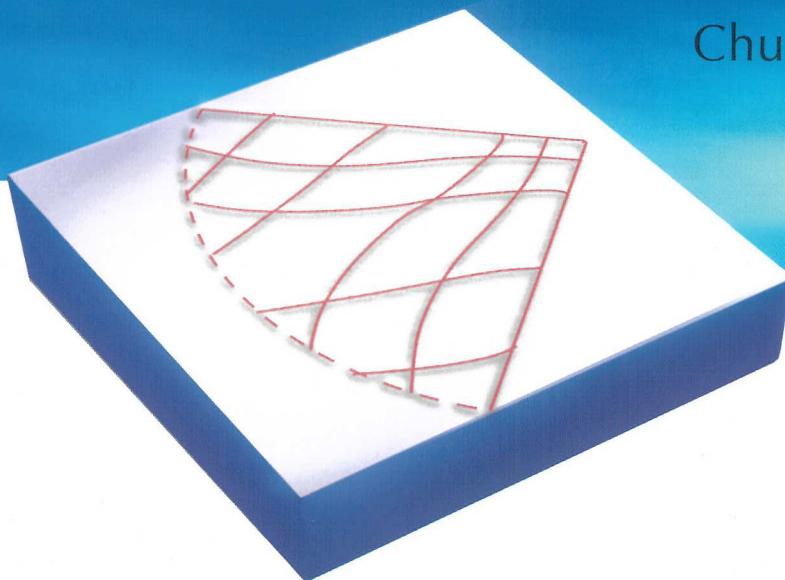


Series in Contemporary Applied Mathematics
CAM 13

Nonlinear Conservation Laws, Fluid Systems and Related Topics

Gui-Qiang Chen
Ta-Tsien Li
Chun Liu

editors



Nonlinear Conservation Laws, Fluid Systems and Related Topics

Series in Contemporary Applied Mathematics CAM

Honorary Editor: Chao-Hao Gu (*Fudan University*)

Editors: P. G. Ciarlet (*City University of Hong Kong*),
Ta-Tsien Li (*Fudan University*)

1. Mathematical Finance —— Theory and Practice
(Eds. Yong Jiongmin, Rama Cont)
2. New Advances in Computational Fluid Dynamics
—— Theory, Methods and Applications
(Eds. F. Dubois, Wu Huamo)
3. Actuarial Science —— Theory and Practice
(Eds. Hanji Shang, Alain Tosseti)
4. Mathematical Problems in Environmental Science and Engineering
(Eds. Alexandre Ern, Liu Weiping)
5. Ginzburg-Landau Vortices
(Eds. Haïm Brezis, Ta-Tsien Li)
6. Frontiers and Prospects of Contemporary Applied Mathematics
(Eds. Ta-Tsien Li, Pingwen Zhang)
7. Mathematical Methods for Surface and Subsurface Hydrosystems
(Eds. Deguan Wang, Christian Duquennoi, Alexandre Ern)
8. Some Topics in Industrial and Applied Mathematics
(Eds. Rolf Jeltsch, Ta-Tsien Li, Ian H. Sloan)
9. Differential Geometry: Theory and Applications
(Eds. Philippe G. Ciarlet, Ta-Tsien Li)
10. Industrial and Applied Mathematics in China
(Eds. Ta-Tsien Li, Pingwen Zhang)
11. Modeling and Dynamics of Infectious Diseases
(Eds. Zhien Ma, Yicang Zhou, Jianhong Wu)
12. Multi-scale Phenomena in Complex Fluids: Modeling, Analysis
and Numerical Simulations
(Eds. Tomas Y. Hou, Chun Liu, Jianguo Liu)
13. Nonlinear Conservation Laws, Fluid Systems and Related Topics
(Eds. Gui-Qiang Chen, Ta-Tsien Li, Chun Liu)

Series in Contemporary Applied Mathematics CAM 13

Nonlinear Conservation Laws, Fluid Systems and Related Topics

editors

Gui-Qiang Chen

Northwestern University, USA

Ta-Tsien Li

Fudan University, China

Chun Liu

Pennsylvania State University, USA



Higher Education Press



World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Gui-Qiang Chen
Department of Mathematics
Northwestern University
Evanston, IL 60208-2730
USA

Ta-Tsien Li
School of Mathematical Sciences
Fudan University
Shanghai, 200433
China

Chun Liu
Department of Mathematics
The Penn State University
University Park, PA 16802
USA

Editorial Assistant: Zhou Chun-Lian

图书在版编目 (CIP) 数据

非线性守恒律、流体力学方程组及相关主题 = Nonlinear Conservation Laws, Fluid Systems and Related Topics: 英文 / 陈贵强, 李大潜, 柳春主编. — 北京: 高等教育出版社, 2009.3
(现代应用数学丛书)

ISBN 978-7-04-024944-6

I. 非… II. ①陈… ②李… ③柳… III. ①非线性-能量守恒定律-研究-英文 ②流体力学-方程组-研究-英文 IV. 0411 035

中国版本图书馆 CIP 数据核字 (2009) 第 021412 号

Copyright © 2009 by

Higher Education Press

4 Dewai Dajie, Beijing 100120, P. R. China, and

World Scientific Publishing Co Pte Ltd

5 Toh Tuch Link, Singapore 596224

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the Publisher.

ISBN 978-7-04-024944-6

Printed in P. R. China

Preface

This book is a collection of lecture notes mainly from the short courses given in the 2007 Shanghai Summer School on Nonlinear Conservation Laws, Fluid Systems and Related Topics at Fudan University, July 5–August 4, 2007. There were more than 130 participants, including graduate students, postdoctors and junior faculty members from more than 30 universities in China and USA.

This summer school provided an occasion for a series of courses (25–26 hours each) by four distinguished contributors of this volume, Denis Serre (ENS-Lyon, France), Xiaoming Wang (Florida State University, USA), Tong Yang (CUHK, Hong Kong), and Yuxi Zheng (Penn State, USA), and a series of invited lectures by distinguished speakers including Jerry Bona (UIC, USA), Hongqiu Chen (The University of Memphis, USA), Emmanuele DiBenedetto (Vanderbilt University, USA), Willi Jäger (University of Heidelberg, Germany), Fanghua Lin (NYU, USA), Tai-Ping Liu (Stanford University, USA), Yuejun Peng (Université Blaise Pascal, France), Weiwei Wang (Shanghai Jiao Tong University, PRC), and Ping Zhang (Chinese Academy of Sciences, PRC), besides the editors of this volume.

This volume comprises five chapters, ranging from the mathematical theory and numerical approximation of both incompressible and compressible fluid flows, kinetic theory and conservation laws, to statistical theories for fluid systems, with expectation to lead the readers from the basics to the frontiers of the current research in these areas.

Chapter 1 is an introduction to the theory of incompressible inviscid flows with emphasis on classical results and recent developments. Chapter 2 is an introduction to one-dimensional hyperbolic systems of conservation laws with emphasis on theory, numerical approximation, and discrete shock profiles. Chapter 3 is an introduction to the kinetic theory, conservation laws and their intrinsic connections. Chapter 4 is an introduction to elementary statistical theories with applications to various fluid systems. Chapter 5 is an introduction to the Euler equations for compressible fluids in two space dimensions with emphasis on the self-similar isentropic irrotational case. These topics are naturally interrelated and represent a cross-section of the most significant recent

advances and current trends in nonlinear conservation laws, fluid systems and related topics.

The editors would like to express their sincere thanks to all the authors in this volume for their contributions and to all the participants in the Summer School. Zhiqiang Wang and Chunlian Zhou deserve our special thanks for their prompt and effective assistance to make the Summer School run smoothly. The editors are grateful to Fudan University, the Mathematical Center of Ministry of Education of China, the National Natural Science Foundation of China (NSFC) and the Institut Sino-Francais de Mathématiques Appliquées (ISFMA) for their help and support. Finally, the editors wish to thank Tianfu Zhao (Senior Editor, Higher Education Press) for his patience and professional assistance.

Gui-Qiang Chen, Ta-Tsien Li, Chun Liu

March 2008

Contents

Preface

<i>Thomas Y. Hou, Xinwei Yu:</i> Introduction to the Theory of Incompressible Inviscid Flows	1
<i>Denis Serre:</i> Systems of Conservation Laws. Theory, Numerical Approximation and Discrete Shock Profiles.....	72
<i>Seiji Ukai, Tong Yang:</i> Kinetic Theory and Conservation Laws: An Introduction	126
<i>Xiaoming Wang:</i> Elementary Statistical Theories with Applications to Fluid Systems	230
<i>Yuxi Zheng:</i> The Compressible Euler System in Two Space Dimensions	301

This page intentionally left blank

Introduction to the Theory of Incompressible Inviscid Flows*

Thomas Y. Hou

*Applied and Computational Mathematics, Caltech,
Pasadena, USA*

E-mail: hou@acm.caltech.edu

Xinwei Yu

*Department of Mathematics, UCLA,
Los Angeles, USA*

E-mail: xinweiyu@math.ucla.edu

Abstract

In this chapter, we consider the 3D incompressible Euler equations. We present classical and recent results on the issue of global existence/finite time singularity. We also introduce the theories of lower dimensional model equations of the 3D Euler equations and the vortex patch problem.

1 Introduction

The goal of these lecture notes is to introduce to the readers classical results as well as recent developments in the theory of 3D incompressible Euler equations. We will focus on the global existence/finite time singularity issue. We will start with the basic properties of the incompressible fluid flows, and then discuss the local and global well-posedness of the incompressible Euler equations. Of particular interest is the global existence or possible finite time blow-up of the 3D incompressible Euler equation. This is one of the most outstanding open problems in the past century. Here, we carefully examine the nature of the nonlinear vortex stretching term for the 3D Euler equation as well as several model problems for the 3D Euler equation. We put extra effort in taking into account the local geometrical properties and possible depletion of nonlinearity. By going through the nonlinear analysis of various fluid models,

*The first author is partly supported by an NSF grant DMS-0713670 and an FRG grant DMS-0353838. The second author is partly supported by an NSF grant DMS-0354488.

we can gain valuable insights into the fluid dynamic problems being studied. Through the analysis, we can also learn how various functional analysis and PDE techniques are being used for realistic applications, and what are their strengths and limitations. We especially emphasize the interplay between the physical and geometric properties of the fluid flows and modern nonlinear PDE techniques. By going through these analyses systematically, we can have a good understanding of the state of the art of nonlinear PDE methods and their applications to fluid dynamics problems.

This chapter is organized as follows:

1. Introduction
2. Derivation and Exact Solutions
3. Local Well-posedness of the 3D Euler Equation
4. The BKM Blow-up Criterion
5. Recent Global Existence Results
6. Lower Dimensional Models for the 3D Euler Equation
7. Vortex Patch

2 Derivation and exact solutions

2.1 Derivation of the Euler equations

The equation that governs the evolution of inviscid and incompressible flow is the Euler equation. Here we first derive the 3D Euler equation briefly. For more detailed derivations, the readers should consult other textbooks in fluid mechanics, such as Chorin-Marsden [12], Lamb [31], Marchioro-Pulvirenti [36], or Lopes Filho-Nussenzveig Lopes-Zheng [33].

We consider a domain Ω which is filled with a fluid, such as water. In classical continuum mechanics, the fluid can be seen as consisting of infinitesimal particles. At each time t , each particle has a one-to-one correspondence to the coordinates $x = (x_1, x_2, x_3) \in \Omega$. The fluid can be described by its density ρ , velocity $\mathbf{u} = (u_1, u_2, u_3)$ and pressure p at each such point $x \in \Omega$. Under the above assumptions, we can denote the position of any particle at time t by $X(\alpha, t)$ which starts at the position $\alpha \in \Omega$ at $t = 0$. Its evolution is governed by the following differential equation:

$$\begin{aligned} \frac{dX(\alpha, t)}{dt} &= \mathbf{u}(X(\alpha, t), t), \\ X(\alpha, 0) &= \alpha. \end{aligned} \tag{2.1}$$

To study the dynamics of the fluid, we must establish relations between ρ , \mathbf{u} and p . We do this by considering two basic mechanical rules: the conservation of mass, and the conservation of momentum.

The *conservation of mass* claims that, for any fixed region $W \subseteq \Omega$ which does not change with time,

$$\frac{d}{dt} \int_W \rho(x, t) dx = - \int_{\partial W} \rho(x, t) \mathbf{u}(x, t) \cdot \mathbf{n}(x, t) d\sigma \quad (2.2)$$

for all time t , where $\mathbf{n}(x, t)$ is the outer unit normal vector to ∂W , and $d\sigma$ is the area unit on ∂W . Using the Gauss theorem we arrive at

$$\frac{d}{dt} \int_W \rho(x, t) dx = - \int_W \nabla \cdot (\rho(x, t) \mathbf{u}(x, t)) dx$$

which implies

$$\int_W (\rho_t + \nabla \cdot (\rho \mathbf{u})) dx = 0.$$

If we assume the continuity of the integrand $\rho_t + \nabla \cdot (\rho \mathbf{u})$, by the arbitrariness of W , we get

$$\rho_t + \nabla \cdot (\rho \mathbf{u}) = 0. \quad (2.3)$$

Since otherwise, there would be a point x_0 such that the integrand is not 0. Without loss of generality, we assume $(\rho_t + \nabla \cdot (\rho \mathbf{u}))(x_0) > 0$. Then by continuity, there is $r > 0$ such that $\rho_t + \nabla \cdot (\rho \mathbf{u}) > 0$ for any $x \in B(x_0, r)$. This leads to a contradiction by taking $W = B(x_0, r)$. Equation (2.3) is called the *continuity equation*.

Let J be the determinant of the Jacobian matrix, $\frac{\partial X}{\partial \alpha}$. It can be proved by direct calculations (the reader should try to prove this as an exercise, see also Chorin-Marsden [12]) that

$$\frac{dJ}{dt} = (\nabla \cdot \mathbf{u}) J, \quad J(0) = 1.$$

We assume that the flow is incompressible. Incompressibility implies that the flow is volume preserving. Using the above equation one can show that the velocity is divergence-free, i.e.

$$\nabla \cdot \mathbf{u} = 0. \quad (2.4)$$

In this case, we have the determinant of the Jacobian matrix, J , to be identically equal to one, i.e. $J \equiv 1$. If the initial density is constant, i.e. $\rho(x, 0) \equiv \rho_0$, equation (2.3) implies that density is constant globally, i.e.

$$\rho(x, t) \equiv \rho_0.$$

Remark 2.1.

1. The above derivation of the mass conservation equation is under the assumption that ρ , \mathbf{u} and ∂W are all smooth enough, e.g., C^1 .
2. One can also derive (2.3) in a Lagrangian way, i.e., by considering an evolving region Ω_t that is a collection of particles. See e.g. Lopes Filho-Nussenzveig Lopes-Zheng [33].
3. Yet another way is through the variational formulation. See e.g. Marchioro-Pulvirenti [36].

The *conservation of momentum* means

$$\frac{d}{dt} \int_{\Omega_t} \rho \mathbf{u} \, dx = \mathbf{F}(\Omega_t), \quad (2.5)$$

where $\mathbf{F}(\Omega_t)$ is the force acting on Ω_t . Here $\Omega_t \equiv \cup_{\alpha \in \Omega_0} X(\alpha, t)$ for some $\Omega_0 \subseteq \Omega$ is a collection of particles that is carried by the flow. We first assume that the interaction in the fluid is local, i.e., all the forces between points inside Ω_t cancel each other by Newton's third law. This assumption implies

$$\mathbf{F}(\Omega_t) = \int_{\partial\Omega_t} \mathbf{f} \, d\sigma$$

for some \mathbf{f} . Our second assumption is that the fluid is ideal, which means that $\mathbf{f} = -p\mathbf{n}$, where \mathbf{n} is the unit outer normal to $\partial\Omega_t$. Now the momentum relation becomes

$$\frac{d}{dt} \int_{\Omega_t} \rho \mathbf{u} \, dx = \int_{\partial\Omega_t} -p\mathbf{n} \, d\sigma = - \int_{\Omega_t} \nabla p \, dx,$$

where the second equality follows from the Gauss theorem

$$\int_{\Omega} \partial_i f \, dx = \int_{\partial\Omega} f n_i \, d\sigma.$$

To derive a pointwise equation similar to (2.3), we need to put the $\frac{d}{dt}$ inside the integration in the term

$$\frac{d}{dt} \int_{\Omega_t} \rho \mathbf{u} \, dx.$$

Note that since $\Omega_t = X(\Omega_0, t)$ depends on t , it is not the same as

$$\int_{\Omega_t} (\rho \mathbf{u})_t \, dx.$$

Instead of naïvely putting the differentiation inside, we proceed as follows. We first change variables from the Eulerian variable x to the Lagrangian variable α . Since the flow is incompressible, the determinant of the Jacobian matrix is equal to one, i.e., $\det(X_\alpha) = 1$. Thus we have

$$\begin{aligned} \frac{d}{dt} \int_{\Omega_t} \rho \mathbf{u} \, dx &= \frac{d}{dt} \int_{\Omega_0} \rho(X(\alpha, t), t) \mathbf{u}(X(\alpha, t), t) \, d\alpha \\ &= \int_{\Omega_0} \frac{d}{dt} \rho(X, t) \mathbf{u}(X, t) + \rho(X, t) \frac{d}{dt} \mathbf{u}(X, t) \, d\alpha \\ &= \int_{\Omega_0} (\rho_t + \mathbf{u} \cdot \nabla \rho) \mathbf{u} + \rho (\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u}) \, d\alpha \\ &= \int_{\Omega_0} \rho (\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u}) \, d\alpha \\ &= \int_{\Omega_t} \rho (\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u}) \, dx, \end{aligned}$$

where the first equality follows from the fact that the flow map $\alpha \mapsto X(\alpha, t)$ is one-to-one and has Jacobian 1, and the fourth equality follows from (2.3) and the incompressibility condition. Now we have

$$\int_{\Omega_t} \rho (\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u}) \, dx = - \int_{\Omega_t} \nabla p \, dx.$$

Finally, by the arbitrariness of Ω_t , we get

$$\rho (\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u}) = -\nabla p. \quad (2.6)$$

by an argument that is similar to the one leading to (2.3). (2.6) is the *balance of momentum*.

If we further assume that the flow has constant initial density, then we have $\rho(x, t) \equiv \rho_0$, and equation (2.6) is equivalent to:

$$\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla p,$$

where p is the “rescaled” pressure p/ρ_0 .

Under these assumptions, we obtain the 3D Euler equation as follows:

$$\begin{aligned} \mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} &= -\nabla p, \\ \nabla \cdot \mathbf{u} &= 0. \end{aligned} \quad (2.7)$$

In the remaining part of this lecture note, we will focus on (2.7).

2.2 The Vorticity-Stream function formulation

2.2.1 Vorticity

We consider the Taylor expansion of the velocity $\mathbf{u}(x, t)$ at some point x .

$$\begin{aligned}\mathbf{u}(x + h, t) &= \mathbf{u}(x, t) + \nabla \mathbf{u} \cdot h + O(h^2) \\ &= \mathbf{u}(x, t) + \frac{\nabla \mathbf{u} + \nabla \mathbf{u}^t}{2} h + \frac{\nabla \mathbf{u} - \nabla \mathbf{u}^t}{2} h + O(h^2) \\ &\equiv \mathbf{u}(x, t) + S(x, t)h + \Omega(x, t)h + O(h^2),\end{aligned}$$

where S is symmetric and Ω is anti-symmetric. In 3D, it is easy to see that there is a vector ω such that

$$\Omega(x, t)h = \frac{1}{2}\omega(x, t) \times h.$$

This implies that locally, the flow is rotating around an axis $\xi(x, t) \equiv \frac{\omega(x, t)}{|\omega(x, t)|}$. The vector field $\omega(x, t)$ is called “vorticity”. And it is easy to check that

$$\omega(x, t) = \nabla \times \mathbf{u}(x, t).$$

2.2.2 Vorticity-Stream function formulation

By taking $\nabla \times$ on both sides of the 3D Euler equation (2.7), we have

$$\omega_t + \mathbf{u} \cdot \nabla \omega = \omega \cdot \nabla \mathbf{u} = S \cdot \omega. \quad (2.8)$$

which is the vorticity formulation. The last equality follows from the fact that

$$\Omega \cdot \omega = \frac{1}{2}\omega \times \omega \equiv 0,$$

since by definition we have

$$\frac{1}{2}\omega \times h \equiv \Omega \cdot h$$

for any vector h . Now there are two unknowns ω and \mathbf{u} , so we have to find the relation between them to close the system. This relation is the so-called Biot-Savart law:

$$\mathbf{u}(x) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{x - y}{|x - y|^3} \times \omega(y) dy. \quad (2.9)$$

Note that we need $u(x)$ to vanish at ∞ for the above formula to hold. To derive the Biot-Savart law, first define a vector valued function Ψ , called “stream function”, such that

$$-\Delta \Psi = \omega.$$

Now it is easy to check that

$$\mathbf{u} = \nabla \times \Psi$$

satisfies

$$\nabla \times \mathbf{u} = \omega.$$

(Hint: Use the identity

$$-\nabla \times (\nabla \times) + \nabla(\nabla \cdot) = \Delta,$$

and then try to show

$$\|\nabla(\nabla \cdot \Psi)\|_{L^2}^2 = 0$$

using the same identity. Details are left as exercises. Or see Bertozzi-Majda [35]).

Now the Biot-Savart law (2.9) follows from the formula

$$\Psi = \frac{1}{4\pi} \int \frac{1}{|x-y|} \omega(y) dy,$$

where $\frac{1}{4\pi} \frac{1}{|x|}$ is the fundamental solution for the Poisson equation

$$-\Delta u = f$$

in 3D.

Besides (2.8), another important form of the vorticity evolution is the “stretching formula”.

$$\omega(X(\alpha, t), t) = \nabla_\alpha X(\alpha, t) \omega_0(\alpha), \quad (2.10)$$

where $\omega_0(\alpha) = \omega(X(\alpha, 0), 0) = \omega(\alpha, 0)$, and X is defined by (2.1). To prove it, just differentiate both sides with respect to time, which yields

$$\begin{aligned} \omega_t + \mathbf{u} \cdot \nabla \omega &= \nabla_\alpha \mathbf{u}(X(\alpha, t), t) \omega_0(\alpha) \\ &= \nabla \mathbf{u} \cdot (\nabla_\alpha X \cdot \omega_0) \\ &= \nabla \mathbf{u} \cdot \omega(x, t), \end{aligned}$$

which is just (2.8). One catch: this “proof” actually uses the uniqueness of the solution to the system (2.8), (2.9).

For the convenience of future references, we will denote the differentiation in time along the Lagrangian trajectory as $\frac{D}{Dt}$, which has the property:

$$\frac{D}{Dt} w = w_t + \mathbf{u} \cdot \nabla w.$$

$\frac{D}{Dt}$ is also called material derivative.

2.2.3 2D Euler equations

In some physical cases, such as the flow passing around a cylinder with infinite length, we can assume that $u_3 \equiv 0$ and \mathbf{u}, p depend on x_1, x_2 only. In this case, the Euler equations (2.7) remains the same form, but the vorticity-stream function form reduces to

$$\omega_t + \mathbf{u} \cdot \nabla \omega = 0 \quad (2.11)$$

and

$$\mathbf{u}(x) = \frac{1}{2\pi} \int \frac{(x-y)^\perp}{|x-y|^2} \omega(y) dy, \quad (2.12)$$

where ω is a short-hand for ω_3 .

One important difference between 2D and 3D Euler equations is that, the right hand side is 0 in (2.11), which means the vorticity is conserved along Lagrangian trajectory pathes. This point can be illustrated more clearly by looking at the “stretching formula” in 2D, which is

$$\omega(X(\alpha, t), t) = \omega_0(\alpha). \quad (2.13)$$

This difference plays an important role in the theory of 2D Euler equations, which is far more complete than its 3D counterpart.

2.3 Conserved quantities

2.3.1 Local conserved quantities

First we consider those quantities that are carried by a collection of flow particles.

Let C_0 be a closed curve in \mathbb{R}^3 . We define

$$C_t = \cup_{\alpha \in C_0} X(\alpha, t)$$

and the circulation

$$\Gamma_{C_t} \equiv \oint_{C_t} \mathbf{u} \cdot ds.$$

Theorem 2.2 (Kelvin’s Circulation Theorem). $\Gamma_{C_t} \equiv \Gamma_{C_0}$.

Proof. We first prove the following.

$$\frac{d}{dt} \int_{C_t} \mathbf{u} \cdot ds = \int_{C_t} \frac{D\mathbf{u}}{Dt} \cdot ds.$$

To prove it, let $\alpha(\beta)$ be a parametrization of the loop C_0 , with $0 \leq \beta \leq 1$. Then C_t is parametrized as $X(\alpha(\beta), t)$. Thus

$$\begin{aligned} \frac{d}{dt} \int_{C_t} \mathbf{u} \cdot ds &= \frac{d}{dt} \int_0^1 \mathbf{u}(X(\alpha(\beta), t), t) \cdot \frac{\partial}{\partial \beta} X(\alpha(\beta), t) d\beta \\ &= \int_0^1 \frac{D\mathbf{u}}{Dt}(X(\alpha(\beta), t), t) \cdot \frac{\partial}{\partial \beta} X(\alpha(\beta), t) d\beta \\ &\quad + \int_0^1 \mathbf{u}(X(\alpha(\beta), t), t) \cdot \frac{\partial}{\partial \beta} \mathbf{u}(X(\alpha(\beta), t), t) d\beta, \end{aligned}$$

where we have used the relation

$$\frac{\partial X}{\partial t}(\alpha, t) = \mathbf{u}(X(\alpha, t), t).$$

Note that the first term is just

$$\int_{C_t} \frac{D\mathbf{u}}{Dt} \cdot ds,$$

we just need to show that the second term is 0. This is easy, since we have

$$\int_0^1 \mathbf{u} \cdot \frac{\partial}{\partial \beta} \mathbf{u} ds = \frac{1}{2} \int_0^1 \frac{\partial}{\partial \beta} (\mathbf{u} \cdot \mathbf{u}) ds = 0,$$

which follows from the fact that C_t is a close loop.

Now we prove the circulation theorem. We have

$$\frac{d}{dt} \int_{C_t} \mathbf{u} \cdot ds = \int_{C_t} \frac{D\mathbf{u}}{Dt} \cdot ds = - \int_{C_t} \nabla p \cdot ds = - \int_{C_t} p_s ds = 0$$

since C_t is closed. This ends the proof. \square

Next let C_0 be a general curve and $C_t = X(C_0, t)$. Then as long as the flow is still regular, C_t is still a curve in \mathbb{R}^3 . C_t is called a vortex line if the following is satisfied

$$C_0 \text{ is tangent to } \omega_0(\alpha) \text{ at any } \alpha \in C_0. \quad (2.14)$$

One can verify that as long as (2.14) is satisfied, the same tangency condition is satisfied at every moment t , i.e.,

$$C_t \text{ is tangent to } \omega(x, t) \text{ at any } x \in C_t.$$

A collection of vortex lines is called a “vortex tube”. One readily sees that vorticity is always tangent to the side surface of a vortex tube.

The above properties make vortex tube/line very important objects in the theories/numerical simulations/physical experiments of the 3D Euler equation, as we will reveal later in this lecture note.

2.3.2 Global conserved quantities

The most well-known global conserved quantities are the following (we will indicate the dimension and region/manifold, \mathbb{T}^d stands for d -dimensional periodic torus):

1. The integral of velocity (\mathbb{R}^d and \mathbb{T}^d , $d = 2, 3$).

$$\frac{d}{dt} \int \mathbf{u} \, dx = 0.$$

2. Kinetic energy (\mathbb{R}^d , \mathbb{T}^d , smooth bounded domain, $d = 2, 3$).

$$\frac{d}{dt} \int |\mathbf{u}|^2 \, dx = 0.$$

Remark 2.3. In the \mathbb{R}^d case, caution must be taken. We actually need that the kinetic energy $\int |\mathbf{u}|^2 \, dx$ to be finite. In 3D this requirement is reasonable, while in 2D it is not.

3. Center of vorticity (\mathbb{R}^2 , if $\mathbf{u}\omega$ decays fast enough at ∞).

$$\bar{x} = \int_{\mathbb{R}^2} x\omega \, dx = \text{const.}$$

4. Moment of inertia (\mathbb{R}^2 , if $\mathbf{u}\omega$ decays fast enough at ∞).

$$I = \int_{\mathbb{R}^2} |x|^2 \omega \, dx = \text{const.}$$

5. Functions of vorticity ($d = 2$).

$$\int_{\Omega_t} f(\omega) \, dx = \int_{\Omega_0} f(\omega_0) \, d\alpha$$

for any measurable f and material domain Ω_t . In particular, we see that the L^p norm of ω is conserved for $1 \leq p \leq \infty$.

6. Other quantities.

$$\begin{aligned} & \int_{\mathbb{R}^3} x \times \omega \, dx, \\ & \int_{\mathbb{R}^3} x \times (x \times \omega) \, dx; \end{aligned}$$

helicity

$$\int_{\mathbb{R}^3} \mathbf{u} \cdot \omega \, dx;$$

and spirality

$$\omega \cdot \gamma,$$

where $\gamma = \mathbf{u} + \nabla\phi$ with ϕ solving

$$\frac{D}{Dt}\phi = -|\mathbf{u}|^2/2 + p.$$

This quantity is conserved along particle trajectories.

2.4 Special flows

2.4.1 Axisymmetric flow

In this subsection we introduce the axisymmetric flow, i.e., when written in cylindrical coordinates $x_1 = r \cos \theta$, $x_2 = r \sin \theta$ and $x_3 = z$, the velocity \mathbf{u} and the pressure p depend only on r and z . Unlike the 2D Euler equations, this particular flow retains some 3D characters and is often referred to as the $2\frac{1}{2}$ -D equations.

We introduce the cylindrical frame of reference:

$$\begin{aligned} e_r &= (\cos \theta, \sin \theta, 0), \\ e_\theta &= (-\sin \theta, \cos \theta, 0), \\ e_z &= (0, 0, 1), \end{aligned}$$

and can easily rewrite the 3D Euler equations in the new frame, with $\mathbf{u} = \mathbf{u}(r, z)$ and $p = p(r, z)$, as

$$\mathbf{u}_t + (\mathbf{u} \cdot \tilde{\nabla})\mathbf{u} + B = -\tilde{\nabla}p, \quad (2.15)$$

where

$$\tilde{\nabla} = (\partial_r, 0, \partial_z)$$

and

$$B = \frac{u^\theta}{r}(-u^\theta, u^r, 0).$$

We leave the details (which can be found in e.g. Lopes Filho-Nussenzveig Lopes-Zheng [33]) for this system to the reader as exercises.

1. Derive equations (2.15).
2. Prove that, in the moving frame (e_r, e_θ, e_z) , we have

$$\begin{aligned} \omega &= \omega^r e_r + \omega^\theta e_\theta + \omega^z e_z \\ &\equiv (-\partial_z u^\theta) e_r + (\partial_z u^r - \partial_r u^z) e_\theta + \left(\partial_r u^\theta + \frac{u^\theta}{r} \right) e_z. \end{aligned}$$

3. When $u^\theta \equiv 0$, (2.15) becomes axisymmetric flows without swirl.
Prove that the equations are

$$\begin{aligned} \left(\partial_t + \mathbf{u} \cdot \tilde{\nabla} \right) \mathbf{u} &= -\tilde{\nabla}p, \\ \tilde{\nabla} \cdot (r\mathbf{u}) &= 0. \end{aligned} \quad (2.16)$$

Furthermore, one can reduce the equation into the $r-z$ plane which is 2D. Prove that the equation for ω^θ (note that $\omega^r = \omega^z = 0$) is

$$(\partial_t + \mathbf{u} \cdot \nabla) \left(\frac{\omega^\theta}{r} \right) = 0.$$

2.4.2 Radially (circularly) symmetric flow

In the 2D case, we consider $\omega_0 \equiv \omega_0(r)$ which is circularly symmetric. Then by exploring the invariance of the Laplacian we easily see that ψ defined by

$$-\Delta\psi = \omega_0$$

is also a circularly symmetric function. Thus

$$\mathbf{u} = \nabla^\perp \psi$$

is always tangent to the contours $\omega_0 \equiv \text{const}$. One can easily verify that

$$\omega \equiv \omega_0, \quad \mathbf{u} \equiv \mathbf{u}_0$$

is a steady solution for the 2D Euler equations. The velocity is explicitly given as

$$\mathbf{u} = \frac{x^\perp}{r^2} \int_0^r s\omega(s) \, ds, \quad (2.17)$$

where $r = |x|$. These stationary solutions are called Rankine vortices. The reader can try to derive the “radial_symmetric_biot_savart law” (2.17) as an exercise (Hint: it is easier to start from the stream function Ψ).

Now consider the special case, where ω_0 is supported in $B_R \equiv \{x \mid |x| \leq R\}$, with $\int_{B_R} \omega_0 = 0$. Then it is easy to see that \mathbf{u} is also supported in B_R . Such a vortex is called a confined eddy. The importance of this observation can be seen from the following property:

The superposition of two disjoint confined eddies is still a solution.

This gives us a way to construct very complicated exact solutions to the 2D Euler equations.

2.4.3 Jets and strains

Let $D(t)$ be any family of symmetric and trace-free matrices that smoothly depends on t , and let ω solves

$$\begin{aligned} \frac{d\omega}{dt} &= D(t)\omega, \\ \omega(0) &= \omega_0. \end{aligned}$$

We introduce

$$\mathbf{u} = \frac{1}{2}\omega \times \mathbf{x} + D(t)\mathbf{x}.$$

It is easy to check that we can define p such that \mathbf{u} solves the 3D Euler equations in the whole space. One thing that worths noting is that, the velocity we defined above is growing unboundedly at ∞ and is thus non-physical.

It is illustrating to study some special cases.

1. Jet. Take $\omega_0 = 0$ thus $\omega \equiv 0$. Note that we can write $D(t)$ to be diagonal:

$$D(t) = \begin{bmatrix} -\gamma_1 & 0 & 0 \\ 0 & -\gamma_2 & 0 \\ 0 & 0 & \gamma_1 + \gamma_2 \end{bmatrix}$$

and get

$$\mathbf{u} = (-\gamma_1 x_1, -\gamma_2 x_2, (\gamma_1 + \gamma_2)x_3).$$

2. Swirling jet. We take $\omega_0 = (0, 0, a)$ and get

$$\omega = (0, 0, ae^{(\gamma_1 + \gamma_2)t}),$$

and

$$\mathbf{u} = \left(-\gamma_1 x_1 - \frac{1}{2}a(t)x_2, -\gamma_2 x_2 + \frac{1}{2}a(t)x_1, (\gamma_1 + \gamma_2)x_3 \right).$$

3. Strain. We take $\omega_0 = 0$ and $\gamma_1 = -\gamma_2 = \gamma$,

$$\mathbf{u} = (-\gamma x_1, \gamma x_2, 0).$$

3 Local well-posedness of the 3D Euler equation

First we consider the local well-posedness for classical solutions. By classical solutions we mean solutions such that (2.7) holds in the classical sense, i.e., all the derivatives are in the classical sense, the multiplications are pointwise, and the equalities hold everywhere. Our main goal in this section is to prove the following:

Theorem 3.1. *If the initial velocity $\mathbf{u}_0 \in H^m \cap C^2$ for some $m > 2+d/2$, then there is $T > 0$ such that there is a unique solution $\mathbf{u} \in H^m \cap C^2$ in $[0, T]$.*

To do this, we use the standard technique of mollifiers. In short, we approximate (2.7) by a sequence of equations that can be shown to admit global smooth solutions, and then establish the local in time existence by taking limit.

3.1 Analytical preparations

3.1.1 Sobolev spaces

The Sobolev spaces $H^k, k \in \mathbb{Z}, k \geq 0$ is defined as

$$H^k(\mathbb{R}^d) = \left\{ f(x) \mid \sum_{|\alpha| \leq k} \|\partial^\alpha f\|_{L^2}^2 < \infty \right\},$$

where α is a multi-index $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$. $|\alpha| \equiv \sum \alpha_i$ and $\partial^\alpha \equiv \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$. H^k is a Banach space with norm

$$\|f\|_{H^k} = \left(\sum_{|\alpha| \leq k} \|\partial^\alpha f\|_{L^2}^2 \right)^{1/2}.$$

If we consider the Fourier transform of f , we have

$$\|f\|_{H^k} = \left(\sum_{|\alpha| \leq k} \|\xi^\alpha \hat{f}\|_{L^2}^2 \right)^{1/2},$$

where $\xi^\alpha \equiv \xi_1^{\alpha_1}, \dots, \xi_d^{\alpha_d}$. Now by some simple algebra we can obtain the following equivalent norm

$$\|f\|_{H^k} \sim \left\| \langle \xi \rangle^k \hat{f} \right\|_{L^2} \sim \left\| (1 - \Delta)^{k/2} f \right\|_{L^2},$$

where $\langle \xi \rangle \equiv (1 + |\xi|^2)^{1/2}$, and Δ is the Laplacian.

The point in writing the H^k norm this way is that, now we can take k to be any real number instead of non-negative integers. Usually, when k is not an integer, we replace it by s .

The following theorem is used extensively in PDE researches.

Theorem 3.2. *The space $C_0^\infty(\mathbb{R}^d)$ is dense in $H^s(\mathbb{R}^d)$.*

The most important property of the Sobolev spaces is the embedding theorems. We will not prove these theorems here, interested readers can look up the proof in e.g. Adams [1], which is a classic and not very hard to read.

Before introducing the theorems, we first recall what ‘‘embedding’’ means. Consider two Banach spaces X and Y , with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. Assume that there is a third space Z which is dense in both X and Y . We say X is embedded in Y , if there is a constant C such that

$$\|\cdot\|_Y \leq C \|\cdot\|_X.$$

This means that all the elements in X is also in Y . Furthermore, we say X is compactly embedded in Y , if X is embedded in Y , and any bounded subset of X (in the X norm) is precompact in Y (with respect to the Y norm). That is, if $\{x_n\} \subset X$ is uniformly bounded, then there is a subsequence which is Cauchy in Y . We denote embedding by \hookrightarrow .

Theorem 3.3 (Embeddings for H^s). *Let $H^s(\mathbb{R}^d)$ be the Sobolev space. We have*

$$H^{s+k} \hookrightarrow C^k$$

for all $s > d/2$ and $k \in \mathbb{Z}$, nonnegative.

3.1.2 Hodge decomposition and the Leray projection

We denote by $H^s(\mathbb{R}^d)$ the Sobolev spaces, and let $V^s \subset H^s(\mathbb{R}^d; \mathbb{R}^d)$ be the subspace of divergence-free vector fields.

Lemma 3.4 (Hodge decomposition). *Let \mathbf{u} be a vector field with components in $L^2(\mathbb{R}^d)$. There exists a unique decomposition $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, where \mathbf{u}_1 is divergence-free and \mathbf{u}_2 is a gradient. Furthermore \mathbf{u}_1 and \mathbf{u}_2 are orthogonal in L^2 . We denote by P the projection $L^2(\mathbb{R}^d; \mathbb{R}^d) \mapsto V^0$ which maps \mathbf{u} to \mathbf{u}_1 , then P commutes with derivatives, convolution and is also a map from H^s to V^s .*

Proof. First we solve

$$\Delta\phi = \nabla \cdot \mathbf{u}.$$

Thus

$$\phi = \Delta^{-1}(\nabla \cdot \mathbf{u}) + H,$$

where H is a harmonic function and Δ^{-1} is the convolution with the Green's function of the Laplacian in \mathbb{R}^d . Now define

$$\mathbf{u}_2 = \nabla\phi = (\nabla^2\Delta^{-1}) \cdot \mathbf{u} + \nabla H.$$

By going to the Fourier space, it is easy to see that the first term is in L^2 . To make $\mathbf{u}_2 \in L^2$, we must have $\nabla H \in L^2$, which means it must vanish at ∞ . But since each entry of ∇H is harmonic, we see that this implies that $\nabla H \equiv 0$.

Now we have

$$P = (I - \nabla^2\Delta^{-1}) \cdot . \quad (3.1)$$

It is easy to check the commutativity properties. \square

This operator P is often referred to as the *Leray projection operator*.

3.1.3 The Aubin-Lions lemma

For evolution PDEs, generally one can not treat time and space as equal, so one need compactness results that has different requirement in space and time. A standard result is the Aubin-Lions lemma.

First we prove a technical lemma. Let $X \hookrightarrow Y \hookrightarrow Z$ be Banach spaces that have embedding relations as indicated. Recall that $X \hookrightarrow Y$ is compact means that for any $\{f_n\}$ that is uniformly bounded in X , there is a subsequence that is convergent in the norm of Y .

Lemma 3.5. *Assume that $X \hookrightarrow Y$ is compact, then for every $\eta > 0$ there exists a constant $C_\eta > 0$ such that*

$$\|v\|_Y \leq \eta \|v\|_X + C_\eta \|v\|_Z$$

for every $v \in X$.

