(I) &:= \frac{1}{2}\langle(-i\varepsilon\nabla - \mathbf{A})\partial_t u^\varepsilon, (-i\varepsilon\nabla - \mathbf{A})u^\varepsilon\rangle + \frac{1}{2}\langle(-i\varepsilon\nabla - \mathbf{A})u^\varepsilon, (-i\varepsilon\nabla - \mathbf{A})\partial_t u^\varepsilon\rangle,\\ (II) &:= \langle\partial_t u^\varepsilon, Vu^\varepsilon\rangle + \langle u^\varepsilon, V\partial_t u^\varepsilon\rangle.\end{aligned}$$

由 $(-i\varepsilon\nabla - \mathbf{A})$ 是自共轭算子，我们有

$$\begin{aligned}(I) =&\frac{1}{2}\langle(-i\varepsilon\nabla - \mathbf{A})\left(\frac{1}{2i\varepsilon}(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon + \frac{V}{i\varepsilon}u^\varepsilon\right), (-i\varepsilon\nabla - \mathbf{A})u^\varepsilon\rangle\\ &+ \frac{1}{2}\langle(-i\varepsilon\nabla - \mathbf{A})u^\varepsilon, (-i\varepsilon\nabla - \mathbf{A})\left(\frac{1}{2i\varepsilon}(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon + \frac{V}{i\varepsilon}u^\varepsilon\right)\rangle\\ =&\frac{1}{2}\langle\frac{1}{2i\varepsilon}(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon + \frac{V}{i\varepsilon}u^\varepsilon, (-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon\rangle\\ &+ \frac{1}{2}\langle(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon, \frac{1}{2i\varepsilon}(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon + \frac{V}{i\varepsilon}u^\varepsilon\rangle\\ =&\frac{1}{2i\varepsilon}\langle Vu^\varepsilon, (-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon\rangle - \frac{1}{2i\varepsilon}\langle(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon, Vu^\varepsilon\rangle.\end{aligned}$$

类似的，

$$\begin{aligned}(II) =&\langle\frac{1}{2i\varepsilon}(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon + \frac{V}{i\varepsilon}u^\varepsilon, Vu^\varepsilon\rangle + \langle Vu^\varepsilon, \frac{1}{2i\varepsilon}(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon + \frac{V}{i\varepsilon}u^\varepsilon\rangle\\ =&\frac{1}{2i\varepsilon}\langle(-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon, Vu^\varepsilon\rangle - \frac{1}{2i\varepsilon}\langle Vu^\varepsilon, (-i\varepsilon\nabla - \mathbf{A})^2 u^\varepsilon\rangle = -(I).\end{aligned}$$

因此我们有 $(I) + (II) = 0$，也就说明了 $\mathcal{E}(t) = \mathcal{E}(0)$。 □

# 参考文献

[1] JIN S. Numerical methods for hyperbolic systems with singular coefficients: well-balanced scheme, Hamiltonian preservation and beyond[C]//Proc. of the 12th International Conference on Hyperbolic Problems: Theory, Numerics, Applications, Univeristy of Maryland, College Park: Proceedings of Symposia in Applied Mathematics: vol. 67: 1. American Mathematical Society, 2009: 93-104.

[2] JIN S, WEN X. Hamiltonian-preserving schemes for the Liouville equation with discontinuous potentials[J]. Communications in Mathematical Sciences, 2005, 3(3): 285-315.

[3] JIN S, ZHOU Z. A semi-Lagrangian time splitting method for the Schrödinger equation with vector potentials[J]. Communications in Information and Systems, 2013, 13: 247-289.

[4] BAO W, JIN S, MARKOWICH P. Time-splitting spectral approximations for the Schrödinger equation in the semiclassical regime[J]. J. Compt. Phys., 2002, 175: 487-524.

[5] GARDINER C. Handbook of stochastic methods: for physics, chemistry and the natural sciences[M]. Springer-Verlag, 1985.

[6] KLOEDEN P, PLATEN E. Numerical Solution of Stochastic Differential Equations[M]. Springer-Verlag, 1999.

[7] KARATZAS I, SHREVE S. Brownian Motion and Stochastic Calculus[M]. Springer-Verlag, 1988.

[8] OKSENDAL B. Stochastic differential equations. An introduction with applications[M]. Springer-Verlag, 1998.

[9] FISHMAN G. Monte Carlo: Concepts, Algorithms, and Applications[M]. New York: Springer-Verlag, 1996.

[10] LOH W L, et al. On Latin hypercube sampling[J]. The annals of statistics, 1996, 24(5): 2058-2080.

[11] STEIN M. Large sample properties of simulations using Latin hypercube sampling[J]. Technometrics, 1987, 29(2): 143-151.

[12] FOX B L. Strategies for Quasi-Monte Carlo: vol. 22[M]. Springer Science & Business Media, 1999.

[13] NIEDERREITER H. Random number generation and quasi-Monte Carlo methods[M]. SIAM, 1992.

[14] NIEDERREITER H, HELLEKALEK P, LARCHER G, et al. Monte Carlo and quasi-Monte Carlo methods 1996: proceedings of a conference at the University of Salzburg, Austria, July 9-12, 1996: vol. 127[M]. Springer Science & Business Media, 2012.

[15] GILES M B. Multilevel Monte Carlo methods[J]. Acta Numerica, 2015, 24: 259.

[16] LIU W K, BELYTSCHKO T, MANI A. Probabilistic finite elements for nonlinear structural dynamics[J]. Computer Methods in Applied Mechanics and Engineering, 1986, 56(1): 61-81.

[17] LIU W K, BELYTSCHKO T, MANI A. Random field finite elements[J]. International journal for numerical methods in engineering, 1986, 23(10): 1831-1845.

[18] ZHANG D. Stochastic methods for flow in porous media: coping with uncertainties[M]. Academic press, 2001.

[19] SHINOZUKA M, DEODATIS G. Response variability of stochastic finite element systems [J]. Journal of Engineering Mechanics, 1988, 114(3): 499-519.

[20] YAMAZAKI F, MEMBER A, SHINOZUKA M, et al. Neumann expansion for stochastic finite element analysis[J]. Journal of Engineering Mechanics, 1988, 114(8): 1335-1354.

[21] DEODATIS G. Weighted integral method. I: stochastic stiffness matrix[J]. Journal of Engineering Mechanics, 1991, 117(8): 1851-1864.

[22] DEODATIS G, SHINOZUKA M. Weighted integral method. II: response variability and reliability[J]. Journal of Engineering Mechanics, 1991, 117(8): 1865-1877.

[23] XIU D, KARNIADAKIS G. The Wiener-Askey polynomial chaos for stochastic differential equations[J]. SIAM J. Sci. Comput., 2002, 24(2): 619-644.

[24] GHANEM R G, SPANOS P D. Stochastic Finite Elements: A Spectral Approach[M]. New York: Springer Verlag, 1991.

[25] WIENER N. The homogeneous chaos[J]. Amer. J. Math., 1938, 60: 897-936.

[26] CHORIN A J. Gaussian fields and random flow[J]. Journal of Fluid Mechanics, 1974, 63(01): 21-32.

[27] ORSZAG S A, BISSONNETTE L. Dynamical Properties of Truncated Wiener-Hermite Expansions[J]. The Physics of Fluids, 1967, 10(12): 2603-2613.

[28] XIU D, KARNIADAKIS G E. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos[J]. Computer methods in applied mechanics and engineering, 2002, 191(43): 4927-4948.

[29] XIU D, KARNIADAKIS G E. Modeling uncertainty in flow simulations via generalized polynomial chaos[J]. Journal of computational physics, 2003, 187(1): 137-167.

[30] BABUSKA I, TEMPONE R, ZOURARIS G E. Galerkin finite element approximations of stochastic elliptic partial differential equations[J]. SIAM Journal on Numerical Analysis, 2004, 42(2): 800-825.

[31] LE MAITRE O, NAJM H, GHANEM R, et al. Multi-resolution analysis of Wiener-type uncertainty propagation schemes[J]. Journal of Computational Physics, 2004, 197(2): 502-531.

[32] WAN X, KARNIADAKIS G E. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations[J]. Journal of Computational Physics, 2005, 209(2): 617-642.

[33] XIU D, HESTHAVEN J S. High-order collocation methods for differential equations with random inputs[J]. SIAM Journal on Scientific Computing, 2005, 27(3): 1118-1139.