Proof. The proof is standard. We prove by contradiction. Assume there is a $\eta > 0$ and a sequence $\{v^n\} \subset X$ such that

$$\|v^n\|_Y > \eta \|v^n\|_X + n \|v^n\|_Z,$$

then by taking $w^n \equiv v^n / \|v^n\|_X$ we see that the same inequality holds for w^n . Now w^n is bounded in X , which means there is a subsequence, still denote as w^n , such that

$$w^n \rightarrow w \in Y$$

in Y . Note that $\|w^n\|_Y \leq C \|w^n\|_X \leq C$ by the embedding assumption and the fact that $\|w^n\|_X = 1$. Now divide both sides of the equation for w^n by n , we have

$$w^n \rightarrow 0 \text{ in } Z.$$

But on the other hand, we have

$$w^n \rightarrow w \neq 0$$

in Y and thus we have a contradiction, since the embedding, convergence in Y to some limit implies convergence in Z to the same limit. \square

Lemma 3.6 (Aubin-Lions). *Suppose that $X \hookrightarrow Y$ is compact. Let $T > 0$. Let $\{u^n\}$ be a bounded sequence in $L^\infty([0, T]; X)$. Suppose this sequence is equicontinuous as Z -valued functions defined on $[0, T]$. Then the same sequence is precompact in $C([0, T]; Y)$.*

Proof. First, it follows directly from Lemma 3.5 that each u^n is in $C([0, T]; Y)$. Second, by the conditions in the Lemma we see that we can use the Arzela-Ascoli lemma on $C([0, T]; Z)$ and see that u^n is precompact in it. Finally, still by Lemma 3.5 we see that u^n is precompact in $C([0, T], Y)$. \square

Remark 3.7. A comparison with the Arzela-Ascoli lemma in analysis is helpful. There we basically have a sequence that is uniformly bounded and equicontinuous in $C([0, T], Y)$ for some Y . Here the boundedness condition, which is usually easier to establish, is strengthened, while the harder condition equicontinuity is weakened.

3.1.4 Calculus inequalities

Let u and v be in $H^m(\mathbb{R}^d)$ with $m \in \mathbb{N}$.

Lemma 3.8.

1. If u and v are bounded and continuous then there exists a constant $C > 0$ such that

$$\|uv\|_{H^m} \leq C (\|u\|_{L^\infty} \|D^m v\|_{L^2} + \|v\|_{L^\infty} \|D^m u\|_{L^2}).$$

2. If u, v and ∇u are bounded and continuous then there exists a constant $C > 0$ such that

$$\begin{aligned} \sum_{0 \leq |\alpha| \leq m} \|D^\alpha (uv) - u D^\alpha v\|_{L^2} &\leq C (\|\nabla u\|_{L^\infty} \|D^{m-1} v\|_{L^2} \\ &\quad + \|v\|_{L^\infty} \|D^m u\|_{L^2}). \end{aligned}$$

Proof. First we prove 1. It is enough to prove that

$$\|D^\alpha u D^\beta v\|_{L^2} \leq C (\|u\|_{L^\infty} \|D^m v\|_{L^2} + \|v\|_{L^\infty} \|D^m u\|_{L^2}),$$

where in the RHS (right hand side) we actually define

$$\|D^m v\|_{L^2}^2 = \sum_{|\alpha|=m} \|D^\alpha v\|_{L^2}^2,$$

while in the LHS (left hand side) α, β are multi-indices with $|\alpha| + |\beta| = m$.

We illustrate the idea of the proof by considering the scalar case. We estimate

$$\|u'v'\|_{L^2} = \left(\int (u'v')^2 dx \right)^{1/2},$$

where $\alpha = \beta = 1$ and $m = 2$. By Hölder's inequality, we have

$$\|u'v'\|_{L^2} \leq \|u'\|_{L^4} \|v'\|_{L^4}.$$

Next we establish the Gagliardo-Nirenberg inequality

$$\|D^i u\|_{L^{2r/i}} \leq c_r \|u\|_{L^\infty}^{1-i/r} \|D^r u\|_0^{i/r}$$

with $0 \leq i \leq r$. In our case, $i = 1, r = 2$, the Gagliardo-Nirenberg inequality reduces to

$$\|u'\|_{L^4} \leq c \|u\|_{L^\infty}^{1/2} \|u''\|_0^{1/2}. \quad (3.2)$$

The proof is easy. We have

$$\begin{aligned} \|u'\|_{L^4}^4 &= \int (u')^4 dx \\ &= \int (u')^3 du \\ &\leq c \left| \int u (u')^2 u'' dx \right| \\ &\leq c \left| \int u^2 (u'')^2 dx \right|^{1/2} \left| \int (u')^4 dx \right|^{1/2} \\ &\leq c \|u\|_{L^\infty} \|u''\|_{L^2} \|u'\|_{L^4}^2 \end{aligned}$$

which proves (3.2).

Now we have

$$\|u'v'\|_{L^2} \leq c \|u\|_{L^\infty}^{1/2} \|u''\|_{L^2}^{1/2} \|v\|_{L^\infty}^{1/2} \|v''\|_{L^2}^{1/2}.$$

By using Young's inequality

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

where $p, q > 0$ with $\frac{1}{p} + \frac{1}{q} = 1$ we finish the proof.

The general cases of 1 and 2 are left as exercises. \square

3.1.5 Gronwall's inequality

In dealing with evolution equations, we need to estimate various quantities. In doing so we often end up with inequalities like

$$X(t) \leq a(t) + \int_0^t b(s)X(s) ds,$$

where $X(t)$ is the non-negative quantity we need to estimate, and $a(t)$, $b(t) \geq 0$ with $a(t)$ differentiable. The trick in getting an estimate for X is the following. We also assume that everything is continuous.

Fix $\varepsilon > 0$, let $Y^\varepsilon(t)$ satisfy

$$Y^\varepsilon(t) = a(t) + \varepsilon + \int_0^t b(s)Y^\varepsilon(s) ds,$$

then it is easy to see that $Y^\varepsilon(t)$ is differentiable, and satisfies

$$(Y^\varepsilon)'(t) = a'(t) + b(t)Y^\varepsilon(t), \\ Y^\varepsilon(0) = a(0) + \varepsilon,$$

which gives

$$Y^\varepsilon(t) = (a(0) + \varepsilon) e^{\int_0^t b(s) ds} + \int_0^t a'(s) e^{\int_s^t b(\tau) d\tau} ds.$$

Now by arbitrariness of ε we get what we need, as long as we have

$$X(t) \leq Y^\varepsilon(t)$$

for any $\varepsilon > 0$. To show this, consider $W \equiv Y^\varepsilon - X$, which satisfies

$$W(t) \geq \varepsilon + \int_0^t b(s)W(s) ds, \\ W(0) = \varepsilon > 0.$$

By the continuity of W and the condition $b(s) \geq 0$ it is easy to see that $W(t) \geq \varepsilon$ for all $t > 0$. Thus we proved the following lemma.

Lemma 3.9 (Grönwall's lemma). *If $X(t), a(t), b(t) \geq 0$ are continuous, $a(t)$ differentiable, with*

$$X(t) \leq a(t) + \int_0^t b(s)X(s) ds,$$

then we can estimate $X(t)$ by

$$X(t) \leq a(0)e^{\int_0^t b(s) ds} + \int_0^t a'(s)e^{\int_s^t b(\tau) d\tau} ds.$$

3.2 Properties of mollifiers

Definition 3.10. Let $\rho \in C_0^\infty(\mathbb{R}^d)$ be any radial function, i.e., $\rho(x)$ depends only on $|x|$. We choose $\rho \geq 0$ with $\int_{\mathbb{R}^d} \rho dx = 1$. For any $\varepsilon > 0$, define

$$\rho_\varepsilon(x) = \varepsilon^{-d} \rho(x/\varepsilon).$$

Then we call the family $\{\rho_\varepsilon\}$ a family of mollifiers.

In the following, we will denote

$$M^\varepsilon f = (\rho_\varepsilon * f)(x)$$

for any function f .

Next we develop some main properties of the mollification operator M^ε .

Lemma 3.11. *For any function f such that $M^\varepsilon f$ is well-defined, we have*

1. $M^\varepsilon f$ is smooth, i.e., C^∞ .
 2. For all $f \in C^0(\mathbb{R}^d)$, we have $M^\varepsilon f \rightarrow f$ uniformly on any compact set Ω , and
- $$\|M^\varepsilon f\|_{L^\infty} \leq \|f\|_{L^\infty}.$$
3. $M^\varepsilon D^\alpha = D^\alpha M^\varepsilon$ for any multi-index α .
 4. For all $f \in L^p$, $g \in L^q$ with $1/p + 1/q = 1$,

$$\int_{\mathbb{R}^d} (M^\varepsilon f) g \, dx = \int_{\mathbb{R}^d} f (M^\varepsilon g) \, dx.$$

5. For all $f \in H^s(\mathbb{R}^d)$, $M^\varepsilon f$ converges to f in H^s and the rate of convergence in the H^{s-1} norm is $O(\varepsilon)$.
6. For all $f \in H^s(\mathbb{R}^d)$, $k \in \mathbb{Z}^+ \cup \{0\}$, and $\varepsilon > 0$, we have

$$\begin{aligned} \|M^\varepsilon f\|_{s+k} &\leq \frac{c_{sk}}{\varepsilon^k} \|f\|_s, \\ \|M^\varepsilon D^k f\|_{L^\infty} &\leq \frac{c_k}{\varepsilon^{d/2+k}} \|f\|_{L^2}. \end{aligned}$$

Proof. 1–4 are easy and omitted. Interested readers can try to prove them or check Bertozzi-Majda [35]. To prove 5 and 6, it is important to know the representation of M^ε in the Fourier space:

$$\widehat{M^\varepsilon f}(\xi) = \hat{\rho}(\varepsilon \xi) \hat{f}(\xi).$$

Note that by construction

$$\hat{\rho}(0) = \int \rho \, dx = 1.$$

As $\varepsilon \rightarrow 0$, for any ξ , we have

$$\hat{\rho}(\varepsilon \xi) \sim 1 + O(\varepsilon).$$

It is clear now that why we can expect $M^\varepsilon f \rightarrow f$ at all.

Another key factor in proving 5 and 6 is the Fourier side characterization of $H^s(\mathbb{R}^d)$. Recall that

$$|\widehat{\nabla f}(\xi)| = c |\xi| |\hat{f}(\xi)|,$$

where c depends on the definition of Fourier transforms, e.g., if we define

$$\hat{f}(\xi) = \int e^{-i\xi \cdot x} f(x) dx,$$

then $c = 1$. The particular value of c is not important here. In the following, we will just take $c = 1$. Now $f \in H^s$ is equivalent to

$$\langle \xi \rangle^s \hat{f}(\xi) \in L^2,$$

where $\langle \xi \rangle \equiv (1 + |\xi|^2)^{1/2}$.

With the above understanding, 5 and 6 are easy to prove. For example, we prove the second estimate in 6. For any multi-index α with $|\alpha| = k$, we have

$$\begin{aligned} |(M^\varepsilon D^\alpha f)(x)| &= c \left| \int e^{i\xi \cdot x} \hat{\rho}(\varepsilon \xi) \xi^\alpha \hat{f}(\xi) d\xi \right| \\ &\leq c \int_{\mathbb{R}^d} |\hat{\rho}(\varepsilon \xi)| |\xi|^k |\hat{f}(\xi)| d\xi \\ &\lesssim \|f\|_{L^2} \left(\int_{\mathbb{R}^d} |\hat{\rho}(\varepsilon \xi)|^2 |\xi|^{2k} d\xi \right)^{1/2} \\ &= \|f\|_{L^2} \left(\int_{\mathbb{R}^d} |\hat{\rho}(\eta)|^2 |\eta|^{2k} d\eta \right)^{1/2} \varepsilon^{-k-d/2} \\ &\lesssim \varepsilon^{-k-d/2} \|f\|_{L^2}, \end{aligned}$$

where $\eta \equiv \varepsilon \xi$ and note that the integration is over \mathbb{R}^d , thus the factor $\varepsilon^{-d/2}$. The integral on $\hat{\rho}$ is bounded since $\rho \in C_0^\infty \subset H^k$ is a fixed function.

The other inequalities in 5 and 6 can be proved similarly and are left to the readers. \square

3.3 Global existence of the mollified equation

We consider the mollified equations:

$$\begin{aligned} \partial_t u^\varepsilon + M^\varepsilon (((M^\varepsilon u^\varepsilon) \cdot \nabla) (M^\varepsilon u^\varepsilon)) &= -\nabla p^\varepsilon, \\ \nabla \cdot u^\varepsilon &= 0, \\ u^\varepsilon(x, 0) &= u_0(x), \end{aligned} \tag{3.3}$$

or, by using the Leray projection operator,

$$\begin{aligned} \partial_t u^\varepsilon + P(M^\varepsilon (((M^\varepsilon u^\varepsilon) \cdot \nabla) (M^\varepsilon u^\varepsilon))) &= 0, \\ P u^\varepsilon &= u^\varepsilon, \\ u^\varepsilon(x, 0) &= u_0(x), \end{aligned} \tag{3.4}$$

where u^ε denotes the solution and is not necessarily of the form $M^\varepsilon v$ for some v . We will prove the global existence (i.e., for all time $t \in \mathbb{R}^+$) of the mollified 3D Euler equations. Our strategy is to prove local existence by treating (3.4) as an ODE in some Banach space, and then extend the existence time to ∞ . In the following of this section, we will omit the superscript ε and denote u^ε by u .

Lemma 3.12. *Let $m \in \mathbb{N}$. Then for every $u_0 \in V^m$ and $\varepsilon > 0$ there exists $T^\varepsilon > 0$ and a solution $u^\varepsilon \in C^1([0, T^\varepsilon]; V^m)$ to the problem (3.4), or equivalently, (3.3).*

Proof. Let

$$F_\varepsilon(u) = -P(M^\varepsilon(((M^\varepsilon u) \cdot \nabla)(M^\varepsilon u))).$$

Then (3.4) becomes

$$\frac{du^\varepsilon}{dt} = F_\varepsilon(u^\varepsilon),$$

which is an ODE in a Banach space. The only thing we need to check before applying the Picard iteration to get local in time existence is that

1. $F_\varepsilon : V^m \mapsto V^m$, and
2. F_ε is locally Lipschitz in V^m .

For the first claim, we have the following estimate:

$$\begin{aligned} \|F_\varepsilon(u)\|_{H^m} &\leq \|M^\varepsilon(((M^\varepsilon u) \cdot \nabla)(M^\varepsilon u))\|_{H^m} \\ &\leq C \|M^\varepsilon(\nabla \cdot (M^\varepsilon u \otimes M^\varepsilon u))\|_{H^m} \\ &\leq \frac{C}{\varepsilon} \|M^\varepsilon u \otimes M^\varepsilon u\|_{H^m} \\ &\leq \frac{C}{\varepsilon^{3/2}} \|u\|_{H^m}^2, \end{aligned}$$

where we have used the calculus inequalities (see Lemma 2.1.8) and the following properties of the mollifiers: $\|M^\varepsilon Df\|_{H^m} \leq C \|f\|_{H^m} / \varepsilon$, $\|M^\varepsilon u\|_{L^\infty} \leq C \|u\|_{H^m} / \varepsilon^{d/2}$, which follows from Lemma 2.1.11 (6).

Next we show that F_ε is Lipschitz. Let v_1 and v_2 belong to V^m , then

$$\begin{aligned} \|F_\varepsilon(v_1) - F_\varepsilon(v_2)\|_{H^m} &\leq \frac{C}{\varepsilon} (\|M^\varepsilon v_1 \otimes M^\varepsilon(v_1 - v_2)\|_{H^m} \\ &\quad + \|M^\varepsilon v_2 \otimes M^\varepsilon(v_1 - v_2)\|_{H^m}) \end{aligned}$$

by adding and subtracting $M^\varepsilon(((M^\varepsilon v_1) \cdot \nabla)(M^\varepsilon v_2))$. By using the calculus inequality again (Lemma 2.1.8), we can bound the RHS by

$$\frac{C}{\varepsilon^{3/2}} (\|v_1\|_{H^m} + \|v_2\|_{H^m}) \|v_1 - v_2\|_{H^m} \leq C_\varepsilon \|v_1 - v_2\|_{H^m}$$

since $\|v_i\|_{H^m}$ ($i=1,2$) is bounded and ε is finite. This proves the local Lipschitz condition of F_ε . \square

To extend the existence time to infinity we need to show that the Lipschitz constant

$$\frac{C}{\varepsilon^{3/2}} (\|v_1\|_{H^m} + \|v_2\|_{H^m})$$

depends only on ε and initial conditions. We only need to show that for any solution u , $\|u\|_{H^m}$ is bounded by the H^m norm of the initial value u_0 .

First, by integration by parts, it is easy to see that

$$\|u\|_{L^2} \leq \|u_0\|_{L^2}.$$

The remaining is done by the following lemma:

Lemma 3.13. *Let $m \in \mathbb{N}$ and $u \in C^1([0, T); V^m)$ be a solution of the mollified 3D Euler equations (3.4). Then*

$$\|u\|_{H^m} \leq \|u_0\|_{H^m} e^{C \int_0^t \|\nabla M^\varepsilon u\|_{L^\infty} dt}.$$

Proof. Let α be a multi-index, with $|\alpha| \leq m$. Applying D^α to both sides of (3.3), multiplying them by $D^\alpha u$ and integrating, we have

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \int |D^\alpha u|^2 dx &= - \int D^\alpha u \cdot (M^\varepsilon D^\alpha (((M^\varepsilon u) \cdot \nabla) (M^\varepsilon u))) \\ &= - \int D^\alpha M^\varepsilon u \cdot D^\alpha (((M^\varepsilon u) \cdot \nabla) (M^\varepsilon u)) dx \\ &= - \int D^\alpha M^\varepsilon u \cdot D^\alpha (((M^\varepsilon u) \cdot \nabla) (M^\varepsilon u)) \\ &\quad + \int D^\alpha M^\varepsilon u \cdot (((M^\varepsilon u) \cdot \nabla) D^\alpha M^\varepsilon u) dx, \end{aligned}$$

where the term involving the pressure vanishes after integrated by parts due to the incompressibility condition, and the last equality comes from the following argument:

$$\begin{aligned} \int D^\alpha M^\varepsilon u \cdot (((M^\varepsilon u) \cdot \nabla) D^\alpha M^\varepsilon u) dx &= \frac{1}{2} \int (M^\varepsilon u) \cdot \nabla (|D^\alpha M^\varepsilon u|^2) dx \\ &= 0 \end{aligned}$$

via integration by parts due to the incompressibility condition.

Now we sum over all $0 \leq |\alpha| \leq m$. Using the calculus inequality, we have

$$\begin{aligned} &\frac{d}{dt} \|u\|_{H^m}^2 \\ &\leq C \|u\|_{H^m} \sum_{|\alpha| \leq m} \|D^\alpha (((M^\varepsilon u) \cdot \nabla) (M^\varepsilon u)) - ((M^\varepsilon u) \cdot \nabla) D^\alpha M^\varepsilon u\|_{L^2} \\ &\leq C \|u\|_{H^m} (\|\nabla M^\varepsilon u\|_{L^\infty} \|D^{m-1} D M^\varepsilon u\|_{L^2} + \|D^m M^\varepsilon u\|_{L^2} \|\nabla M^\varepsilon u\|_{L^\infty}) \\ &\leq C \|\nabla M^\varepsilon u\|_{L^\infty} \|u\|_{H^m}^2. \end{aligned}$$

To finish the proof, we just need to apply the standard Gronwall's inequality from Lemma 3.9. \square

3.4 Local existence of the Euler equations

Now we are ready to give the local existence theorem.

Theorem 3.14. *Let $u_0 \in V^m$ for $m \geq 4$. There exists $T_0 = T_0(\|u_0\|_{H^m}) > 0$ such that for any $T < T_0$, there exists a unique solution $u \in C^1([0, T]; V^m)$ of the 3D incompressible Euler equations with u_0 as initial data.*

Proof. By Lemma 3.13 we have

$$\frac{d}{dt} \|u^\varepsilon\|_{H^m}^2 \leq C \|\nabla M^\varepsilon u^\varepsilon\|_{L^\infty} \|u^\varepsilon\|_{H^m}^2.$$

Note that $m \geq 4 > 3/2 + 1$, by Theorem 3.3, H^m is embedded into C^1 , which means $\|\nabla M^\varepsilon u^\varepsilon\|_{L^\infty} \leq \|M^\varepsilon u^\varepsilon\|_{C^1} \lesssim \|M^\varepsilon u^\varepsilon\|_{H^m} \leq \|u^\varepsilon\|_{H^m}$. Thus we have

$$\frac{d}{dt} \|u^\varepsilon\|_{H^m} \leq C \|u^\varepsilon\|_{H^m}^2$$

and the constant C here is independent of ε . Therefore we see that our u^ε is uniformly bounded in $L^\infty([0, T]; H^m)$ by

$$\frac{\|u_0\|_{H^m}}{1 - CT \|u_0\|_{H^m}}$$

for any $T < T_0 \equiv (C \|u_0\|_{H^m})^{-1}$. To apply the Lions-Aubin lemma we need to show that u^ε is Lipschitz in t in some larger space, which we take to be H^{m-1} . In fact we have

$$\begin{aligned} \|\partial_t u\|_{H^{m-1}} &= \|F_\varepsilon(u^\varepsilon)\|_{H^{m-1}} \\ &\leq C \|\nabla \cdot (M^\varepsilon u^\varepsilon \otimes M^\varepsilon u^\varepsilon)\|_{H^{m-1}} \\ &\leq C \|M^\varepsilon u^\varepsilon \otimes M^\varepsilon u^\varepsilon\|_{H^m} \\ &\leq C \|M^\varepsilon u^\varepsilon\|_{L^\infty} \|u^\varepsilon\|_{H^m} \\ &\leq C \|u^\varepsilon\|_{H^m}^2, \end{aligned}$$

where we have used the calculus inequality (Lemma 2.1.8) and the Sobolev embedding theorem. Thus we see that u^ε is Lipschitz in t wrt H^{m-1} -norm.

We fix $R_k > 0$ and use Lemma 3.6 (The reason we need this step is that we need $H^m \hookrightarrow H^{m-1}$ to be compact, which will not hold for unbounded regions, as can be seen by taking $X = H^1(\mathbb{R})$, $Y = L^2(\mathbb{R})$ and $f_n(x) = f(x - n)$ for some $f \in H^1$). Obviously $\{f_n\}$ is bounded

in X but not convergent in Y .) with $X = H^m(B(0, R_k))$, $Y = Z = H^{m-1}(B(0, R_k))$. Taking $R_k \rightarrow \infty$ and using a diagonal argument we see that u^ε has a subsequence, which we do not relabel, that is strongly convergent in $C([0, T]; H_{loc}^{m-1}(\mathbb{R}^3))$. Denote the limit by u . Moreover, since $m \geq 4 > 3/2 + 2$, we see that the convergence also holds in $C([0, T]; C_{loc}^1(\mathbb{R}^3))$.

We rewrite the equation as

$$u^\varepsilon = u_0 + \int_0^t F^\varepsilon(u^\varepsilon) ds.$$

It is easy to see that

$$u = u_0 + \int_0^t F(u) ds,$$

where $F(u) \equiv P(u \cdot \nabla u)$. Thus we have further that

$$u \in C^1([0, T]; C_{loc}^1(\mathbb{R}^3)),$$

which implies that we can legitimately differentiate with respect to t . Now taking d/dt on both side, we see that u satisfies

$$\begin{aligned} u_t + P(u \cdot \nabla u) &= 0, \\ \nabla \cdot u &= 0, \\ u(\cdot, 0) &= u_0, \\ |u| &\rightarrow 0 \text{ as } |x| \rightarrow \infty. \end{aligned}$$

The final step for existence is to recover the pressure. This follows directly from the Leray decomposition.

Now we show the uniqueness. Suppose that there are two solutions u_1 and u_2 , then we immediately have

$$(u_1 - u_2)_t + P(u_1 \cdot \nabla u_1 - u_2 \cdot \nabla u_2) = 0$$

with $u_1 - u_2 = 0$ at $t = 0$. Multiply to $u_1 - u_2$ and integrate, we can easily derive

$$\frac{d}{dt} \|u_1 - u_2\|_{L^2}^2 \leq C(\|u_1\|_{H^m} + \|u_2\|_{H^m}) \|u_1 - u_2\|_{L^2}^2$$

by the calculus inequalities. Then by using Gronwall's inequality, we see that the only solution is $u_1 - u_2 \equiv 0$. This ends the proof for uniqueness. \square

4 The BKM blow-up criterion

4.1 The Beale-Kato-Majda criterion

One of the important points that should be noted is that the above existence result is local in time, meaning that the solution may cease to be in H^m (also known as (aka) blow-up) in some finite time. Thus it is important to have some quantities to indicate such a blow-up. One of them is the quantity

$$\int_0^T \|\omega(\cdot, s)\|_{L^\infty} ds$$

proposed by T. Beale, T. Kato and A. Majda.

By the same method used in the last section, we can have the following bound:

$$\|u(\cdot, t)\|_{H^m} \leq C e^{c \int_0^t \|\nabla u\|_{L^\infty} ds} \|u_0\|_{H^m}.$$

So it is clear that as long as $\|\nabla u\|_{L^\infty}$ is uniformly bounded in some time interval $(0, T)$, then the solution exists upto T . In fact this is what Ebin, Fischer and Marsden proved in their 1972 paper [23]. Thus the key is to bound $\|\nabla u\|_{L^\infty}$ by $\|\omega\|_{L^\infty}$ at the same time t . Recall the 3D Biot-Savart law

$$u(x) = \int K(x - y)\omega(y) dy,$$

where $K(z)$ is the matrix kernel

$$K(z) = \frac{1}{|z|^3} \begin{pmatrix} 0 & -z_3 & z_2 \\ z_3 & 0 & -z_1 \\ -z_2 & z_1 & 0 \end{pmatrix}.$$

If we differentiate under the integration formally, we would have

$$\nabla u(x) = \int \nabla K(x - y)\omega(y) dy. \quad (4.1)$$

The operator $\nabla K*$ in fact has nice properties. To see this, we recall a theorem from Stein [44], which is also called the Calderon-Zygmund Lemma.

Theorem 4.1. *Let $K \in L^2(\mathbb{R}^d)$. We suppose:*

1. *The Fourier transform of K is essentially bounded*

$$|\hat{K}(x)| \leq B.$$

2. *K is C^1 outside the origin and*

$$|\nabla K(x)| \leq B/|x|^{d+1}.$$

For $f \in L^1 \cap L^p$, let us set

$$(Tf)(x) = \int_{\mathbb{R}^d} K(x-y)f(y) dy.$$

Then there exists a constant A_p , so that

$$\|T(f)\|_p \leq A_p \|f\|_p, \quad 1 < p < \infty.$$

One can thus extend T to all of L^p by continuity. The constant A_p depends only on p, B , and the dimension n . In particular, it does not depend on the L^2 norm of K .

The remark following the theorem in Stein [44] claims that the assumption $K \in L^2$ can be safely dropped in practice.

Now it is easy to check that our kernel ∇K satisfies the conditions in the theorem, thus the L^p norm of ∇u is thus bounded by the L^p -norm of ω . But here what we need is a L^∞ bound. The key lies in the following lemma. It will also become clear that the formal differentiation in (4.1) is “almost legitimate”.

Lemma 4.2. *Let u and ω be related with the Biot-Savart law, and $u \in H^3(\mathbb{R}^3)$, then*

$$\|\nabla u\|_{L^\infty} \leq C (1 + \ln^+ \|u\|_{H^3} + \ln^+ \|\omega\|_{L^2}) (1 + \|\omega\|_{L^\infty}). \quad (4.2)$$

Proof. By the Biot-Savart law, $u = K * \omega$, where K is a matrix-valued singular kernel, homogeneous of degree -2 , behaves like $O(|x|^{-2})$ at ∞ . Since $u \in H^3(\mathbb{R}^3)$, we have $\omega \in H^2(\mathbb{R}^3)$ and thus in $C^{0,\gamma}(\mathbb{R}^3)$ for some $0 < \gamma < 1$ by the Sobolev embedding theorems. Now we compute ∇u .

$$\begin{aligned} & \partial_{x_j} u(x) \\ &= \int_{\mathbb{R}^3} K(y) \partial_{x_j} \omega(x-y) dy \\ &= - \int_{\mathbb{R}^3} K(y) \partial_{y_j} \omega(x-y) dy \\ &= - \lim_{\delta \rightarrow 0} \int_{|y| \geq \delta} K(y) \partial_{y_j} \omega(x-y) dy \\ &= \lim_{\delta \rightarrow 0} \left(\int_{|y| \geq \delta} \partial_{y_j} K(y) \omega(x-y) dy - \int_{|y|=\delta} K(y) \omega(x-y) \frac{-y_j}{\delta} dy \right) \\ &= p v \int_{\mathbb{R}^3} \partial_{y_j} K(y) \omega(x-y) dy + \lim_{\delta \rightarrow 0} \int_{|z|=1} K(z) \omega(x-\delta z) z_j dz \\ &= p v \int_{\mathbb{R}^3} \partial_{y_j} K(y) \omega(x-y) dy + C_j \cdot \omega(x), \end{aligned}$$

where $C_j = \int_{|z|=1} K(z) z_j \, dz$ is a matrix. Here $\text{p.v.} \int f dx$ stands for principle value integral. Note that we can also write $C_j \cdot \omega$ as $c_j \times \omega$ for some c_j defined as $\int_{|z|=1} \frac{z}{|z|^3} z_j \, dz$. The above computation shows that, for our purpose, it is enough to estimate the formal ∇u as given in (4.1).

Now to estimate $\|\nabla u\|_{L^\infty}$ by ω , we only need to bound the principal value integral

$$\text{p.v.} \int \nabla K(y) \omega(x-y) \, dy.$$

Note that for any $a < b$, we have the important cancellation property

$$\int_{a \leq |y| \leq b} \nabla K(y) \, dy = 0.$$

Fix $x \in \mathbb{R}^3$ and $0 < \delta < \varepsilon \leq R < \infty$, we have

$$\begin{aligned} & \left| \int_{|y| \geq \delta} \nabla K(y) \omega(x-y) \, dy \right| \\ & \leq \left| \int_{\delta \leq |y| \leq \varepsilon} \nabla K(y) (\omega(x-y) - \omega(x)) \, dy \right| \\ & \quad + \left| \int_{\varepsilon \leq |y| \leq R} \nabla K(y) \omega(x-y) \, dy \right| + \left| \int_{|y| \geq R} \nabla K(y) \omega(x-y) \, dy \right| \\ & \leq C \|\omega\|_{C^{0,\gamma}} \int_{\delta \leq |y| \leq \varepsilon} |y|^{\gamma-3} \, dy + C \|\omega\|_{L^\infty} \int_{\varepsilon \leq |y| \leq R} |y|^{-3} \, dy \\ & \quad + C \|\omega\|_{L^2} \left(\int_{|y| \geq R} |y|^{-6} \, dy \right)^{1/2} \\ & \leq C \|u\|_{H^3} \varepsilon^\gamma + C \|\omega\|_{L^\infty} \ln(R/\varepsilon) + CR^{-3/2} \|\omega\|_{L^2}. \end{aligned}$$

Finally, taking $R^{3/2} = \|\omega\|_{L^2}$, and $\varepsilon = 1$ if $\|u\|_{H^3} \leq 1$ and $(\|u\|_{H^3})^{-1/\gamma}$ otherwise, we get the desired estimate. \square

The main result is almost straightforward now.

Theorem 4.3 (Beale, Kato, Majda 1984). *Let $u_0 \in V^m$ with $m \geq 4$. Let $u \in C^1([0, T); V^m)$ be a solution of the 3D incompressible Euler equations (2.7) with initial data u_0 . Let $\omega = \nabla \times u$ be the associated vorticity. Then T is the maximum time for u to be in the above function class if and only if*

$$\int_0^T \|\omega\|_{L^\infty} \, dt = \infty.$$

Proof. The “if” part is obvious. Since $\int_0^T \|\omega\|_{L^\infty} dt = \infty$, necessarily $\|\omega\|_{L^\infty} \rightarrow \infty$ at $t \rightarrow T$. Then $\|u\|_{W^{1,\infty}} \rightarrow \infty$ as $t \rightarrow T$ and u can not be in V^m for $m \geq 4$ by the embedding theorems.

Now we deal with the “only if” part. First, as we have shown at the beginning of this subsection,

$$\|u\|_{H^m} \leq C e^{C \int_0^T \|\nabla u\|_{L^\infty} dt} \|u_0\|_{H^m}.$$

Furthermore, by applying the same method to the vorticity equation, we can easily derive

$$\|\omega\|_{L^2} \leq \|\omega_0\| e^{C \int_0^t \|\nabla u\|_{L^\infty} ds}.$$

Substituting the above two inequalities into (4.2) in Lemma 4.2 gives

$$\|\nabla u\|_{L^\infty} \leq C \left(1 + (1 + \|\omega\|_{L^\infty}) \int_0^T \|\nabla u\|_{L^\infty} dt \right).$$

From this we have the estimate

$$\|\nabla u\|_{L^\infty} \leq \|\nabla u_0\|_{L^\infty} e^{C \int_0^T \|\omega\|_{L^\infty} dt}$$

by the Grönwall’s lemma 3.9. This ends the proof. \square

Remark 4.4. An immediate result of applying the Beale-Kato-Majda criterion is this. There is no finite-time blow-up in 2D Euler equations.

4.2 Improvements of the BKM criterion

During the more than 20 years following the BKM criterion, there are several improvements ([7, 8, 9, 42, 43], to name a few). In particular, in Chae [9], the condition of $\int_0^T \|\omega\|_\infty dt = \infty$ is sharpened to

$$\int_0^T \|\tilde{\omega}(t)\|_{\dot{B}_{\infty,1}^0}^2 dt = \infty,$$

where for any fixed orthonormal frame (e_1, e_2, e_3) ,

$$\tilde{\omega} = \omega^1 e_1 + \omega^2 e_2$$

is the projection of the vorticity in the plane of $e_1 - e_2$. The Besov space $\dot{B}_{\infty,1}^0$ is defined as f such that

$$\sum_{j \in \mathbb{Z}} \|\varphi_j * f\|_{L^\infty} < \infty,$$

where the Schwarz function $\varphi \in \mathcal{S}$ satisfying

1. $\text{Supp } \hat{\varphi} \subset \{\xi \in \mathbb{R}^d \mid \frac{1}{2} \leq |\xi| \leq 2\}$, (note this is why we can not take $\varphi \in C_0^\infty$).
2. $\hat{\varphi}(\xi) \geq C > 0$ if $\frac{2}{3} < |\xi| < \frac{3}{2}$.
3. $\sum_{j \in \mathbb{Z}} \hat{\varphi}_j(\xi) = 1$ where $\hat{\varphi}_j = \hat{\varphi}(2^{-j}\xi)$.

We present the main idea of the proof here. The key to the proof is to bound the growth of $\omega^3 \equiv \omega \cdot e_3$ by $\tilde{\omega} = \omega^1 e_1 + \omega^2 e_2$.

Recall that the evolution of ω satisfies

$$\omega_t + u \cdot \nabla \omega = S \cdot \omega,$$

where $S = \frac{1}{2} (\nabla u + \nabla u^t)$. Dot product with e_3 , we have

$$\frac{D(\omega^3)}{Dt} = \omega \cdot S \cdot e_3.$$

Now we estimate the right hand side. We have (since this estimate is independent of time, we omit t)

$$\begin{aligned} & \omega \cdot S \cdot e_3 \\ &= \frac{1}{4\pi} p v \int \frac{\omega(x) \times \omega(x+y)}{|y|^3} \cdot e_3 - 3 \frac{y \times \omega(x+y)}{|y|^5} \cdot e_3 (y \cdot \omega(x)) \, dy \\ &= \frac{1}{4\pi} p v \int \left\{ \frac{\tilde{\omega}(x) \times \tilde{\omega}(x+y)}{|y|^3} \cdot e_3 - 3 \frac{y \times \tilde{\omega}(x+y)}{|y|^5} \cdot e_3 y_3 \omega_3(x) \right. \\ &\quad \left. - 3 \frac{y \times \tilde{\omega}(x+y)}{|y|^5} \cdot e_3 (y \cdot \tilde{\omega}(x)) \right\} \, dy \\ &= \tilde{\omega} \cdot \mathcal{P}(\tilde{\omega}) \cdot e_3 + \omega^3 e_3 \cdot \mathcal{P}(\tilde{\omega}) \cdot e_3, \end{aligned}$$

where \mathcal{P} is the matrix valued singular integral operator defined by

$$\mathcal{P}(\omega) = S = \frac{1}{2} (\nabla u + \nabla u^t)$$

for ω and u related by the Biot-Savart law. This operator \mathcal{P} is known to be bounded on $\dot{B}_{\infty,1}^0$. This combined with the fact that $\dot{B}_{\infty,1}^0 \hookrightarrow L^\infty$ yields

$$\begin{aligned} \|\omega \cdot S \cdot e_3\|_{L^\infty} &\lesssim \|\omega^3\|_{L^\infty} \|\mathcal{P}(\tilde{\omega})\|_{L^\infty} + \|\tilde{\omega}\|_{L^\infty} \|\mathcal{P}(\tilde{\omega})\|_{L^\infty} \\ &\leq \|\omega^3\|_{L^\infty} \|\mathcal{P}(\tilde{\omega})\|_{\dot{B}_{\infty,1}^0} + \|\tilde{\omega}\|_{L^\infty} \|\mathcal{P}(\tilde{\omega})\|_{\dot{B}_{\infty,1}^0} \\ &\lesssim \|\omega^3\|_{L^\infty} \|\tilde{\omega}\|_{\dot{B}_{\infty,1}^0} + \|\tilde{\omega}\|_{\dot{B}_{\infty,1}^0}^2. \end{aligned}$$

Then it is easy to get

$$\|\omega^3\|_{L^\infty} \leq \left(\|\omega_0^3\|_{L^\infty} + \int_0^t \|\tilde{\omega}\|_{\dot{B}_{\infty,1}^0}^2 \, ds \right) \exp \left(C \int_0^t \|\tilde{\omega}(s)\|_{\dot{B}_{\infty,1}^0} \, ds \right)$$

by integrating the equation for ω^3 along one particle trajectory $X(\alpha, t)$, and then applying the Grönwall's lemma.

Finally, using the Cauchy-Schwarz inequality, and the embedding $\dot{B}_{\infty,1}^0 \hookrightarrow L^\infty$ again, we have

$$\begin{aligned} \int_0^T \|\omega\|_{L^\infty} dt &\leq \int_0^T \|\tilde{\omega}\|_{L^\infty} dt + \int_0^T \|\omega^3\|_{L^\infty} dt \\ &\leq \sqrt{T} A_T + [\|\omega_0^3\|_{L^\infty} + C A_T^2] T \exp(C\sqrt{T} A_T), \end{aligned}$$

where $A_T \equiv \left(\int_0^T \|\tilde{\omega}\|_{\dot{B}_{\infty,1}^2}^2 dt \right)^{1/2}$. This ends the proof for the necessity part. The sufficient part is trivial from the embedding $H^m \hookrightarrow \dot{B}_{\infty,1}^0$ for $m > 5/2$.

This result is sharper than the BKM criterion, but its disadvantage is that it is not as applicable to numerical simulations as the BKM one. For example, it is not always as easy to measure the Besov norm as the L^∞ norm accurately in numerical computations.

5 Recent global existence results

In this chapter we review some recent results which are in the same line with the BKM criterion. Due to the limited scope of this lecture note, we will not be able to cover all relevant results in this area, even for those results that are related to the Beale-Kato-Majda criterion.

5.1 Sufficient conditions by Constantin-Fefferman-Majda

In 1996, Constantin-Fefferman-Majda [14] proposed an non-blow-up condition based on the BKM criterion. To understand the main idea, we recall the BKM criterion: If $\int_0^T \|\omega(\cdot, t)\|_{L^\infty} dt < \infty$, then no blow-up can happen in $[0, T]$. This implies that one should investigate the vorticity magnitude $|\omega(x, t)|$.

The first step would naturally be deriving the evolution equation for this quantity. This equation is derived in Constantin [13]. It is

$$\frac{D}{Dt} |\omega| = \alpha(x, t) |\omega|, \quad (5.1)$$

where

$$\begin{aligned} \alpha(x, t) &\equiv \xi(x, t) \cdot \nabla u(x, t) \cdot \xi(x, t) \\ &= \xi(x, t) \cdot S(x, t) \cdot \xi(x, t), \end{aligned}$$

where $S(x, t)$ is the symmetric part of ∇u and $\xi(x, t) = \frac{\omega(x, t)}{|\omega(x, t)|}$ is the direction of $\omega(x, t)$.

Remark 5.1. Note that ξ is well defined only for those points where $\omega(x, t) \neq 0$. For those points where $\omega(x, t) = 0$, $\omega(x, t)$ will always be 0 as long as the flow is not singular, along the trajectory path of the same point, forward and backward in time. This can be seen from the formula

$$\omega(X(\alpha, t), t) = \nabla_\alpha X \cdot \omega(\alpha, 0)$$

and the fact that $\nabla_\alpha X$ is non-singular as long as the flow is not singular. So at those points where vorticity vanishes, one can reasonably define $\alpha(x, t) = 0$.

(5.1) can be derived by applying the inner product of the vorticity equation (2.8) with ξ , and using the fact that $\partial_{x_j} \xi \cdot \xi = 0$ since $\xi \cdot \xi = 1$. The proof is left as an exercise.