[34] MATHELIN L, HUSSAINI M Y, ZANG T A. A stochastic collocation algorithm for uncertainty analysis[J]. 2003.

[35] TATANG M A, PAN W, PRINN R G, et al. An efficient method for parametric uncertainty analysis of numerical geophysical models[J]. Journal of Geophysical Research: Atmospheres, 1997, 102(D18): 21925-21932.

[36] BABUŠKA I, NOBILE F, TEMPONE R. A stochastic collocation method for elliptic partial differential equations with random input data[J]. SIAM Journal on Numerical Analysis, 2007, 45(3): 1005-1034.

[37] XIU D. Efficient collocational approach for parametric uncertainty analysis[J]. Commun. Comput. Phys, 2007, 2(2): 293-309.

[38] XIU D. Fast numerical methods for stochastic computations: a review[J]. Communications in computational physics, 2009, 5(2-4): 242-272.

[39] JIN S, QI P. $\ell^1$-error estimates on the immersed interface upwind scheme for linear convection equations with piecewise constant coefficients: A simple proof[J]. Science China Mathematics, 2013.

[40] JIN S, WEN X. Hamiltonian-preserving schemes for the Liouville equation of geometrical optics with partial transmissions and reflections[J]. SIAM J. Num. Anal., 2006, 44: 1801-1828.

[41] JIN S, XIU D, ZHU X. Asymptotic-preserving methods for hyperbolic and transport equations with random inputs and diffusive scalings[J/OL]. J. Comput. Phys., 2015, 289: 35-52. http://dx.doi.org/10.1016/j.jcp.2015.02.023. DOI: 10.1016/j.jcp.2015.02.023.

[42] DUTT A, ROKHLIN V. Fast Fourier transforms for nonequispaced data[J]. SIAM J. Sci. Comput., 1993, 14: 1368-1393.

[43] GREENGARD L, LEE J Y. Accelerating the nonuniform fast Fourier transform[J]. SIAM Rev., 2004, 46: 443-454.

[44] BIJL H, LUCOR D, MISHRA S, et al. Uncertainty Quantification in Computational Fluid Dynamics: vol. 92[M]. Springer, 2013.

[45] GUNZBURGER M D, WEBSTER C G, ZHANG G. Stochastic finite element method for partial differential equations with random input data[J]. Acta Numer., 2014, 23: 521-650.

[46] MAITRE O L, KNIO O. Spectral Methods for Uncertainty Quantification[M]. New York: Springer, 2010.

[47] PETTERSSON M, IACCARINO G, NORDSTRÖM J. Polynomial Chaos Methods for Hyperbolic Differential Equations[M]. Switzerland: Springer, 2015.

[48] TRYOEN J, LE MAÎTRE O, NDJINGA M, et al. Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems[J]. J. Comput. Phys., 2010, 229(18): 6485-6511.

[49] XIU D. Numerical Methods for Stochastic Computations[M]. Princeton University Press, 2010.

[50] MOTAMED M, NOBILE F, TEMPONE R. A stochastic collocation method for the second order wave equation with a discontinuous random speed[J]. Numer. Math, 2012, 123: 493-536.

[51] ZHOU T, TANG T. Convergence Analysis for Spectral Approximation to a Scalar Transport Equation with a Random Wave Speed[J]. Journal of Computational Mathematics, 2012, 30(6): 643-656.

[52] LEVEQUE R. Finite Volume Methods for Hyperbolic Problems[M]. Cambridge University Press, 2002.

[53] KURGANOV A, TADMOR E. New High-Resolution Central Schemes for Nonlinear Conservation Laws and Convection-Diffusion Equations[J]. Journal of Computational Physics, 2000, 160(1): 241-282.

[54] NESSYAHU H, TADMOR E. Non-oscillatory central differencing for hyperbolic conservation laws[J]. 1990.

[55] CHOI H, LIU J G. The reconstruction of upwind fluxes for conservation laws: its behavior in dynamic and steady state calculations[J]. Journal of Computational Physics, 1998, 144(2): 237-256.

[56] GOTTLIEB D, XIU D. Galerkin method for wave equations with uncertain coefficients[J]. Comm. Comp. Phys., 2008, 3: 505-518.

[57] TANG T, ZHOU T. Convergence Analysis for Stochastic Collocation Methods to Scalar Hyperbolic Equations with a Random Wave Speed[J]. Communications in Computational Physics, 2010.

[58] CANUTO C, QUARTERONI A. Approximation results for orthogonal polynomials in Sobolev spaces[J]. Mathematics of Computation, 1982, 38(157): 67-86.

[59] JIN S, NOVAK K A. A Semiclassical Transport Model for Thin Quantum Barriers[J]. Multiscale Model Simul, 2006, 5(4): 1063-1086.

[60] DESPRES B, POETTE G, LUCOR D. Robust uncertainty propagation in systems of conservation laws with the entropy closure method in Uncertainty Quantication in Computational

Fluid Dynamics[G]//Lect. Notes Comput. Sci. Eng. Vol. 92. Heidelberg: Springer, 2013: 105-149.

[61] HU J, JIN S, XIU D. A stochastic Galerkin method for Hamilton-Jacobi equations with uncertainty[J]. SIAM J. Sci. Comput., 2015, 37: A2246-A2269.

[62] GHANEM R, SPANOS P. Stochastic Finite Elements: a Spectral Approach[M]. Springer-Verlag, 1991.

[63] XIU D. Numerical methods for stochastic computations[M]. Princeton, New Jersey: Princeton Univeristy Press, 2010.

[64] FICHTL E D, PRINJA A K, WARSA J S. Stochastic methods for uncertainty quantification in radiation transport[C]//International Conference on Mathematics, Computational Methods & Reactor Physics (M&C 2009), Saratoga Springs, New York, May 3-7, 2009. 2009.

[65] LARSEN E W, KELLER J B. Asymptotic solution of neutron transport problems for small mean free paths[J]. J. Mathematical Phys., 1974, 15: 75-81.

[66] BENSOUSSAN A, LIONS J L, PAPANICOLAOU G. Asymptotic analysis for periodic structures: vol. 5[M]. North-Holland Publishing Co., Amsterdam-New York, 1978: xxiv+700.

[67] BARDOS C, SANTOS R, SENTIS R. Diffusion approximation and computation of the critical size[J/OL]. Trans. Amer. Math. Soc., 1984, 284(2): 617-649. http://dx.doi.org/10.2307/1999099. DOI: 10.2307/1999099.

[68] LARSEN E W, MOREL J E. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. II[J]. J. Comput. Phys., 1989, 83(1): 212-236.

[69] LARSEN E W, MOREL J E, MILLER W F, Jr. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes[J]. J. Comput. Phys., 1987, 69(2): 283-324.

[70] GOLSE F, JIN S, LEVERMORE C D. The convergence of numerical transfer schemes in diffusive regimes. I. Discrete-ordinate method[J/OL]. SIAM J. Numer. Anal., 1999, 36(5): 1333-1369. http://dx.doi.org.ezproxy.library.wisc.edu/10.1137/S0036142997315986. DOI: 10.1137/S0036142997315986.

[71] JIN S. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations [J/OL]. SIAM J. Sci. Comput., 1999, 21(2): 441-454. http://dx.doi.org.ezproxy.library.wisc.edu/10.1137/S1064827598334599. DOI: 10.1137/S1064827598334599.

[72] JIN S, PARESCHI L, TOSCANI G. Uniformly accurate diffusive relaxation schemes for multiscale transport equations[J]. SIAM J. Numer. Anal., 2000, 38(3): 913-936.

[73] KLAR A. An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit[J/OL]. SIAM J. Numer. Anal., 1998, 35(3): 1073-1094 (electronic). http://dx.doi.org.ezproxy.library.wisc.edu/10.1137/S0036142996305558. DOI: 10.1137/S0036142996305558.

[74] LEMOU M, MIEUSSENS L. A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit[J/OL]. SIAM J. Sci. Comput., 2008, 31(1): 334-368. http://dx.doi.org.ezproxy.library.wisc.edu/10.1137/07069479X. DOI: 10.1137/07069479X.

[75] JIN S. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review[J]. Lecture Notes for Summer School on "Methods and Models of Kinetic Theory" (M&MKT), Porto Ercole (Grosseto, Italy), 2010.