Next we recall that

$$\nabla u = pv \int_{\mathbb{R}^3} \nabla K(x - y) \omega(y) dy + C\omega(x),$$

where C is a third order tensor $C = [C_1, C_2, \dots, C_d]$ where

$$C_j = \int_{|z|=1} K(z) z_j dz$$

as defined in the proof to Lemma 4.2. Note that, since $C_j \omega = c_j \times \omega$ for some $c_j \equiv \int_{|z|=1} \frac{z}{|z|^3} z_j dz$,

$$\xi \cdot (C\omega) \cdot \xi = 0.$$

Now it is easy to get

$$\alpha(x, t) = \frac{3}{4\pi} pv \int_{\mathbb{R}^3} (\hat{y} \cdot \xi(x)) \det(\hat{y}, \xi(x+y), \xi(x)) |\omega(x+y)| \frac{dy}{|y|^3}, \quad (5.2)$$

where $\hat{y} = y/|y|$ is the direction of y , and $\det(a, b, c)$ is the determinant of the matrix with columns a, b, c in that order. The constant $\frac{3}{4\pi}$ will have no effect in the following argument, and will thus be neglected from now on.

The main idea of Constantin-Fefferman-Majda's argument comes from the following observation. Consider the 2D Euler equations. We know that no blow-up can ever occur. Put into the framework of (5.1) and (5.2), we see that the reason can be interpreted as the fact that for 2D flows, $\xi(x+y) = \xi(x) = e_3$ for all x and y , which means $\alpha(x, t) \equiv 0$. This implies that, if the orientation of the vorticity vectors varies only mildly, there would be no blow-up. Thus comes the following theorem. First we give some definitions.

Definition 5.2 (Smoothly directed). We say a set W_0 is *smoothly directed* if there exists $\rho > 0$ and r , $0 < r \leq \frac{\rho}{2}$ such that the following three conditions are satisfied.

First, for every $q \in W_0^* \equiv \{q \in W_0; |\omega_0(q)| \neq 0\}$ and all time $t \in [0, T)$, the function $\xi(\cdot, t)$ has a Lipschitz extension (denoted by the same letter) to the Euclidean ball of radius 4ρ centered at $X(q, t)$, denoted as $B_{4\rho}(X(q, t))$, and

$$M = \lim_{t \rightarrow T} \sup_{q \in W_0^*} \int_0^t \|\nabla \xi(\cdot, t)\|_{L^\infty(B_{4\rho}(X(q, t)))}^2 dt < \infty.$$

Secondly,

$$\sup_{B_{3r}(W_t)} |\omega(x, t)| \leq m \sup_{B_r(W_t)} |\omega(x, t)|$$

holds for all $t \in [0, T)$ with $m \geq 0$ constant. Here

$$W_t \equiv X(W_0, t).$$

Thirdly, for all $t \in [0, T)$,

$$\sup_{B_{4\rho}(W_t)} |u(x, t)| \leq U.$$

Theorem 5.3 (Constantin-Fefferman-Majda 1996). *Assume W_0 is smoothly directed. Then there exists $\tau > 0$ and Γ such that*

$$\sup_{B_r(W_t)} |\omega(x, t)| \leq \Gamma \sup_{B_\rho(W_{t_0})} |\omega(x, t_0)|$$

holds for any $0 \leq t_0 < T$ and $0 \leq t - t_0 \leq \tau$.

Noticing that, in (5.2), $\alpha(x, t)$ would also be zero when $\xi(x + y) = -\xi(x)$. This inspires the following pair of definition and theorem.

Definition 5.4. W_0 is said to be regularly directed, if there exists $\rho > 0$ such that

$$\sup_{q \in W_0^*} \int_0^T K_\rho(X(q, t)) dt < \infty,$$

where

$$K_\rho(x) = \int_{|y| \leq \rho} (\hat{y} \cdot \xi(x)) \det(\hat{y}, \xi(x + y), \xi(x)) |\omega(x + y)| \frac{dy}{|y|^3}.$$

Theorem 5.5 (Constantin-Fefferman-Majda 1996). *Assume W_0 regularly directed. Then there exists a constant Γ such that*

$$\sup_{q \in W_0} |\omega(X(q, t), t)| \leq \Gamma \sup_{q \in W_0} |\omega_0(q)|$$

holds for all $t \in [0, T]$.

Remark 5.6. An easy corollary to either theorem is that, there will be no blow-up up to time T .

The remaining of this subsection is devoted to the proof of Theorem 5.3. As will be seen during the proof, proving Theorem 5.5 is quite easy and will thus be omitted.

We decompose

$$\alpha(x) = \alpha_{in}(x) + \alpha_{out}(x),$$

where

$$\alpha_{in}(x) = pv \int \chi\left(\frac{|y|}{\rho}\right) (\hat{y} \cdot \xi(x)) \det(\hat{y}, \xi(x+y), \xi(x)) |\omega(x+y)| \frac{dy}{|y|^3}$$

and

$$\alpha_{out}(x) = \int \left(1 - \chi\left(\frac{|y|}{\rho}\right)\right) (\hat{y} \cdot \xi(x)) \det(\hat{y}, \xi(x+y), \xi(x)) |\omega(x+y)| \frac{dy}{|y|^3}$$

with $\chi(r)$ being a smooth non-negative function satisfying $\chi(r) = 1$ for $r \leq 1/2$ and 0 for $r \geq 1$. Then, recalling $\omega(x) = \nabla \times u(x)$ and $\xi(x+y) |\omega(x+y)| = \omega(x+y)$, we can do integration by parts in α_{out} and get

$$|\alpha_{out}(x)| \lesssim \rho^{-1} \int_{|y| \geq \rho/2} |u(x+y)| \frac{dy}{|y|^3}.$$

Then by Cauchy-Schwarz and the conservation of $\int |u|^2 dx$, we easily reach

$$|\alpha_{out}(x)| \lesssim C\rho^{-5/2} \|u_0\|_{L^2},$$

which remains bounded.

To estimate α_{in} , denote

$$G_\rho(x) = \sup_{|y| \leq \rho} |\nabla \xi(x+y)|.$$

Observe that $\det(\hat{y}, \xi(x+y), \xi(x)) = \hat{y} \cdot (\xi(x+y) \times \xi(x)) = \hat{y} \cdot ((\xi(x+y) - \xi(x)) \times \xi(x))$ which is bounded by $G_\rho(x) |y|$. Thus we have

$$|\alpha_{in}(x)| \leq G_\rho(x) I(x)$$

with

$$I(x) \equiv \int \chi\left(\frac{|y|}{\rho}\right) |\omega(x+y)| \frac{dy}{|y|^2}.$$

Next we split $I = I_1 + I_2$, where

$$I_1(x) = \int \chi\left(\frac{|y|}{\delta}\right) \chi\left(\frac{|y|}{\rho}\right) |\omega(x+y)| \frac{dy}{|y|^2}$$

and

$$I_2(x) = \int \left[1 - \chi\left(\frac{|y|}{\delta}\right) \right] \chi\left(\frac{|y|}{\rho}\right) |\omega(x+y)| \frac{dy}{|y|^2}$$

with $\delta \leq \rho/2$. Clearly we get

$$|I_1(x)| \leq C\delta\Omega_\delta,$$

where

$$\Omega_\delta(x) = \sup_{|y| \leq \delta} |\omega(x+y)|$$

by evaluating the integration through polar coordinates. To estimate I_2 , we replace $|\omega(x+y)|$ by $\xi(x+y) \cdot \omega(x+y) = \xi(x+y) \cdot (\nabla \times u(x+y))$ and invoke integration by parts, which gives

$$I_2(x) = \int u(x+y) \cdot \left\{ \nabla \times \left[\xi(x+y) \frac{1}{|y|^2} \chi\left(\frac{|y|}{\rho}\right) \left(1 - \chi\left(\frac{|y|}{\delta}\right) \right) \right] \right\} dy.$$

By putting $\nabla \times$ on each of the four terms, we decompose I_2 into four terms as follows:

$$I_2(x) = A + B + D + E.$$

It is easy to see that

$$|A| \leq CG_\rho(x) \int_{|y| \leq \rho} |u(x+y)| \frac{dy}{|y|^2},$$

$$|B| \leq C \int |u(x+y)| \left[1 - \chi\left(\frac{|y|}{\delta}\right) \right] \chi\left(\frac{|y|}{\rho}\right) \frac{dy}{|y|^3},$$

$$|D| \leq \frac{C}{\rho} \int_{|y| \leq \rho} |u(x+y)| \frac{dy}{|y|^2}$$

and

$$|E| \leq \frac{C}{\delta} \int_{\frac{\delta}{2} \leq |y| \leq \delta} |u(x+y)| \frac{dy}{|y|^2}.$$

If we denote

$$U_\rho(x) = \sup_{|y| \leq \rho} |u(x+y)|,$$

then we can easily estimate

$$|A| \leq C\rho U_\rho(x) G_\rho(x), \\ |D|, |E| \leq CU_\rho(x)$$

and

$$|B| \leq CU_\rho(x) \log\left(\frac{\rho}{\delta}\right).$$

Putting them together, we have

$$|\alpha(x)| \leq A_\rho(x) \left[1 + \log \left(\frac{\rho}{\delta} \right) \right] + G_\rho(x) \delta \Omega_\delta(x),$$

where

$$A_\rho(x) = C\rho^{-5/2} \|u_0\|_{L^2} + CG_\rho(x)U_\rho(x)(1 + \rho G_\rho(x)).$$

Studying what we have for a while, we see that if we can replace $\Omega_\delta(x)$ by $|\omega(x)|$, then by taking $\delta = |\omega(x)|^{-1}$, we will have

$$\int_0^T |\alpha| dt \leq \int_0^T G_\rho(x)^2 dt < \infty$$

by the smoothly directness of our set W_0 , since we have U_ρ to be bounded all the time. And this will effectively end the proof. So the final step should be to relate $\Omega_\delta(x)$ with $|\omega(x)|$, although the final proof doesn't go along the idea described above for technical reasons.

Consider a bunch of trajectories $X(q, t)$ and a neighborhood

$$\mathcal{B}_{4\rho} \equiv \{(x, t) : 0 \leq t < T, \exists q \in W_0, |X(q, t) - x| \leq 4\rho\}.$$

By the smoothly directness,

$$\sup_{(x,t) \in \mathcal{B}_{4\rho}} |u(x, t)| \leq U < \infty$$

and

$$M = \lim_{t \rightarrow T} \sup_{q \in W_0^*} \int_0^t G_{4\rho}^2(X(q, s)) ds < \infty.$$

Now define

$$B_r(W_t) = \{x; \exists q \in W_0, |x - X(q, t)| \leq r\}$$

with $2r \leq \rho$.

Let

$$\tau = \frac{r}{4U}$$

be a (possibly very short) time interval. Denote

$$w_r(t) = \sup_{B_r(W_t)} |\omega(x, t)|.$$

By assumption

$$w_{3r}(t) \leq mw_r(t).$$

Now consider $x \in B_r(W_t)$ for some $t < T$. The Lagrangian trajectory passing through x at time t is denoted $X(q', t)$. Note that q' may not be

in W_0 . Nevertheless, if $r \leq \frac{\rho}{2}$ and $0 \leq t-s \leq \tau$ then $X(q', s) \in B_{2r}(W_s)$, i.e.,

$$|X(q, s) - X(q', s)| \leq 2r \leq \rho$$

for some $q \in W_0$. Then it follows that

$$G_\rho(X(q', s)) \leq G_{4\rho}(X(q, s))$$

and

$$|\alpha(X(q', s))| \leq A_{4\rho}(X(q, s)) \left[1 + \log \frac{\rho}{\delta} \right] + G_{4\rho}(X(q, s)) \delta \Omega_\delta(X(q', s)).$$

Denoting

$$\mathcal{A}(s) = \sup_{q \in W_0^*} A_{4\rho}(X(q, s)),$$

$$\mathcal{G}(s) = \sup_{q \in W_0^*} G_{4\rho}(X(q, s)).$$

Then integrating (5.1) would give us

$$|\omega(X(q', t))| \leq K e^{\int_{t_0}^t \{ \mathcal{A}(s)[1 + \log(\rho/\delta)] + \mathcal{G}(s) \delta \Omega_\delta(X(q', s)) \} ds},$$

where

$$K = w_\rho(t_0).$$

Now we choose $\delta \leq r$, then $X(q', s) \in B_{2r}(W_s)$ and by assumption

$$\Omega_\delta(X(q', s)) \leq m w_r(s),$$

which implies

$$w_r(t) \leq K e^{\int_{t_0}^t \{ \mathcal{A}(s)[1 + \log(\rho/\delta)] + m \delta \mathcal{G}(s) w_r(s) \} ds}$$

for any $0 < \delta \leq r$ and $0 \leq t - t_0 \leq \tau$.

To simplify, define

$$A = A(t, t_0) = \int_{t_0}^t \mathcal{A}(s) ds$$

and

$$Q = K \rho \int_0^T \mathcal{G}(s) ds.$$

Let

$$y(t) = \max_{t_0 \leq s \leq t} \left(\frac{w_r(s)}{K} \right)$$

and

$$\frac{\rho}{\delta} = \max \left\{ m y(t) Q, \frac{\rho}{r} \right\}.$$

Then we obtain

$$y(t) \leq \left(\frac{\rho}{\delta}\right)^A e^{1+A}.$$

Finally, we can choose τ such that

$$A(t, t_0) \leq \frac{1}{2}.$$

This can be done since by assumption \mathcal{A} is integrable. Now fix this τ , we have

$$y(t) \leq \max \left\{ me^3 Q; \frac{\rho}{mrQ} \right\} \equiv \Gamma$$

and thus ends the proof.

5.2 Sufficient conditions by Deng-Hou-Yu

The result by Constantin, Fefferman and Majda reveals the subtlety between the smoothness of the vorticity direction field and the accumulation rate of vorticity. But on the other hand, their theorems are not quite applicable to various numerical simulations studying the blow-up issue of the 3D Euler equations in recent years. The most interesting ones among them are Kerr [26, 27, 28, 29] and Pelz [39, 40]. From their observations the following seems to hold for flows that may be singular, i.e., flows that seem to have the critical singular vorticity growth rate $(T-t)^{-1}$ for some $T > 0$ (Note: unforced flows that have higher vorticity growth rate have never been observed):

1. Large vorticity, or more specifically, those $|\omega| \geq c \|\omega\|_{L^\infty}$, are concentrated in small regions of length $O((T-t)^{1/2})$ in the vorticity direction and with cross-section area $O((T-t)^2)$. These regions look like two vortex sheets with thickness $O(T-t)$ meeting at an angle.
2. The vorticity direction field $\xi(x, t)$ looks more regular inside this region than outside, where $\xi(x, t)$ is wildly helical.

Checking these observations against Definition 5.2 and Theorem 5.3 (Note that Definition 5.4 is obviously unverifiable with numerical quantities, so we will not consider Theorem 5.5), we see that the conditions there are not satisfied. The main reason is that, according to numerical simulations, the “smoothly directed” region can never have fixed size, instead is always rapidly shrinking in all three directions. Thus there is a gap between theoretical theorems and numerical observations and leaving Theorem 5.3 unable to explain the numerical results.

In 2005, Deng, Hou and Yu [19] made a first step in filling this gap. The key is to focus on one vortex line and study its local stretching

behaviors. Before introducing the main result, we introduce some notations.

Denote by $\Omega(t)$ the maximum vorticity magnitude at time t . Let L_t be a family of vortex line segments and $L(t)$ be the length of L_t . Denote $U_\xi(t) \equiv \max_{x,y \in L_t} |(\mathbf{u} \cdot \xi)(x,t) - (\mathbf{u} \cdot \xi)(y,t)|$, $U_n(t) \equiv \max_{L_t} |\mathbf{u} \cdot \mathbf{n}|$ where \mathbf{n} is the normal of the curve L_t , i.e., $\frac{\partial}{\partial s}\xi = (\xi \cdot \nabla)\xi \equiv \kappa \mathbf{n}$ where κ is the curvature, and $M(t) \equiv \max \left(\|\nabla \cdot \xi\|_{L^\infty(L_t)}, \|\kappa\|_{L^\infty(L_t)} \right)$. We also define $X(a, t_1, t_2)$ as follows:

$$\frac{dX(\alpha, t_1, t)}{dt} = \mathbf{u}(X(\alpha, t_1, t), t); \quad X(\alpha, t_1, t_1) = \alpha.$$

It is related to the usual flow map $X(q, t)$ as follows:

$$X(q, t_2) = X(X(q, t_1), t_1, t_2)$$

for any q, t_1, t_2 .

Now the main theorem reads:

Theorem 5.7 (Deng-Hou-Yu, 2005). *Assume there is a family of vortex line segments L_t and $T_0 \in [0, T]$, such that $X(L_{t_1}, t_1, t_2) \supseteq L_{t_2}$ for all $T_0 < t_1 < t_2 < T$. We also assume that $\Omega(t)$ is monotonically increasing and $\|\omega(t)\|_{L^\infty(L_t)} \geq c_0 \Omega(t)$ for some $c_0 > 0$ when t is sufficiently close to T . Furthermore, we assume that*

1. $[U_\xi(t) + U_n(t)M(t)L(t)] \lesssim (T-t)^{-\alpha}$ for some $\alpha \in (0, 1)$,
2. $M(t)L(t) \leq C_0$, and
3. $L(t) \gtrsim (T-t)^\beta$ for some $\beta < 1 - \alpha$.

Then there will be no blow-up in the 3D incompressible Euler flow up to time T .

Remark 5.8. Note that the conditions 1–3 are inspired by the numerical observations. In Kerr's computations, the velocity blows up like $O((T-t)^{-1/2})$, which gives $\alpha = 1/2$. On the other hand, $M(t) = (T-t)^{-1/2}$. If we take $L(t) = (T-t)^{1/2}$, then the second condition is satisfied, but it would just violate the third condition. Thus Kerr's computations fall into the critical case of our theorem.

Remark 5.9. In a follow-up paper [21], Deng, Hou and Yu improved the above result and obtained non-blowup conditions for the critical case $\beta = 1 - \alpha$. The new conditions depend on some fine relations among the asymptotic behaviors of the rescaled quantities $(T-t)^\alpha [U_\xi(t) + U_n(t) M(t)L(t)]$, $(T-t)^{\alpha-1} L(t)$ and the bound C_0 . In [25], Hou and Li repeated Kerr's computations using a pseudo-spectral method with

resolution up to $1536 \times 1024 \times 3072$ up to $T = 19$, beyond the singularity time $T_c = 18.7$ predicted by Kerr. They found that there is a tremendous dynamic depletion of the vortex stretching term. The velocity field is found to be bounded, and the maximum vorticity does not grow faster than doubly exponential in time. The fact that velocity is bounded allows us to apply the non-blowup conditions of [22], which provides further theoretical evidence of the non-blowup of the Euler equations with Kerr's initial data.

We give a simple proof of the non-blowup result of Deng-Hou-Yu.

First we investigate the incompressibility condition of vorticity. $\nabla \cdot \omega = 0$. It is easy to see that

$$\frac{\partial |\omega|}{\partial s}(x, t) = -(\nabla \cdot \xi(x, t)) |\omega|(x, t),$$

where s is the arc length of the vortex line containing (x, t) , so that $\frac{\partial}{\partial s} = \xi \cdot \nabla$. This implies that for any two points $x, y \in L_t$, as long as $|\int_x^y \nabla \cdot \xi ds| \leq M(t)L(t) \leq C$, we have

$$e^{-M(t)L(t)} \leq \frac{|\omega(y, t)|}{|\omega(x, t)|} \leq e^{M(t)L(t)}. \quad (5.3)$$

Next we study the relation between vorticity magnitude and vortex line stretching. Recall that

$$\omega(X(\alpha, t), t) = \nabla_\alpha X(\alpha, t) \cdot \omega_0(\alpha).$$

Multiplying both side by $\xi(X(\alpha, t), t)$ we have

$$|\omega(X(\alpha, t), t)| = \xi(X(\alpha, t), t) \cdot \nabla_\alpha X(\alpha, t) \cdot \xi(\alpha) |\omega_0(\alpha)|.$$

Noticing

$$\xi(X(\alpha, t), t) = \frac{\partial X}{\partial s}$$

along the vortex line at time t , and similarly

$$\xi(\alpha) = \frac{\partial \alpha}{\partial \beta},$$

where β is the arc length parameter at time 0. Substituting these relations in, we have

$$\begin{aligned} |\omega(X(\alpha, t), t)| &= \frac{\partial X(\alpha, t)}{\partial s} \cdot \nabla_\alpha X(\alpha, t) \cdot \frac{\partial \alpha}{\partial \beta} |\omega_0(\alpha)| \\ &= \frac{\partial X}{\partial s} \cdot \frac{\partial X}{\partial \beta} |\omega_0(\alpha)| \\ &= \left(\frac{\partial X}{\partial s} \cdot \frac{\partial X}{\partial s} \right) \frac{\partial s}{\partial \beta} |\omega_0(\alpha)| \\ &= \frac{\partial s}{\partial \beta} |\omega_0(\alpha)|, \end{aligned}$$

since $\frac{\partial X}{\partial s} = \xi$ is a unit vector. It is easy to generalize the above result to prove that

$$\frac{\partial s}{\partial \beta}(X(\alpha, t_1, t), t) = \frac{|\omega(X(\alpha, t_1, t), t)|}{|\omega(\alpha, t_1)|}.$$

Now we have the relations between any two points on L_t , and between vortex line stretching and growth of vorticity magnitude. A third ingredient is the evolution equation of s_β . It is easy to see that s_β is governed by the same equation as $|\omega|$ in (5.1).

$$\begin{aligned} \frac{D}{Dt} s_\beta &= \xi \cdot \nabla \mathbf{u} \cdot \xi \ s_\beta \\ &= [(\xi \cdot \nabla)(\mathbf{u} \cdot \xi) - u \cdot (\xi \cdot \nabla)\xi] s_\beta \\ &= (\mathbf{u} \cdot \xi)_\beta - \kappa (\mathbf{u} \cdot \mathbf{n}) s_\beta, \end{aligned}$$

where we have used $\xi \cdot \nabla \xi = \partial_s \xi = \kappa \mathbf{n}$ by the Frènet relationship. Integrating it along L_t and in time, we easily get the estimate

$$l(t_2) \leq l(t_1) + \int_{t_1}^{t_2} U_\xi \ d\tau + \int_{t_1}^{t_2} M(\tau) U_n(\tau) l(\tau) \ d\tau,$$

where l_t is a segment of L_t such that $l_{t_2} = X(l_{t_1}, t_1, t_2)$, and $l(t)$ is the arclength of l_t .

Next we will show how $l(t_2)/l(t_1)$ is related to the vorticity growth:

$$\begin{aligned} e^{-(M(t)l(t)+M(t_1)l(t_1))} \frac{|\omega(X(\alpha', t_1, t), t)|}{|\omega(\alpha', t_1)|} &\leq \frac{l(t)}{l(t_1)} \\ &\leq e^{(M(t)l(t)+M(t_1)l(t_1))} \frac{|\omega(X(\alpha', t_1, t), t)|}{|\omega(\alpha', t_1)|}. \end{aligned} \quad (5.4)$$

The proof of (5.4) is not difficult. Let β denote the arc length parameter at time t_1 . Denote by l_t the vortex line segment from 0 to β , and use s as the arc length parameter at time t . Now by the mean value theorem, we have (β is the arclength variable at t_1)

$$\frac{l(t)}{l(t_1)} = \frac{\int_0^\beta s_\beta(\eta) \ d\eta}{\beta} = s_\beta(\eta') = \frac{|\omega(X(\alpha'', t_1, t), t)|}{|\omega(\alpha'', t_1)|}$$

for some α'' on the same vortex line. Now the inequality (5.4) follows from (5.3).

Now putting the three ingredients together, we get an estimate for the vorticity magnitude.

$$\Omega_l(t_2) \leq e^{C_0} \Omega_l(t_1) \left[1 + \frac{1}{l(t_1)} \int_{t_1}^{t_2} (U_\xi(\tau) + M(\tau) U_n(\tau) l(\tau)) \ d\tau \right], \quad (5.5)$$

where $\Omega_l(t)$ denotes the maximum vorticity magnitude along l_t .

Now we start the proof of Theorem 5.7 itself. The idea is the following. Note that the above inequality actually controls the growth rate of vorticity. So we can expect to prove non-blow up if $l(t_1)$ does not shrink to zero too fast. If we assume, in the same spirit as those by Constantin-Fefferman-Majda, that $l(t) > c > 0$ for some fixed c , then effectively we have

$$\Omega(t_2) \leq e^{C_0} \Omega(t_1)$$

and obviously no blow-up can happen. Now we illustrate the proof along this simple idea.

We prove by contradiction. First, by translating the initial time we can assume that the assumptions hold in $[0, T)$. Define

$$r \equiv (R/c_0) + 1,$$

where $R \equiv e^{2C_0}$. Recall that C_0 is the bound of $M(t)L(t)$, and c_0 is the lower bound of $\Omega_L(t)/\Omega(t)$, where $\Omega_L(t) \equiv \|\omega(\cdot, t)\|_{L^\infty(L_t)}$.

If there is a finite time blow-up at time T , then we must have

$$\int_0^T \Omega(t) dt = \infty$$

and necessarily $\Omega(t) \nearrow \infty$ as $t \nearrow T$. Take $t_1, t_2, \dots, t_n, \dots$ such that

$$\Omega(t_{k+1}) = r\Omega(t_k).$$

Since $\Omega(t)$ is monotone by assumption, and T is the smallest time that $\int_0^T \Omega(t) dt = \infty$, we have $t_n \nearrow T$ as $n \rightarrow \infty$.

Now we choose $l_{t_2} = L_{t_2}$. By assumptions on L_t , we have $l_{t_1} \subset L_{t_1}$ such that $X(l_{t_1}, t_1, t_2) = l_{t_2}$. And furthermore, by using (5.4), we obtain

$$l(t_1) \geq l(t_2) \frac{1}{R} \frac{\Omega_L(t_1)}{\Omega_L(t_2)} \geq l(t_2) \frac{c_0}{R^2} \frac{1}{r} \gtrsim (T - t_2)^\beta,$$

where the hidden constant in \gtrsim is independent of time. Now plugging this into (5.5) we have, after some algebra,

$$\Omega(t_2) \leq (r - 1)\Omega(t_1) + \frac{C}{(1 - \alpha)c_0} \frac{\Omega(t_1)}{(T - t_2)^\beta} (T - t_1)^{1-\alpha}.$$

Recalling $\Omega(t_2) = r\Omega(t_1)$, we have

$$r \leq (r - 1) + C \frac{(T - t_1)^{1-\alpha}}{(T - t_2)^\beta},$$

which gives

$$(T - t_2) \leq C(T - t_1)^{1+2\delta}$$

with

$$\delta \equiv \frac{1-\alpha}{\beta} - 1,$$

which is positive by assumption. By taking t_1 close enough to T , we can cancel C and have

$$(T - t_2) \leq (T - t_1)^{1+\delta}.$$

Next do the same thing for all pairs (t_n, t_{n+1}) , (note that $(T - t_n)^\delta < (T - t_1)^\delta \leq C^{-1}$) we have

$$(T - t_{k+1}) \leq (T - t_k)^{1+\delta} \leq (T - t_1)^{(1+\delta)^k} \leq (T - t_1)(T - t_1)^{\delta k} \quad (5.6)$$

if we take $T - t_1 < \frac{1}{T}$.

Now we study $\int_0^T \Omega(t) dt = \infty$. By assumption that $\Omega(t)$ is monotone, we have

$$\Omega(t_1) \sum_{k=1}^{\infty} r^k (t_{k+1} - t_k) = \sum_{k=1}^{\infty} \Omega(t_{k+1}) (t_{k+1} - t_k) \geq \int_{t_1}^T \Omega(t) dt = \infty,$$

which implies

$$\begin{aligned} (r-1) \sum_{l=0}^{\infty} r^l (T - t_{l+1}) &= \sum_{l=0}^{\infty} (r^{l+1} - r^l) (T - t_{l+1}) \\ &= \sum_{l=0}^{\infty} \sum_{k=l+1}^{\infty} (r^{l+1} - r^l) (t_{k+1} - t_k) \\ &= \sum_{k=1}^{\infty} \sum_{l=0}^{k-1} (r^{l+1} - r^l) (t_{k+1} - t_k) \\ &= \sum_{k=1}^{\infty} (r^k - 1) (t_{k+1} - t_k) \\ &= \infty. \end{aligned}$$

All the equalities are legitimate since all the terms in the summations are positive (Fubini's theorem).

On the other hand, from (5.6), we obtain

$$\infty = \sum_{k=0}^{\infty} r^k (T - t_{k+1}) \leq (T - t_1) \sum_{k=0}^{\infty} [r(T - t_1)^\delta]^k < \infty,$$

if we choose t_1 close to T so that $r(T - t_1)^\delta < 1$. Therefore, we reach a contradiction. Thus, we obtain

$$\int_{t_1}^T \Omega(t) dt < \infty.$$

By the BKM criterion, we conclude that there is no finite time blow-up up to T .

6 Lower dimensional models for the 3D Euler equations

6.1 1-D model

In 1985, P. Constantin, P. Lax and A. Majda proposed the following 1-D model of the 3D Euler equations:

$$\omega_t = H\omega \cdot \omega,$$

where H is the Hilbert transform:

$$Hf = p v \int_{\mathbb{R}} \frac{f(y)}{x - y} dy.$$

The relation to the 3D Euler equations is the following. In 3D Euler equation, the evolution of the vorticity magnitude is governed by the following equation:

$$\frac{D}{Dt} |\omega| = T(\omega) |\omega|,$$

where T is a Calderon-Zygmund operator with a convolution kernel that is homogeneous of degree $-d$ where d is the dimension. In 1-D, only one such singular integral kernel exists, i.e., the Hilbert transform.

This simplified model can be explicitly solved. To solve it, we first get familiar with some properties of the Hilbert transform.

Lemma 6.1. *The Hilbert transform has the following properties:*

1. H is bounded from H^m to H^m for all $m \geq 0$.
2. $H(Hf) = -f$.
3. $H(fg) = f(Hg) + g(Hf) + H(Hf \cdot Hg)$.

Proof. Properties (1) and (2) follow immediately from the fact that

$$\widehat{Hf}(\xi) = sgn(\xi) \hat{f}(\xi).$$

For property (3), we check

$$\begin{aligned} \widehat{H(fg)} - H(\widehat{Hf} \cdot Hg) &= \int_{-\infty}^{\infty} sgn(\xi) \hat{f}(\eta) \hat{g}(\xi - \eta) d\eta \\ &\quad - \int_{-\infty}^{\infty} sgn(\xi) sgn(\eta) sgn(\xi - \eta) \hat{f}(\eta) \hat{g}(\xi - \eta) d\eta \\ &= \int_{-\infty}^{\infty} sgn(\xi) (1 - sgn(\eta) sgn(\xi - \eta)) \hat{f}(\eta) \hat{g}(\xi - \eta) d\eta \\ &= \int_{-\infty}^{\infty} (sgn(\xi - \eta) + sgn(\eta)) \hat{f}(\eta) \hat{g}(\xi - \eta) d\eta \\ &= \widehat{f(Hg)} + \widehat{g(Hf)}, \end{aligned}$$

and this ends the proof. \square

Now we set out to find the explicit solutions. We define

$$z(x, t) = H\omega(x, t) + i\omega(x, t).$$

By Lemma 6.1, the equation for z is

$$\frac{dz}{dt} = \frac{1}{2}z^2$$

whose explicit solution is

$$z(t) = \frac{2z_0}{2 - z_0 t},$$

which implies

$$\omega(x, t) = \frac{4\omega_0(x)}{(2 - tH\omega_0)^2 + t^2\omega_0^2(x)}.$$

It is obvious that $\omega(x, t)$ will blow-up at points with $\omega_0(x) = 0$ but $H\omega_0 > 0$.

6.2 The 2-D QG equation

The 2D QG equation (see Pedlosky [41]) is given by

$$\frac{D\theta}{Dt} \equiv \theta_t + u \cdot \nabla \theta = 0, \quad (6.1)$$

where $\theta(x, t)$ is a scalar, and u is defined by

$$\begin{aligned} (-\Delta)^{1/2} \psi &= -\theta, \\ u &= \nabla^\perp \psi. \end{aligned}$$

Here $(-\Delta)^{1/2}$ is defined by

$$(-\Delta)^{1/2} \psi = \int e^{2\pi i x \cdot \xi} 2\pi |\xi| \hat{\psi}(\xi) d\xi$$

if

$$\psi = \int e^{2\pi i x \cdot \xi} \hat{\psi}(\xi) d\xi.$$

The 2D QG equation (aka surface-quasi-geostrophic equations, SQG) describes the variation of the density variation θ at the surface of the earth. The name θ , usually represents temperature, is chosen because in the case the ideal gas, the density variation is proportional to the temperature.

To get an explicit form of the formula for ψ in the space variable x instead of the Fourier modes ξ , we use the following lemma:

Lemma 6.2. Denote

$$h_a(x) = \frac{\Gamma(a/2)}{\pi(a/2)} |x|^{-a},$$

then we have

$$\hat{h}_a = h_{N-a}$$

for $0 < \Re(a) < N$, where N is the dimension of the space. Γ is the Gamma function, defined as

$$\Gamma(s) = \int_0^\infty e^{-t} t^{s-1} dt.$$

Proof. See e.g. Thomas Wolff [46]. □

By the above lemma we easily derive

$$\psi(x) = - \int_{\mathbb{R}^2} \frac{\theta(x+y)}{|y|} dy.$$

Thus we get

$$u(x) = \int_{\mathbb{R}^2} \frac{y^\perp}{|y|^2} \theta(x+y) dy.$$

If we define “vorticity”

$$\omega(x) = \nabla^\perp \theta,$$

we obtain by differentiating (6.1) that

$$\frac{D\omega}{Dt} = \nabla u \cdot \omega,$$

from which we can derive

$$\begin{aligned} \frac{D|\omega|}{Dt} &= \frac{1}{2} \xi (\nabla u + \nabla u^T) \xi |\omega| \\ &\equiv S(x, t) |\omega| \\ &= \int_{\mathbb{R}^2} \frac{(\hat{y} \cdot \xi^\perp(x)) (\xi(x+y) \cdot \xi^\perp(x))}{|y|^2} |\omega(x+y)| dy |\omega| \\ &= \int_{\mathbb{R}^2} \frac{(\hat{y} \cdot \xi(x)) \det(\xi(x+y), \xi(x))}{|y|^2} |\omega(x+y)| dy |\omega|, \end{aligned}$$

where

$$\xi(x, t) \equiv \frac{\omega(x, t)}{|\omega(x, t)|}$$

as long as it is well-defined, and $\hat{y} = y/|y|$. Note that those points with $\omega(x, t) = 0$ is transported by the flow, since $\omega = 0$ implies $\nabla\theta = 0$ and

$$\begin{aligned}\nabla\theta(X(q, t)) &= \nabla_x\theta_0(q) \\ &= (\nabla_q X)^{-1} \cdot \nabla_q\theta_0(q),\end{aligned}$$

which means $\nabla_q\theta_0(q) = 0 \Leftrightarrow \nabla_x\theta(X(q, t)) = 0$. So those points where ξ is not well-defined are not important to the stretching.

Recall that for the evolution of the vorticity magnitude in 3D Euler, we have

$$\frac{D|\omega|}{Dt} = \alpha(x, t)|\omega|,$$

where

$$\alpha(x, t) = \frac{3}{4\pi} \int_{\mathbb{R}^3} \frac{(\hat{y} \cdot \xi(x)) \det(\hat{y}, \xi(x+y), \xi(x))}{|y|^3} |\omega(x+y)| dy.$$

We see that $S(x, t)$ and $\alpha(x, t)$ indeed share very similar cancellation properties. Thus the 2D QG equation can be viewed as a 2D model of the 3D Euler equation, especially in the vorticity form.

There are several other similarities between 2D QG and 3D Euler. For example, the levelsets of $\theta(x, t)$, which are lines that are always tangent to $\omega(x, t)$ so can be defined as “vortex lines”, are carried by the flow, similar to the vortex lines in the 3D Euler dynamics. For more comparison between 2D QG and 3D Euler equations, as well as other properties of the 2D QG equations, see Constantin-Majda-Tabak [15], or the book by Majda-Bertozzi [35].

Remark 6.3. Note that in the 2D QG equation, we no longer have the property

$$\frac{1}{2} (\nabla u - \nabla u^T) \omega = 0$$

as in the 3D Euler case. This implies that, the “vorticity” here doesn’t satisfy

$$\frac{D\omega}{Dt} = \frac{1}{2} (\nabla u + \nabla u^T) \omega$$

as in the Euler case. Only the evolution of the vorticity magnitude $|\omega|$ satisfies the same equation as in the 3D Euler equation.

6.2.1 Existence and blow-up criteria

By the same technique as in Chapter 2, we can prove the local in time existence and blow-up criterion.

Theorem 6.4 (Constantin-Majda-Tabak [15]). *If the initial value $\theta_0(x)$ belongs to the Sobolev space $H^k(\mathbb{R}^2)$ for some integer $k \geq 3$, then there is a smooth solution $\theta(x, t) \in H^k(\mathbb{R}^2)$ for the 2D QG equation for each time t , in a sufficiently small time interval $[0, T^*)$, where T^* is characterized by*

$$\|\theta(\cdot, t)\|_k \nearrow \infty \text{ as } t \nearrow T^*$$

and can be estimated from below by

$$T^* \gtrsim \frac{1}{1 - \|\theta_0\|_k}.$$

We can also apply the technique for the BKM criterion in Chapter 4 to obtain similar blow-up criteria:

Theorem 6.5 (Constantin-Majda-Tabak [15]). *Consider the unique smooth solution of the 2D QG equations with initial data $\theta_0(x) \in H^k(\mathbb{R}^2)$ for some $k \geq 3$. Then the following are equivalent:*

1. *The time interval $[0, T^*)$ for some $T^* < \infty$ is maximal for the solution to be in $H^k(\mathbb{R}^2)$.*
2. *The vorticity magnitude accumulates so rapidly that*

$$\int_0^T \|\omega(\cdot, t)\|_{L^\infty} dt \nearrow \infty \text{ as } T \nearrow \infty.$$

3. *Let $S^*(t) \equiv \max_{x \in \mathbb{R}^2} S(x, t)$, then*

$$\int_0^{T^*} S^*(t) dt = \infty.$$

There are, though, properties that seem to hold only in the 2D QG case. For example, when we assume that there is a smooth curve $x(t)$, such that each point $(x(t), t)$ is an isolated maximum of $|\omega(x, t)|$, we can have the following result:

$$\frac{d}{dt} \|\omega(\cdot, t)\|_{L^\infty} = S(x(t), t) \|\omega(\cdot, t)\|_{L^\infty}.$$

To prove it, let $q(t)$ be the Lagrange marker of the points $(x(t), t)$, i.e.,

$$X(q(t), t) = x(t),$$

then we have

$$\begin{aligned}
 \frac{d}{dt} \|\omega(\cdot, t)\|_{L^\infty} &= \frac{d}{dt} |\omega(x(t), t)| \\
 &= \frac{d}{dt} |\omega(X(q(t), t), t)| \\
 &= \frac{D}{Dt} |\omega|(x(t), t) + \nabla_x |\omega| \cdot \nabla_q X \cdot \dot{q} \\
 &= S(x(t), t) |\omega(x(t), t)| \\
 &= S(x(t), t) \|\omega(\cdot, t)\|_{L^\infty}.
 \end{aligned}$$

Note that $\nabla_x |\omega| = 0$ by our assumption that $x(t)$ is an isolated maximum.

The above result implies that, under the assumption on $x(t)$, we can just consider $S(x, t)$ for the particular point $(x(t), t)$ instead of the maximum of $S(x, t)$ over the whole space. The assumption on $x(t)$ is very likely to hold in practical cases according to various numerical results, see e.g. Constantin-Majda-Tabak [15].

Remark 6.6. It is claimed in Constantin-Majda-Tabak [15] that the assumption on $x(t)$ can be dropped with a more lengthy proof, while that proof is omitted.

6.2.2 Global existence result by Constantin-Majda-Tabak

In their 1994 paper [15], Constantin, Majda and Tabak studied the evolution of the vorticity magnitude both numerically and theoretically, concluded that when the vorticity direction $\xi(x, t)$ varies not too fast in space, there can be no finite time blow-up, i.e., the classical solution exists globally in time.

To understand the basic idea, we recall the evolution equation for $|\omega|$:

$$\frac{D|\omega|}{Dt} = S(x, t) |\omega|,$$

where

$$S(x, t) = \int_{\mathbb{R}^2} \frac{(\hat{y} \cdot \xi^\perp(x)) (\xi(x+y) \cdot \xi^\perp(x))}{|y|^2} |\omega(x+y)| dy.$$

In general, since $S(x, t) = T\omega$ with T being a singular integral operator, $\|S(\cdot, t)\|_{L^\infty}$ can not be bounded by $\|\omega(\cdot, t)\|_{L^\infty}$. Even if it can, the right hand side would be quadratic and give us a finite time blow-up. But if we make assumptions on $\xi(x+y)$, the situation would be different. We illustrate this through several examples.

Example 6.7 (Constantin-Majda-Tabak [15]). We consider the classical frontogenesis with trivial topology. Let

$$x_2 = f(x_1)$$

be a smooth curve in the plane, we study the possibility that the solution $\theta(x, t)$ develops a sharp front along this curve, through the simplified ansatz

$$\theta(x, t) = F\left(\frac{x_2 - f(x_1)}{\delta(t)}\right),$$

where $F(s)$ is a smooth function on \mathbb{R} , with the properties that $F(s) = 1$ for $s \geq 3$, $F(s) = 0$ for $s \leq 1$ and $F'(s) \geq 0$ for all s . Assume that

$$\delta(t) \rightarrow 0, \quad \text{as } t \rightarrow T^*$$

for some $T^* < \infty$.

We can plug the formula for θ into the 2D QG equation and get

$$F' \left[\frac{d}{dt} \left(\frac{1}{\delta(t)} \right) + \mathbf{u} \cdot \begin{pmatrix} f'(x_1) \\ 1 \end{pmatrix} \left(\frac{1}{\delta(t)} \right) \right] = 0.$$

Since obviously $\|\omega\|_{L^\infty}(t) \sim 1/\delta(t)$, we have the estimate

$$\frac{d}{dt} (\log \|\omega\|_{L^\infty}(t)) \lesssim \|u\|_{L^\infty}(t).$$

It can be shown that for 2D QG equation

$$\|u\|_{L^\infty}(t) \lesssim \log \|\omega\|_{L^\infty}. \quad (6.2)$$

We see that the growth rate of the maximum vorticity is at most double exponential, and there will be no finite time blow-up.

The last thing is to prove the estimate (6.2), which first appears in Cordoba [17].

Recall that, for the 2D QG equation, we have

$$u = (-\Delta)^{-1/2} \omega = \int \frac{1}{|y|} \omega(x+y) dy.$$

Now let $r > 0$ fixed, large enough, $\rho \in (0, r)$ to be specified later, and χ be the standard cut-off function, we decompose u into 3 terms as follows

$$|u(x)| = U_{in}(x) + U_{med}(x) + U_{out}(x),$$

where

$$\begin{aligned} U_{in}(x) &= \int \chi\left(\frac{|x|}{\rho}\right) \frac{1}{|y|} \omega(x+y) dy \\ &\leq \|\omega\|_{L^\infty} \rho \end{aligned}$$

by simply using polar coordinates. For U_{med} , we have

$$\begin{aligned} U_{med}(x) &= \int \chi\left(\frac{|x|}{r}\right) \left(1 - \chi\left(\frac{|x|}{\rho}\right)\right) \frac{1}{|y|} \omega(x+y) dy \\ &= \int \chi\left(\frac{|x|}{r}\right) \left(1 - \chi\left(\frac{|x|}{\rho}\right)\right) \frac{1}{|y|} \nabla^\perp \theta(x+y) dy \\ &\lesssim \int_{2r \geq |y| \geq \rho/2} \frac{1}{|y|^2} \theta(x+y) dy + \frac{1}{\rho} \int_{2\rho \geq |y| \geq \rho/2} \frac{\theta(x+y)}{|y|} dy \\ &\quad + \frac{1}{r} \int_{2r \geq |y| \geq r/2} \frac{1}{|y|} \theta(x+y) dy \\ &\lesssim -\|\theta\|_{L^\infty} (1 + |\log \rho|) = -\|\theta_0\|_{L^\infty} (1 + |\log \rho|) \end{aligned}$$

as long as $\rho < c < 1$ for some fixed constant c . Here we have used the fact that $\nabla \chi\left(\frac{|x|}{\rho}\right) = 0$ for all $|x| \leq \rho/2$ or $|x| \geq 2\rho$ and the maximum of $|\theta|$ is bounded by the initial data.

Now we estimate U_{out} ,

$$\begin{aligned} U_{out}(x) &= \int \left(1 - \chi\left(\frac{|x|}{r}\right)\right) \frac{1}{|y|} \nabla^\perp \theta(x+y) dy \\ &\lesssim \frac{1}{r} \int_{2r \geq |y| \geq r/2} \frac{1}{|y|} \theta(x+y) dy + \int_{|y| \geq r/2} \theta(x+y) \frac{dy}{|y|^2} \\ &\equiv I + II. \end{aligned}$$

I is obviously bounded by some constant since $\|\theta\|_\infty \leq \|\theta_0\|_\infty$. For II , we use the Cauchy-Schwarz inequality and the fact that the L^2 norm of θ is conserved. We get

$$II \lesssim r^{-1} \|\theta_0\|_{L^2},$$

which is also bounded by a constant.

Finally, if $\|\omega\|_{L^\infty} \leq e$, (6.2) trivially holds. If not, taking $\rho = \|\omega\|_{L^\infty}^{-1}$ immediately gives the desired estimate.

We look at another example, the singular thermal ridge.

Example 6.8 (Constantin-Majda-Tabak [15]). The assumptions are similar to the previous example, the only difference is that $F(s) = 0$ for both $s \geq 3$ and $s \leq 1$, with $F'(s) > 0$ for $1 < s < 2$, $F'(s) < 0$ for $2 < s < 3$. There can be no finite time blow-up for these ridges either. The proof is similar to that in the last example and is omitted.

The above two examples imply that, for θ whose levelsets form simple geometries, there may be no finite time blow-up. To quantify what we mean by “simple geometry”, we use the direction of the “vorticity vectors” $\xi = \omega / |\omega|$. The precise statement of the theorem is the following (Constantin-Majda-Tabak [15]):

Definition 6.9. A set Ω_0 is *smoothly directed* if there exists $\rho > 0$ such that

$$\sup_{q \in \Omega_0} \int_0^T |u(X(q, t), t)|^2 dt < \infty$$

and

$$\sup_{q \in \Omega_0^*} \int_0^T \|\nabla \xi(\cdot, t)\|_{L^\infty(B_\rho(X(q, t), t))} dt < \infty,$$

where $B_\rho(x)$ is the ball of radius ρ centered at x and

$$\Omega_0^* = \{q \in \Omega_0 \mid |\omega_0(q)| \neq 0\}.$$

We use the following notations:

$$\Omega_t = X(\Omega_0, t),$$

$$O_T(\Omega_0) = \{(x, t) \mid x \in \Omega_t, 0 \leq t \leq T\}.$$

Theorem 6.10. Assume Ω_0 is smoothly directed, then

$$\sup_{O_T(\Omega_0)} |\nabla \theta(x, t)| < \infty,$$

i.e., there can be no blow-up in $O_T(\Omega_0)$.

Definition 6.11. We say that the set Ω_0 is *regularly directed* if there exists $\rho > 0$ such that

$$\sup_{q \in \Omega_0^*} \int_0^T K_\rho(X(q, t)) dt < \infty,$$

where

$$K_\rho(x) = \int_{|y| \leq \rho} |\hat{y} \cdot \xi^\perp(x)| |\xi(x + y) \cdot \xi^\perp(x)| |\omega(x + y)| \frac{dy}{|y|^2}.$$

Theorem 6.12. Assume that Ω_0 is regularly directed, then

$$\sup_{O_T(\Omega_0)} |\omega(x, t)| < \infty.$$

The proofs to these theorems are similar to the ones in the global existence results by Constantin-Fefferman-Majda for the 3D Euler equations, only less technical. The main difference is that here we have a conserved quantity θ , whose L^p norm is conserved for all $1 \leq p \leq \infty$. This simplifies the proof a lot. First, $S(x, t)$ is bounded by

$$|S(x, t)| \leq C [G(t) |u(x, t)| + (\rho G(t) + 1) (G(t) \|\theta\|_{L^\infty} + \rho^{-2} \|\theta\|_{L^2})],$$

where $G(t) \equiv \sup_{|y|, \rho} |\nabla \xi(x + y)|$ for some fixed $\rho > 0$, via similar estimates as in Chapter 2. Next we integrate the above in time and use the Cauchy-Schwarz inequality. For details see Constantin-Majda-Tabak [15].

Remark 6.13. The reader may notice that our condition on the maximum velocity, i.e., L^2 -integrable in time, is much weaker than the one in the 3D Euler case, i.e., L^∞ bounded. This is because, in 2D QG, we have $\omega = (\partial_2, -\partial_1)\theta$ with θ being bounded. For 3D Euler, we have $\omega = \nabla \times u$ and we do not have *a priori* bound on u . Thus in the case of the 3D Euler equation, we have a term

$$G(t)^2 U(t),$$

which will not be integrable if $U(t) \equiv \|u\|_{L^\infty}(t)$ is not bounded in addition.

6.2.3 Global existence result by Cordoba and Fefferman

The results by Constantin-Majda-Tabak claim that, as long as the direction field of the levelsets is smooth enough locally around the maximum stretching point, there can be no finite time blow-up in the 2D QG equations. This leaves one candidate for finite-time blow-up in their numerical simulations, i.e., the “hyperbolic saddle” situation. In fact, they performed detailed numerical experiments and found that the maximum vorticity can be fitted by $1/(8.25 - t)^{1.7}$, which suggests a finite time blow-up. In 1997, Ohkitani and Yamada re-did the simulations and pushed further to higher resolutions ([38]), and found that the same result can be fitted as well by double exponential growth, indicating that no finite time blow-up can occur, at least up to the time of their computations. Subsequently, Constantin-Nie-Schörgofer [16]) found that the double exponential is in several aspects a better fit, suggesting that no finite-time blow-up can occur. Around the same time, D. Cordoba [17] proved that under some mild assumptions, the hyperbolic saddles will not cause a finite time blow-up, instead the growth of $|\omega|$ is bounded by quadruple exponential. The proof is technical and we will not reproduce it here.

In 2002, D. Cordoba and C. Fefferman [18] considered a case that covers most of the scenarios considered by Constantin-Majda-Tabak and the hyperbolic saddle case by Cordoba, which they called “semi-uniform collapse”, and obtained the numerically observed double exponential growth by clever estimates. We will recap their work here.

Assume that there is an interval $[a, b]$ such that

$$\theta(x_1, \phi_\rho(x_1, t), t) = G(\rho)$$

for $x_1 \in [a, b]$, where $x_2 = \phi_\rho(x_1, t)$ is a level curve of θ , $\phi_\rho \in C^1([a, b] \times [0, T])$ for some alleged blow-up time T . By a “semi-uniform” collapse we mean that the level sets are almost parallel to each other (Note that the sharpening front and ridge in Examples 6.7 and 6.8 satisfy that the

level curves are exactly parallel to each other). More specifically, if we denote

$$\delta(x_1, t) \equiv |\phi_\rho(x_1, t) - \phi_{\rho'}(x_1, t)|,$$

then δ satisfies

$$\min_{[a,b]} \delta(x_1, t) \geq c \max_{[a,b]} \delta(x_1, t).$$

By this assumption, we always have

$$|\omega(x_1, \phi_\rho(x_1, t), t)| \sim \frac{1}{\delta(x'_1, t)}$$

for any $x_1, x'_1 \in [a, b]$.

From this observation, it is enough to consider

$$I = \frac{d}{dt} \left(\int_a^b [\phi_{\rho_2}(x_1, t) - \phi_{\rho_1}(x_1, t)] dx_1 \right)$$

since the quantity being differentiated is comparable to $|\omega|^{-1}$. (Note that, since different level curves will never cross, the sign of the difference is fixed.)

We compute

$$\frac{d}{dt} \phi_\rho(x_1, t)$$

for some fixed ρ . First note that, the curve $(x_1, \phi_\rho(x_1, t))$ is transported by the flow, since it always parametrized the level curve $\theta = G(\rho)$. So we have

$$\frac{d}{dt} \phi_\rho(x_1, t) = u_2 - u_1 \frac{\partial \phi_\rho}{\partial x_1} = \frac{d}{dx_1} \psi(x_1, \phi_\rho(x_1, t), t),$$

where $\psi = (-\Delta)^{-1/2} \theta$ so that $u = \nabla^\perp \psi$. The first equality can be seen by drawing a picture and studying the difference between ϕ_ρ at t and $t + \delta t$, or go through the argument using the QG equation as in Cordoba-Fefferman [18].

Now it is immediate that

$$\begin{aligned} I &= \psi(b, \phi_{\rho_2}(b, t), t) - \psi(a, \phi_{\rho_2}(a, t), t) \\ &\quad + \psi(a, \phi_{\rho_1}(a, t), t) - \psi(b, \phi_{\rho_2}(b, t), t). \end{aligned}$$

Let

$$A(t) \equiv \frac{1}{b-a} \int_a^b [\phi_{\rho_2}(x_1, t) - \phi_{\rho_1}(x_1, t)] dx_1,$$

we have

$$\left| \frac{d}{dt} A(t) \right| \lesssim \sup_{[a,b]} |\psi(x_1, \phi_{\rho_2}(x_1, t), t) - \psi(x_1, \phi_{\rho_1}(x_1, t), t)|.$$

Finally we prove a general estimate

$$|\psi(z_1, t) - \psi(z_2, t)| \lesssim |z_1 - z_2| \log |z_1 - z_2|.$$

Obviously, that will end the proof, and bound the maximum growth by some double exponential.

Recall that

$$\psi(x, t) = (-\Delta)^{-1/2} \theta = - \int \frac{\theta(x + y)}{|y|} dy.$$

Taking $\tau = |z_1 - z_2|$ we have

$$\begin{aligned} \psi(z_1) - \psi(z_2) &= \int \theta(y) \left(\frac{1}{|y - z_1|} - \frac{1}{|y - z_2|} \right) dy \\ &= \int_{|y-z_1| \leq 2\tau} + \int_{2\tau < |y-z_1| \leq k} + \int_{k < |y-z_1|} \\ &= I_1 + I_2 + I_3, \end{aligned}$$

where $k > 2\tau$ is some constant.

Now trivially,

$$|I_1| \leq C\tau.$$

For I_2 , by the mean value theorem

$$\left| \frac{1}{|y - z_1|} - \frac{1}{|y - z_2|} \right| = \tau \left| \nabla \frac{1}{|y - z'|} \right|$$

for some z' lying on the line segment connecting z_1 and z_2 , thus we can further bound it by

$$\tau \max_s \frac{1}{|y - s|^2},$$

where the maximum is taken over the line connecting z_1 and z_2 . Now it is clear that

$$|I_2| \leq C\tau |\log \tau|.$$

I_3 is also trivially bounded by $C\tau$ using the conservation of the L^2 norm of θ , and the mean value theorem.

This ends the proof.

6.2.4 Final remarks about the QG equation

The global existence/blow-up issue for the 2D quasi-geostrophic equation is still open today, and solving it would for sure shed light on and help solving the same problem for the 3D Euler equations. A recent progress is Deng-Hou-Li-Yu [22], where the authors applied the method

developed in their papers dealing with the 3D Euler equations [19, 21], and obtained triple exponential growth bound for $\|\nabla\theta\|_\infty$ under very mild conditions. Furthermore, under slightly stronger conditions the authors show that the growth rate of $\|\nabla\theta\|_\infty$ can be bounded by double exponential, which is the real growth rate observed in numerical computations. High resolution numerical computations carried out by the authors suggest that these conditions are indeed satisfied in the 2D QG flow. This observation suggests that these conditions may have touched the essence of the QG dynamics. The authors are currently making an effort to further investigate this problem.

7 Vortex patch

A vortex patch is a bounded, simply connected, open material domain \mathcal{D}_t such that the vorticity is constant inside it and 0 elsewhere. It is a special case of the $L^1 \cap L^\infty$ weak solutions. Here we will describe the problem without using the general weak solution formalism.

7.1 The contour dynamics equation (CDE)

By definition and our expectation that the vorticity will be conserved along particle trajectories (should check that they really exist), it is (hopefully) enough to derive an equation that governs the evolution of the boundary.

Assume that the solution do behave this way, i.e., the vorticity at any time t is ω_0 in some smooth region $\mathcal{D}(t)$ and 0 outside, where $\mathcal{D}(t)$ is smoothly parametrized by t . Then the velocity is

$$u(x, t) = \frac{\omega_0}{2\pi} \int_{\mathcal{D}(t)} \frac{(x - y)^\perp}{|x - y|^2} dy.$$

By the Divergence Theorem we can rewrite it as a contour integral

$$u(x, t) = \frac{\omega_0}{2\pi} \int_{\partial\mathcal{D}(t)} \log|x - y| n^\perp(y) dS(y),$$

where $n(y)$ is the unit outer normal vector. Note that in our setting, $\mathcal{D}(t)$ and then $\omega(x, t)$ is determined by the evolution of the boundary $\partial\mathcal{D}(t)$. If we parametrize it by $x = x(s, t)$, we have

$$\frac{\partial x(s, t)}{\partial t} = -\frac{\omega_0}{2\pi} \int_{\partial\mathcal{D}(t)} \log|x(s, t) - x(s', t)| \frac{\partial x}{\partial s'}(s', t) ds'.$$

This is called the CDE (contour dynamics equation). It can be checked that as long as the boundary remains smooth enough, $\omega(x, t)$ defined by the CDE is a weak solution.

In [34], A. Majda observed that, $Y = \frac{\partial x}{\partial s}$ satisfies an evolution equation very similar to the 1-D model:

$$\frac{DY}{Dt} = (M(Y)) Y,$$

where $M(Y)$ is a matrix whose entries are Cauchy integrals on a curve, i.e., a generalization of the 1D Hilbert transform. He further conjectured that a finite time singularity would form from smooth initial data.

In 1991, J.-Y. Chemin [10] proved that in fact the above resemblance is just superficial. The evolution of the vortex patch boundary behaves much better than the 3D Euler equations. Namely, the boundary will remain in $C^{1,\mu}$ if it is started in this function class. In [4], A. Bertozzi and P. Constantin give an alternative proof that is easier to understand. We will present this proof in the next subsection.

7.2 Levelset formulation and global existence

Let $0 < \mu < 1$ and \mathcal{D} be a simply connected, bounded and open subset of the plane whose boundary is $C^{1,\mu}$ smooth, i.e., for any $x^0 \in \partial\mathcal{D}$ there exists a ball $B(x^0; r_0)$ and a $C^{1,\mu}$ function $\varphi : \mathbb{R} \mapsto \mathbb{R}$ such that, after a rotation,

$$\partial\mathcal{D} \cap B(x^0, r_0) = \{x \in B(x^0, r_0) \mid x_2 = \varphi(x_1)\}.$$

Now we introduce the levelset formulation. Let $\varphi \in C^{1,\mu}(\mathbb{R}^2)$ be such that

$$\mathcal{D} = \{x \mid \varphi(x) > 0\}$$

and $|\nabla \varphi| \geq c > 0$ on the boundary. By the implicit function theorem we see that $\partial\mathcal{D}$ defined by $\varphi = 0$ is indeed $C^{1,\mu}$. Thus to establish the long time existence, we only need to show the existence of $C^{1,\mu}$ function $\varphi(x, t)$ such that $\mathcal{D}(t) = \{x \mid \varphi(x, t) > 0\}$ and $\nabla \varphi(x, t)$ is bounded below by $c > 0$ uniformly in t .

It is easy to see that the evolution of $\varphi(x, t)$ should be governed by

$$\varphi_t + u \cdot \nabla \varphi = 0 \tag{7.1}$$

and thus

$$\frac{D}{Dt} \nabla^\perp \varphi \equiv \nabla u \cdot \nabla^\perp \varphi,$$

which looks similar to the 3D Euler equation.

We need to show two things, first $\|\nabla^\perp \varphi\|_{C^{0,\mu}}$ is bounded above, second $|\nabla^\perp \varphi| = |\nabla \varphi|$ is bounded below at $\varphi = 0$.

Proposition 7.1. Let u be the velocity field associated to a vortex patch. Denote

$$\sigma(z) = \begin{pmatrix} \frac{2z_1 z_2}{|z|^2} & \frac{z_2^2 - z_1^2}{|z|^2} \\ \frac{z_2^2 - z_1^2}{|z|^2} & -\frac{2z_1 z_2}{|z|^2} \end{pmatrix}.$$

Then

$$\nabla u(x) = \frac{\omega_0}{2\pi} p v \int_{\mathcal{D}} \frac{\sigma(x-y)}{|x-y|^2} dy + \frac{\omega_0}{2} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \chi_{\mathcal{D}}(x).$$

Proof. The proof is straightforward, similar to those in Chapter 4. \square

First we need to notice some properties of $\sigma(x-y)$.

1. It is smooth outside of the origin and homogeneous of degree 0.
2. It is symmetric with respect to reflection about the origin, i.e., $\sigma(z) = \sigma(-z)$.
3. It has mean 0 on the unit circle.
4. By (2) and (3), it has mean 0 on any half circle centered at 0.

By (1) the kernel in the integral is a singular integral kernel. But one important difference with the 3D Euler or other model equations (1D Constantin-Lax-Majda Model, 2D QG) is that, this singular integral kernel is acting on a characteristic function instead of $\nabla^\perp \varphi$, thus it can be expected to behave much better than the 3D Euler equation.

To see this point, we consider a naïve approach. Instead of the technical $C^{1,\gamma}$, suppose we would like to prove that the level set equation (7.1) is well-posed in C^1 . For this purpose, it is enough to prove that $\|\nabla u\|_{L^\infty}$ is bounded. We have

$$\nabla u(x) = \frac{\omega_0}{2\pi} p v \int_{\mathcal{D}} \frac{\sigma(x-y)}{|x-y|^2} dy + \frac{\omega_0}{2} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \chi_{\mathcal{D}}(x).$$

So it is enough to prove that

$$p v \int_{\mathcal{D}} \frac{\sigma(x-y)}{|x-y|^2} dy$$

remains bounded for all x . Suppose that we only need to worry about the integral

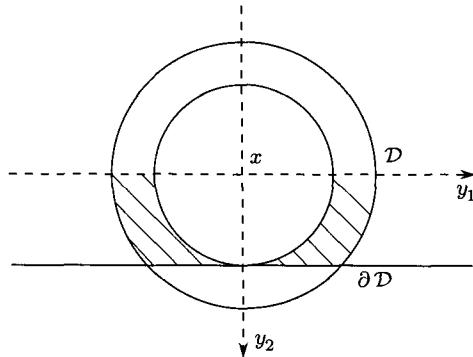
$$I(x) \equiv p v \int_{\mathcal{D} \cap B(x, \delta)} \frac{\sigma(x-y)}{|x-y|^2} dy$$

for some $\delta > 0$. Obviously when $d(x) \equiv \text{dist}(x, \partial\mathcal{D}) \geq \delta$, $I(x) = 0$. On the other hand, when $d(x) < \delta$, we need subtle cancellations. To get

some insight, assume that locally $\partial\mathcal{D}$ is $x_2 = 0$, and $x = (0, x_2) \in \mathcal{D}$ with $\delta > x_2 > 0$. By the properties of σ , we see that σ has mean 0 on semi-circles. This implies that,

$$I(x) = \int_{\mathcal{D}_{\text{eff}}} \frac{\sigma(x - y)}{|x - y|^2} dy,$$

where $\mathcal{D}_{\text{eff}} \equiv \mathcal{D} \cap (B(x, \delta) \setminus B(x, d(x)) \cap \{0 < y_2 < d(x)\})$ is illustrated in the following figure by the shaded area:



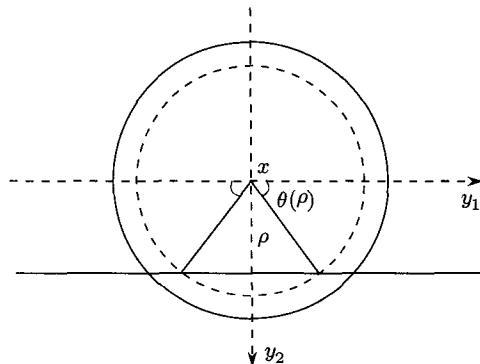
Using Polar coordinates, we have

$$|I(x)| \leq 2 \int_{d(x)}^{\delta} \frac{1}{\rho^2} \theta(\rho) \rho d\rho,$$

where $\theta(\rho)$ is the size of the angle interval corresponding to the curve

$$\{y = (y_1, y_2) \mid |y - x| = \rho, 0 < y_1, 0 < y_2 < d(x)\}.$$

See the following figure.



By the inequality

$$\arcsin t \leq \frac{\pi}{2} t$$

for $t \in [0, 1]$, we have

$$t \leq \sin \frac{\pi}{2} t \leq \frac{\pi}{2} \sin t,$$

which implies

$$\theta(\rho) \leq \frac{\pi}{2} \frac{d(x)}{\rho}.$$

Now it is easy to see that

$$|I(x)| \leq C \int_{d(x)}^{\delta} \frac{d(x)}{\rho^2} d\rho \leq C \left(1 - \frac{d(x)}{\delta} \right) \leq C$$

is bounded. Thus $\|\nabla u\|_{L^\infty}$ is bounded and φ stays in C^1 .

The above “proof” is easy, but there are several un-bridgeable gaps in the argument. The major one is the following. Recall that we assumed $\partial\mathcal{D}$ to be straight when estimating the integral. In fact it can be at most as smooth as φ , i.e., C^1 , and our argument breaks down when the boundary is only C^1 . It turns out that, to get a good estimate on ∇u , we need the boundary to be at least $C^{1,\mu}$ with some $\mu > 0$. But then we need to prove that φ stays in $C^{1,\mu}$ instead of C^1 , which means that it is not enough to estimate $\|\nabla u\|_{L^\infty}$. Thus the real proof is much more complicated although the main idea is the same as the one presented above. Now we turn to the real proof.

The next Proposition is very important.

Proposition 7.2. We have

$$\nabla u(x) \nabla^\perp \varphi(x) = \frac{\omega_0}{2\pi} p v \int_{\mathcal{D}} \frac{\sigma(x-y)}{|x-y|^2} (\nabla^\perp \varphi(x) - \nabla^\perp \varphi(y)) dy.$$

Proof. First we observe that

$$\frac{\sigma(z)}{|z|^2} = \nabla (\nabla^\perp \log |z|).$$

Thus

$$\left(\frac{\sigma(x-y)}{|x-y|^2} \cdot \nabla^\perp \varphi(y) \right)_i = \nabla \cdot ((\nabla^\perp \log |z|)_i \cdot \nabla^\perp \varphi(y)).$$

Now if we consider the i -th component of the integral and omit the subscript i ,

$$\begin{aligned}
& pv \int_{\mathcal{D}} \frac{\sigma(x-y)}{|x-y|^2} \nabla^{\perp} \varphi(y) dy \\
&= \lim_{\delta \rightarrow 0} \int_{\mathcal{D} \cap \{|x-y| \geq \delta\}} \nabla \nabla^{\perp} \log |x-y| \cdot \nabla^{\perp} \varphi(y) dy \\
&= \lim_{\delta \rightarrow 0} \int_{\mathcal{D} \cap \{|x-y| \geq \delta\}} \nabla \cdot (\nabla^{\perp} \log |x-y| \cdot \nabla^{\perp} \varphi(y)) dy \\
&= - \lim_{\delta \rightarrow 0} \int_{\mathcal{D} \cap \{|x-y|=\delta\}} \frac{(x-y)^{\perp}}{|x-y|^2} \left(\nabla^{\perp} \varphi(y) \cdot \frac{x-y}{\delta} \right) dS(y) \\
&= -\pi \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \chi_{\mathcal{D}}(x) \nabla^{\perp} \varphi(x).
\end{aligned}$$

Then the proposition is straightforward. \square

We denote

$$|f|_{\mu} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^{\mu}}$$

to be the $C^{0,\mu}$ semi-norm.

Proposition 7.3. There exists a constant $C = C(\mu)$ such that

$$|\nabla u(\cdot) \nabla^{\perp} \varphi(\cdot)|_{\mu} \leq C (1 + \|\nabla u\|_{L^\infty}) |\nabla \varphi|_{\mu}.$$

Proof. Let $x, h \in \mathbb{R}^2$. We estimate

$$\begin{aligned}
& \frac{2\pi}{\omega_0} |(\nabla u \cdot \nabla^{\perp} \varphi)(x+h) - (\nabla u \cdot \nabla^{\perp} \varphi)(x)| \\
&\leq \left| pv \int_{\mathcal{D}} \frac{\sigma(x+h-y)}{|x+h-y|^2} (\nabla^{\perp} \varphi(x+h) - \nabla^{\perp} \varphi(y)) dy \right| \\
&\quad + \left| pv \int_{\mathcal{D}} \frac{\sigma(x-y)}{|x-y|^2} (\nabla^{\perp} \varphi(x) - \nabla^{\perp} \varphi(y)) dy \right| \\
&\leq \left| pv \int_{\mathcal{D} \cap \{|x-y| \leq 2|h|\}} \frac{\sigma(x+h-y)}{|x+h-y|^2} (\nabla^{\perp} \varphi(x+h) - \nabla^{\perp} \varphi(y)) dy \right|
\end{aligned}$$

$$\begin{aligned}
& + \left| \int_{D \cap \{|x-y| > 2|h|\}} \frac{\sigma(x+h-y)}{|x+h-y|^2} (\nabla^\perp \varphi(x+y) - \nabla^\perp \varphi(y)) dy \right| \\
& + \left| p v \int_{D \cap \{|x-y| \leq 2|h|\}} \frac{\sigma(x-y)}{|x-y|^2} (\nabla^\perp \varphi(x) - \nabla^\perp \varphi(y)) dy \right| \\
& + \left| \int_{D \cap \{|x-y| > 2|h|\}} \left(\frac{\sigma(x-y)}{|x-y|^2} - \frac{\sigma(x+h-y)}{|x+h-y|^2} \right) \right. \\
& \quad \left. (\nabla^\perp \varphi(x+h) - \nabla^\perp \varphi(y)) dy \right| \\
& \equiv I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

We estimate them one by one.

For I_1 , we use the fact that $\varphi \in C^{1,\mu}$ and get

$$I_1 \leq C |h|^\mu |\nabla \varphi|_\mu.$$

For I_2 , by Cotlar's lemma, we have

$$I_2 \leq C (\|\nabla u\|_\infty + 1).$$

For I_3 , Similar to I_1 , we have

$$I_3 \leq C |h|^\mu |\nabla \varphi|_\mu.$$

Lastly, for I_4 , by the mean value theorem, we have

$$I_4 \leq \int_{D \cap \{|x-y| \geq 2|h|\}} |h| \frac{C}{|x-y|^3} |x-y|^\mu |\nabla \varphi|_\mu \leq C |h|^\mu |\nabla \varphi|_\mu. \quad \square$$

Remark 7.4. Cotlar's lemma is the following result:

For any singular integral kernel $K(x)$, define

$$K^\varepsilon(x) \equiv \begin{cases} 0, & |x| \leq \varepsilon, \\ K(x), & |x| > \varepsilon. \end{cases}$$

Then there is a constant $C > 0$, such that for any $\varepsilon > 0$,

$$|K^\varepsilon * f|(x) \leq C (M(K * f)(x) + M(f)(x)),$$

where $M(f)$ denotes the maximal function of f .

$$M(f) \equiv \sup_{r>0} \frac{1}{|B(x,r)|} \int_{B(x,r)} f(y) dy.$$

Thus in particular, if both $K * f$ and f are in L^∞ , then we can replace $M(K * f)(x)$ by $\|K * f\|_{L^\infty}$ and $M(f)$ by $\|f\|_{L^\infty}$. For more about Cotlar's lemma, see e.g. Section 1.7 of Stein [45] or Chapter 7 of Meyer-Coifman [37].

Our next task is to give a upper bound for $\|\nabla u\|_{L^\infty}$. Denote the infimum norm of a function f on $\partial\mathcal{D}$ by

$$|f|_{inf} = \inf_{x \in \partial\mathcal{D}} |\nabla \varphi(x)|.$$

Proposition 7.5. Let u be the velocity and φ be a solution to (7.1). Then there is a constant $C = C(\mu) > 0$ such that

$$\|\nabla u\|_{L^\infty} \leq C |\omega_0| \left(1 + \log \left(\frac{|\nabla \varphi|_\mu}{|\nabla \varphi|_{inf}} \right) \right).$$

Proof. First note that we only need to estimate the principal integral

$$pv \int_{\mathcal{D}} \frac{\sigma(x-y)}{|x-y|^2} dy.$$

Denote

$$\delta = \frac{|\nabla \varphi|_{inf}}{|\nabla \varphi|_\mu}$$

and $d(x) = dist(x, \partial\mathcal{D})$ for any $x \in \mathbb{R}^2$. Intuitively, the main difficulty would come from near the boundary.

First we assume $d(x) \geq \delta$. Take η small enough, we have

$$\begin{aligned} \left| \int_{\mathcal{D} \cap \{|x-y| \geq \eta\}} \frac{\sigma(x-y)}{|x-y|^2} dy \right| &\leq \left| \int_{\mathcal{D} \cap \{\eta \leq |x-y| \leq d(x)\}} \frac{\sigma(x-y)}{|x-y|^2} dy \right| \\ &\quad + \left| \int_{\mathcal{D} \cap \{|x-y| > d(x)\}} \frac{\sigma(x-y)}{|x-y|^2} dy \right| \\ &= \left| \int_{\mathcal{D} \cap \{|x-y| > d(x)\}} \frac{\sigma(x-y)}{|x-y|^2} dy \right| \\ &\leq \left| \int_{\mathcal{D} \cap \{|x-y| > d(x)\}} \frac{1}{|x-y|^2} dy \right| \\ &\leq \left| \int_{d(x) < |x-y| \leq R} \frac{1}{|x-y|^2} dy \right|, \end{aligned}$$

where πR^2 is the area of \mathcal{D} , which is conserved by the incompressibility of the flow. The last inequality can be readily checked by using Polar

coordinates and the fact that $\frac{1}{r^2}$ is monotonically decreasing in r . The proof is left as an exercise. Now it is easy to see that the integral is bounded by what we want.

Now for the case $d(x) < \delta$. Again taking η small enough, we have

$$\begin{aligned} \left| \int_{\mathcal{D} \cap \{|x-y| \geq \eta\}} \frac{\sigma(x-y)}{|x-y|^2} dy \right| &\leq \left| \int_{\mathcal{D} \cap \{\eta \leq |x-y| \leq d(x)\}} \frac{\sigma(x-y)}{|x-y|^2} dy \right| \\ &\quad + \left| \int_{\mathcal{D} \cap \{d(x) \leq |x-y| < \delta\}} \frac{\sigma(x-y)}{|x-y|^2} dy \right| \\ &\quad + \left| \int_{\mathcal{D} \cap \{|x-y| \geq \delta\}} \frac{\sigma(x-y)}{|x-y|^2} dy \right|. \end{aligned}$$

We know that the first integral vanishes due to symmetry, and the third term can be estimated as in the $d(x) \leq \delta$ case.

For the second one, we denote

$$S = \{d(x) \leq |x-y| \leq \delta\}$$

and study its special geometrical properties. The heuristic is the following. Assume that the boundary is a straight line, then we try to bound the integral by estimating the area of the integration in which the integral doesn't vanish. This is where the regularity of the boundary comes into play. To make the above idea rigorous, we denote by \tilde{x} the point on $\partial\mathcal{D}$ such that $d(x, \tilde{x}) = d(x)$. Let \mathcal{L} be the line through x in the direction that is tangent to $\partial\mathcal{D}$ at \tilde{x} . Then the annulus $\{d(x) \leq |x-y| \leq \delta\}$ is divided into two half annuli. Denote the one containing \tilde{x} by A_s and the other by A_l . First note that the integration on A_l vanishes. So

$$\begin{aligned} \left| \int_{S \cap \mathcal{D}} \frac{\sigma(x-y)}{|x-y|^2} dy \right| &= \left| \int_{(A_s \cap \mathcal{D}) \cup (A_l \cap \mathcal{D}^c)} \frac{\sigma(x-y)}{|x-y|^2} dy \right| \\ &\leq \int_{(A_s \cap \mathcal{D}) \cup (A_l \cap \mathcal{D}^c)} \frac{C}{|x-y|^2} dy. \end{aligned}$$

Note that S should more and more resembles a half-annulus as $d(x) \rightarrow 0$. So our integral should vanish. We estimate the area of $S_\epsilon \equiv (A_s \cap \mathcal{D}) \cup (A_l \cap \mathcal{D}^c)$. Write it in polar coordinates and denote by $H(E_\rho)$ the 1-D Hausdorff measure of

$$\{\theta \in (0, 2\pi] \mid (\rho, \theta) \in S_\epsilon\}$$

for $d(x) \leq \rho \leq \delta$. By the Geometric lemma that will be proved later,

$$H(E_\rho) \leq C \left(\frac{d(x)}{\rho} + \left(\frac{\rho}{\delta} \right)^\mu \right)$$

and then the result is straightforward.

Now we prove the Geometric Lemma.

Lemma 7.6 (Geometric Lemma). *We have*

$$H(E_\rho) \leq 2\pi \left[(1 + 2^\mu) \frac{d(x_0)}{\rho} + 2^\mu \left(\frac{\rho}{\delta} \right)^\mu \right]$$

for all $\rho \geq d(x_0)$, $1 > \mu > 0$ and x_0 so that

$$d(x_0) < \delta = \left(|\nabla \phi|_{inf} / |\nabla \varphi|_\mu \right)^{1/\mu}.$$

Proof. Let

$$\begin{aligned} S_\rho(x_0) &= \{z \mid |z| = 1, x = x_0 + \rho z \in \mathcal{D}\}, \\ \Sigma(x_0) &= \{z \mid |z| = 1, \nabla_x \varphi(\tilde{x}) \cdot z \geq 0\}, \end{aligned}$$

where $\tilde{x} \in \partial\mathcal{D}$ such that $|x_0 - \tilde{x}| = d(x_0)$. This point exists since the boundary is $C^{1,\mu}$. Then we have

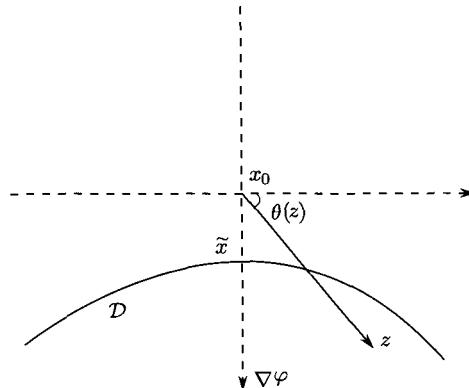
$$E_\rho = [S_\rho \setminus \Sigma_\rho] \cup [\Sigma_\rho \setminus S_\rho].$$

The readers should draw a picture to see what E_ρ looks like (there are two cases, $x_0 \in \mathcal{D}$ and $x_0 \notin \mathcal{D}$). Note that since $\varphi(x) > 0$ for $x \in \mathcal{D}$, the direction of $\nabla \varphi$ at \tilde{x} should be pointing inward instead of outward.

We use polar coordinates and denote the angle for a point z in E_ρ by $\theta(z)$, with $\theta(z)$ defined by

$$\sin \theta(z) = \frac{\nabla \varphi(\tilde{x}) \cdot z}{|\nabla \varphi(\tilde{x})| \cdot |z|}.$$

See the following illustration.



Thus we have

$$\sin \theta(z) = \frac{\nabla \varphi(\tilde{x}) \cdot (\tilde{x} - x_0)}{|\nabla \varphi(\tilde{x})| \rho} + \frac{\nabla \varphi(\tilde{x}) \cdot (x_0 + \rho z - \tilde{x})}{|\nabla \varphi(\tilde{x})| \rho}.$$

Now in the RHS z is in the unit circle.

For any $z \in E_\rho(x_0)$, we can see that either $\sin \theta(z) > 0$ and $\varphi(x_0 + \rho z) < 0$ or $\sin \theta(z) < 0$ and $\varphi(x_0 + \rho z) > 0$. In either case, noting that $\varphi(\tilde{x}) = 0$ and $\nabla \varphi(\tilde{x}) \parallel (x_0 - \tilde{x})$, we have

$$|\sin \theta(z)| \leq \frac{d(x_0)}{\rho} + \left| \frac{\nabla \varphi(\tilde{x}) \cdot (x_0 + \rho z - \tilde{x})}{|\nabla \varphi(\tilde{x})| \rho} - \frac{\varphi(x_0 + \rho z) - \varphi(\tilde{x})}{|\nabla \varphi(\tilde{x})| \rho} \right|.$$

Since $-\frac{\varphi(x_0 + \rho z)}{|\nabla \varphi(\tilde{x})| \rho}$ is always of the same sign as $\sin \theta(z)$, so adding it will only increase the absolute value. Now by the mean value theorem we have

$$|\varphi(x) - \varphi(y) - \nabla \varphi(y) \cdot (x - y)| \leq |\nabla \varphi|_\mu |x - y|^{1+\mu},$$

which gives

$$\begin{aligned} |\sin \theta(z)| &\leq \frac{d(x_0)}{\rho} + \frac{|\nabla \varphi|_\mu |x_0 + \rho z - \tilde{x}|^{1+\mu}}{\rho |\nabla \varphi|_{inf}} \\ &\leq \frac{d(x_0)}{\rho} + \frac{|\nabla \varphi|_\mu}{\rho |\nabla \varphi|_{inf}} [d(x_0) + \rho]^{1+\mu} \\ &\leq \frac{d(x_0)}{\rho} + 2^\gamma \frac{|\nabla \varphi|_\mu}{\rho |\nabla \varphi|_{inf}} [d(x_0)^{1+\mu} + \rho^{1+\mu}], \end{aligned}$$

where the last inequality comes from the Jensen's inequality applying to the convex function $x^{1+\mu}$ for positive x .