[76] HU J, JIN S. A stochastic Galerkin method for the Boltzmann equation with uncertainty[J]. J. Comp. Phys., 2016, 315: 150-168.

[77] JIN S, LIU L. An Asymptotic-Preserving Stochastic Galerkin Method for the Semiconductor Boltzmann Equation with Random Inputs and Diffusive Scalings[J]. SIAM Multiscale Modeling and Simulation, 2017, 15: 157-183.

[78] JIN S, LU H. An Asymptotic-Preserving Stochastic Galerkin Method for the Radiative Heat Transfer Equations with Random Inputs and Diffusive Scalings[J]. J. Comp. Phys., 2017, 334: 182-206.

[79] JIN S, TANG M, HAN H. On a uniformly second order numerical method for the one-dimensional discrete-ordinate transport equation and its diffusion limit with interface[J]. Networks and Heterogeneous Media, 2009, 4: 35-65.

[80] JIN S, LI Q, WANG L. The uniform convergence of generalized polynomial chaos based methods for multiscale transport equation with random scattering[J]. preprint,

[81] LIU J G, MIEUSSENS L. Analysis of an asymptotic preserving scheme for linear kinetic equations in the diffusion limit[J/OL]. SIAM J. Numer. Anal., 2010, 48(4): 1474-1491. http://dx.doi.org/10.1137/090772770. DOI: 10.1137/090772770.

[82] LIU T P, YU S H. Boltzmann equation: micro-macro decompositions and positivity of shock profiles[J/OL]. Comm. Math. Phys., 2004, 246(1): 133-179. http://dx.doi.org/10.1007/s00220-003-1030-2. DOI: 10.1007/s00220-003-1030-2.

[83] KLAR A, SCHMEISER C. Numerical passage from radiative heat transfer to nonlinear diffusion models[J/OL]. Math. Models Methods Appl. Sci., 2001, 11(5): 749-767. http://dx.doi.org/10.1142/S0218202501001082. DOI: 10.1142/S0218202501001082.

[84] BENNOUNE M, LEMOU M, MIEUSSENS L. Uniformly stable numerical schemes for the Boltzmann equation preserving the compressible Navier-Stokes asymptotics[J/OL]. J. Comput. Phys., 2008, 227(8): 3781-3803. http://dx.doi.org/10.1016/j.jcp.2007.11.032. DOI: 10.1016/j.jcp.2007.11.032.

[85] JIN S, LEVERMORE D. The discrete-ordinate method in diffusive regimes[J/OL]. Transport Theory Statist. Phys., 1991, 20(5-6): 413-439. http://dx.doi.org.ezproxy.library.wisc.edu/10.1080/00411459108203913. DOI: 10.1080/00411459108203913.

[86] LEMOU M, MÉHATS F. Micro-Macro Schemes for Kinetic Equations Including Boundary Layers[J]. SIAM J. Sci. Comput., 2012, 34(6): 734-760.

[87] SCULLY M, ZUBAIRY M. Quantum optics[J]. 1997.

[88] AVRON J, HERBST I, SIMON B. Schrödinger operators with magnetic fields. I. General interactions[J]. Duke Math. J., 1978, 45: 847-883.

[89] ENGQUIST B, RUNBORG O. Computational high frequency wave propagation[J]. Acta Numer., 2003, 12: 181-266.

[90] JIN S, OSHER S. A level set method for the computation of multi-valued solutions to quasi-linear hyperbolic PDE's and Hamilton-Jacobi equations[J]. Commun. Math. Sci., 2003, 1: 575-591.

[91] JIN S, LI X. Multi-phase computations of the semiclassical limit of the Schrödinger equation and related problems: Whitham vs Wigner[J]. Phys. D: Nonlinear Phenom., 2003, 182.

[92] JIN S, MARKOWICH P A, SPARBER C. Mathematical and computational methods for semiclassical Schrödinger equations[J]. Acta Numer., 2011, 20: 121-209.

[93] HELLER E. Time dependent approach to semiclassical dynamics[J]. J. Chem. Phys., 1975, 62: 1544-1555.

[94] POPOV M M. A new method of computation of wave fields using Gaussian beams[J]. Wave Motion, 1982, 4: 85-97.