Now the estimate is easy to see by the fact that $\arcsin t \leq \frac{\pi}{2}t$ for $t \in [0, 1]$. Since we are estimating the absolute value of $\sin \theta$ over $[0, 2\pi]$, the factor should be $\frac{\pi}{2} \cdot 4 = 2\pi$. This completes the proof. \square

Finally we take the dynamics into account.

Proposition 7.7. If the initial data $\varphi_0 \in C^{1,\mu}(\mathbb{R}^2)$, such that $\mathcal{D}_0 = \{\varphi_0(x) > 0\}$ is simply connected and bounded. And $|\nabla \varphi_0| \geq C > 0$ on the boundary $\partial \mathcal{D}_0$, then the following priori estimates holds:

1. $\|\nabla \varphi(\cdot, t)\|_{L^\infty} \leq \|\nabla \varphi_0\|_{L^\infty} \exp \left(\int_0^t \|\nabla u(\cdot, s)\|_{L^\infty} ds \right);$
2. $|\nabla \varphi(\cdot, t)|_{inf} \geq |\nabla \varphi_0|_{inf} \exp \left(- \int_0^t \|\nabla u(\cdot, s)\|_{L^\infty} ds \right);$
3. $|\nabla \varphi(\cdot, t)|_\mu \leq |\nabla \varphi_0|_\mu \exp \left((C_0 + \mu) \int_0^t \|\nabla u(\cdot, s)\|_{L^\infty} ds \right).$

Proof. Let $X = X(\alpha, t)$ denote a particle trajectory and

$$Y(\alpha, t) = \nabla^\perp \varphi(X(\alpha, t), t).$$

Then we have

$$\frac{d}{dt} Y(\alpha, t) = \nabla u(X(\alpha, t), t) Y(\alpha, t)$$

and therefore

$$\left| \frac{d}{dt} \log |Y(\alpha, t)| \right| \leq \|Du(\cdot, t)\|_{L^\infty}.$$

Now by Gronwall's lemma we have

$$e^{-\int_0^t \|\nabla u\|_{L^\infty} ds} \leq \frac{|Y(\alpha, t)|}{|\nabla^\perp \varphi_0(\alpha)|} \leq e^{\int_0^t \|\nabla u\|_{L^\infty} ds},$$

which proves both (1) and (2).

For (3), we write the integral formulation of the equation for $\nabla^\perp \varphi$:

$$\nabla^\perp \varphi(x, t) = \nabla^\perp \varphi_0(X(x, -t)) + \int_0^t (\nabla u \nabla^\perp \varphi)(X(x, s-t), s) ds.$$

And we estimate

$$\begin{aligned} & |\nabla^\perp \varphi(x+h, t) - \nabla^\perp \varphi(x, t)| \\ & \leq |\nabla^\perp \varphi_0(X(x+h, -t)) - \nabla^\perp \varphi_0(X(x, -t))| \\ & + \left| \int_0^t ((\nabla u \nabla^\perp \varphi)(X(x+h, s-t), s) - (\nabla u \nabla^\perp \varphi)(X(x, s-t), s)) ds \right| \\ & \leq |\nabla^\perp \varphi_0|_\mu \|\nabla X(\cdot, -t)\|_{L^\infty}^\mu |h|^\mu \\ & + \int_0^t |\nabla u \nabla^\perp \varphi(\cdot, s)|_\mu \|\nabla X(\cdot, s-t)\|_{L^\infty}^\mu |h|^\mu ds. \end{aligned}$$

For the evolution of ∇X , we have

$$\frac{d}{dt} \nabla X(z, -t) = -\nabla u(X(z, -t), -t) \nabla X(z, -t).$$

Now by Gronwall's lemma we have

$$\|\nabla X(\cdot, s-t)\|_{L^\infty} \leq \exp \left(\int_s^t \|\nabla u(\cdot, s')\|_{L^\infty} ds' \right).$$

Plug it into the inequality above, we get the estimate in (3). \square

Finally we put everything together, and obtain the following theorem:

Theorem 7.8. *Given $\omega_0 \neq 0$, \mathcal{D}_0 a simply connected, bounded, $C^{1,\mu}$ smooth domain with $0 < \mu < 1$, and a function $\varphi_0 \in C^{1,\mu}(\mathbb{R}^2)$ such that $\mathcal{D}_0 = \{\varphi_0 > 0\}$, $|\nabla \varphi_0|_{inf} \geq C > 0$, then the solution φ belongs to $C^{1,\mu}$ for all time. Furthermore, there exists a constant $C > 0$, which depends only on the initial data such that*

1. $\|\nabla u(\cdot, t)\|_{L^\infty} \leq \|\nabla u_0\|_{L^\infty} e^{Ct}$,
2. $|\nabla \varphi(\cdot, t)|_\mu \leq |\nabla \varphi_0|_\mu \exp((C_0 + \mu) e^{Ct})$,
3. $\|\nabla \varphi(\cdot, t)\|_{L^\infty} \leq \|\nabla \varphi_0\|_{L^\infty} \exp(e^{Ct})$,
4. $|\nabla \varphi(\cdot, t)|_{inf} \geq |\nabla \varphi_0|_{inf} \exp(-e^{Ct})$.

Proof. We have

$$\log |\nabla \varphi|_{inf} \geq \log |\nabla \varphi_0|_{inf} - C \int_0^t \left(1 + \log \frac{|\nabla \varphi|_\mu}{|\nabla \varphi|_{inf}} \right) ds$$

after taking logarithm on both sides of estimate (2) in Proposition 5.2.7. Similarly we obtain

$$\log |\nabla \varphi|_\mu \leq \log |\nabla \varphi_0|_\mu + (C_0 + \mu) \int_0^t \left(1 + \log \frac{|\nabla \varphi|_\mu}{|\nabla \varphi|_{inf}} \right) ds.$$

Combining these two, we have

$$\log \frac{|\nabla \varphi|_\mu}{|\nabla \varphi|_{inf}} \leq \log \frac{|\nabla \varphi_0|_\mu}{|\nabla \varphi_0|_{inf}} + (C_0 + \mu + 1) \int_0^t \left(1 + \log \frac{|\nabla \varphi|_\mu}{|\nabla \varphi|_{inf}} \right) ds.$$

By Gronwall's lemma, we easily get

$$\log \frac{|\nabla \varphi|_\mu}{|\nabla \varphi|_{inf}} \leq C e^{Ct}.$$

This also provides a bound for ∇u in (1) from Proposition 5.2.5. The others are straightforward by using the estimate on ∇u given by Property (1). \square

Remark 7.9. The problem of global existence of vortex patch with boundary only C^1 or worse is still open. In [6], J. Carrillo and J. Soler showed numerically that, for initial boundary that is only Lipschitz continuous, the evolution develops cusps from corners.

References

- [1] R.A. Adams, Sobolev Spaces, Academic Press, 1975.
- [2] V.I. Arnold and B. Khesin, Topological Methods in Hydrodynamics, Applied Mathematical Sciences 125, Springer-Verlag, 1998.
- [3] J. Bergh, J. Löfström, Interpolation Spaces: An Introduction, Springer-Verlag, 1976.
- [4] A. Bertozzi and P. Constantin, Global regularity for vortex patches, Comm. Math. Phys., 152 (1993), 19–26.
- [5] R.E. Caflisch and O.F. Orellana, Long Time Existence for a Slightly Perturbed Vortex Sheet, Comm. Pure Appl. Math., 39 (1986), 807–838.
- [6] J.A. Carrillo, J. Soler, On the Evolution of an Angle in a Vortex Patch, Journal of Nonlinear Science, 10 (2000), no. 1, 23–47.
- [7] D. Chae, On the well-posedness of the Euler equations in the Triebel-Lizorkin spaces, Comm. Pure Appl. Math., 55 (2002), 654–678.
- [8] D. Chae, On the Euler equations in the critical Triebel-Lizorkin spaces, Arch. Ration. Mech. Anal., 170 (2003), 185–210.
- [9] D. Chae, Remarks on the blow-up criterion of the three-dimensional Euler equations, Nonlinearity, 18 (2005), 1021–1029.
- [10] J.Y. Chemin, Sur le mouvement des particules d'un fluide parfait incompressible bidimensionnel, Invent. Math., 103 (1991), 599–629.
- [11] A. Chorin, Vorticity and Turbulence, Applied Mathematical Sciences 103, Springer-Verlag, 1994.
- [12] A.J. Chorin, J.E. Marsden, A Mathematical Introduction to Fluid Mechanics, 2nd ed., Texts in Applied Mathematics 4, Springer-Verlag, 1990.
- [13] P. Constantin, Geometric statistics in turbulence, SIAM Review, 36 (1994), 73–98.
- [14] P. Constantin, C. Fefferman and A. Majda, Geometric constraints on potentially singular solutions for the 3-D Euler equations, Comm. PDE, 21 (1996), no. 3&4, 559–571.
- [15] P. Constantin, A.J. Majda and E. Tabak, Formation of strong fronts in the 2-D quasigeostrophic thermal active scalar, Nonlinearity 7 (1994), 1495–1533.
- [16] P. Constantin, Q. Nie and N. Schörghofer, Nonsingular Surface-Quasi-Geostrophic Flow, Phys. Lett. A, 241 (1998), 168–172.

- [17] D. Cordoba, Nonexistence of simple hyperbolic blow-up for the quasi-geostrophic equation, *Annals of Mathematics*, 148 (1998), 1135–1152.
- [18] D. Cordoba and C. Fefferman, Growth of solutions for QG and 2D Euler equations, *J. AMS*, 15 (2002), no. 3, 665–670.
- [19] J. Deng, T.Y. Hou and X. Yu, Geometric Properties and Non-blowup of 3-D Incompressible Euler Flow, *Comm. PDE.*, 30 (2005), no. 1, 225–243.
- [20] J. Deng, T.Y. Hou and X. Yu, A Level Formulation for the 3D Incompressible Euler Equations, *Methods and Applications of Analysis*, 12 (2005), no. 4, 427–440.
- [21] J. Deng, T.Y. Hou and X. Yu, Improved Geometric Conditions for Non-blowup of the 3D Incompressible Euler Equation, *Comm. PDE.*, 31 (2006), no. 2, 293–306.
- [22] J. Deng, T.Y. Hou, R. Li and X. Yu, Level Set Dynamics and Non-blowup of the 2D Quasi-geostrophic Equation, accepted by *Methods and Applications of Analysis*, 2008.
- [23] D.G. Ebin, A.E. Fischer and J.E. Marsden, Diffeomorphism groups, hydrodynamics and relativity. In Vanstone, J., editor, *Proc. of the 13th Biennial Seminar of Canadian Mathematical Congress*, 135–279, Montreal.
- [24] J. Goodman, T.Y. Hou, J. Lowengrub, The Convergence of the Point Vortex Method for the 2-D Euler Equations, *Comm. Pure Appl. Math.*, 43 (1990), 415–430.
- [25] T.Y. Hou, R. Li, Dynamic Depletion of Vortex Stretching and Non-Blowup of the 3-D Incompressible Euler Equations, *J. Nonlinear Science*, 16 (2006), no. 6, 639–664.
- [26] R.M. Kerr, Evidence for a singularity of the three-dimensional incompressible Euler equations, *Phys. Fluids A*, 5 (1993), 1725–1746.
- [27] R.M. Kerr, The role of singularities in Euler, in *Small-scale structure in hydro and magnetohydrodynamic turbulence*, eds. Pouquet, A., Sulem, P. L., Lecture Notes, Springer-Verlag, 1995.
- [28] R.M. Kerr, Euler Singularities and Turbulence, in *19th ICTAM Kyoto' 96*, Eds. Tatsumi, T., Watanabe, E., Kambe, T., Elsevier Science, 1997.
- [29] R.M. Kerr, The Outer Regions in Singular Euler, in *Fundamental Problematic Issues in Turbulence*, eds. Tsinober and Gyr, Birkhäuser, 1998.
- [30] H. Kozono, Y. Taniuchi, Bilinear estimates in BMO and to the Navier-Stokes equations, *Math. Z.*, 235 (2000), 173–194.

- [31] H. Lamb, *Hydrodynamics*, 6th ed., Dover, 1945.
- [32] P.L. Lions, *Mathematical Topics in Fluid Mechanics*, Vol.1. Incompressible Models, Oxford Lecture Series in Mathematics and its Applications 3, Oxford Science Publications, 1996.
- [33] M.C. Lopes Filho, H.J. Nussenzveig Lopes and Y. Zheng, Weak solutions to the equations of incompressible, ideal flow. Text of mini-course for the 22nd Brazilian Colloquium of Mathematics, 1999. Full text at <http://www.ime.unicamp.br/~mlopes/publications.html>.
- [34] A. Majda, Vorticity and the mathematical theory of Incompressible fluid flow, *Comm. Pure. Appl. Math.*, 39 (1986), 187–220.
- [35] A.J. Majda, A.L. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2002.
- [36] C. Marchioro, M. Pulvirenti, *Mathematical Theory of Incompressible Nonviscous Fluids*, Chapter 6. Applied Mathematical Sciences 96. Springer-Verlag, 1994.
- [37] Y. Meyer, R. Coifman, *Wavelets: Calderón-Zygmund operators and multilinear operators*, Cambridge University Press, 1997.
- [38] K. Ohkitani and M. Yamada, Inviscid and inviscid-limit behavior of a surface quasigeostrophic flow, *Phys. Fluids* 9 (1997), no. 4, 876–882.
- [39] R.B. Pelz, Locally self-similar, finite-time collapse in a high-symmetry vortex filament model, *Physical Review E*, 55 (1997), no. 2, 1617–1620.
- [40] R.B. Pelz, Symmetry and the hydrodynamic blow-up problem, *J. Fluid Mech.*, 444 (2001), 299–320.
- [41] J. Pedlosky, *Geophysical Fluid Dynamics*, 345–368, 653–670, Springer-Verlag, 1987.
- [42] F. Planchon, An extension of the Beale-Kato-Majda criterion for the Euler equations, *Comm. Math. Phys.*, 232 (2003), 319–326.
- [43] G. Ponce, Remarks on a paper by J.T. Beale, T. Kato and A. Majda, *Comm. Math. Phys.*, 98 (1985), 349–353.
- [44] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, 1970.
- [45] E. Stein, *Harmonic Analysis*, Princeton University Press, 1993.
- [46] T. Wolff, *Lecture Notes on Harmonic Analysis*, University Lecture Series, 29, AMS, 2003.

Systems of Conservation Laws. Theory, Numerical Approximation and Discrete Shock Profiles*

Denis Serre

*École Normale Supérieure de Lyon,
UMPA (UMR 5669 CNRS)
46, allée d'Italie, F-69364
Lyon, cedex 07, France
E-mail: serre@umpa.ens-lyon.fr*

1 Hyperbolic systems of conservation laws

A first-order system of conservation laws is a system of n partial differential equations in n unknowns u_1, \dots, u_n that are functions of space $x = (x_1, \dots, x_d)$ and time t . It writes

$$\partial_t u_j + \operatorname{div}_x f_j(u) = 0, \quad j = 1, \dots, n,$$

where the *fluxes* f_j^α are given smooth functions over the space of states \mathcal{U} . The latter is in general a convex set of \mathbb{R}^n with a non-void interior.

In the sequel, we shall consider only the case of one space variable ($d = 1$), for which we rewrite

$$\partial_t u + \partial_x f(u) = 0. \quad (1.1)$$

The flux is thus a smooth map $f : \mathcal{U} \rightarrow \mathbb{R}^n$.

A typical example is gas dynamics, which writes

$$\begin{aligned} \partial_t \rho + \partial_x(\rho v) &= 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2 + p(\rho, e)) &= 0, \\ \partial_t \left(\frac{1}{2} \rho v^2 + \rho e \right) + \partial_x \left(\left(\frac{1}{2} \rho v^2 + \rho e + p \right) v \right) &= 0. \end{aligned} \quad (1.2)$$

Hereabove, ρ, v, e denote the mass density, the velocity and the specific internal energy. The pressure p is determined through an equation of

*The author thanks the organizers and Fudan University for their warm hospitality.

state $(\rho, e) \mapsto p(\rho, e)$ that characterizes the nature of the gas. For instance $p = \frac{2}{5}\rho e$ is an acceptable relation for the air in ordinary conditions. The state u has components

$$u_1 = \rho, \quad u_2 = \rho v, \quad u_3 = \frac{1}{2}\rho v^2 + \rho e.$$

The domain \mathcal{U} is defined by

$$u_1 \geq 0, \quad u_1 u_3 \geq \frac{1}{2}u_2^2.$$

About the terminology. The words ‘conservation laws’ are justified by the fact that if one has a reasonable solution, say a field of bounded variations satisfying (1.1) in the distributional sense, then one has

$$\frac{d}{dt} \int_{x_1}^{x_2} u(x, t) dx + f(u(x_2, t)) - f(u(x_1, t)) = 0, \quad \text{a.e. } t > 0.$$

In particular, if u has constant limits u_{\pm} as $x \rightarrow \pm\infty$, then

$$\int_{\mathbb{R}} (u(x, t) - a(x)) dx + t(f(u_+) - f(u_-)) = 0.$$

The terminology thus comes from the case where a tends to a constant state \bar{u} at infinity, for which we have^①

$$\int_{\mathbb{R}} (u(x, t) - \bar{u}) dx = \int_{\mathbb{R}} (a(x) - \bar{u}) dx. \quad (1.3)$$

If $\bar{u} = 0$, the total mass is thus *conserved*.

1.1 The Cauchy problem: classical solutions

The Cauchy problem consists in solving (1.1) in $(0, T) \times \mathbb{R}$ under the condition (initial data) that u is prescribed at initial time:

$$u(0, x) = a(x), \quad (1.4)$$

where $a : \mathbb{R} \rightarrow \mathcal{U}$ is a given function with either smoothness or integrability properties.

1.1.1 Hyperbolicity

Let us denote $A(u) := Df(u)$ the Jacobian matrix of f . We say that the system (1.1) is *hyperbolic* if $A(u)$ is diagonalisable with real eigenvalues. For a linear system ($f(u) = Au$), this is the condition under which the Cauchy problem is well-posed in Sobolev spaces $H^s(\mathbb{R})$. We shall see that the situation is more intricate in the nonlinear case.

^①When the Cauchy problem for a first-order system (1.1) is well-posed, the values at spatial infinity do not change with time.

Exercise. Compute the Jacobian matrix for gas dynamics. Show that the system (1.2) is hyperbolic if, and only if, the equation of state $p = p(\rho, e)$ satisfies

$$p \frac{\partial p}{\partial e} + \rho^2 \frac{\partial p}{\partial \rho} > 0.$$

The basic example of a hyperbolic system is that of a *scalar* equation, which means that $n = 1$. Then $A(u) = f'(u)$ is 1×1 , thus diagonal. The best-known scalar equation is that of Burgers:

$$\partial_t u + \partial_x (u^2/2) = 0. \quad (1.5)$$

Hyperbolic systems cover a wide variety of applications:

- Compressible fluid dynamics,
- Electromagnetism (Maxwell's equations),
- Magnetohydrodynamics,
- Electrophoresis,
- Chromatography,
- Traffic flow,
- Elastodynamics,
- Singular limit of dispersive waves,
- Einstein equation of general relativity,
-,

as long as the diffusive or dispersive effects can be neglected.

1.1.2 Entropies

When a system (1.1) has a physical meaning, it is often compatible with an additional scalar conservation law

$$\partial_t \eta(u) + \partial_x q(u) = 0, \quad (1.6)$$

where the functions $\eta, q : \mathcal{U} \rightarrow \mathbb{R}$ are smooth and η is strongly convex, in the sense that its Hessian matrix $D^2\eta(u)$ at u is positive definite for every $u \in \mathcal{U}$. We say that η is a *convex entropy* and that q is its *entropy flux*.

The compatibility between (1.6) and (1.1) means that for every differentiable field $u : \mathbb{R} \rightarrow \mathcal{U}$, one has

$$\partial_t \eta(u) + \partial_x q(u) = d\eta(u) \cdot (\partial_t u + \partial_x f(u)).$$

In other words, an entropy-entropy flux pair is a solution of the linear differential system

$$dq(u) = d\eta(u)A(u), \quad (1.7)$$

which can be written componentwise as

$$\frac{\partial q}{\partial u_i} = \sum_{j=1}^n \frac{\partial \eta}{\partial u_j} \frac{\partial f_j}{\partial u_i}, \quad i = 1, \dots, n.$$

Differentiating (1.7), we find that $A(u)$ is self-adjoint with respect to the scalar product induced by $D^2\eta(u)$:

$$\langle D^2\eta(u)X, A(u)Y \rangle = \langle D^2\eta(u)Y, A(u)X \rangle, \quad X, Y \in \mathbb{R}^n.$$

This identity is at the basis of the symmetrization result of Godunov and Friedrichs: there exist two symmetric matrices $S_0(u), S(u)$, depending smoothly on u , with S_0 positive definite, such that the system (1.1) rewrites

$$S_0(u)\partial_t u + S(u)\partial_x u = 0.$$

We say that (1.1) is *symmetrizable* in Friedrichs sense.

Warning. Our terminology “entropy” differs from that employed in Physics. For gas dynamics, our entropy will be $\eta = -\rho s$ with $s = s(\rho, e)$ the physical entropy. In particular η is convex in u if, and only if, s is a concave function of the quantities

$$\frac{1}{\rho}, \quad v, \quad \frac{1}{2}v^2 + e.$$

Exercise. Show that $\eta = -\rho s(\rho, e)$ is an entropy of gas dynamics whenever s is a solution of the transport equation

$$p \frac{\partial s}{\partial e} + \rho^2 \frac{\partial s}{\partial \rho} = 0.$$

1.1.3 Local well-posedness in $H^s(\mathbb{R}^d)$

The best result for classical solutions of the Cauchy problem is stated in Sobolev spaces. More precisely, we ask the derivatives to be in a Sobolev space, thus allowing the data and the solution to be non-zero at infinity. Because we deal with nonlinear systems, the well-posedness in Sobolev spaces H^s requires that s be large enough.

Theorem 1.1. *We assume that the system (1.1) is symmetrizable in Friedrichs sense.*

Let $s > 3/2$ be a real number, and let K be a compact subset of \mathcal{U} . Let $a : \mathbb{R} \rightarrow K$ be an initial data such that $a' \in H^{s-1}(\mathbb{R})^n$.

Then there exists a time $T > 0$, and a unique classical solution u of the Cauchy problem (1.1,1.4) on $(0, T) \times \mathbb{R}$. The solution has the property that

$$\partial_t u, \partial_x u \in L^\infty(0, T; H^{s-1}).$$

The theorem above has a counterpart in several space dimensions, but with $\nabla_x a \in H^{s-1}(\mathbb{R}^d)$ and $s > 1 + d/2$. This condition and the Sobolev embedding ensure that the data and the solution are of class C^1 . Whence the terminology of “classical solution”. For a full proof, see either of [4, 7, 23].

We warn the reader that the existence time T of a classical solution depends on the data a . Thus the theorem cannot be used repeatedly to build a global solution.

This theorem applies in particular to systems (1.1) endowed with a strongly convex entropies, since such systems are symmetrizable.

1.2 The Cauchy problem: weak solutions

We show now that we may not expect a global-in-time classical solution of the Cauchy problem, for general data. We thus introduce a weakened notion of solution, which refers to the theory of distributions. This is coherent with the physical meaning of the equations. We then explain why the physically relevant solutions should display an irreversibility property. This is reminiscent to the second principle of thermodynamics.

1.2.1 Break-down of smooth solutions

Let us consider for instance the Burgers equation. There are at least two ways to see that the classical solution breaks down in finite time. For this to happen, we only need that a' takes a negative value somewhere.

The method of characteristics. Given a base point $x_0 \in \mathbb{R}$ we define a characteristic curve by solving the ODE

$$\frac{dx}{dt} = u(x, t), \quad x(0) = x_0.$$

An elementary calculus gives the following results (**Exercise:** fill the details). The characteristic curve is the straight line $t \mapsto x_0 + ta(x_0)$, on which $u \equiv a(x_0)$.

Assume that a is not non-decreasing. There exist two points x_0, y_0 such that $(y_0 - x_0)(a(y_0) - a(x_0)) < 0$. The intersection (x, t) of the characteristics issued from x_0 and y_0 occurs at some point (x, t) with $t > 0$. We then have the contradiction $a(x_0) = u(x, t) = a(y_0)$.

Blow-up of derivatives. We can also calculate the x -derivative $v = \partial_x u$ along a characteristics. Differentiating (1.5), we have

$$\frac{dv}{dt} = (\partial_t + u\partial_x)v = -v^2.$$

This is a Riccati equation, of which the solution blows up at a positive time if the initial data $v(0) = a'(x_0)$ is negative.

Exercise. Prove that the maximal classical solution is defined on the strip $(0, T^*) \times \mathbb{R}$, where $T^* = +\infty$ if a is non-decreasing, and

$$-\frac{1}{T^*} = \inf_{x \in \mathbb{R}} a'(x)$$

otherwise.

1.2.2 Weak solutions

Since classical solutions are not global in general, although gas does flow ..., we need to accept solutions with less regularity. Typically, we consider fields $u(x, t)$ that are bounded measurable. Therefore $f(u)$ is also bounded and measurable, and the derivatives $\partial_t u$ and $\partial_x f(u)$ make sense, at least as distributions. The system (1.1) has to be rewritten by introducing test functions: We say that u is a *weak solution* of (1.1) over a domain $\Omega \in \mathbb{R} \times (0, +\infty)$ if there holds

$$\int_{\Omega} (u_j \partial_t \phi + f_j(u) \partial_x \phi) dx dt = 0, \quad (1.8)$$

for every $j = 1, \dots, n$ and every test function $\phi \in \mathcal{D}(\Omega)$.

For the Cauchy problem, we have a refined definition:

Definition 1.2. We say that u is a *weak solution* of (1.1, 1.4) over a strip $(0, T) \times \mathbb{R}$ if there holds

$$\int_0^T \int_{\mathbb{R}} (u_j \partial_t \phi + f_j(u) \partial_x \phi) dx dt + \int_{\mathbb{R}} a(x) \phi(x, 0) dx = 0, \quad (1.9)$$

for every $j = 1, \dots, n$ and every test function $\phi \in \mathcal{D}(\mathbb{R} \times (-\infty, T))$.

It is clear, from integration by parts, that a classical solution is also a solution in this weak sense. Conversely, a weak solution of class C^1 is also a classical solution. It is not hard to prove a little bit more, that a continuous field u that is piecewise C^1 , is a classical solution if, and only if, it is a weak solution. We can view (1.9) as the most natural way to express the conservation laws, that is the underlying physical principles, for non-smooth flows.

1.2.3 The Rankine-Hugoniot condition

Immediately next to the piecewise- C^1 fields come the piecewise continuous ones. Thus let us consider a domain Ω , divided into two pieces Ω_{\pm} by a smooth curve Γ , and a weak solution of (1.1) u , such that its restrictions to each Ω_{\pm} is of class C^1 up to Γ . Each of these restrictions has limits along Γ , which we denote by u_{\pm} . When g is a function over \mathcal{U} , we define

$$[g(u)] := g(u_+) - g(u_-),$$

the *jump* of $g(u)$ across Γ . In the calculation below, we also denote ν the unit normal vector to Γ , with a given orientation.

Since u is a weak solution in each Ω_{\pm} , where it is smooth, it is a classical solution away from Γ . Then we may integrate by parts (**Exercise:** fill the details) in each Ω_{\pm} , to compute the left-hand side of (1.8). There remains the identity

$$\int_{\Gamma} ([u_j]\nu_t + [f_j(u)]\nu_x)\phi \, ds = 0.$$

Since the test function is arbitrary, this amounts to writing

$$[u_j]\nu_t + [f_j(u)]\nu_x = 0,$$

or in vectorial form,

$$[u]\nu_t + [f(u)]\nu_x = 0.$$

Since $[u] \neq 0$ by assumption, we see that $\nu_x \neq 0$, which means that the curve Γ can be parametrized by the time: $t \mapsto (X(t), t)$. Then the ratio $-\nu_t/\nu_x$ is nothing but the slope X' . Finally, we obtain the *Rankine-Hugoniot* relation

$$[f(u)] = \frac{dX}{dt}[u]. \tag{1.10}$$

These calculations can be made in the reverse order, and one obtains the following important result:

Proposition 1.3. *Let Ω and Γ (a smooth curve) be as above, and let $u : \Omega \rightarrow \mathcal{U}$ be a field such that the restriction of u to each of Ω_{\pm} is of class C^1 and extends as a C^1 -field up to Γ .*

Then u is a solution of the system (1.1) in Ω if, and only if,

- *It is a classical solution away from Γ ,*
- *It satisfies the Rankine-Hugoniot relation (1.10) across Γ .*

Example. For a scalar equation, one may rewrite (1.10) as

$$\frac{dX}{dt} = \frac{[f(u)]}{[u]},$$

which shows that the slope of Γ is $f'(\bar{u})$ for some \bar{u} in the interval of extremities u_{\pm} . For the Burgers equation, we simply have

$$\frac{dX}{dt} = \frac{u_+ + u_-}{2}.$$

Convention. Since a discontinuity curve is parametrized by the time, we shall always choose the \pm sides in the natural way:

$$u_-(X(t), t) = \lim_{x \uparrow X(t)} u(x, t), \quad u_+(X(t), t) = \lim_{x \downarrow X(t)} u(x, t).$$

This amounts to choose the orientation of ν so that $\nu_x > 0$.

1.2.4 Non-uniqueness of weak solutions

The extension of the notion of solution resolves the lack of solution that we encountered in our study of classical solutions. However it introduces spurious, unphysical solutions. In particular, we have way too many solutions, typically an infinity, to the Cauchy problem.

To see this, we again consider the Burgers equation (1.5). We content ourselves with the null initial data $a \equiv 0$. Of course, there is a solution $u \equiv 0$, which is the physically relevant one. However, we can use discontinuities to build non-trivial solutions. For instance, the following definition yields a solution, for every choice of $b, c \in \mathbb{R}$ such that $b < 0 < c$:

$$u(x, t) := \begin{cases} 0, & x < bt, \\ 2b, & bt < x < (b+c)t, \\ 2c, & (b+c)t < x < ct, \\ 0, & ct < x, \end{cases}$$

since such a u is constant off lines, across which it satisfies the Rankine-Hugoniot condition.

Exercise. Build more general piecewise constant solutions to this Cauchy problem.

1.2.5 Entropy admissibility condition

Since we have replaced a non-existence trouble by a non-uniqueness one, we need to make a step backward and restrict the notion of weak solution. The clue is that, despite the apparent reversibility of systems (1.1), which is invariant under the space-time reversal $(x, t) \mapsto (X - x, T - t)$, the second principle of thermodynamics tells us that non-smooth flows of gas dynamics are irreversible. This is exactly saying that not all the discontinuities described by (1.10) are admissible.

To be more explicit, let $(u_-, u_+; s)$ be a triple that satisfies the Rankine-Hugoniot relation:

$$[f(u)] = s[u]. \quad (1.11)$$

Then

$$u(x, t) := \begin{cases} u_-, & x < st, \\ u_+, & st < x \end{cases}$$

defines a weak solution of (1.1). However, since (1.11) is perfectly symmetric in u_\pm , we may exchange the role of u_- and u_+ in our construction, and we obtain another solution v of (1.1). Irreversibility is that at least one of u and v is physically irrelevant.

One thus needs a criterion in order to select the admissible discontinuities, presumably in the form of an inequality, not symmetric in u_\pm .

When the system is compatible with (1.6), where η is strongly convex, we remark that the Rankine-Hugoniot condition for (1.6)

$$[q(u)] = s[\eta(u)]$$

is not compatible with (1.10) in general. Because the elimination of s yields

$$[\eta(u)] [f(u)] = [q(u)] [u], \quad (1.12)$$

which is an equation in \mathbb{R}^n , with the obvious solution $u_+ = u_-$. It often happens that it has no other solution. For instance (1.12) gives, for the Burgers equation, $[u]^4 = 0$, from which we have $u_+ = u_-$.

Exercise. Prove that in the scalar case, with $f'' > 0$ and $\eta'' > 0$, (1.12) implies $u_+ = u_-$.

The calculation above tells that a discontinuous solution u of (1.1) cannot satisfy simultaneously (1.6) across discontinuities. Whence the

idea to replace (1.6) by an inequality, in the sense of distributions:

$$\partial_t \eta(u) + \partial_x q(u) \leq 0. \quad (1.13)$$

It is important in this condition that we have chosen the pair (η, q) such that η is strongly convex. If it was concave, we should change the sense of the inequality in (1.13).

Since (1.6), hence (1.13), is automatically satisfied whenever u is a classical solution, (1.13) serves only across shocks. It can be reinterpreted as a jump inequality

$$[q(u)] \leq \frac{dX}{dt} [\eta(u)], \quad (1.14)$$

where we recall our convention of the \pm sides and our definition $[g(u)] = g(u_+) - g(u_-)$.

Let us take the example of the Burgers equation, with $f(u) = \eta(u) = u^2/2$, thus $q(u) = u^3/3$. We already know that $s = (u_+ + u_-)/2$. Then (1.14) tells that

$$\left(\frac{u_+^2 + u_+ u_- + u_-^2}{3} - s \frac{u_+ + u_-}{2} \right) [u] \leq 0.$$

Since the parenthesis equals $[u]^2/12$, this amounts to saying $u_+ \leq u_-$. This is the admissibility condition that we were looking for.

Exercise. More generally, in the scalar case with a convex flux f , show that a discontinuity is admissible if and only if $u_+ \leq u_-$.

Terminology. The condition (1.13) is the Lax *entropy inequality*. Admissible discontinuities are called *shocks*, especially when (1.14) is a strict inequality. We shall see other selection criteria in the sequel, which are not all equivalent. Thus the notion of shock might differ, depending on which criterion we adopt to select admissible solutions. These criteria are however equivalent for many reasonable systems and in particular for scalar equations with convex fluxes.

1.2.6 The viscosity approach

A way to justify (1.13) is to say that a system like (1.1) describes only an idealized physical process, which would be better represented by the parabolic system of conservation laws

$$\partial_t u + \partial_x f(u) = \epsilon \partial_x^2 u, \quad (1.15)$$

where $\epsilon > 0$ is a small number. One may also have $\partial_x(B(u)\partial_x u)$ instead of $\partial_x^2 u$ in the right-hand side. Then $B(u) \in \mathbf{M}_n(\mathbb{R})$ is called the *viscosity tensor*.

What we expect is that the Cauchy problem (1.15, 1.4) admits a unique smooth solution u^ϵ , which converges boundedly almost everywhere to a field $u(x, t)$ as $\epsilon \rightarrow 0+$. If this is true, then one can pass to the limit in the sense of distributions in (1.15), and we find that u is a weak solution of (1.1). More precisely, we have

$$\int_0^T \int_{\mathbb{R}} (u_j^\epsilon \partial_t \phi + f_j(u^\epsilon) \partial_x \phi + \epsilon u_j^\epsilon \partial_x^2 \phi) dx dt + \int_{\mathbb{R}} a_j(x) \phi(x, 0) dx = 0$$

for every test function ϕ . Passing to the limit, we obtain that u is a weak solution of the Cauchy problem (1.1, 1.4).

We now multiply (1.15) (with u^ϵ) by $d\eta(u^\epsilon)$. We obtain

$$\partial_t \eta(u^\epsilon) + \partial_x q(u^\epsilon) = \epsilon \partial_x^2 \eta(u^\epsilon) - \epsilon D^2 \eta(u^\epsilon) : \partial_x u^\epsilon \otimes \partial_x u^\epsilon.$$

Since η is convex, this implies

$$\partial_t \eta(u^\epsilon) + \partial_x q(u^\epsilon) \leq \epsilon \partial_x^2 \eta(u^\epsilon).$$

Passing again to the limit as above, we obtain the entropy inequality (1.13) in the sense of distributions.

1.2.7 The scalar case

In the scalar case, every function η is an entropy, thus every convex function is a convex entropy. This yields as many entropy inequalities as there are convex functions. And all these inequalities must be written simultaneously. Since the set of convex funtions is a convex cone spanned by the affine functions and by the so-called *Kružkov entropies*

$$\eta_k(u) := |u - k|, \quad q_k(u) = (f(u) - f(k))\text{sign}(u - k),$$

we find that a discontinuity $(u_-, u_+; s)$ is admissible if, and only if

Rankine-Hugoniot. One has $s = [f(u)]/[u]$,

Oleinik condition. Either $u_- < u_+$ and the graph of the restriction of f to the interval $[u_-, u_+]$ is above its chord, or $u_+ < u_-$ and the graph of the restriction of f to the interval $[u_+, u_-]$ is below its chord.

We leave the proof of that as an **Exercise**.

We also have a well-posedness result, due to S. Kružkov, for which we refer to [7, 23]:

Theorem 1.4. Assume that the flux $f : \mathbb{R} \rightarrow \mathbb{R}$ is of class C^1 . Let $a \in L^\infty(\mathbb{R})$ be given. Then there exists a unique bounded solution of the Cauchy problem satisfying the entropy inequality (1.13). It satisfies

$$\sup_{x \in \mathbb{R}, t > 0} u(x, t) = \sup_{x \in \mathbb{R}} a(x), \quad \inf_{x \in \mathbb{R}, t > 0} u(x, t) = \inf_{x \in \mathbb{R}} a(x).$$

If b is another bounded data, with associated solution v , then one has the contraction property, for all $A < B$ and $0 < t < (B - A)/M$,

$$\int_{A+Mt}^{B-Mt} |u(x, t) - v(x, t)| dx \leq \int_A^B |a(x) - b(x)| dx, \quad (1.16)$$

where M is the supremum of f' over the interval

$$[\inf_{x \in \mathbb{R}} a(x), \sup_{x \in \mathbb{R}} a(x)].$$

The Cauchy problem is thus well-understood in the scalar case, and even in several space dimensions ($n = 1$ and $d \geq 1$). The situation is however much more open in case of systems ($n \geq 2$), for we have not any more a comparison or a contraction principle.

1.3 Shock waves

In this section, we investigate in more details the admissible discontinuities, and we introduce new admissibility criteria: the Lax shock inequality and the existence of viscous shock profile.

1.3.1 The Hugoniot locus

To begin with, we describe the set defined by the Rankine-Hugoniot condition (1.10), called the *Hugoniot locus* \mathcal{H} . Since it is an equation in \mathbb{R}^n with $2n + 1$ parameters $(u_-, u_+; s) \in \mathcal{U} \times \mathcal{U} \times \mathbb{R}$, we expect that the Hugoniot locus be a piecewise smooth manifold of dimension $n + 1$. We observe that it contains the trivial elements $(w, w; s)$ where $w \in \mathcal{U}$ and $s \in \mathbb{R}$ are arbitrary. This exhausts \mathcal{H} in the neighbourhood of $X_0 = (w_0, w_0; s_0)$, whenever the map

$$(w, z; s) \mapsto H(w, z; s) := f(z) - f(w) - s(z - w)$$

is a submersion at X_0 , which means that $DH(X_0)$ is onto. Since

$$D_s H(X_0) = 0, \quad D_z H(X_0)Z = (df(w_0) - s)Z,$$

and

$$D_w H(X_0)Z = -(df(w_0) - s)Z,$$

we find that \mathcal{H} is smooth at X_0 when s is not an eigenvalue of $df(w_0)$.

We already know that the eigenvalues of $df(w)$ are real numbers. From now on, we shall assume that they are *simple* (we say that the system (1.1) is *strictly hyperbolic*). We arrange the eigenvalues in increasing order

$$\lambda_1(w) < \cdots < \lambda_n(w).$$

We denote by $r_j(w)$ an eigenvector:

$$df(w)r_j(w) = \lambda_j(w)r_j(w).$$

We also denote $(\ell_1(w), \dots, \ell_n(w))$ a dual basis: each ℓ_j is a differential form and one has

$$\ell_j(w)r_k(w) = \delta_j^k.$$

At a point X_0 with $s_0 = \lambda_j(w_0)$, there is a bifurcation. Lax showed that \mathcal{H} is locally the union of two smooth manifolds. The first one is the trivial one ($z = w$). The second one is parametrized by (w, s) with $z - w \sim (s - \lambda_j(w))r_j(w)$. It can be shown actually that at fixed w_0 , the curve $s \mapsto z$ is tangent at second order to the integral curve of r_j , the solution of the differential equation

$$\frac{dZ}{ds} = r_j(Z), \quad Z(s_0) = w_0.$$

To see this, we use the Taylor formula

$$f(v) - f(u) = \left(\int_0^1 df((1-\tau)u + \tau v) d\tau \right) (v - u) =: A(u, v)(v - u).$$

Since $A(u, u) = df(u)$ has simple real eigenvalues, the eigenvalues of $A(u, v)$ are still simple and real for u and v close to each other. We denote them $\Lambda_k(u, v)$. These are smooth functions in a neighbourhood of the diagonal, such that $\Lambda_k(u, u) = \lambda_k(u)$. Likewise, we have eigenfields $R_k(u, v)$, with $R_k(u, u) = r_k(u)$.

The Rankine-Hugoniot condition rewrites as

$$(A(w, z) - sI_n)[u] = 0.$$

Away from the diagonal, the Hugoniot locus is thus described as the union of sets \mathcal{H}_k , defined by

$$s = \Lambda_k(w, z), \quad z - w \parallel R_k(w, z).$$

There remains to solve the equation

$$z - w = \alpha R_k(w, z).$$

To this end, we define $N(w, z; \alpha) := z - w - \alpha R_k(w, z)$. We have $N(w_0, w_0; 0) = 0$. Since $D_s N(w_0, w_0; 0) = I_n$, this function is a submersion, and its zero set is thus locally a submanifold of codimension n , thus of dimension $n + 1$. Its tangent space at $(w_0, w_0; 0)$ is given by

$$dz - dw = (d\alpha)R_k(w_0, w_0) = (d\alpha)r_k(w_0).$$

At fixed $w \equiv w_0$, this means that the Hugoniot locus is a curve $\alpha \mapsto (w_0, z(\alpha); s(\alpha))$ with

$$z = w_0 + \alpha r_k(w_0) + O(\alpha^2), \quad s = \Lambda_k(w_0, z).$$

A refined estimate is that

$$z - w_0 = \alpha r_k \left(\frac{w_0 + z}{2} \right) + O(\alpha^3).$$

This curve is denoted by $\mathcal{H}_k(w_0)$. We say that such a triple $(w_0, z; s)$ is a k -discontinuity.

1.3.2 Genuine nonlinearity

Let us investigate the inequality (1.14) when $(u_+, s) \in H_k(u_-)$. We again express the jumps $[q(u)]$ and $[\eta(u)]$ with the help of the Taylor formula. With the results of the previous paragraph, we obtain

Lemma 1.5. *Across a k -discontinuity, we have*

$$[q(u)] - s[\eta(u)] = \frac{\alpha^3}{12} (d\lambda_k r_k) D^2 \eta(r_k, r_k) + O(\alpha^4).$$

It is time to introduce the following notion:

Definition 1.6. We say that the k -th characteristic field (λ_k, r_k) is *genuinely nonlinear* at u_- if $d\lambda_k r_k \neq 0$ at u_- .

When the k -th field is genuinely nonlinear (in short GNL), we can normalize r_k by $d\lambda_k r_k = 1$.

Under the genuine nonlinearity assumption, we see that

$$[q(u)] - s[\eta(u)] \sim \frac{\alpha^3}{12} D^2 \eta(r_k, r_k),$$

so that, for small α , the sign of $[q(u)] - s[\eta(u)]$ is that of α . A small k -discontinuity is thus admissible for the entropy inequality if, and only if, α is negative.

1.3.3 The Lax shock inequality

Let us continue our study of small k -shocks under GNL assumption. We have

$$s = \lambda_k(u_-) + \alpha + O(\alpha^2) = \lambda_k(u_+) - \alpha + O(\alpha^2).$$

With α negative and small, this gives

$$\lambda_k(u_+) < s < \lambda_k(u_-), \quad (1.17)$$

while non-admissible small k -discontinuities satisfy the opposite inequalities. Therefore (1.17) is equivalent to $\alpha < 0$, thus to (1.13) for small k -discontinuities. Of course, since $[u]$ is small and the eigenvalues are simple, we also have

$$\lambda_{k-1}(u_-) < s < \lambda_{k+1}(u_+). \quad (1.18)$$

Definition 1.7. Let $(u_-, u_+; s)$ satisfy the Rankine-Hugoniot condition. We say that it is a k -shock if it satisfies also the inequalities (1.17, 1.18).

These inequalities are called the *Lax shock inequalities*. For small k -discontinuities, and if the k -th field is GNL, they are equivalent to the entropy inequality.

Shock vs entropy inequalities. As mentioned above, both admissibility conditions (shock/entropy inequalities) are equivalent when the discontinuity has a small strength, provided that the system admits a convex entropy, and that the k -th characteristic field is GNL. However they may not be equivalent when one of these assumptions is dropped, for instance the genuine nonlinearity or the smallness of the strength. It may even happen that one condition makes sense while the other one does not. The shock inequality makes sense even when the system does not admit a convex entropy, while the entropy condition does not need that the solution be piecewise smooth.

1.3.4 Viscous shock profiles

A third criterion consists in going back to the parabolic system (1.15) and looking for a travelling wave

$$u^\epsilon(x, t) := U\left(\frac{x-st}{\epsilon}\right).$$

Such a u^ϵ is a solution of (1.15) whenever U satisfies the ODE

$$U'' = (f(U) - sU)'.$$
(1.19)

If U has limits u_{\pm} at $\pm\infty$, then u^ϵ converges boundedly almost everywhere towards

$$u(x, t) := \begin{cases} u_-, & x < st, \\ u_+, & st < x, \end{cases}$$

which will thus be declared admissible. The function U is called the *viscous shock profile* (VSP) of $(u_-, u_+; s)$. Integrating once (1.19), we obtain the first-order ODE

$$U' = f(U) - sU - q,$$

where q is a constant of integration. This one can be calculated by letting $x \rightarrow \pm\infty$; we obtain on the one hand $q = f(u_-) - su_-$ and on the other hand $q = f(u_+) - su_+$. In particular, a necessary condition for the existence of such a U is the Rankine-Hugoniot condition (1.10). Finally, we obtain the ODE, called the *profile equation*

$$U' = f(U) - f(u_-) - s(u - u_-). \quad (1.20)$$

With a more general viscosity tensor $B(u)$, the profile equation is

$$B(U)U' = f(U) - f(u_-) - s(u - u_-). \quad (1.21)$$

Since the calculations of Section 1.2.6 apply to our u^ϵ , we deduce that if the discontinuity $(u_-, u_+; s)$ admits a shock profile, then it satisfies the entropy inequality (1.14).

1.4 The Riemann problem

The system (1.1), as well as various admissibility criteria, is invariant under the rescaling $(x, t) \mapsto (\mu x, \mu t)$. If the initial data is of the form

$$a(x) = \begin{cases} a_-, & x < 0, \\ a_+, & x > 0, \end{cases}$$

we therefore expect that the solution be homogenous of degree zero: $u(x, t) = V(x/t)$. This must be so if the solution is unique. Finding V when a_- and a_+ are given is the *Riemann problem*. Its solution is used to design some efficient numerical schemes, among which the Godunov scheme (see Section 2.2.4) and the Glimm scheme.

To solve the Riemann problem, we need first to know what are elementary centered waves. We already have encountered the shock waves. We have yet to discover the rarefaction waves and the contact discontinuities.

1.4.1 Rarefaction waves

The rarefaction waves are smooth solutions of the form $u(x, t) = V(x/t)$. They arise when a field is GNL. For such a solution, we have

$$u_t = -\frac{x}{t^2}V', \quad f(u)_x = \frac{1}{t}(f(V))'.$$

Therefore the system (1.1) reduces to

$$\frac{d}{d\xi}f(V(\xi)) - \xi \frac{dV}{d\xi} = 0,$$

that is

$$(df(V(\xi)) - \xi) \frac{dV}{d\xi} = 0. \quad (1.22)$$

With $V' \neq 0$, this means that ξ is an eigenvalue:

$$\xi = \lambda_k(V(\xi)). \quad (1.23)$$

In addition,

$$V'(\xi) \parallel r_k(V(\xi)). \quad (1.24)$$

Differentiating (1.23), we obtain

$$1 = d\lambda_k(V)V'.$$

Using (1.24), we deduce that the k -th characteristic field must be GNL. We then have

$$\frac{dV}{d\xi} = r_k(V). \quad (1.25)$$

An integral curve of (1.25) is called a *k -rarefaction curve*. Two states $V(\xi_1)$ and $V(\xi_2)$ on the same curve can be linked by a rarefaction wave in the wedge $t\xi_1 < x < t\xi_2$. Because of (1.23), we see that the state with the smallest value of $\lambda_k(V)$ is at left and the other one is at right.

1.4.2 Contact discontinuities

When a field is not genuinely non linear, it may happen the opposite:

Definition 1.8. The k -th characteristic field is said to be *Linearly degenerate* if $d\lambda_k r_k \equiv 0$.

When a field is linearly degenerate (LD), the eigenfield r_k does not have a canonical normalization, contrary to the GNL case.

The important point is that linear degeneracy implies a coincidence between rarefaction (notice that rarefaction waves do not exist in this case) and discontinuities, as well as some kind of reversibility:

Theorem 1.9. Assume that the k -th field is LD. Let γ be a k -rarefaction curve. Then

1. λ_k is constant along γ ,
2. If $u_{\pm} \in \gamma$ and $s = \lambda_k|_{\gamma}$, then $(u_-, u_+; s)$ satisfies the Rankine-Hugoniot relation (1.10),
3. The triple $(u_-, u_+; s)$ also satisfies the identity $[q(u)] = s[\eta(u)]$.

In particular, γ is contained (and in general equals to) the u -projection of $\mathcal{H}_k(u_-)$, for every of its points u_- .

Such triples $(u_-, u_+; s)$ are called *contact discontinuities*. They are always declared admissible. It is clear from above that the reversed triple $(u_+, u_-; s)$ is admissible too: contact discontinuities are reversible.

1.4.3 The theorem of Lax

The general solution of the Riemann problem is made of $n + 1$ constant states $u_0 = a_-, u_1, \dots, u_n = a_+$, separated by simple (or composite) waves. Typically, u_{j-1} is separated from u_j by a j -th wave. If the j -th field is LD, the wave is a CD and u_{j-1}, u_j must belong to the same integral curve of r_j . If the j -th field is GNL, the wave is a rarefaction or a shock. When a_- and a_+ are far apart, or when the fields are neither LD nor GNL, then one can find either composite waves, where shocks are embedded in rarefactions, or shocks that are not Lax shocks. Terminology lists under-compressive and over-compressive shocks besides Lax shocks.

Let us assume for instance that the k -th field is GNL at u_- . Then the set of states u_+ (at right) that can be reached from u_- (at left) through a rarefaction is the forward part of the integral curve of r_k starting from u_- . Call it $R_k(u_-)$. The set of states u_+ (at right) that can be reached from u_- (at left) is the backward part ($s < \lambda_k(u_-)$) of the Hugoniot curve $\mathcal{H}_k(u_-)$. Call it $S_k(u_-)$; it is tangent at second order to $R_k(u_-)$. Therefore the union $W_k^f(u_-) = S_k(u_-) \cup R_k(u_-)$ is locally a C^2 -curve, tangent to r_k at u_- . It is the k -th *forward wave curve*, named that way because the left state u_- is specified. If we fix u_+ instead and consider the set of states u_- that can be connected to u_+ , we find the *backward wave curve* $W_k^b(u_+)$. By definition, we have

$$(u_+ \in W_k^f(u_-)) \iff (u_- \in W_k^b(u_+)).$$

For an LD field, the forward and backward wave curves coincide and consist in the integral curve of r_k .

Solving the Riemann problem amounts to find the collection of intermediate states u_1, \dots, u_{n-1} such that

$$u_1 \in W_1^f(u_0), \dots, u_j \in W_j^f(u_{j-1}), \dots, u_n \in W_n^f(u_{n-1}). \quad (1.26)$$

Using a parametrization of the wave curves, and the fact that they are tangent to r_k at their base point, Lax established the following fundamental result. Again, we refer to [7, 23] for a full proof.

Theorem 1.10. *Assume that the characteristic fields are either GNL or LD. let $a_- \in \mathcal{U}$ be given. Then there exists two neighbourhoods $\mathcal{V} \subset \mathcal{W}$ of a_- such that, if $a_+ \in \mathcal{V}$, then there exists a unique solution of the Riemann Problem in the form of (1.26), where all the waves take values in \mathcal{W} .*

We point out that a solution to the RP still satisfies $u = U(x/t)$ with

$$\frac{d}{d\xi} f(U) = \xi \frac{dU}{d\xi}$$

in the distributional sense. In particular

$$\frac{d}{d\xi} (f(U) - \xi U) = -U$$

shows that $\xi \mapsto f(U) - \xi U$ is Lipschitz continuous, despite the fact that the solution itself may be discontinuous. This remark is at the basis of the Godunov scheme.

1.5 Existence of viscous shock profiles

We now consider whether or not a given discontinuity $(u_-, u_+; s)$ admits a shock profile. The profile equation, here with a general viscous tensor B , is

$$B(U)U' = f(U) - f(u_-) - s(U - u_-). \quad (1.27)$$

We recall that we have assumed the RH condition (1.10). Our first assumption about B is that it is invertible, although in realistic examples (like gas dynamics), it would not be. We shall also need an assumption that makes (1.15) a parabolic system that stabilizes the constant states, see (**Stab**) below. It will follow from dissipativeness, a rather natural assumption.

The profile equation can be recast as an ODE $U' = G(U; s)$, where we know that $G(u_\pm; s) = 0$. Since we look for a heteroclinic orbit from u_- to u_+ (meaning that $U(\pm\infty) = u_\pm$), a shock profile exists if, and only if, the unstable manifold $W^u(u_-)$ and the stable manifold $W^s(u_+)$, at u_\pm respectively, of this dynamical system have a non-trivial intersection. As a matter of fact, every value $U(x)$ of a shock profile belongs to this intersection, and conversely, if a is in this intersection, then the solution of the Cauchy problem

$$U' = G(U; s), \quad U(0) = a$$

is a shock profile, because $\lim_{x \rightarrow \pm\infty} U(x) = u_\pm$ by assumption.

Structural stability. Assume that s is not an eigenvalue of either $A(u_-)$ or $A(u_+)$. Thus $\lambda_j(u_+) < s < \lambda_{j+1}(u_+)$ and $\lambda_{k-1}(u_-) < s < \lambda_k(u_-)$ for some j, k . If $B \equiv I_n$, then the dimensions of $W^u(u_-)$ and $W^s(u_+)$ are respectively $n - k + 1$ and j . It is well-known that the intersection persists under small perturbations of the data (one says that this intersection is *structurally stable*) if, and only if, the tangent spaces to these manifolds add up to \mathbb{R}^n ; this property is called *transversality*. Since the intersection of these tangent spaces must contain U' , thus must be of dimension one at least, this implies that the sum $(n - k + 1) + j$ be at least $n + 1$. Whence the necessary condition for this structural stability: $j \geq k$. The limit case $j = k$ is nothing but the Lax shock condition. Other cases where $j > k$ are called *over-compressive* shocks. When $j < k$, the transversality always fails and the shock is called *under-compressive*.

The discussion above remains valid whenever the system (1.1) admits a strongly convex entropy η and the tensor B is *dissipative*, in the sense that

$$X \mapsto D^2\eta(U)(B(U)X, X) \quad (1.28)$$

is positive definite for every U .

Since small shocks for GNL fields are Lax shocks, we shall only consider the case $j = k$. More generally, we shall discuss the case where $u_+ \in \mathcal{H}_k(u_-)$.

Exercise. Assume that B is dissipative for some U and that s is not an eigenvalue of $A(u_-)$. Prove that $D_U G(u_-; s)$ does not have a purely imaginary eigenvalue. One says that u_- is a *hyperbolic* fixed point of the ODE. Here *hyperbolic* is in the sense of dynamical systems; it has nothing to do with hyperbolic PDEs.

1.5.1 The scalar case

In the scalar case, B must be positive in order that (1.15) be parabolic. Then

$$G(u; s) = \frac{f(u) - f(u_-) - s(u - u_-)}{B(u)}.$$

Every solution of the ODE is monotonous and tends at $\pm\infty$ towards consecutive zeroes of $G(\cdot; s)$. A shock profile exists if, and only if, $G(\cdot; s)$ is strictly of the sign of $u_+ - u_-$ over the interval between u_- and u_+ . This amounts to saying that the Oleinik condition is satisfied in a strict sense, with the graph of f strictly above or below its chord.

Therefore, in the scalar case, the existence of a shock profile is, apart from borderline cases, equivalent to the entropy criterion. For instance, if f is convex, both criteria give the same constraint $u_+ < u_-$.

1.5.2 Reduction to a center manifold (bifurcation analysis)

We now develop a strategy for proving the existence of discrete shock profiles associated to discontinuities of small strength. More precisely, we look for profiles of small amplitude, although we do not exclude that large amplitude profiles could exist for some small shocks in pathological situations.

The idea is to augment the profile equation (1.27) with the obvious one $s' = 0$. Thus we deal with the dynamical system

$$\begin{pmatrix} U \\ s \end{pmatrix}' = H(U; s) := \begin{pmatrix} G(U; s) \\ 0 \end{pmatrix}. \quad (1.29)$$

This is the reason why we kept trace of s in the abstract form $U' = G(U; s)$ of the profile equation.

In the following analysis, we choose an index $1 \leq k \leq n$ and we keep u_- fixed. Then we investigate the flow of (1.29) in a small enough neighbourhood \mathcal{V} of

$$X_- := \begin{pmatrix} u_- \\ \lambda_k(u_-) \end{pmatrix}.$$

To begin with, we notice that the rest points ($H(X) = 0$) in \mathcal{V} fall into two categories:

1. The pairs $X = (u_-; s)$ for every s close to $\lambda_k(u_-)$,
2. The points $(u_+; s)$ corresponding to triples $(u_-, u_+; s)$ on the Hugoniot curve $\mathcal{H}_k(u_-)$.

We point out that these two curves are transversal to each other since their tangents at X_- are linearly independent.

We now make a reduction to the *center manifold* \mathcal{M}_- of (1.29) at the point X_- . This is a smooth manifold with several properties, among which the local invariance under the flow of the ODE (1.29). For a thorough account of what is a center manifold and how one proves its existence, we refer to [5]. The reason why we use it is that \mathcal{M}_- contains every trajectory that is globally defined and is contained in \mathcal{V} . In particular, it contains

- The rest points in \mathcal{V} ,
- The homoclinic and heteroclinic orbits that remain in \mathcal{V} .

According to the latter point, \mathcal{M}_- contains all the shock profiles that are associated to discontinuities $(u_-, u_+; s)$ as long as they are entirely contained in \mathcal{V} (in particular the final state u_+ belongs to \mathcal{V}).

The center manifold is not always unique, but it has some amount of uniqueness since it necessarily contains some specific orbits (see above).

Its dimension c and its tangent space T at X_- are given by the linearized system

$$\begin{pmatrix} U \\ s \end{pmatrix}' = DH(X_-) \begin{pmatrix} U \\ s \end{pmatrix}. \quad (1.30)$$

The tangent space T is nothing but the central invariant subspace of $DH(X_-)$, that is the sum of the characteristic spaces associated to the eigenvalues with zero real part. Whence $c = \dim T$ is the sum of the multiplicities of these eigenvalues.

Thus let us investigate the spectrum of

$$\begin{aligned} DH(X_-) &= \begin{pmatrix} D_U G(X_-) & D_s G(X_-) \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} B(u_-)^{-1}(df(u_-) - \lambda_k(u_-)I_n) & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

We shall assume in the sequel that the state u_- is linearly stable for (1.15), which means

(Stab) There exists a number $\theta > 0$ such that for every $\xi \in \mathbb{R}$, the eigenvalues of $\xi^2 B(u_-) + i\xi df(u_-)$ have a real part larger than or equal to $\theta\xi^2$.

Exercise. Prove that if B is dissipative with respect to a strongly convex entropy, then **(Stab)** is fulfilled.

With **(Stab)**, we now that $B(u_-)^{-1}(df(u_-) - sI_n)$ is almost a hyperbolic matrix: its only purely imaginary eigenvalue is $\mu = 0$. Actually, one can prove a little bit more:

Lemma 1.11. *Under the stability assumption **(Stab)**, one has*

- *for every $k = 1, \dots, n$, $\ell_k Br_k > 0$ at u_- ,*
- *and $\mu = 0$ is a simple eigenvalue of $B(u_-)^{-1}(df(u_-) - \lambda_k(u_-))$.*

Under **(Stab)**, the dimension of the center manifold is thus $c = 2$. Moreover, since the tangent space at X_- is spanned by the vectors

$$\begin{pmatrix} r_k(u_-) \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

we can take $X \mapsto y := (\ell_k(u_-)(U - u_-), s)$ as local coordinates on \mathcal{M}_- . The flow of (1.29) over the center manifold can be rewritten as the flow of a tangent vector field h :

$$y' = h(y) := \begin{pmatrix} \ell_k(u_-)G(u(y); y_2) \\ 0 \end{pmatrix}. \quad (1.31)$$

Because s is constant along trajectories, the second component h_2 vanishes identically.

We are now in a very simple situation. All the interesting dynamics (small and global trajectories) near X_- is contained in the surface \mathcal{M}_- , and the dynamics on \mathcal{M}_- is horizontal ($y_2 = s$ constant). On every horizontal line $y_2 = s$, the dynamics is given by $y'_1 = h_1(y_1, s)$, an autonomous differential equation on an interval. Between two consecutive zeroes of $h_1(\cdot, s)$, there is a heteroclinic orbit, which is nothing but a shock profile.

There remains to describe the zero set of h and of $h_1(\cdot, s)$, and to check the sign of the latter function between consecutive zeroes. The zero set is precisely the set of rest points of (1.29), already described. At fixed s , the zeroes of $h_1(\cdot, s)$ are therefore of two types:

- either $y_1 = 0$, corresponding to the pair (u_-, s) ,
- or the non-zero values y_1 corresponding to the states $u_+ \neq u_-$ for which $(u_-, u_+; s)$ is in the Hugoniot set.

1.5.3 Lax shocks

Let us assume in this paragraph that the k -th field is GNL. We recall that the k -th Hugoniot curve at u_- has the property that

$$u = u_- + \alpha r_k(u_-) + O(\alpha^2) \quad (\text{thus } y_1 \sim \alpha), \quad s = \lambda_k(u_-) + \alpha + O(\alpha^2).$$

Its tangent is therefore transversal to the s -axis. In a neighbourhood of X_- , this curve intersects a horizontal line $y_2 = s$ in exactly one point which we denote $(u_+(s), s)$. Thus the ODE $y'_1 = h_1(y_1, s)$ has exactly two fixed points, namely 0 and $\ell_k(u_-)(u_+(s) - u_-)$. Therefore there exists one, and only one, heteroclinic orbit of (1.27) between u_- and $u_+(s)$.

When $s \neq \lambda_k(u_-)$, a Taylor expansion using the fact that $u - u_- \sim y_1 r_k(u_-)$ on \mathcal{M}_- gives

$$h_1(y_1, s) \sim (\lambda_k(u_-) - s)y_1.$$

In other words,

$$\frac{\partial h_1}{\partial y_1}(0, s) = \lambda_k(u_-) - s.$$

Therefore $(0, s)$ is an unstable rest point of (1.31) if and only if $\lambda_k(u_-) < s$.

The orbit mentioned above thus goes from u_- to $u_+(s)$ if and only if $\lambda_k(u_-) < s$. Since a discontinuity $(u_-, u_+(s); s)$ is either a Lax shock or an anti-Lax one (the latter terminology means that $(u_+, u_-; s)$ is a Lax shock), we have the following conclusion.

Theorem 1.12. Assume that u_- is linearly stable (property **(Stab)**) for the parabolic equation (1.15). Let us assume also that the k -th characteristic field is genuinely nonlinear at u_- .

Then there are neighbourhoods $\mathcal{V}_- \subset \mathcal{W}_-$ of u_- such that, for every triple $(u_-, u_+; s)$ satisfying the Rankine-Hugoniot condition and $u_+ \in \mathcal{V}_-$, there exists a viscous shock profile from u_- (at left) to u_+ (at right), entirely contained in \mathcal{W}_- , if and only if, this discontinuity is a Lax shock.

Among all fields taking values in \mathcal{W}_- , this profile is unique, up to the shift $U \mapsto U(\cdot - \xi_0)$.

Remarks.

- Under the assumptions of Theorem 1.12, and if the discontinuity $(u_-, u_+; s)$ does not admit a profile, then it is not a Lax shock. Since the k -th field is GNL and $(u_+, u_-; s)$ is in \mathcal{H}_k with $u_+ \in \mathcal{V}_-$, the latter discontinuity is a Lax shock and therefore it admits a viscous shock profile.
- An important fact is that the existence or the non-existence of a viscous shock profile does not really depend on the choice of a tensor B satisfying **(Stab)**. What could depend on this choice is the size of the neighbourhood \mathcal{V}_- .

Exercise. What can be said if $d\lambda_k r_k$ vanishes at u_- , but the second derivative $d(d\lambda_k r_k)r_k$ is non-zero?

1.5.4 Under-compressive shocks

Let $(u_-, u_+; s)$ satisfy the Rankine-Hugoniot condition, and let us assume that

$$\lambda_{k-1}(u_-) < s < \lambda_k(u_-), \quad \lambda_{k-1}(u_+) < s < \lambda_k(u_+) \quad (1.32)$$

for some index k . We say that the discontinuity is an under-compressive shock with *defect index one*.

Since the stable manifold $W^s(u_+)$ has dimension $k - 1$ and the unstable $W^u(u_-)$ has dimension $n - k + 1$, which sum up to only n , and since the intersection is invariant by the flow of (1.27), it is unlikely that these manifolds intersect. Therefore there does not exist a heteroclinic orbit from u_- to u_+ in general. The same is true from u_+ to u_- .

Exercise. Figure out what happens for a differential equation in the plane, where the vector field has two hyperbolic zeroes. The cases saddle-sink or spring-saddle correspond to Lax shocks, while the case saddle-saddle corresponds to the under-compressive situation.

Let us assume however that the map $(a, b; \sigma) \mapsto f(b) - f(a) - \sigma(b - a)$ is a submersion at $(u_-, u_+; s)$. Then the Hugoniot locus is locally a submanifold of dimension $n + 1$. It is now a generic fact that the set \mathcal{P} of under-compressive shocks $(a, b; \sigma)$ for which there exists a viscous profile from a to b is a submanifold of codimension one, thus a manifold of dimension n . In favourable cases, \mathcal{P} can be parametrized by either a or b : for every a in some open region, there exists a state $b = \beta(a)$ and a velocity $s = \sigma(a)$, such that $(a, b; s)$ is an undercompressive shock admitting a viscous shock profile. The maps β, σ are smooth.

We warn the reader that, contrary to the Lax case, the manifold \mathcal{P} does depend on the choice of B . It varies smoothly, in general, when B changes. Thus it becomes crucial to have a physically relevant viscosity tensor.

2 Finite difference schemes

We now turn to the numerical analysis of first-order systems of conservation laws (1.1). Given an initial data (1.4), we wish to compute an accurate approximation of the solution of the Cauchy problem. We warn the reader that since we do not have yet shown that the Cauchy problem is well-posed, except in either the scalar case or locally in time for smooth data, it is hard to say that an approximate solution given by a numerical scheme is accurate. The accuracy of a numerical scheme in presence of shocks is an outstanding problem, which has not been fully resolved so far. This problem is at the origin of some questions addressed in the next Chapter.

At the beginning, we give ourselves a mesh length $\Delta x > 0$ and a time step $\Delta t > 0$. We then approximate the space time domain $(0, +\infty) \times \mathbb{R}$ by the grid of points $(t_m := m\Delta t, x_j := j\Delta x)$ for $m \in \mathbb{N}$ and $j \in \mathbb{Z}$. We also use the points $x_{j+1/2}$ defined in the same way.

A discretization of the initial data provides initial values u_j^0 . Several choices are possible. For instance, we could take

$$u_j^0 := \int_{x_{j-1/2}}^{x_{j+1/2}} a(x) dx, \quad (2.1)$$

although it can be easier to set $u_j^0 := a(x_j)$ if a is continuous. We denote $U^0 := (u_j^0)_{j \in \mathbb{Z}}$. It is an element of $\ell^\infty(\mathbb{Z})$ if $a \in L^\infty(\mathbb{R})$. More generally, we have $U^0 \in \ell^p$ provided $a \in L^p$ and we make the choice (2.1). When $p = 2$, (2.1) amounts to projecting orthogonally over the subspace of piecewise constant elements of $L^2(\mathbb{R})$.

An *explicit* numerical scheme is a mapping $H_\Delta : \ell^\infty \rightarrow \ell^\infty$. We use the scheme to define vectors $U^m = (u_j^m)_{j \in \mathbb{Z}}$ inductively by

$$U^{m+1} = H_\Delta(U^m).$$

If the scheme is appropriately chosen, we expect that u_j^m is a good approximation of $u(t_m, x_j)$ where u is the supposed-to-be solution of the Cauchy problem (1.1,1.4). By a solution, we mean an admissible one, with respect to appropriate entropy conditions.

We warn the reader that since we expect the solution to have discontinuities, a pointwise convergence of the approximate solution towards u might be too ambitious. We should merely ask for a boundedly almost everywhere convergence. This requires to extend the approximate solution to the whole domain $(0, \infty) \times \mathbb{R}$. This is usually done by interpolation. For instance, we may define $u^{\Delta x} \equiv u_j^m$ over the cell $(x_{j-1/2}, x_{j+1/2}) \times (t_m, t_{m+1})$.

2.1 Conservative schemes

Since the system (1.1) commutes with translations, we ask that a scheme has the same property. Also, for a scheme to be of practical interest, we wish that the component $H_\Delta(U)_j$ depends only on finitely many components of U . Therefore a scheme has the general form

$$H_\Delta(U)_j = h(u_{j-p}, u_{j-p+1}, \dots, u_{j+q}).$$

We say that such a scheme is a $(p + q + 1)$ -point scheme. For instance, a scheme

$$H_\Delta(U)_j = h(u_{j-1}, u_j, u_{j+1})$$

is a 3-point scheme.

We warn the reader that the function h depends also on Δx and Δt . See the examples given in Section 2.2.

The induction defined by a scheme writes

$$u_j^{m+1} = h(u_{j-p}^m, u_{j-p+1}^m, \dots, u_{j+q}^m), \quad (2.2)$$

where the initial data u_j^0 is computed from a , as explained above.

Another natural requirement is *conservativity*, in order to mimic the conservation property of (1.1). We therefore ask that there is a function F of $p + q$ arguments in \mathcal{U} , called the *numerical flux*, such that

$$h(v_{-p}, \dots, v_q) = v_0 + \frac{\Delta t}{\Delta x} (F(v_{-p}, \dots, v_{q-1}) - F(v_{1-p}, \dots, v_q)).$$

Then the scheme rewrites in the natural way

$$\frac{u_j^{m+1} - u_j^m}{\Delta t} + \frac{f_{j+1/2}^m - f_{j-1/2}^m}{\Delta x} = 0, \quad (2.3)$$

with

$$f_{j+1/2}^m := F(u_{j+1-p}^m, \dots, u_{j+q}^m).$$

For instance, in a 3-point scheme, one has $f_{j+1/2}^m = F(u_j^m, u_{j+1}^m)$.

Once again, the numerical flux may depend on Δx and/or Δt . In practice, it depends only on the ratio $\lambda := \Delta t / \Delta x$, in order to reflect the scale invariance of the PDEs (1.1) under the dilations $(x, t) \mapsto (\mu x, \mu t)$.

2.1.1 Consistency

For a scheme to be *consistent* with (1.1), it is natural to assume that the numerical flux equals f on the diagonal:

$$F(v, \dots, v) = f(v). \quad (2.4)$$

For a consistent scheme, we have the Lax-Wendroff Theorem:

Theorem 2.1. *Assume that the finite difference scheme (2.3) is consistent. Let $a \in L^\infty(\mathbb{R})$ be given. Given a sequence $\epsilon_k \rightarrow 0$ and a number $\lambda > 0$, denote u^{ϵ_k} the approximate solution associated to $\Delta x = \epsilon_k$ and $\Delta t = \lambda \epsilon_k$ (one interpolates u^{ϵ_k} in order to have it defined on $(0, \infty) \times \mathbb{R}$).*

Let us assume that the sequence u^{ϵ_k} converges boundedly almost everywhere towards a field u . Then u is a weak (i.e. distributional) solution of the Cauchy problem (1.1, 1.4).

One may ask whether the following partial converse of Theorem 2.1 holds: if the Cauchy problem (1.1, 1.4) admits a smooth solution u over $(0, T) \times \mathbb{R}$, then the approximate solution converges towards u as $\Delta x \rightarrow 0$. It turns out that the answer is negative in general under the assumption of consistency only. Such a statement needs an extra assumption, of stability. This is a well-known general fact in numerical analysis, at least in the realm of linear evolution problems.

2.1.2 Order of accuracy

Let us assume that u is a smooth solution of (1.1). We know that such solutions exist for rather general smooth data, at least locally in time (Theorem 1.1). Let us fix the grid ratio $\lambda = \Delta t / \Delta x$. We Taylor expand the following expression in terms of Δt :

$$u(x, t + \Delta t) - h(u(x - p\Delta x, t), \dots, u(x + q\Delta x, t)).$$

If u was also a solution of the difference scheme, this should be zero. Since the scheme is consistent and u is smooth, it is certainly an $O(\Delta t^2)$. We say that the scheme is of order ℓ at least if this expression is an $O(t^{\ell+1})$.

This notion of accuracy refers only to the approximation of smooth solutions. In presence of shocks, the situation is not so nice in practice. One observes that second-order or higher-order schemes generate wild oscillations around discontinuities, a kind of *Gibbs phenomenon*. For

this reason, second-order schemes are usually completed by *flux limiters* which have the role to cancel such oscillations. The price to pay is the loss of accuracy: the location of the shock waves is computed at first order only. Besides, flux limiters destroy the abstract form (2.3) and it becomes almost impossible to make a theoretical analysis of the scheme under consideration.

Numerical viscosity. Let us consider a first-order scheme in conservative form (2.3). If u is a smooth solution, we have

$$\begin{aligned} u(x, t + \Delta t) &= u(x, t) + \Delta t \partial_t u(x, t) + \frac{\Delta t^2}{2} \partial_t^2 u(x, t) + O(\Delta t^3) \\ &= u(x, t) - \Delta t \partial_x f(u(x, t)) \\ &\quad + \frac{\Delta t^2}{2} \partial_x (df(u(x, t)) \partial_x f(u(x, t))) + O(\Delta x^3). \end{aligned}$$

Likewise, we have

$$\begin{aligned} F_{j+1/2} - F_{j-1/2} &= F(u(x_{j+1-p}), \dots, u(x_{j+q})) - F(u(x_{j-p}), \dots, u(x_{j+q-1})) \\ &= \Delta x \sum_{k=1-p}^q d_k F(u, \dots, u) \partial_x u \\ &\quad + \Delta x^2 \sum_k (k - 1/2) d_k F(u, \dots, u) \partial_x^2 u \\ &\quad + \frac{\Delta x^2}{2} \sum_{k,l} (k + l - 1) D_{k,l}^2 F(u, \dots, u) (\partial_x u, \partial_x u) + O(\Delta x^3) \\ &= \Delta x df(u(x)) \partial_x u + \Delta x^2 \sum_k (k - 1/2) d_k F(u, \dots, u) \partial_x^2 u \\ &\quad + \frac{\Delta x^2}{2} \sum_{k,l} (k + l - 1) D_{k,l}^2 F(u, \dots, u) (\partial_x u, \partial_x u) + O(\Delta x^3), \end{aligned}$$

where we have denoted $d_k F$ the differential of $F(u_{1-p}, \dots, u_q)$ with respect to u_k , and we have used the identity

$$\sum_{k=1-p}^q d_k F(u, \dots, u) = df(u),$$

which follows from consistency.

In conclusion, we obtain

$$u(x, t + \Delta t) - h(u(x - p\Delta x, t), \dots, u(x + q\Delta x, t)) = \lambda \frac{\Delta x^2}{2} D + O(\Delta x^3)$$

with

$$\begin{aligned} D := & \lambda \partial_x (df(u) \partial_x f(u)) - \sum_k (2k-1) d_k F(u, \dots, u) \partial_x^2 u \\ & + \sum_{k,l} (k+l-1) D_{k,l}^2 F(u, \dots, u) (\partial_x u, \partial_x u). \end{aligned}$$

We use again consistency to obtain

$$\sum_{k,l=1-p}^q D_{k,l}^2 F(u, \dots, u) = D^2 f(u).$$

This allows us to simplify the formula for D :

$$\begin{aligned} D = & \partial_x \left(\lambda df(u) \partial_x f(u) + \sum_k (2k-1) d_k F(u, \dots, u) \partial_x u \right) \\ := & -\partial_x (B(u) \partial_x u). \end{aligned} \tag{2.5}$$

The tensor B , given by

$$B(u) = -\lambda df(u)^2 - \sum_k (2k-1) d_k F(u, \dots, u),$$

is called the *numerical viscosity*.

The fact that the scheme be of order one tells us that B does not vanish. We point out that if v is a smooth solution of the second-order system in conservation form

$$\partial_t v + \partial_x f(v) = \Delta x \partial_x (B(v) \partial_x v), \tag{2.6}$$

instead of (1.1), then

$$v(x, t + \Delta t) - h(v(x - p\Delta x, t), \dots, v(x + q\Delta x, t)) = O(\Delta x^3).$$

We point out that this v does depend on Δx and is expected to tend towards u as $\Delta x \rightarrow 0$. The numerical scheme thus approximates (2.6) in a better way than (1.1). One says that (2.6) is the *equivalent equation* of the difference scheme.

Recall that for the Cauchy problem for (2.6) being well-posed, one needs that the spectrum of $B(u)$ be of non-negative real part. This comes naturally as a necessary condition for the stability of the difference scheme.

2.1.3 Linearized L^2 -stability

When approximating smooth solutions, it is useful to make a linear analysis, by linearizing both the system (1.1) and the difference scheme. We are thus led to the study of the linear system

$$\partial_t u + A(\bar{u}) \partial_x u = 0, \quad (2.7)$$

together with the linear scheme

$$u_j^{m+1} = u_j^m + \lambda \sum_{k=1-p}^q d_k F(\bar{u}, \dots, \bar{u})(u_{j+k-1}^m - u_{j+k}^m). \quad (2.8)$$

The Cauchy problem for (2.7) is well-posed in every space $L^p(\mathbb{R})^n$ under hyperbolicity. Actually the system can be recast as a list of decoupled transport equations and the solution can be computed explicitly. The situation is not so nice in several space dimensions, where we do have well-posedness in L^2 , but not in L^p for $p \neq 2$. For this reason, we shall work only in L^2 within this section.

The scheme (2.8) defines a linear operator $S_\Delta : \ell^2 \rightarrow \ell^2$. The approximate solution at time t_m is given by

$$U^m = (S_\Delta)^m U^0.$$

Recall that the ratio λ is kept fixed. Let us say that $\Delta t = 1/N$ for some large integer N . Then the approximate solution at time $t = 1$ is $(S_\Delta)^N U^0$. If the difference scheme converges in this linear setting, then the Principle of Uniform Boundedness (sometimes called the Banach-Steinhaus Theorem) implies that $\|(S_\Delta)^N\|_{\mathcal{L}(\ell^2)}$ remains bounded as $N \rightarrow +\infty$. In other words, a necessary condition for convergence is the (linear) *stability*:

$$\exists C < \infty \quad \text{s.t.} \quad \|(S_\Delta)^N\|_{\mathcal{L}(\ell^2)} < C \quad \text{for} \quad N\Delta t \sim 1. \quad (2.9)$$

We notice that the stability of the scheme implies

$$\exists C < \infty \quad \text{s.t.} \quad \rho(S_\Delta) \leq 1 + C\Delta t, \quad (2.10)$$

as $\Delta t \rightarrow 0$, where ρ denotes the spectral radius of the linear operator S_Δ in ℓ^2 .

Thanks to linearity and translation invariance, we can perform a Fourier transform, a continuous one for (2.7) and a discrete one for (2.8):

$$\hat{U}^m(\xi) := \sum_{j \in \mathbb{Z}} e^{-ij\xi} u_j^m.$$

We find that

$$\hat{U}^m(\xi) = \sigma(\xi)^m \hat{U}^0, \quad \forall \xi \in \mathbb{R},$$

for some matrix $\sigma(\xi) \in \mathbf{M}_n(\mathbb{C})$ defined by

$$\sigma(\xi) = I_n + \lambda \sum_{k=1-p}^q \left(e^{i(k-1)\xi} - e^{ik\xi} \right) d_k F(\bar{u}, \dots, \bar{u}).$$

The linear stability thus amounts to saying that

$$\exists C < \infty \quad \text{s.t.} \quad \|\sigma(\xi)^N\|_{\ell^2} < C, \quad \forall \xi \in \mathbb{R} \quad \text{and} \quad N\Delta t \sim 1. \quad (2.11)$$

This requires

$$\exists C < \infty \quad \text{s.t.} \quad \rho(\sigma(\xi)) \leq 1 + C\Delta t, \quad \forall \xi \in \mathbb{R}. \quad (2.12)$$

It may happen that $\sigma(\xi)$ does not depend at all upon Δt , but only on the grid ratio λ . In this particular case (the Godunov scheme for instance), then the stability condition becomes that the set of matrices $\sigma(\xi)^m$ is uniformly bounded in $\xi \in \mathbb{R}/2\pi\mathbb{Z}$ and $m \in \mathbb{N}$. In this situation, (2.12) is equivalent to

$$\rho(\sigma(\xi)) \leq 1, \quad \forall \xi \in \mathbb{R}. \quad (2.13)$$

We warn the reader that (2.11) is a necessary and sufficient condition for linearized stability, but (2.12) or (2.13) is only a necessary condition.