[95] RALSTON J. Gaussian beams and thepropagation of singularities[J]. Studies in partial differential equations, 1982, 23: 206-248.

[96] JIN S, WU H, YANG X. Gaussian beam methods for the Schrödinger equation in the semiclassical regime: Lagrangian and Eulerian formulations[J]. Commun. Math. Sci., 2008, 6: 995-1020.

[97] TANUSHEV N. Superpositions and higher order Gaussian beams[J]. Commun. Math. Sci., 2008, 6: 449-475.

[98] JIN S, WU H, YANG X. Semi-Eulerian and high order Gaussian beam methods for the Schrödinger equation in the semiclassical regime[J]. Commun. Comput. Phys., 2011, 9: 668-687.

[99] LIU H, RUNBORG O, Tanushev. Error estimates for Gaussian beams[J]. Math. Comp., 2013, 82: 919-952.

[100] ZHOU Z. Numerical approximation of the Schrödinger equation with the electromagnetic field by the Hagedorn wave packets[J]. J. Comput. Phys., 2014, 272: 386-40.

[101] HAGEDORN G. Raising and lowering operators for semi-classical wave packets[J]. Ann. Phys., 1998, 269: 77-104.

[102] GRADINARU V, HAGEDORN G. Convergence of a semiclassical wavepacket based time-splitting for the Schrödinger equation[J]. Numer. Math., 2013, 126: 1-21.

[103] RUSSO G, SMEREKA P. The Gaussian wave packet transform: Efficient computation of the semi-classical limit of the Schrödinger equation. Part 1-Formulation and the one dimensional case[J]. J. Comput. Phys., 2013, 233: 192-209.

[104] RUSSO G, SMEREKA P. The Gaussian wave packet transform: Efficient computation of the semi-classical limit of the Schrödinger equation. Part 2. Multidimensional case[J]. J. Comput. Phys., 2014, 257: 1022-1038.

[105] BAO W, JIN S, MARKOWICH P. Numerical studies of time-splitting spectral discretizations of nonlinear Schrödinger equations in the semiclassical regime[J]. SIAM J. Sci. Compt., 2003, 25: 27-64.

[106] SHEN J, TANG T. Spectral and High-Order Methods with Applications[M]. Beijing: Science Press, 2006.

[107] SÜLI E, WARE A. A spectral method of characteristics for hyperbolic problems[J]. SIAM J. Appl. Math., 1991, 28: 423-445.

[108] CAPUTO M. Diffusion of fluids in porous media with memory[J]. Geothermics, 1999, 28(1): 113-130.

[109] DEL-CASTILLO-NEGRETE D, CARRERAS B, LYNCH V. Fractional diffusion in plasma turbulence[J]. Phys. Plasmas, 2004, 11(8): 3854-3864.

[110] DEL-CASTILLO-NEGRETE D, CARRERAS B, LYNCH V. Nondiffusive transport in plasma turbulence: a fractional diffusion approach[J]. Phys. Rev. Lett., 2005, 94(6): 065003.

[111] METZLER R, KLAFTER J. The random walk's guide to anomalous diffusion: a fractional dynamics approach[J]. Phys. Rep., 2000, 229(1): 1-77.

[112] ZASLAVSKY G. Chaos, fractional kinetics, and anomalous transport[J]. Phys. Rep., 2002, 371(6): 461-580.

[113] ALLEN M, CAFFARELLI L, VASSEUR A. Porous Medium Flow with both a Fractional Potential Pressure and Fractional Time Derivative[J]. arXiv preprint, 2015.

[114] CAO J, XU C. A high order schema for the numerical solution of the fractional ordinary differential equations[J]. J. of Comput. Phys., 2013, 238(C): 154-168.

[115] LIN Y, XU C. Finite difference/spectral approximations for the time-fractional diffusion equation[J]. J. Comput. Phys., 2007.

[116] ZHAO X, SUN Z, KARNIADAKIS G. Second-order approximations for variable order fractional derivatives: Algorithms and applications[J]. J. Comput. Phys., 2015, 293(C): 184-200.