Small frequencies. Let us Taylor expand $\sigma(\xi)$ about $\xi = 0$:

$$\sigma(\xi) = I_n - i\xi\lambda df(\bar{u}) + \lambda\xi^2 \sum_{k=1-p}^q (k-1/2)d_k F(\bar{u}, \dots, \bar{u}) + O(\xi^3). \quad (2.14)$$

The spectrum of $\sigma(\xi)$ obeys a Taylor expansion, which can be found from (2.14) by using the Implicit Function Theorem. For each j , there is a smooth eigenvalue $\xi \mapsto \Lambda_j(\xi)$, such that

$$\Lambda_j(\xi) = 1 - i\xi\lambda\lambda_j + \lambda\xi^2 \ell_j \left(\sum_{k=1-p}^q (k-1/2)d_k F(\bar{u}, \dots, \bar{u}) \right) r_j + O(\xi^3),$$

where we recall that $\ell_j(u)$ and $r_j(u)$ are the eigenform and the eigenvector of $df(u)$, respectively, associated to $\lambda_j(u)$ and normalized by $\ell_j r_j = 1$.

The modulus of Λ_j equals

$$1 + \xi^2 \left(\lambda^2 \lambda_j^2 + \lambda \sum_{k=1-p}^q (2k-1)\ell_j(\bar{u})d_k F(\bar{u}, \dots, \bar{u})r_j(\bar{u}) \right) + O(\xi^3).$$

The necessary condition (2.12) thus implies

$$\lambda \lambda_j^2 + \sum_{k=1-p}^q (2k-1) \ell_j(\bar{u}) d_k F(\bar{u}, \dots, \bar{u}) r_j(\bar{u}) \leq 0, \quad \forall j = 1, \dots, n. \quad (2.15)$$

This is equivalent to saying that

$$\ell_j(\bar{u}) B(\bar{u}) r_j(\bar{u}) \geq 0, \quad \forall j = 1, \dots, n. \quad (2.16)$$

Exercise. In general, the inequality (2.16) is independent of the positivity of the spectrum of $B(u)$. Prove however that this positivity implies (2.16) when both matrices $df(u)$ and $B(u)$ are symmetric. More generally, assume that (1.1) admits a strongly convex entropy η , and that B is η -dissipative in the sense of (1.28). Prove that $D^2\eta(r_j, \cdot) = D^2\eta(r_j, r_j)\ell_j$. Deduce that (2.16) holds true.

The asymptotic analysis at small frequency has the interest of relying the numerical viscosity B to the linearized stability of the scheme. However it is far from satisfactory in general. As we shall see in the examples below, the stronger constraints that stability imposes often come because of not-to-small values of ξ .

2.1.4 The Courant-Friedrichs-Lowy condition

Since λ is positive, the necessary condition (2.15) tells us that the sum over k has to be negative. Actually, it tells us more than that: λ has to be small enough. Imagine for instance that the derivatives $d_k F$ of the numerical flux are proportional to λ^{-1} . This sounds reasonable since both quantities have the dimension of a velocity. Then (2.15) provides an upper bound for λ , which is proportional to the inverse of the spectral radius $\rho(df(u))$. An explicit condition relating $df(u)$ and λ will be given on specific examples.

The condition (2.15), expressed as an upper bound of $\lambda = \Delta t / \Delta x$, is called the *Courant-Friedrichs-Lowy condition* (CFL). It tells us that Δt must be smaller than a given constant times Δx . When $\Delta t / \Delta x$ is too large, violent instabilities take place, in general in the form of wild oscillations. The numerical solution $u^{\Delta x}$ then does not converge at all and it is impossible to guess what the exact solution looks like.

A practical explanation of (CFL), viewed as a limitation of λ , is given by the propagation with finite velocity. Say for instance that there is a finite constant V such that $\rho(df(u)) < V$ for every relevant value u of the state. Let us consider an initial data a that is constant ($\equiv \bar{u}$) outside a compact interval $[-L, L]$. Then it can be proved that the solution of

the Cauchy problem has the property

$$\text{Supp}[u(\cdot, t) - \bar{u}] \subset [-L - Vt, L + Vt]. \quad (2.17)$$

This estimate is accurate in the sense that for most initial data of this form, the solution does vary (*i.e.* is not constant) in the domain $-L + t\lambda_1(\bar{u}) < x < L + t\lambda_n(\bar{u})$.

On another hand, the numerical solution is such that $u^{\Delta x}(x, t) \equiv \bar{u}$ when $x > L + pt/\lambda + O(\Delta t)$ and also when $x < -L - qt/\lambda + O(\Delta t)$. If $u^{\Delta x}$ is going to converge pointwise towards u , then one needs obviously that

$$-L - qt/\lambda \leq -L + t\lambda_1(\bar{u}), \quad L + t\lambda_n(\bar{u}) \leq L + pt/\lambda$$

for positive t , that is

$$\lambda\lambda_1(\bar{u}) \geq -q, \quad \lambda\lambda_n(\bar{u}) \leq p. \quad (2.18)$$

For instance, in the case of a three-point scheme ($p = q = 1$), one finds the well-known CFL condition

$$\lambda\rho(df(\bar{u})) \leq 1, \quad \forall \bar{u}. \quad (2.19)$$

2.1.5 Entropy-consistent schemes

Let us assume that the system (1.1) admits an entropy-flux pair (η, q) with as usual $D^2\eta > 0$. One may wonder whether the limit of approximate solutions, assuming that it exists, satisfies the entropy inequality (1.13). The way to ensure that is to ask the scheme to be *entropy-consistent*.

We say that a conservative difference scheme is entropy-consistent if there exists a numerical entropy flux $Q = Q(u_{1-p}, \dots, u_q)$, consistent in the sense that $G(u, \dots, u) \equiv q(u)$ and such that, whenever

$$v := u_0 + \lambda(F(u_{-p}, \dots, u_{q-1}) - F(u_{1-p}, \dots, u_q)),$$

one has

$$\eta(v) \leq \eta(u_0) + \lambda(G(u_{-p}, \dots, u_{q-1}) - G(u_{1-p}, \dots, u_q)). \quad (2.20)$$

This inequality is the discrete counterpart of (1.13). The Lax-Wendroff Theorem 2.1 extends to the context of entropy-consistent schemes, in the sense that if $u^{\Delta x}$ converges boundedly almost everywhere, then its limit is not only a weak solution, but it is an entropy solution: It satisfies (1.13) in the distributional sense.

Entropy consistency provides a non-linear form of stability. If a is constant ($\equiv \bar{u}$) outside of a compact interval, we may assume (up to the

addition of an affine function to η) that $\eta(\bar{u}) = 0$ and that η is positive otherwise. Then

$$\sum_{j \in \mathbb{Z}} \eta(u_j^0)$$

is finite. Because of (2.20), this remains true for the approximate solution:

$$\sum_{j \in \mathbb{Z}} \eta(u_j^0) \leq \sum_{j \in \mathbb{Z}} \eta(u_j^0) \leq \Delta x \int_{\mathbb{R}} \eta(a(x)) dx.$$

This provides an *a priori* estimate of $u^{\Delta x}$ in either some Lebesgue space $L^\infty(0, +\infty; L^p(\mathbb{R}))$ or an Orlicz space.

By a linearization procedure, it can be proved that an entropy-consistent scheme is linearly L^2 -stable. Such a scheme does have numerical viscosity: it cannot be second-order accurate or more.

2.2 Examples

2.2.1 The naive centered scheme

The simplest way to approximate (1.1) is to replace the time derivative by a backward difference and space derivative by a centered difference:

$$\partial_t u \mapsto \frac{u_j^{m+1} - u_j^m}{\Delta t}, \quad \partial_x f(u) \mapsto \frac{f(u_{j+1}^m) - f(u_{j-1}^m)}{2\Delta x}.$$

This yields a three-point scheme with the numerical flux

$$F(u_0, u_1) = \frac{1}{2}(f(u_0) + f(u_1)).$$

We find easily the numerical viscosity

$$B(u) = -\frac{1}{2}(df(u))^2.$$

Since $\ell_j Br_j = -\lambda_j^2/2$ is negative, the linearized stability condition is violated. The centered scheme suffers violent instabilities of Hadamard type, which make it useless in the approximation of the Cauchy problem. This instability was observed by von Neumann in the very first attempt to calculate solutions of gas dynamics.

Exercise. Let us approximate the time derivative by a centered difference too:

$$\partial_t u \mapsto \frac{u_j^{m+1} - u_j^{m-1}}{2\Delta t}.$$

This is the *leap-frog*^① scheme. Show that it is linearly stable under the CFL condition (2.19). You are warned that the leap-frog scheme involves two time steps instead of one. Thus you must rewrite it in the form

$$\begin{pmatrix} U^{m+1}(\xi) \\ U^m(\xi) \end{pmatrix} = \Sigma(\xi) \begin{pmatrix} U^m(\xi) \\ U^{m-1}(\xi) \end{pmatrix}.$$

2.2.2 The Lax-Friedrichs scheme

The Lax-Friedrichs scheme uses the approximation

$$\partial_t u \mapsto \frac{1}{\Delta t} \left(u_j^{m+1} - \frac{u_{j+1}^m + u_{j-1}^m}{2} \right),$$

still with the centered difference for space derivative. It thus writes

$$u_j^{m+1} = \frac{1}{2}(u_{j+1}^m + u_{j-1}^m) + \frac{\lambda}{2}(f(u_{j-1}^m) - f(u_{j+1}^m)).$$

It is a three-point scheme (although u_0 is not present in H_Δ) with numerical flux

$$F_{LF}(u_0, u_1) = \frac{1}{2\lambda}(u_0 - u_1) + \frac{1}{2}(f(u_0) + f(u_1)).$$

The numerical viscosity is

$$B_{LF}(u) = \frac{1}{2}(\lambda^{-2}I_n - (df(u))^2).$$

The stability condition (2.15) at small frequencies thus writes as (2.19).

Let us now investigate the more precise condition (2.13). We first compute the matrix $\sigma(\xi)$:

$$\sigma_{LF}(\xi) = (\cos \xi)I_n - i\lambda(\sin \xi)df(\bar{u}).$$

For the spectral radius of $\sigma_{LF}(\xi)$ to be less than one at every $\xi \in \mathbb{R}$, it is necessary and sufficient to have (2.15). Therefore this CFL condition ensures the linearized stability of the Lax-Friedrichs scheme.

In practice, one observes acceptable numerical results with the Lax-Friedrichs scheme under the CFL condition. The drawback is that the shock waves are smeared because of a rather high numerical viscosity. Understanding this phenomenon is one task of the next chapter.

The Lax-Friedrichs scheme is entropy-consistent, with the numerical entropy flux

$$Q_{LF} = \frac{1}{2\lambda}(\eta(u_0) - \eta(u_1)) + \frac{1}{2}(q(u_0) + q(u_1)).$$

^①The English translation for the French expression *sauté-mouton*, despite the facts that Brittons are fond of lamb meat and Frenchies are fond of frogs.

Exercise. Because the formula $u_j^{m+1} = h(u_{j-1}^m, u_{j+1}^m)$ does not involve u_j^m itself, one can express $u_j^{m+2} = \hat{h}(u_{j-2}^m, u_j^m, u_{j+2}^m)$. Show that \hat{h} defines a conservative difference scheme. Write explicitly its numerical flux \hat{F} , then the numerical viscosity \hat{B} .

2.2.3 The Lax-Wendroff scheme

The Lax-Friedrichs scheme is only first order, since its numerical viscosity is non-zero. We now present a second-order scheme, due to Lax and Wendroff. It has several variants, depending on the choice for $A_{j\pm 1/2}^m$ in the formula below:

$$\begin{aligned} u_j^{m+1} &= u_j^m + \frac{\lambda}{2}(f(u_{j-1}^m) - f(u_{j+1}^m)) \\ &\quad + \frac{\lambda^2}{2} \left(A_{j+1/2}^m(f(u_{j+1}^m) - f(u_j^m)) + A_{j-1/2}^m(f(u_{j-1}^m) - f(u_j^m)) \right). \end{aligned}$$

The purpose of the matrix $A_{j+1/2}^m$ is to approximate $df(u)$ at the grid point $(x_{j+1/2}, t_m)$. Convenient choices for $A_{j+1/2}^m$ are

$$df \left(\frac{u_{j+1}^m + u_j^m}{2} \right), \quad \frac{1}{2}(df(u_{j+1}^m) + df(u_j^m)), \quad A(u_j^m, u_{j+1}^m).$$

What we always need is that

$$(u_j = u_{j+1}) \implies (A_{j+1/2} = df(u_j)).$$

Exercise. Write the numerical flux F_{LW} . Check that the numerical viscosity vanishes identically. Interestingly enough, this property does not depend upon the choice of $A_{j\pm 1/2}^m$. At last, show that the Lax-Wendroff scheme is linearly stable under the CFL condition (2.19).

The Lax-Wendroff is not entropy-consistent, since the numerical viscosity vanishes identically.

2.2.4 The Godunov scheme

The Godunov scheme is a little bit more elaborated. To some extent, it is a finite volume scheme. Once $(u_j^m)_{j \in \mathbb{Z}}$ has been computed, one defines $u^{\Delta x}$ at time $t_m = m\Delta t$ by

$$u(x, t^m) := u_j^m \text{ over } (x_{j-1/2}, x_{j+1/2}).$$

Thus u is piecewise constant at time t^m . This allows us to solve explicitly the Cauchy problem for (1.1) over a time interval $(t^m, t^m + T)$ by

gluing the solutions of the Riemann problems between consecutive state (u_j^m, u_{j+1}^m) . These solutions agree as long as the waves emanating from the discontinuity do not reach the walls of the cell $(j\Delta x, (j+1)\Delta x)$. Since the waves typically travel at a velocity $\lambda_k(\bar{u})$ for some k and some \bar{u} , the time interval during which we may glue the Riemann problems is at least

$$\frac{\Delta x}{2 \max \rho(df(\bar{u}))}.$$

This can be improved, by remarking that we shall need only to know the values of the solution on the vertical lines defined by $x = x_j$ for $j \in \mathbb{Z}$. Thus it is sufficient that this value is unchanged, even though consecutive Riemann problems interact. It is thus enough that the waves emanating from the points $x_{j\pm 1}$ do not reach the line $x = x_j$. This must be true on a time interval twice as big as our first estimate above. We thus allow a time interval

$$\Delta t = \frac{\Delta x}{\max \rho(df(\bar{u}))}.$$

In other words, the Godunov scheme can be used under the CFL condition (2.19).

We now construct the values u_j^{m+1} . To do so, we consider the solution U^m constructed above, at time $t_{m+1} - 0$. Then we make an average in each cell $(x_{j-1/2}, x_{j+1/2})$:

$$u_j^{m+1} := \frac{1}{\Delta x} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} U^m(y, t_{m+1}) dy.$$

Integrating over the domain $(x_{j-1/2}, x_{j+1/2}) \times (t_m, t_{m+1})$ the conservation law (1.1) satisfied by U^m , we obtain

$$u_j^{m+1} = u_j^m + \lambda(F_{j-1/2}^m - F_{j+1/2}^m),$$

where the numerical flux is given by

$$F_{j+1/2}^m = F_G(u_j^m, u_{j+1}^m), \quad F_G(a, b) := f(R(a, b; x/t = 0)).$$

Hereabove, we have denoted $R(a, b; x/t)$ the solution of the Riemann problem between the left state a and the right state b . We recall that this solution depends only on x/t .

We point out that there may be an ambiguity in the value $R(a, b; 0)$, in the case where the Riemann problem admits a discontinuity across $x = 0$. However this is harmless because then the Rankine-Hugoniot condition $[f(u)] = 0$ tells us that the value $f(R(a, b; 0))$ is not ambiguous.

Stability analysis. At a linearized level, $F_G(a, b)$ is replaced by $f(\bar{u}) + df(\bar{u})_+(a - \bar{u}) + df(\bar{u})_-(b - \bar{u}) = df(\bar{u})_+a + df(\bar{u})_-b$, where $df(\bar{u})_{\pm}X$ are the projections of $df(\bar{u})X$ on the stable/unstable subspaces of $df(\bar{u})$. We therefore have $d_0 F_G(\bar{u}, \bar{u}) = df(\bar{u})_+$ and $d_1 F_G(\bar{u}, \bar{u}) = df(\bar{u})_-$. Whence the formula

$$\sigma_G(\xi) = I_n + \lambda(e^{-i\xi} - 1)df(\bar{u})_+ + \lambda(1 - e^{i\xi})df(\bar{u})_-.$$

The eigenvalues of $\sigma_G(\xi)$ are the numbers

$$1 + 2(1 - \cos \xi)\lambda|\lambda_j|(\lambda|\lambda_j| - 1), \quad j = 1, \dots, n.$$

One deduces that the condition (2.13) of linearized stability is equivalent to the CFL condition (2.19).

Once again, the Godunov scheme gives an acceptable approximation whenever the CFL condition holds for relevant states. Shocks are smeared because of the numerical viscosity. We observe however that the viscous tensor

$$B(u) = |df(u)|(I_n - \lambda|df(u)|) \quad (\text{with } |df(u)| := df(u)_+ - df(u)_-)$$

vanishes in the direction of the kernel of $df(u)$. This suggests that steady discontinuities are not smeared. This point will be studied in the next Chapter.

The Godunov scheme is entropy-consistent, with numerical entropy flux

$$Q_G = q(R(u_0, u_1; 0)).$$

We point out that this flux is not that well defined if the Riemann problem admits a steady discontinuity. In this case, Q_G should be multi-valued:

$$Q_G(u_0, u_1) = [q(R(u_0, u_1; 0+)), q(R(u_0, u_1; 0-))].$$

2.3 Schemes for scalar equations

We have seen in Section 1.2.7 that for scalar equations ($n = 1$), the Cauchy problem is well-posed in the L^∞ class. Actually, the estimate (1.16) shows that well-posedness holds true in $L^1(\mathbb{R})$ too. We thus have an alternative to the L^2 -theory. In particular, we may expect to work directly at the non-linear level.

The well-posedness of the Cauchy problem is a consequence of the following properties:

Comparison. If $a \leq b$ almost everywhere, then the entropy solutions u and v evolve accordingly: $u \leq v$,

Conservation. If $b - a \in L^1(\mathbb{R})$, then $v(\cdot, t) - u(\cdot, t) \in L^1(\mathbb{R})$ and

$$\int_{\mathbb{R}} (v(x, t) - u(x, t)) dx = \int_{\mathbb{R}} (b(x) - a(x)) dx, \quad \forall t > 0.$$

Exercise. Show that the comparison and conservation imply together the *contraction*: If $b - a \in L^1(\mathbb{R})$, then

$$\int_{\mathbb{R}} |v(x, t) - u(x, t)| dx \leq \int_{\mathbb{R}} |b(x) - a(x)| dx, \quad \forall t > 0.$$

At the discrete level, we should like to preserve the above properties of comparison and conservation. They imply a form of nonlinear stability. Together with the consistency, we expect that they yield convergence results.

2.3.1 Monotone schemes

We say that a conservative difference scheme for a scalar equation is *monotone* if it satisfies the following comparison principle: Given two approximate solutions, if $u_k^m \leq v_k^m$ for a given m and every k , then $u_j^{m+1} \leq v_j^{m+1}$. This amounts to saying that

H_Δ is a non-decreasing function of each of its arguments.

Exercise. Show that, under the CFL condition $\lambda|f'| < 1$, the Lax-Friedrichs and the Godunov schemes are monotone.

Exercise. Show that the Lax-Wendroff scheme cannot be monotone.

The drawback of monotone schemes is that they may not be high order accurate:

Proposition 2.2. *In the scalar case, consider a monotone scheme. Let us assume the consistency and the CFL condition (2.19).*

Then the scheme has positive numerical viscosity, unless the flux F depends either only on u_0 or only on u_1 , and the CFL condition is an equality.^①

Proof. Denoting $a_k := d_k F(u, \dots, u)$, the numerical viscosity is

$$b(u) = -\lambda f'(u)^2 - \sum_{1-p}^q (2k-1)a_k = -\lambda \left| \sum_{1-p}^q a_k \right|^2 - \sum_{1-p}^q (2k-1)a_k.$$

^①This combination of borderline properties is very much unlikely.

Because of monotonicity, one has

$$a_1 \leq a_2 \leq \cdots \leq a_q \leq 0 \leq a_{1-p} \leq \cdots \leq a_0, \quad 1 + \lambda(a_1 - a_0) \geq 0.$$

Finally, the consistency and the CFL condition write

$$\lambda \left| \sum_{1-p}^q a_k \right| \leq 1.$$

We thus have

$$b(u) \geq - \left| \sum_{1-p}^q a_k \right| - \sum_{1-p}^q (2k-1)a_k.$$

There remains to show that both

$$\pm \sum_{1-p}^q a_k - \sum_{1-p}^q (2k-1)a_k$$

are positive, unless the equality case. This follows from the expressions

$$\sum_{1-p}^q a_k - \sum_{1-p}^q (2k-1)a_k = 2 \sum_{1-p}^q (1-k)a_k$$

and

$$-\sum_{1-p}^q a_k - \sum_{1-p}^q (2k-1)a_k = -2 \sum_{1-p}^q k a_k,$$

where the right-hand sides are sums of non-negative terms.

The equality case is left to the reader. \square

2.3.2 Kuznetsov's error estimate

The monotone schemes have been widely used in the 1980's because of the following convergence result:

Theorem 2.3. *Let $a \in \bar{u} + L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ be a given initial data for a scalar conservation law (1.1). Let $u^{\Delta x}$ be an approximate solution, provided by a monotone difference scheme with a fixed grid ratio λ . We assume that the scheme is consistent with (1.1).*

Then $u^{\Delta x}$ converges boundedly almost everywhere towards the Kružkov solution u of the Cauchy problem.

Suppose in addition that the total variation of a is finite. Then one has the estimate

$$\|u^{\Delta x}(\cdot, t) - u(\cdot, t)\|_{L^1} \leq C\sqrt{t\Delta x} TV(a). \quad (2.21)$$

Comments.

- In the estimate above, the constant C depends only on the Lipschitz constant of f and on that of the numerical flux F .
- When a is not of bounded variations, one still has an estimate, but the right-hand side involves the modulus

$$\omega(h) := \|a(\cdot + h) - a\|_{L^1},$$

and the dependence in t and Δx depends on the behaviour of ω near the origin.

- If $f'' > 0$ (genuine nonlinearity, then $\sqrt{t\Delta x}$ can be replaced by $t^{1/4}\sqrt{\Delta x}$. Then the estimates is valuable whenever $t\Delta x \ll 1$.

3 Discrete shock profiles

So far, our analysis of the consistency of conservative difference schemes has been limited to the case where the underlying solution is smooth, thus slowly varying. This limit manifests itself in (2.4), where we underestimate that $u_{j-p}^n, \dots, u_{j+q}^n$ are close to a constant state v . This is by no means correct in presence of a jump. When the solution of the Cauchy problem admits a discontinuity from u_- to u_+ across a line $x = X(t)$, we expect that $u_{j-p}^n \sim u_-$ and $u_{j+q}^n \sim u_+$ when the mesh $(n\Delta t, j\Delta x)$ is somewhere close to the shock front. In this situation, we need another notion of consistency, which tells us how the approximate solution varies in a region of width $O(\Delta x)$ around the front. The appropriate concept is that of *Discrete shock profile*.

3.1 DSPs and conservation

We wish to mimic the notion of viscous shock profile. Let $(u_-, u_+; s)$ be a triplet satisfying (1.10). A discrete shock profile would be a travelling wave

$$U\left(\frac{x-st}{\epsilon}\right), \quad U(\pm\infty) = u_\pm,$$

defined at every grid point $(t, x) = (n\Delta t, \Delta x)$. Therefore the argument $(x-st)/\epsilon$ runs over the additive subgroup $\mathbb{Z} + \eta\mathbb{Z}$, where we have defined

$$\eta := s \frac{\Delta t}{\Delta x} = s\lambda.$$

In addition, the wave has to be an exact solution of the numerical scheme (as a viscous profile was an exact solution of the parabolic conservation law (1.15)). Whence the definition:

Definition 3.1. A *Discrete shock profile* (DSP) is a function

$$U : \mathbb{Z} + \eta\mathbb{Z} \rightarrow \mathcal{U},$$

satisfying

$$\lim_{y \rightarrow \pm\infty} U(y) = u_{\pm}$$

and the *profile equation*

$$\begin{aligned} U(y - \eta) &= U(y) + \lambda \{ F(U(y - p), \dots, U(y + q - 1)) \\ &\quad - F(U(y - p + 1), \dots, U(y + q)) \}, \quad \forall y \in \mathbb{Z} + \eta\mathbb{Z}. \end{aligned} \quad (3.1)$$

We immediately see a new feature in this beginning theory. The subgroup $\mathbb{Z} + \eta\mathbb{Z}$ can be discrete (when η is rational) or dense (when η is irrational). Thus one speaks of the *rational case* and of the *irrational case*.

Rational case. When $\eta \in \mathbb{Q}$, we write $\eta = \frac{m}{\ell}$ as an irreducible fraction. Then the domain $\mathbb{Z} + \eta\mathbb{Z}$ equals $\ell^{-1}\mathbb{Z}$. We ask that the profile equation (3.1) be satisfied for every $y \in \ell^{-1}\mathbb{Z}$. We shall often treat this equation as a dynamical system for a diffeomorphism.

Irrational case. When $\eta \notin \mathbb{Q}$, the domain is dense in \mathbb{R} . It becomes natural to look for a *continuous* travelling wave U , defined over the whole line \mathbb{R} . Then the equation (3.1) has to be satisfied for every $y \in \mathbb{R}$. This is a much more difficult situation from the analytical point of view.

We notice that we pass continuously from one case to the other one. We may either keep the triplet $(u_-, u_+; s)$ fixed and let vary the aspect ratio λ of the grid, or keep the grid fixed and let the triplet vary, as it would happen generically along a curved shock. We therefore expect that the qualitative results be the same in both cases. Sadly, we shall see that this is not true at all.

The profile equation can be integrated once, as in the viscous case, because of conservativeness. For instance, in the irrational case, we have

$$U(y) - U(y - \eta) = \frac{d}{dy} \int_{y-\eta}^y U(\xi) d\xi,$$

while

$$\begin{aligned} F(U(y - p + 1), \dots, U(y + q)) - F(U(y - p), \dots, U(y + q - 1)) \\ = \frac{d}{dy} \int_y^{y+1} F(U(\xi - p), \dots, U(\xi + q - 1)) d\xi. \end{aligned}$$

The profile equation is thus equivalent to the consistency of

$$y \mapsto \int_{y-\eta}^y U(\xi) d\xi - \lambda \int_y^{y+1} F(U(\xi-p), \dots, U(\xi+q-1)) d\xi.$$

From the condition at infinity, together with (2.4), we see that this constant must equal both of $\eta u_\pm - \lambda f(u_\pm)$. This in turn implies the Rankine-Hugoniot condition. In conclusion, the Rankine-Hugoniot condition is a necessary condition for the existence of a DSP (as it was for the existence of a viscous profile). Finally, we have the integrated equation

$$\int_{y-\eta}^y U(\xi) d\xi - \lambda \int_y^{y+1} F(U(\xi-p), \dots, U(\xi+q-1)) d\xi = \lambda(su_- - f(u_-)).$$

Exercise. Assume that the scheme is entropy-consistent. Prove that (1.13) is a necessary condition for the existence of a DSP.

In the rational case, the integrated profile equation involves finite sums instead of integrals. We leave it as an exercise.

3.1.1 The function Y

Let U be a DSP. In the rational case, we allow the domain of definition to be an arbitrary set \mathcal{D} with the property that $\mathcal{D} + \ell^{-1}\mathbb{Z} = \mathcal{D}$. We sometimes write $(U; \mathcal{D})$ to recall what is the domain of definition of U . In this case, each of the restriction of U to translates $d + \ell^{-1}\mathbb{Z}$, where $d \in \mathcal{D}$ is given, is a discrete shock profile. Taking two base points d, d' that are not congruent modulo ℓ^{-1} amounts to comparing two distinct DSPs for the same shock. We shall see later on that this is relevant at least for small Lax shocks.

Let us define the following function

$$Y(x; h) := \sum_{y \in x + \mathbb{Z}} (U(y+h) - U(y)), \quad \forall x, x+h \in \mathcal{D}.$$

Hereabove we assume that U has finite total variation in order to ensure the convergence of the series. This is true for instance in the scalar case, where one can prove the existence of a monotonous DSP with domain of definition $\mathcal{D} = \mathbb{R}$; see below.

Of course, the series

$$\sum_{y \in x + \mathbb{Z}} |U(y)|, \quad \sum_{y \in x + \mathbb{Z}} |U(y+h)|$$

do not converge, especially because $u_+ \neq u_-$. Therefore one must take care with resummation. For instance,

$$\begin{aligned} Y(x; h+1) - Y(x; h) &= \sum_{y \in x+\mathbb{Z}} (U(y+h+1) - U(y+h)) \\ &= \sum_{y \in x+h+\mathbb{Z}} (U(y+1) - U(y)) = u_+ - u_-! \end{aligned}$$

This identity does not involve the fact that U is a DSP. It only uses the limits of U at $\pm\infty$. On the contrary, the identity below, which at first glance looks very similar to the former, is a consequence of (3.1):

$$\begin{aligned} Y(x; h-\eta) - Y(x; h) &= \sum_{y \in x+\mathbb{Z}} (U(y+h-\eta) - U(y+h)) \\ &= \sum_{y \in x+h+\mathbb{Z}} (G(y+h) - G(y+h+1)), \end{aligned}$$

where

$$G(y) := \lambda F(U(y-p), \dots, U(y+q-1)).$$

We therefore have

$$\begin{aligned} Y(x; h-\eta) - Y(x; h) &= G(-\infty) - G(+\infty) = \lambda(f(u_-) - f(u_+)) \\ &= \eta(u_- - u_+), \end{aligned}$$

where we have used (2.4) and (1.10).

Combining the above calculations, we deduce

Theorem 3.2. *Let us assume that the DSP $U : \mathcal{D} \rightarrow \mathcal{U}$ has bounded variations. Then the function Y satisfies*

$$Y(x; h) - Y(x; k) = (h - k)[u], \quad \forall h, k \in \mathcal{D}.$$

In particular, if $\mathcal{D} = \mathbb{R}$ (for instance in the irrational case), we have

$$Y(x; h) = h[u].$$

3.1.2 Scalar case: monotone schemes

We consider in this paragraph the situation for a scalar equation, for which we employ a monotone scheme. For instance, it could be the Lax-Friedrichs scheme or the Godunov scheme, under the CFL condition. It is not difficult to show that a monotone scheme satisfies a comparison principle, as well as L^1 -contraction: If (u_j^n) and (v_j^n) are approximate solutions governed by the scheme, we have

- If $u_j^n \leq v_j^n$ for all $j \in \mathbb{Z}$, then $u_j^{n+1} \leq v_j^{n+1}$.

- If $u^n - v^n \in \ell^1$, then

$$\sum_{j \in \mathbb{Z}} |v_j^{n+1} - u_j^{n+1}| \leq \sum_{j \in \mathbb{Z}} |v_j^n - u_j^n|.$$

These properties have the consequences that whenever U is a discrete shock profile (with \mathcal{D} minimal), it is monotonous. Using then the function Y , we deduce

Theorem 3.3. *Consider a monotone conservative scheme for a scalar conservation law. Then every DSP (with minimal domain) is monotonous and Lipschitz continuous:*

$$|U(y) - U(z)| \leq |(y - z)[u]|.$$

The existence of a DSP for a given shock can be obtained in two steps. When dealing with a strictly monotone scheme in an interval that contains u_\pm , ordering arguments were employed by G. Jennings [13] to prove that every Lax shock with a rational η admits a one-parameter family of DSPs, whatever the strength $|u^r - u^l|$ of the shock. For every u^* taken in (u^l, u^r) (or (u^r, u^l)), there exists a unique DSP with $u_0 = u^*$. As mentioned above, this DSP is itself strictly monotone.

Jennings claimed that this result could be extended to irrational values of η . However, his density argument did not contain any detail. The question has been therefore considered as open for a long time. The gap was filled recently in [25], using the function Y described above. The Lipschitz estimate in Theorem 3.3 provides a compactness argument. This method allows to relax also the strict monotonicity. In particular, it handles the case of the Godunov scheme, for which Jennings' proof was powerless even in the rational case. We now have an as general as possible existence and uniqueness theorem, since only the monotonicity of the scheme over (u^-, u^+) is required. The final result is

Theorem 3.4. *Let us consider a scalar conservation law and assume that the conservative difference scheme is monotone (not necessarily strictly) in some interval I . Then every shock $(u^-, u^+; s)$ satisfying the Oleinik condition with strict inequalities admits a continuous DSP, defined on the whole line \mathbb{R} .*

In the rational case, this result tells us that the shock admits a continuum of DSPs. For this reason, we often speak of *continuous* DSPs; even though this terminology is a bit paradoxical.

An explicit DSP. In general, it is rather difficult to provide DSPs in closed form. However there is a special case where this is possible. Let

us consider the scalar equation

$$\partial_t u + \partial_x f(u) = 0, \quad f(u) := -\frac{2}{\lambda} \log \cosh \frac{u}{2},$$

which we approximate through the Lax-Friedrichs scheme. Using the Hopf-Cole transformation, P. Lax found the following formula for the DSP:

$$U(y) = \log \frac{a^{y+1} + 1}{a^{y-1} + 1},$$

where a is the unique root of

$$2a^{-\eta} = a^{-1} + a, \quad a \neq 1.$$

This travelling wave connects the states $u = 0$ and $u = 2 \log a$, in an order that depends on the position of a with respect to 1, that is of the sign of s (still, $\eta := s\lambda$). With the addition of a constant to U , this formula provides a DSP for every Oleinik shock of our conservation law.

We point out that the domain of definition of U is the whole line, as expected from Theorem 3.4. This justifies the fact that, even in the rational case, we look for continuous DSPs, at least in the case of Lax shocks.

3.2 Existence theory for rational η

We discuss in this section the tools for the existence of DSPs in the rational case. They follow the ideas used for VSPs, borrowed from dynamical systems theory. The main modification is that instead of working with vector fields, we work with diffeomorphisms. Thus the relevant theory is that of discrete dynamical systems.

For this procedure to apply, we need that the numerical flux be an invertible function of its extreme arguments. This works for instance for the Lax-Friedrichs scheme, but not for the Godunov scheme.

For a short account of the Center Manifold Theorem, taylored for its application to bifurcation analysis, we refer to [5].

3.2.1 DSPs for small steady Lax shocks

For the sake of simplicity, we assume a three-point scheme. As mentioned above, our flux $F(a, b)$ is invertible with respect to both a and b . The profile equation is integrated once. If $\eta = m/\ell$, this yields

$$\begin{aligned} \sum_{j=0}^{m-1} U\left(y - \frac{j}{\ell}\right) - \lambda \sum_{k=1}^{\ell} F\left(U\left(y - 1 + \frac{k}{\ell}\right), U\left(y + \frac{k}{\ell}\right)\right) \\ = m(u_- - sf(u_-)). \end{aligned}$$

By the CFL condition, we know that $|m| < \ell$, and therefore this equation is equivalent to an induction of the form

$$U(y+1) = \phi \left(U(y-1), \dots, U \left(y + \frac{\ell-1}{\ell} \right) \right).$$

Expanding our unknown as

$$V(y) := \left(U(y-1), \dots, U \left(y + \frac{\ell-1}{\ell} \right) \right),$$

our problem can be recast as finding a heteroclinic orbit from $V_- := (u_-, \dots, u_-)$ to $V_+ := (u_+, \dots, u_+)$ of a dynamical system

$$V \left(y + \frac{1}{\ell} \right) = \Phi(V(y)). \quad (3.2)$$

As in Section 1.5, we have to let the triplet $(u_-, u_+; s)$ vary. However, we need that the integrated profile equation keep a fixed form. In particular, we want to keep a same size ℓn of the unknown. For this reason, we ask that η remains constant. This can be achieved in two ways: Either one keeps a fixed grid and let vary u_- and u_+ *simultaneously*, in such a way that the shock velocity s remains constant. Or one keeps u_- fixed, let vary u_+ and choose the grid ratio according to $\lambda := m/s\ell$. In both cases, we start from a triplet $(u_-, u_-; \lambda_k(u_-))$. When applying a center manifold theorem, we need that the differential of the diffeomorphism (extended to a larger unknown, say (V, u_+)) have a well-identified eigenspace associated to the eigenvalue $\mu = 1$, and no over eigenvalue on the unit circle (this last part is called *non-resonance*). Then the dynamics reduces to a simpler dynamics over the center manifold.

For a Lax shock, this manifold is of dimension $n+1$. Each line of equation $u_+ = \text{cst}$ is invariant under the dynamics. We can therefore reduce the analysis to such lines, on which we find two fixed points, corresponding to the Hugoniot triplets $(u_-, u_+; s)$ and $(u_+, u_+; \lambda_k(u_+))$. The restriction of the diffeomorphism over such a line $\gamma(u_+)$ preserves the orientation. Therefore every point $V_0 \in \gamma(u_+)$ between the fixed points serves to build a DSP, through $V(j) := \Phi^{(j)}(V_0)$. We obtain in this way a continuum of DSPs. In other words, we are in presence of a ‘continuous’ DSP. Mind that such a continuous DSP is far from unique, even up to a shift. For if $\rho = \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing and such that $\rho(y+1/\ell) = \rho(y) + 1/\ell$, and if U is a continuous DSP, then $U \cdot \rho$ is an other one. In practice, we do not distinguish both.

In the construction above, the Lax shock inequalities (1.17, 1.18) guarantee that the profile goes from u_- to u_+ , as in the viscous case. For a complete proof of the following result, we refer to [26]. The result itself is due to Majda and Ralston [19]. See also [20].

Theorem 3.5. Assume that the numerical flux be invertible to its extreme arguments. Let $u_- \in \mathcal{U}$ and λ_0 be given, in such a way that the k -th characteristic field be GNL at u_- , and $\eta := \lambda_k(u_-)\lambda$ be rational, $\eta = m/\ell$. We also assume some linear stability and a non-resonance condition for the scheme (details are omitted).

Then there exist two neighbourhoods $\mathcal{V} \subset \mathcal{V}_1$ of u_- such that, for every entropy-admissible shock $(u_l, u_r; s)$ with $u_{l,r} \in \mathcal{V}$, and every grid ratio λ such that $s\lambda = m/\ell$, there exists a continuous DSP with values in \mathcal{V}_1 . In addition this DSP is unique, up to transformations ρ as described above.

Remarks.

- The smallness assumption, represented by the neighbourhood \mathcal{V} , depends dramatically upon the denominator ℓ . This set shrinks to $\{u_-\}$ as $\ell \rightarrow +\infty$. Therefore this theorem cannot be used to prove an existence result for irrational η 's by passing to the limit from rationals to irrationals.
- Amazingly, the non-resonance condition is not satisfied by the Lax-Friedrichs scheme, for an obvious reason: this scheme acts on the grid points with $j+n$ even on the one hand, and on the grid points with $j+n$ odd on the other hand, independently. To get rid of this decoupling, we can iterate twice the scheme and then restrict to the coarser grid of points with j, n even. This new grid has time and space lengths $2\Delta t$ and $2\Delta x$, respectively (λ is thus unchanged). The scheme then rewrites

$$u_j^{n+2} = u_j^n + \lambda(F(u_{j-2}^n, u_j^n) - F(u_j^n, u_{j+2}^n))$$

with numerical flux

$$\begin{aligned} F_{LF2}(a, b) &= \frac{1}{4\lambda}(a - b) + \frac{1}{4}(f(a) + f(b)) \\ &\quad + \frac{1}{2}f\left(\frac{a+b}{2} + \frac{\lambda}{2}(f(a) - f(b))\right). \end{aligned}$$

For reasonable fluxes, this new scheme is non-resonant.

Other shocks. For shocks with larger amplitude, we cannot write such a general result. What we can do is to figure out the shape of the intersection of the stable manifold of (3.2) at V_+ , and its unstable manifold at V_- . While in the case of a Lax shock the sum of their dimensions exceeds by one the dimension $N = \ell n$ of the ambient space, it equals N for undercompressive shock, as defined by (1.32). Since these manifolds

do not need to have a common tangent vector (contrary to the viscous case), the generic picture is that they intersect transversally along a discrete set. Therefore the situation for under-compressive shocks is completely different from the viscous case: – on the one hand the existence of DSPs is generic instead of exceptional of codimension one, – on the other hand the DSP is not ‘continuous’, but genuinely discrete. It was shown actually in [22] that the number of ‘distinct’ DSP for a shock is even. In particular, it is not unique.

3.2.2 DSPs for steady Lax shocks: the Godunov scheme

The procedure described above does not work for the Godunov scheme because its numerical flux is not invertible with respect to either of its arguments. For non-stationary shocks, the existence of DSPs is an open problem, except in the scalar case where we have Theorem 3.4. However, the case of steady shocks can be treated explicitly. In particular, there is no need of a smallness assumption. Given a steady shock $(u_-, u_+; s = 0)$, we shall make two natural assumptions:

- The Riemann problem with a constant initial data $u(t = 0, x) \equiv a$ admits only the constant solution $u \equiv a$. We point out that this assumption is a direct consequence of the entropy inequality (1.13) if there is a convex entropy.
- The equation $f(v) = f(u_-)$ has only the two solutions u_- and u_+ .