[117] CAO J, XU C. A high order schema for the numerical solution of the fractional ordinary differential equations[J/OL]. J. of Comput. Phys., 2013, 238: 154-168. http://www.sciencedi

rect.com/science/article/pii/S0021999112007449. DOI: http://dx.doi.org/10.1016/j.jcp.2012.12.013.

[118] KUMAR P, AGRAWAL O. An approximate method for numerical solution of fractional differential equations[J/OL]. Signal Process., 2006, 86(10): 2602-2610. http://www.sciencedirect.com/science/article/pii/S0165168406000466. DOI: http://dx.doi.org/10.1016/j.sigpro.2006.02.007.

[119] LIU F, ZHUANG P, ANH V, et al. Stability and convergence of the difference methods for the space-time fractional advection-diffusion equation[J]. Appl. Math. Comput., 2007, 191(1): 12-20.

[120] ZHANG H, LIU F, PHANIKUMAR M S, et al. A novel numerical method for the time variable fractional order mobile-immobile advection-dispersion model[J]. Comput. Math. Appl., 2013, 66(5): 693-701.

[121] LANGLANDS T, HENRY B. The accuracy and stability of an implicit solution method for the fractional diffusion equation[J/OL]. J. Comput. Phys., 2005, 205(2): 719-736. http://www.sciencedirect.com/science/article/pii/S0021999104004887. DOI: http://dx.doi.org/10.1016/j.jcp.2004.11.025.

[122] SUN Z, WU X. A fully discrete difference scheme for a diffusion-wave system[J/OL]. Appl. Numer. Math., 2006, 56(2): 193-209. http://www.sciencedirect.com/science/article/pii/S0168927405000668. DOI: http://dx.doi.org/10.1016/j.apnum.2005.03.003.

[123] LV C, XU C. Improved error estimates of a finite difference/spectral method for time-fractional diffusion equations[J]. Int. J. Numer. Anal. Mod, 2015, 12(2): 384-400.

[124] LIN R, LIU F. Fractional high order methods for the nonlinear fractional ordinary differential equation[J/OL]. Nonlinear Anal., 2007, 66(4): 856-869. http://www.sciencedirect.com/science/article/pii/S0362546X05010503. DOI: http://dx.doi.org/10.1016/j.na.2005.12.027.

[125] XU D. Alternating direction implicit Galerkin finite element method for the two-dimensional time fractional evolution equation[J]. Numer. Math.: Theor., Meth., Appl., 2014, 7(01): 41-57.

[126] GAO G, SUN H. Three-point combined compact alternating direction implicit difference schemes for two-dimensional time-fractional advection-diffusion equations[J]. Commun. Comput. Phy., 2015, 17(02): 487-509.

[127] WANG C, LIU J. Positivity property of second-order flux-splitting schemes for the compressible Euler equations[J]. Discrete Contin. Dyn. Syst. Ser. B, 2003, 3(2): 201-228.

[128] SHU C. A numerical method for systems of conservation laws of mixed type admitting hyperbolic flux splitting[J]. J. Comput. Phys., 1992, 100(2): 424-429.

[129] ZAYERNOURI M, KARNIADAKIS G. Fractional Sturm-Liouville eigen-problems: Theory and numerical approximation[J]. J. Comput. Phys., 2013, 252(C): 495-517.

[130] TSAI R, CHENG L, OSHER S, et al. Fast Sweeping Algorithms for a Class of Hamilton-Jacobi Equations[J]. SIAM J. Numer. Anal., 2003, 41(2): 673-694.

# 致 谢

在美丽安静的交大校园里，我度过了生命中无比珍贵的几载年华，完成了论文和学业，增添了知识和能力，更收获了宝贵的人生财富！回想起一路走过的既艰辛又快乐的历程，几载的成长和进步若仅凭一己之力是不可想象的，感激之情不禁油然而生。

首先要感谢的是我的博士导师金石教授。论文的顺利完成离不开金老师的悉心指导。从论文的选题、开题、写作到最后定稿的各个环节，无不倾注了导师的大量心血。几年来，金老师把我从一个稚嫩的本科生，引入计算数学的大门，逐步引领到科研的最前沿。多次资助我到国外顶尖学府交流访问、参加相关领域顶级的学术会议，使我得到了远多于同龄人的机会。在这过程中，大大地开拓了我的科研视野，获得了和相关领域顶级大师交流的机会，为我今后的科研铺平了道路。从导师那里，我不仅得到科学研究的系统训练，而且还亲身领略到了导师的大家风范，特别是导师高尚的人格、渊博的学识、严谨的治学态度、敏锐的洞察力、不懈探索的精神、诲人不倦的师德，以及谦和的为人处事方式，都令我受益良多，成为我一生受之不尽的宝贵财富。同时我要感谢美国杜克大学的刘建国教授，在金老师的引荐下，我有机会能够与刘老师合作，在刘老师的指导下完成博士论文中的部分内容。在这个过程中，刘老师也教会了我许多东西，他对数学的敏锐洞察力，对于数学的忘我的精神，使我认识到了自身的巨大差距，也使我有了努力的方向。可以说，能遇到金老师和刘老师并成为他们的学生是我一生中最大的幸运。

其次我要感谢交大致远学院和自然科学院的创建者蔡申瓯教授、鄂维南教授以及我的导师金石教授。他们作为国际上相关领域的顶尖学者，心系祖国，非常关心国内科研人才的培养。他们不辞辛苦，经常往返于中美之间，为了能给学生上课、指导学生科研以及学院的相关事务安排。他们还吸引了一系列国内外的知名学者加盟，像季向东教授、王立河教授、何小刚教授等等，本科期间他们均亲力亲为给我们上课，使我在人生中最好的年华能够享受到最好的教育。与此同时，交大数学系和物理系的老师们也对我关怀备至，像王亚光教授、黄建国教授、章璞教授等等，他们讲课风趣幽默、通俗易懂而细致入微，使我们打下了坚实的数学基础。同时我也要感谢在我读博士期间，自然科学研究院和数学系的一批优秀的青年教授，像唐敏特别研究员、应文俊特别研究员、胡丹教授、周栋焯教授、徐振礼教授、张镭特别研究员、李敬来特别研究员、张小群特别研究员、魏星特别研究员等等，他们或者组织我们参加会议和讨论班，或者给我们上相关的课程，平时与我们交流讨论。这些青年教授活力四射，亲切可近而又学识渊博，我从中受益良多，他们就是我最好的榜样、努力的目标。

我还要感谢我的几位师兄师姐：周珍楠、柴利慧、李沁等等，他们有的带领我进行科研，同时以自己的亲身经验引导我，有的耐心的回答我的各种学术和生活上的问题。我也要感谢我博士的同学李新春、黄健超、刘沛、许志钦、张耀宇、肖彦洋、蒋诗晓、杜涛等等，有这样一批优秀的同学我感到非常骄傲，与他们讨论总能使我学到新的东西。同时还有本科的一些同学，虽然我们没在一起读博士，可能方向也各不相同，但是仍然总与我讨论科学上的问题，像郜骋、李赫、李志超谈安迪、马璟琛、赵清宇、赵浩然等等。与他们讨论，我获得了更广阔的视野，了解了各行各业的动态和最前沿。

最后，谨以此文献给我挚爱的双亲，他们在背后的默默支持一直是我前进的动力。从小时候的习惯的培养、兴趣的培养、特长的培养，他们是我最好的启蒙老师。在此祝愿年过六旬的他们身体健康，心情愉快!

# 攻读学位期间发表的学术论文

[1] Jin S, Liu J G, Ma Z. Uniform spectral convergence of the stochastic Galerkin method for the linear transport equations with random inputs in diffusive regime and a micro-macro decomposition based asymptotic preserving method[J]. Research in the Mathematical Sciences, to appear.

[2] Jin S, Ma Z. The discrete stochastic Galerkin method for hyperbolic equations with non-smooth and random coefficients[J]. Journal of Scientific Computing, 2017: 1-25.

[3] Liu J G, Ma Z, Zhou Z. Explicit and implicit TVD schemes for conservation laws with Caputo derivatives[J]. Journal of Scientific Computing, 2016: 1-23.

[4] Ma Z, Zhang Y, Zhou Z. An improved semi-Lagrangian time splitting spectral method for the semi-classical Schrödinger equation with vector potentials using NUFFT[J]. Applied Numerical Mathematics, 2017, 111: 144-159.