The profile equation reduces to $F_{God}(u_j, u_{j+1}) = f(u_-)$, that is $f(R(u_j, u_{j+1}; 0)) = f(u_-)$. By assumption, this means that

$$u_{j+1/2} := R(u_j, u_{j+1}; 0) \in \{u_-, u_+\}, \quad \forall j \in \mathbb{Z}.$$

Lemma 3.6. *If $u_{j-1/2} = u_+$, then $u_{j+1/2} = u_+$. Equivalently, if $u_{j+1/2} = u_-$, then $u_{j-1/2} = u_-$.*

Proof. We proceed *ad absurdum*. Let us assume that $u_{j-1/2} = u_+$ and $u_{j+1/2} = u_-$. This means on the one hand that one can pass from u_+ to u_j with only forward waves (*i.e.* waves with positive velocities); we denote by W this self-similar solution. On the other hand, we pass from u_j to u_- by backward waves; we denote by Z this self-similar solution. Then we construct a solution of the Riemann problem between u_j and itself, using first W , then the steady shock $u_- \mapsto u_+$, and finally Z . This contradicts our uniqueness assumption for $a = u_j$. \square

Lemma 3.6 amounts to saying that there exists an index j_0 such that if $j \leq j_0$, then $u_{j-1/2} = u_-$, while if $j \geq j_0$, then $u_{j+1/2} = u_+$.

Lemma 3.7. *If $j < j_0$, then $u_j = u_-$, while if $j > j_0$, then $u_j = u_+$.*

Proof. If $j < j_0$, then $u_{j \pm 1/2} = u_-$. This means that we can pass from u_- to u_j by forward waves, and from u_j to u_- by backward waves. By the uniqueness assumption, we deduce that the Riemann problem between u_j and itself passes through u_- , and therefore $u_j = u_-$, by uniqueness. \square

There remains to identify u_{j_0} . Since $u_{j_0 - 1/2} = u_-$ and $u_{j_0 + 1/2} = u_+$, we can pass from u_- to u_{j_0} by forward waves, and from u_{j_0} to u_+ by backward waves. The first property is written $u_{j_0} \in \mathcal{W}_+^f(u_-)$, where f means *forward*, and the subscript $+$ means that u_- is at right (!!) of u_{j_0} in this Riemann problem. Likewise, the second property is written $u_+ \in \mathcal{W}_+^b(u_{j_0})$, or equivalently $u_{j_0} \in \mathcal{W}_-^b(u_+)$. The set of DSPs for the steady shock $(u_-, u_+; 0)$ is thus parametrized by a pair (j_0, a) where $j_0 \in \mathbb{Z}$ and a is any point of the intersection

$$\Lambda := \mathcal{W}_+^f(u_-) \cap \mathcal{W}_-^b(u_+).$$

For a Lax shock $(u_-, u_+; 0)$, this intersection is usually a curve with end points u_\pm . We warn the reader that the pairs (j_0, u_-) and $(j_0 + 1, u_+)$ define the same DSP. Therefore the set of DSPs is again a one-parameter set, an infinite ‘periodic’ curve, smooth away from the points (j_0, u_-) .

For instance, let us consider an extreme steady shock, say an n -shock $(u_-, u_+; 0)$. Then every velocity $\lambda_k(u_+)$ are negative. Therefore $\mathcal{W}_-^b(u_+)$ is a neighbourhood of u_+ ; it is a ‘half-space’, bounded by the set of states a such that the Riemann problem between a and u_+ is such that $u \equiv a$ precisely on $x < 0$ and only there. In particular, u_- is a boundary point of $\mathcal{W}_-^b(u_+)$. On the other hand, all velocities $\lambda_k(u_-)$ but the last one are negative. Therefore $\mathcal{W}_+^f(u_-)$ is the part of the n -th wave curve of u_- , made of states for which the n -wave is entirely a forward wave. Typically, it is a ‘half’ of the n -th wave curve, bounded by u_+ since the wave between u_- and u_+ is precisely a steady wave. It is clear on this example that γ reduces to the segment of the n -th wave curve of u_- , of extremities u_- and u_+ .

3.2.3 What can go wrong?

If the realm of DSPs was a perfect world, then there would be some kind of well-posedness, with the following properties:

- For every small Lax shock of a GNL field, there would exists a unique (modulo a diffeomorphism ρ if η is rational) continuous DSP, no matter whether η is rational or not,
- This DSP would be absolutely continuous. In particular, it would have bounded variations,

- As the shock data $(u_-, u_+; s)$ and the grid ratio λ vary, the DSP would vary smoothly.

If all this is true, the function Y varies smoothly with the shock data and λ , thus with η . In particular, Theorem 3.2 would extend to the rational case since irrational numbers are dense. In passing, this would fix the DSP up to a translation. However, the identity $Y(x; h) = h[u]$ would mean that given two genuinely discrete shock profiles U and V over the domain $\ell^{-1}\mathbb{Z}$, the sum of the series

$$\sum_{j \in \mathbb{Z}} (U(j) - V(j)) \quad (3.3)$$

is parallel to $u_+ - u_-$. In practice, there is no reason why this should be true, and it is not too difficult to build counter-examples.

For instance, let us consider the Godunov scheme with a steady shock (thus $\ell = 1$). We have described the profiles in the previous paragraph. The profiles U and V are characterized respectively by the pairs (j_0, a) and (j_1, b) with $j_0, j_1 \in \mathbb{Z}$ and $a, b \in \Lambda$. Then the sum above equals $(j_1 - j_0)[u] + a - b$. If this was parallel to $[u]$ for every choice of U and V , then Λ would be the straight segment $[u_-, u_+]$. It is easy to see that in most cases, this is false. For instance, it fails in gas dynamics, where every shock is an extreme shock. As explained above, Λ is the segment between u_- and u_+ in a wave curve, and this wave curve is never a straight line.

In conclusion, something must go wrong in the theory of DSPs. Either there are some small shocks for which no DSP exists. Or DSPs exist but they lack regularity. Actually, Bressan & Coll. constructed [2, 3] examples where the tail of a DSP oscillates so much that the profile has unbounded variations. Or the DSPs do not depend smoothly enough on the shock data and the grid ratio.

Special cases. After this discussion, one may wonder why everything can go as best as possible in the scalar case. Theorem 3.4 tells us that the continuous DSP always exists, that it is monotone and thus of bounded variation, and uniformly Lipschitz. This seems to contradict the analysis made here, but there is no contradiction at all. In the scalar case, any two numbers are parallel vectors!

Something similar might happen for systems under the following circumstances:

- Every component f_j of the flux, but one, is linear. This applies for instance to the so-called p -system

$$\partial_t u_1 + \partial_x u_2 = 0, \quad \partial_t u_2 + \partial_x p(u_1) = 0,$$

or to the full gas dynamics with the equation of state $p = 2\rho e$ (meaning $\gamma = 3$).

- The scheme is Lax-Friedrichs.

It turns out that under these assumptions, the sum (3.3) for two DSPs is automatically parallel to $[u]$. Thus there is no obstruction to a well-posedness theory for DSPs. This theory remains however fully open.

References

- [1] V.I. Arnold. Geometrical methods in the theory of ordinary differential equations. Grundlehren der mathematischen Wissenschaften **250**. Springer-Verlag, New York (1983).
- [2] P. Baiti, A. Bressan, H.-K. Jenssen. Instability of travelling wave profiles for the Lax-Friedrichs scheme. *Discrete Contin. Dyn. Syst.*, **13** (2005), 877–899.
- [3] P. Baiti, A. Bressan, H.-K. Jenssen. An instability of the Godunov scheme. *Comm. Pure Appl. Math.*, **59** (2006), 1604–1638.
- [4] S. Benzoni-Gavage, D. Serre. Multi-dimensional hyperbolic partial differential equations. First-order systems and applications. Oxford Univ. Press (2007). Oxford, UK.
- [5] A. Bressan. Tutorial on the Center Manifold Theorem. Hyperbolic systems of balance laws. CIME cours (Cetraro 2003). Springer Lect. Notes in Maths **1911**, 327–344. Springer-Verlag, Heidelberg (2007).
- [6] M. Bultelle, M. Grassin, D. Serre. Unstable Godunov discrete profiles for steady shock waves. *SIAM J. Numer. Anal.*, **35** (1998), 2272–2297.
- [7] C. Dafermos. Hyperbolic conservation laws in continuum physics. Grundlehren der mathematischen Wissenschaften, **325**. Springer-Verlag (2000). Heidelberg.
- [8] B. Fiedler, J. Scheurle. Discretization of homoclinic orbits, rapid forcing and “invisible chaos”. *Memoirs of the Amer. Math. Soc.*, **119** (1996), no. 570.
- [9] E. Godlewski, P.-A. Raviart. Numerical approximations of hyperbolic systems of conservation laws. Springer-Verlag, New York (1996).
- [10] S. Godunov. A difference scheme for numerical calculations of discontinuous solutions of the equations of hydrodynamics. *Math. Sb.*, **47** (1959), 271–306.

- [11] H. Fan. Existence and uniqueness of travelling waves and error estimates for Godunov schemes of conservation laws. *Math. Computation*, **70** (1998), 87–109.
- [12] H. Fan. Existence of discrete shock profiles of a class of monotonicity preserving schemes for conservation laws. *Math. Computation*, **67** (2000), 1043–1069.
- [13] G. Jennings. Discrete shocks. *Comm. Pure Appl. Math.*, **27** (1974), 25–37.
- [14] P.D. Lax. Hyperbolic systems of conservation laws (II). *Comm. Pure Appl. Math.*, **10** (1957), 537–566.
- [15] R.J. Leveque. Numerical methods for conservation laws. Birkhäuser, Basel (1990).
- [16] J.-G. Liu, Z. Xin. L^1 -stability of stationary discrete shocks. *Math. Comput.*, **60** (1993), 233–244.
- [17] J.-G. Liu, Z. Xin. Nonlinear stability of discrete shocks for systems of conservation laws. *Arch. Rational Mech. Anal.*, **125** (1994), 217–256.
- [18] T.-P. Liu, H.-S. Yu. Continuum shock profiles for discrete conservation laws. *Comm. Pure Appl. Math.*, **52** (1999), I. Construction, 85–127 & II. Stability, 1047–73.
- [19] A. Majda, J. Ralston. Discrete shock profiles for systems of conservation laws. *Comm. Pure Appl. Math.*, **32** (1979), 445–482.
- [20] D. Michelson. Discrete shocks for difference approximations to systems of conservation laws. *Adv. Appl. Math.*, **5** (1984), 433–469.
- [21] D. Serre. Remarks about the discrete profiles of shock waves. *Matemática Contemporânea*, **11** (1996), 153–170.
- [22] D. Serre. Discrete shock profiles and their stability. *Hyperbolic problems: Theory, Numerics and Applications*, Zurich 1998. M. Fey, R. Jeltsch eds. ISNM **130**, Birkhäuser (1999), 843–854.
- [23] D. Serre. Systems of conservation laws, I. Cambridge Univ. Press. Cambridge (1999).
- [24] D. Serre. Systems of conservation laws, II. Cambridge Univ. Press. Cambridge (2000).
- [25] D. Serre. L^1 -stability of nonlinear waves in scalar conservation laws. *Handbook of Differential Equations*, C. Dafermos, E. Feireisl editors. North-Holland, Amsterdam, 2004.
- [26] D. Serre. Discrete shock profiles: Existence and stability. *Hyperbolic systems of balance laws*. CIME cours (Cetraro 2003). Springer Lect. Notes in Maths **1911**, 79–158. Springer-Verlag, Heidelberg (2007).

- [27] L. Ying. Asymptotic stability of discrete shock waves for the Lax–Friedrichs scheme to hyperbolic systems of conservation laws. Japan J. Indus. Appl. Math., **14** (1997), 437–468.

Kinetic Theory and Conservation Laws: An Introduction

Seiji Ukai

*Department of Mathematics and
Liu Bie Ju Center for Mathematical Sciences
City University of Hong Kong, China
E-mail: mcukai@cityu.edu.hk*

Tong Yang

*Department of Mathematics, City University of Hong Kong
and
Department of Mathematics, Shanghai Jiao Tong University, China
E-mail: matyang@cityu.edu.hk*

Abstract

Both the areas of the kinetic equations and conservation laws are extremely large and contain enormous theories and challenging open problems. Based on this thinking, this notes serves only as a short course aiming at some particular parts in these two areas. In other words, many interesting and important theories and phenomena in these two areas are not covered here. However, we will try to make this notes to be self-contained and cover some selected topics according to the authors' interest and to the length of this course. Therefore, it can be viewed as an introduction to these two areas and somehow it is also expected to lead the readers to the frontier of the research in these areas.

1 Introduction

1.1 Overview

At the normal atmospheric pressure and 0°C temperature, there are about 2.7×10^{19} molecules in a cube of 1cm^3 . Since each molecule depends on seven independent variables in physical space which represent the time, location and velocity, it needs independent variables of order 10^{20} to determine the exact behaviour of the molecules. Hence, in reality, it is impossible to do this in a deterministic way. And the Boltzmann equation is built to serve this purpose through a statistical thinking. In

fact, there are two keystones of the kinetic theory but were established independently and on different physical principles: J. C. Maxwell [73] discovered his distribution function in 1858 based on the statistical argument on the equi-partition of the kinetic energy of gas particles, while it was in 1872 when L. Boltzmann [7] established his equation based on the Newtonian mechanics. The Boltzmann equation is the fundamental equation in the kinetic theory of gases which describes the time evolution of particles in a non-equilibrium rarefied gas. Before this equation, Maxwell derived the formula for the velocity distribution function of gas molecules in an equilibrium state, now called Maxwellian which is a Gaussian depending on the parameters of density, temperature and bulk velocity. It is now well-known that the Maxwellian is indeed built into the Boltzmann equation and is its special stationary solution. In physics, the Maxwellian is a universal distribution function which appears when the gas attains an equilibrium state. However, if an external forcing is exerted on the gas, the non-Maxwellian steady state may persist. This external forcing may be caused through the boundary of the vessel containing the gas, the external force field, the external gas source, and others. One of the illustrative example is the time-periodic solution to the Boltzmann equation.

In [7], Boltzmann deduced from his equation the celebrated H-theorem showing that the mathematical entropy is non-increasing in time which asserts that the second law of thermodynamics based on the Newton mechanics. It was a famous episode in the history of science that the H-theorem raised a long and serious controversy between Boltzmann and his contemporaries. However, it is Boltzmann who was endorsed finally, though more than 100 years later, by Lanford [53] who established the convergence of the Newton equation to the Boltzmann equation and by many people who proved the existence of the global solutions. See also [20] for some details of the controversies in that period. The non-increasing of the H-function also suggests that any solution of the Boltzmann equation, if it exists globally in time and has some nice properties, converges to a uniform Maxwellian as the time goes on. In other words, the H-theorem implies that the Maxwellian is the only possible asymptotically stable stationary solution of the Boltzmann equation. From the physical view point, this may be rephrased as the equilibrium state of the gas is uniquely described by the Maxwellian, not by any other distribution functions.

Despite of its significant implications to physics, the mathematical justification of this statement had not been known until 1932 when Carleman [17, 18] proved that the Cauchy problem for the spatially homogeneous Boltzmann equation has solutions globally in time and that any of the solutions converges to a Maxwellian specified by the initial data as the time goes to infinity. Roughly speaking, the Maxwellian is time

asymptotically stable for any initial perturbation.

The spatially inhomogeneous Boltzmann equation, which is physically more realistic model, was solved much later. The first existence theorem of solutions was established, locally in time by Grad [40] in 1965, and globally in time by Ukai [88] in 1974. In [88], the Cauchy problem of the Boltzmann equation is proved to have solutions on the torus, i.e., under the space-periodic boundary condition, globally in time for initial data close to uniform Maxwellians, and the exponential convergence of the solutions to the relevant Maxwellians was also established. Since then, this result has been extended to the Cauchy problem in the whole space and various initial boundary value problems. We mention, Guo [41], Liu-Yu [70], Liu-Yang-Yu [68], Yang-Zhao [102], Nishida-Imai [75], Palczewski [79], Shizuta [81], Shizuta-Asano [82], Strain-Guo [85], Ukai [89, 91, 96], Ukai-Asano [93, 94], and references therein. They show, among others, the asymptotic stability of the uniform Maxwellian for initial perturbation close to the same Maxwellian.

For general initial data which are not necessarily close to Maxwellians, the well-known renormalized solution, introduced by DiPerna-Lions [31] for the whole space, has also been constructed on the torus and in a bounded domain under some physical boundary conditions by Hamdache [43], Arkeryd-Cercignani [2], and others. Moreover, it has been also shown that there is a time sequence along which the solution converges to a uniform Maxwellian in a weak topology, see [19, 27] for details. Here, the uniqueness of the limit Maxwellian is yet open and the solution is known to satisfy the boundary condition only in inequality.

A recent result presented by Desvillettes-Villani [28] suggests that the global asymptotic stability of the uniform Maxwellian holds in a much stronger sense: They discussed both the torus case and the boundary value problem in a bounded domain with the specular or reverse (bounce-back) reflection boundary condition and showed that any sufficiently smooth global solution converges almost exponentially to a uniform Maxwellian at time infinity. This is a remarkable result because no smallness conditions are imposed on the solutions and initial data, although the existence of such smooth solutions is a big open problem in the present mathematical theory of the Boltzmann equation.

Thus, all the results mentioned above lead to the conclusion that the uniform Maxwellian has a kind of universality in the theory of the Boltzmann equation. However, it should be stressed that both the Maxwellian and H-theorem were originally developed in the force-free space.

Suppose that the gas is exerted by an external forcing through the boundary of the vessel of the gas, by the force field, gas source or others. Then, the stationary state of the gas, if sustained, may be different from the uniform Maxwellian. Mathematically, this raises the stationary problems for the Boltzmann equation in a domain with boundary, in the

external force field, with the inhomogeneous term, or others. Among typical examples are the boundary layers, the flow past an obstacle, the interior problem with forcing and the periodic solution. However, we will not cover this part in this notes. Interested readers can refer to [93, 94] and the book chapter of Ukai-Yang, [97].

In the rest of the introduction we will give an intuitive derivation of the Boltzmann equation and some basic properties of the collision operator and the linearized collision operator. In the next section, we will first present the two classical expansions, that is, Hilbert and Chapman-Enskog expansions. Then a macro-micro decomposition around the local Maxwellian determined by the solution to the Boltzmann equation will be introduced and shown to give a unification of the above two expansions. One of the main implications of these expansions and decomposition is the relation of the Boltzmann equation to the classical systems of the gas dynamics, that is, Euler equations and Navier-Stokes equations. For this reason, it is natural to give a brief introduction of the conservation laws. For this purpose, we will give the mathematical theory of the one dimensional conservation laws which starts from the scalar equation to the well-posedness theory and the vanishing viscosity limit for systems. The main part of the notes is the global existence theory of perturbative solutions in a space which is the intersection of L^2 and L^∞ . Therefore, in Section 4, we will give a detailed description of the spectral analysis of the linearized Boltzmann operator. Based on the spectral properties, the global existence of the solution and the optimal convergence rates to the equilibrium will be proved in this new function space. In addition to the problem without external force, the case with external force will also be discussed accordingly.

1.2 Boltzmann equation

1.2.1 Intuitive derivation

In this subsection, as in [21], we will present an intuitive derivation of the Boltzmann equation which is by no means to be rigorous. For the rigorous derivation of the Boltzmann equation from the Liouville equation through the BBGKY hierarchy, please refer to the classical paper of Lanford, [53].

Assume there are N particles in total under consideration. Each particle is assumed to be an identical ball which is the so called hard-sphere model. The purpose is to derive an equation for the time evolution of a particle distribution under the assumption of the elastic binary collision and molecular chaos.

Let $P^1(t, x, \xi)$ be the probability of single particle at space-time (x, t) with velocity ξ , then the time evolution of $P^1(t, x, \xi)$ satisfies the free

transport equation

$$P_t^1(t, x, \xi) + \xi \cdot \nabla_\xi P^1(t, x, \xi) = 0,$$

if we neglect the collisions between particles. However, the particles collide from time to time and the collisions will influence the evolution of the distribution. If we assume by now that only the binary collisions dominate which is true when $N \rightarrow \infty$ in probability, then the effect of the collision can be estimated as follows. Let $P^2(t, x_1, x_2, \xi_1, \xi_2)$ be the probability of two particles at space-time (x_1, t) and (x_2, t) with velocities ξ_1 and ξ_2 respectively. Denote the diameter of each particle by σ . If we consider the binary collision of the first particle with the other $N - 1$ particles, then the infinitesimal amount is given by

$$(N - 1)\sigma^2 P^2(t, x_1, x_1 + \sigma n, \xi_1, \xi_2)|(\xi_2 - \xi_1) \cdot \omega|d\xi_2 d\omega dt,$$

where ω is a unit vector and $\xi_2 - \xi_1$ is the relative velocity.

If we denote the half sphere for $(\xi_2 - \xi_1) \cdot \omega > 0$ by S^+ and $(\xi_2 - \xi_1) \cdot \omega < 0$ by S^- respectively, then the infinitesimal change of the first particle satisfies

$$dP_t^1(t, x_1, \xi_1) + \xi_1 \cdot \nabla_{\xi_1} P^1(t, x_1, \xi_1)dt = g - l, \quad (1.1)$$

where

$$\begin{aligned} g &= (N - 1)\sigma^2 \int_{\mathbb{R}^3} \int_{S^+} P^2(t, x_1, x_1 + \sigma\omega, \xi_1, \xi_2)|(\xi_2 - \xi_1) \cdot \omega|d\xi_2 d\omega dt, \\ l &= (N - 1)\sigma^2 \int_{\mathbb{R}^3} \int_{S^-} P^2(t, x_1, x_1 + \sigma\omega, \xi_1, \xi_2)|(\xi_2 - \xi_1) \cdot \omega|d\xi_2 d\omega dt. \end{aligned} \quad (1.2)$$

Here, g and l represent the infinitesimal change of the gain and loss of the first particle with velocity ξ_1 at space-time (x, t) . However, the above equation is not that useful when we differentiate both sides by with respect to time. On the other hand, we should impose the assumption of molecular chaos on the binary collision which asserts that the probability of two particles before collision is independent. To apply this to the gain term g , we have rewrite (1.2) as follows by using the assumption of elastic collision which is time reversible. Notice that when $|x_1 - x_2| = \sigma$, we have

$$P^2(t, x_1, x_2, \xi_1, \xi_2) = P^2(t, x_1, x_2, \xi'_1, \xi'_2),$$

where ξ'_1 and ξ'_2 denote the corresponding velocities after the collision which satisfy the conservation of momentum and energy:

$$\xi_1 + \xi_2 = \xi'_1 + \xi'_2, \quad |\xi_1|^2 + |\xi_2|^2 = |\xi'_1|^2 + |\xi'_2|^2.$$

Straightforward calculation yields

$$\xi'_1 = \xi_1 - (\omega \cdot (\xi_1 - \xi_2))\omega, \quad \xi'_2 = \xi_2 + (\omega \cdot (\xi_1 - \xi_2))\omega.$$

Thus, we have

$$\begin{aligned} g &= (N-1)\sigma^2 \int_{\mathbb{R}^3} \int_{S^+} P^{(2)}(t, x_1, x_1 + \sigma\omega, \xi'_1, \xi'_2) |(\xi_2 - \xi_1) \cdot \omega| d\xi_2 d\omega dt \\ &= (N-1)\sigma^2 \int_{\mathbb{R}^3} \int_{S^-} P^{(2)}(t, x_1, x_1 - \sigma\omega, \xi'_1, \xi'_2) |(\xi_2 - \xi_1) \cdot \omega| d\xi_2 d\omega dt, \end{aligned}$$

where we have applied the transform $\omega \rightarrow -\omega$, $x_1 + \sigma\omega \rightarrow x_1 - \sigma\omega$. Since molecular chaos implies that for $(\xi_2 - \xi_1) \cdot \omega < 0$,

$$P^{(2)}(t, x_1, x_1 + \sigma\omega, \xi_1, \xi_2) = P^{(1)}(t, x_1, \xi_1) P^{(1)}(t, x_1 + \sigma\omega, , \xi_2).$$

In summary, by letting $N \rightarrow \infty$ and $\sigma \rightarrow 0$ with $N\sigma^2 \rightarrow \frac{1}{\kappa}$ which is called the Boltzmann-Grad limit, we have from (1.1) the time evolution for $P^1(t, x_1, \xi_1)$ as

$$\begin{aligned} &P_t^{(1)}(t, x_1, \xi_1) + \xi_1 \cdot \nabla_{\xi_1} P^{(1)}(t, x_1, \xi_1) \\ &= \frac{1}{\kappa} \int_{\mathbb{R}^3} \int_{S^-} \{P^{(1)}(t, x_1, \xi'_1) P^{(1)}(t, x_1, \xi'_2) \\ &\quad - P^{(1)}(t, x_1, \xi_1) P^{(1)}(t, x_1, \xi_2)\} |(\xi_2 - \xi_1) \cdot \omega| d\xi_2 d\omega, \end{aligned}$$

which is called the Boltzmann equation for the hard sphere model. In general, for the monatomic gas, the Boltzmann equation takes the form

$$\begin{aligned} &P_t^{(1)}(t, x, \xi) + \xi_1 \cdot \nabla_{\xi_1} P^{(1)}(t, x, \xi) \\ &= \frac{1}{\kappa} \int_{\mathbb{R}^3} \int_{S^-} \{P^{(1)}(t, x_1, \xi'_1) P^{(1)}(t, x_1, \xi'_2) \\ &\quad - P^{(1)}(t, x_1, \xi_1) P^{(1)}(t, x_1, \xi_2)\} q(|\xi_1 - \xi_2|, (\xi_2 - \xi_1) \cdot \omega) d\xi_2 d\omega, \end{aligned}$$

where the kernel $q(|\xi_1 - \xi_2|, (\xi_2 - \xi_1) \cdot \omega)$ called the cross section depends only on the strength of the relative velocity and the interaction angle in different physical considerations as explained below.

1.2.2 Collision invariants and H functional

As a slight generation of the equation derived in the last section, the Boltzmann equation with external force and source takes the form

$$\frac{\partial f}{\partial t} + \xi \cdot \nabla_x f + \frac{1}{m} F \cdot \nabla_{\xi} f = \frac{1}{\kappa} Q(f, f) + S, \quad (t, x, \xi) \in \mathbb{R} \times \Omega \times \mathbb{R}^n. \quad (1.3)$$

Here, $\Omega \subset \mathbb{R}^n$ is the domain containing the gas and $f = f(t, x, \xi)$ is the unknown function representing the probability (mass, number) density of gas particles around position $x = (x_1, \dots, x_n) \in \Omega$ with velocity $\xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$ at time $t \in \mathbb{R}$. Here, we consider the problem in n

dimensional space in order to see how the space dimension plays in the convergence rates to the equilibrium. The linear term $-\xi \cdot \nabla_x f - F \cdot \nabla_\xi f$, i.e., the *transport term*, gives the rate of change of f due to the motion of gas particles in the external force field $F = F(t, x, \xi) = (F_1, \dots, F_n)$, m being the mass of the gas particle. The last term on the right hand side is the rate of change of f due to the external source of gas particles of intensity $S = S(t, x, \xi)$. The term $\kappa^{-1}Q$ gives the rate of change of f due to binary elastic collision of gas particles as explained in the last subsection, where Q is a bilinear operator *collision operator*

$$Q(f, f) = \int_{\mathbb{R}^n \times S^{n-1}} q(v, \theta)(f' f'_* - f f_*) d\xi_* d\omega, \quad (1.4)$$

where

$$f = f(t, x, \xi), \quad f' = f(t, x, \xi'), \quad f'_* = f(t, x, \xi'_*), \quad f_* = f(t, x, \xi_*),$$

with (1.5), i.e.,

$$\xi' = \xi - ((\xi - \xi_*) \cdot \omega)\omega, \quad \xi'_* = \xi_* + ((\xi - \xi_*) \cdot \omega)\omega, \quad (1.5)$$

while

$$\omega \in S^{n-1}, \quad v = |\xi - \xi_*|, \quad \cos \theta = \frac{\xi - \xi_*}{|\xi - \xi_*|} \cdot \omega.$$

Again, ξ, ξ_* can be thought as the velocities of gas particles before collision and ξ', ξ'_* as those after collision, respectively.

The *collision cross section* $q(v, \theta)$ is determined by the interaction law between two colliding particles. A classical example is the hard sphere gas for which, ([40]),

$$q(v, \theta) = v |\cos \theta|, \quad \cos \theta = (\xi - \xi_*) \cdot \omega / v, \quad (1.6)$$

and another well-known example is the inverse power law potential r^{-s} for which ([40])

$$q(v, \theta) = v^\gamma |\cos \theta|^{-\gamma'} q_0(\theta), \quad \gamma = 1 - \frac{2(n-1)}{s}, \quad \gamma' = 1 + \frac{n-1}{s}, \quad (1.7)$$

where $q_0(\theta)$ is a bounded non-negative function which does not vanish near $\theta = \pi/2$. The interaction potential is said hard if $s > 2(n-1)$ and soft if $0 < s < 2(n-1)$, and it is called Maxwellian molecule when $s = 2(n-1)$.

Finally, κ is the *Bunsen number* (the ratio of the mean free path of the gas particle and the characteristic length of the domain of the vessel containing the gas) which plays an important role in the study of

asymptotic relations between the Boltzmann equation and the fluid dynamical equations such as the Euler and Navier-Stokes equations. This will become clear in the next section when we introduce the two classical expansions and the micro-macro decomposition. On the other hand, we will fix $\kappa = 1$ and $m = 1$, without loss of generality, when we consider the well-posedness of the Boltzmann equation in the Sections 4 and 5. Roughly speaking, the mathematical study of the Boltzmann equation is aimed at revealing various interplays between the conservative operator $\partial_t + \xi \cdot \nabla_x f + F \cdot \nabla_\xi f$ and the degenerate dissipative operator $Q(f, f)$ on the velocity variable.

Before going further in the analysis, we briefly introduce three main properties of the operator Q deduced by Boltzmann about the collision invariants and the celebrated H-theorem. The reader is referred to some text books on the Boltzmann equation, e.g. [20, 40] for more details.

Define the inner product

$$\langle f, g \rangle = \int_{\mathbb{R}^n} f(\xi)g(\xi)d\xi.$$

A function $\phi(\xi)$ is called a *collision invariant* if

$$\langle \phi, Q(f, f) \rangle = 0 \quad (\forall f \in C_0^\infty(\mathbb{R}_\xi^n, \mathbb{R}_+)).$$

The first property of Q is:

[P1] Q has $n + 2$ collision invariants,

$$\phi_0(\xi) = 1, \quad \phi_i(\xi) = \xi_i \quad (i = 1, 2, \dots, n), \quad \phi_{n+1}(\xi) = \frac{1}{2}|\xi|^2. \quad (1.8)$$

The proof comes via the integral identity

$$\langle \phi, Q(f, f) \rangle = \frac{1}{2} \int_{\mathbb{R}^n \times \mathbb{R}^n \times S^{n-1}} q(v, \theta) f f_* (\phi' + \phi'_* - \phi - \phi_*) d\xi d\xi_* d\omega, \quad (1.9)$$

which can be obtained by several changes of variables ([17, 20]). Then, this integral is zero independently of the particular function f if and only if

$$\phi' + \phi'_* - \phi - \phi_* = 0. \quad (1.10)$$

Boltzmann is the first who solved this equation, showing that within the space of twice differentiable functions, the most general solution is a linear combination of the collision invariants (1.8). Later, this result was extended to more general function spaces including the spaces of continuous and L^1_{loc} functions, see [20]. Here, we include an elegant proof given by Perthame, [78] as follows.

Assume $\varphi \geq 0$ with $(1 + |\xi|^2)\varphi \in L^1(\mathbb{R}^n)$ satisfying

$$\varphi' \varphi'_* = \varphi \varphi_*, \quad (1.11)$$

for a.e. $(\xi, \xi_*, \omega) \in \mathbb{R}^n \times \mathbb{R}^n \times S^{n-1}$. Notice that here $\varphi = \exp\{\phi\}$ with ϕ given in (1.10). Without loss of generality, we normalize φ by

$$\int_{\mathbb{R}^n} \varphi(\xi) d\xi = 1, \quad \int_{\mathbb{R}^n} \xi \varphi(\xi) d\xi = 0. \quad (1.12)$$

Now by fixing ω and taking the Fourier transform in (1.11), we have

$$\begin{aligned} \hat{\varphi}(\eta) \hat{\varphi}(\eta_*) &= \iint_{\mathbb{R}^n \times \mathbb{R}^n} \varphi(\xi) \varphi(\xi_*) e^{-i\eta \cdot \xi - i\eta_* \cdot \xi_*} d\xi d\xi_* \\ &= \iint_{\mathbb{R}^n \times \mathbb{R}^n} \varphi(\xi') \varphi(\xi'_*) e^{-i\eta \cdot \xi - i\eta_* \cdot \xi_*} d\xi d\xi_* \\ &= \iint_{\mathbb{R}^n \times \mathbb{R}^n} \varphi(\xi) \varphi(\xi_*) e^{-i\eta \cdot \xi' - i\eta_* \cdot \xi'_*} d\xi' d\xi'_* \\ &= \iint_{\mathbb{R}^n \times \mathbb{R}^n} \varphi(\xi) \varphi(\xi_*) e^{-i\eta \cdot \xi - i\eta_* \cdot \xi_*} e^{i((\eta - \eta_*) \cdot \omega)((\xi - \xi_*) \cdot \omega)} d\xi d\xi_*, \end{aligned}$$

where $\hat{\varphi}$ represents the Fourier transform of φ , and we have used the fact that Jacobian of the coordinate transform $(\xi, \xi_*) \rightarrow (\xi', \xi'_*)$ is 1. Since the first term of the above equation is independent of ω , the differentiation of the equation with respect to $\omega \perp (\eta - \eta_*)$ gives

$$\iint_{\mathbb{R}^n \times \mathbb{R}^n} \varphi(\xi) \varphi(\xi_*) e^{-i\eta \cdot \xi - i\eta_* \cdot \xi_*} (\xi - \xi_*) \cdot \omega d\xi d\xi_* = 0.$$

Thus, for any $\eta \neq \eta_*$ and $\omega \in S^{n-1}$ satisfying $\omega \perp (\eta - \eta_*)$, we have

$$(\nabla_\eta - \nabla_{\eta_*}) \hat{\varphi}(\eta) \hat{\varphi}(\eta_*) \perp \omega, \quad i.e., \quad (\nabla_\eta - \nabla_{\eta_*}) \hat{\varphi}(\eta) \hat{\varphi}(\eta_*) // (\eta - \eta_*).$$

By setting $\eta_* = 0$ and using (1.12), we have $\nabla_\eta \hat{\varphi}(\eta) // \eta$, that is, $\hat{\varphi}(\eta) = \psi(|\eta|^2)$ where $\psi(\eta)$ satisfies

$$\eta \psi'(|\eta|^2) \psi(|\eta_*|^2) - \eta_* \psi'(|\eta|^2) \psi'(|\eta_*|^2) // (\eta - \eta_*).$$

And this implies that $\psi'(|\eta|^2) \psi(|\eta_*|^2) = \psi(|\eta|^2) \psi'(|\eta_*|^2)$ which gives $\psi'(|\eta|^2) = c\psi(|\eta|^2)$ for some constant c . This in return gives that $\psi = c_1 e^{c_2 |\xi|^2}$ for some constants $c_1 > 0$ and c_2 and this proves that the only $n + 2$ collision invariants are given by ϕ_α in (1.8).

An important consequence of [P1] is the conservation laws from the Boltzmann equation, which connects the microscopic description in the kinetic theory of the gas and the macroscopic description in fluid mechanics. Since $f(t, x, \xi)$ is the mass density in the (x, ξ) -space, that is,

the *microscopic* mass density in the one-particle phase space, its moments with respect to ξ are *macroscopic* quantities in the usual physical space. The first few moments are

$$\begin{aligned}\rho &= \langle \phi_0, f(t, x, \cdot) \rangle, \\ \rho u_i &= \langle \phi_i, f(t, x, \cdot) \rangle \quad (i = 1, 2, \dots, n), \\ \rho \mathcal{E} &= \langle \phi_{n+1}, f(t, x, \cdot) \rangle.\end{aligned}\quad (1.13)$$

Here, ρ is the macroscopic mass density, $u = (u_1, u_2, \dots, u_n)$ is the macroscopic (bulk) velocity, and \mathcal{E} is the average energy density per unit mass, of the gas. The temperature θ and the pressure p are related to \mathcal{E} by

$$\mathcal{E} = \frac{1}{2}|u|^2 + \frac{n}{2}R\theta. \quad p = R\rho\theta. \quad (1.14)$$

Here, R is the *gas constant* (the Boltzmann constant divided by the mass of the gas particle). The last equation in (1.14) is called the *equation of state* for the ideal gas.

Consider the case $\Omega = \mathbb{R}^n$ and $F = 0, S = 0$. Let f be a smooth solution to (1.3) which vanishes sufficiently rapidly with (x, ξ) . Multiply (1.3) by ϕ_j and integrate it over $\mathbb{R}_x^n \times \mathbb{R}_\xi^n$. By virtue of (1.8) and by integration by parts, we have

$$\frac{d}{dt} \int_{\mathbb{R}^n} \langle \phi_i, f(t, x, \cdot) \rangle dx = 0, \quad i = 0, 1, \dots, n+1,$$

which are, in view of (1.13), the conservation laws of total mass ($i = 0$), total momentum in i -direction, ($i = 1, 2, \dots, n$) and total energy ($i+1$), of the gas.

Notice that the conservation laws hold also for the case $\Omega \neq \mathbb{R}^3$ and $F \neq 0$ by taking into account the boundary conditions $\partial\Omega$ and on the external force F .

The second property of Q is the essence of the H-theorem:

$$[\mathbf{P2}] \quad \langle \log f, Q(f, f) \rangle \leq 0 \quad (\forall f \in C_0^\infty(\mathbb{R}_\xi^n, \mathbb{R}_+)).$$

First, we observe ([20]) that putting $\phi = \log f$ in (1.9) and a simple change of variable give the identity

$$\begin{aligned}-\langle Q(f, f), \log f \rangle &= \frac{1}{4} \int_{\mathbb{R}^n \times \mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \omega) (f' f'_* - f f_*) \log \frac{f' f'_*}{f f_*} d\xi d\xi_* d\omega.\end{aligned}$$

Now [P2] follows from the elementary inequality

$$(a - b)(\log a - \log b) \geq 0 \quad (a, b > 0).$$

Notice that here the equality holds if and only if $a = b$, which then implies the third basic property of Q .

[P3] $Q(f, f) = 0 \Leftrightarrow \langle \log f, Q(f, f) \rangle = 0 \Leftrightarrow f = \mathbf{M}(\xi)$ where

$$\mathbf{M}(\xi) = \mathbf{M}_{[\rho, u, \theta]}(\xi) = \frac{\rho}{(2\pi R\theta)^{n/2}} \exp\left(-\frac{|\xi - u|^2}{2R\theta}\right).$$

\mathbf{M} is the *Maxwellian*, and according to Maxwell, it represents the velocity distribution of the gas in an equilibrium state with the mass density $\rho > 0$, bulk velocity $u = (u_1, u_2, \dots, u_n) \in \mathbb{R}^n$, and temperature $\theta > 0$. Maxwell derived \mathbf{M} on physical arguments but [P3] implies that it is built-in for the Boltzmann equation. Notice that [P3] is equivalent to solving the equation

$$f' f'_* - f f_* = 0,$$

which is reduced to the equation (1.10) by transformation $\phi = \log f$. If (ρ, u, θ) are constants, \mathbf{M} is called a uniform (global, absolute) Maxwellian while if they are functions of (x, t) , \mathbf{M} is called a local Maxwellian. An immediate consequence of [P3] is that the uniform Maxwellian is a stationary solution of (1.3) if the external force and source are absent.

We now recall Boltzmann's H-theorem mentioned which is the most significant consequence of the property [P1–3]. Let f be a density function of a gas. Since it is non-negative, we may define the *H-function*,

$$H(t) = \int_{\Omega \times \mathbb{R}^n} f \log f dx d\xi,$$

which gives, according to Boltzmann [7], the negative of the entropy of the gas. Consider again the case $\Omega = \mathbb{R}^n$, $F = 0$, $S = 0$, and let f be a non-negative smooth solution to (1.3) with rapid decay properties in (x, ξ) . Multiply (1.3) by $\log f$ and integrate in (x, ξ) . Integration by parts, together with [P1], yields,

$$\frac{dH(t)}{dt} + D(t) = 0,$$

where

$$D(t) = - \int_{\mathbb{R}^n} \langle Q(f, f), \log f \rangle dx \quad (1.15)$$

is called the *entropy dissipation integral*. Since $D(t)$ is non-negative by virtue of [P2], we have

$$\frac{dH}{dt} \leq 0. \quad (1.16)$$

This implies that the entropy is increasing with time. Moreover, by virtue of [P3], the equality in (1.16) holds only when f is a Maxwellian,

so that f converges to a Maxwellian as t tends to infinity. This is the celebrated *H-theorem*.

The non-negativity of the integral D in (1.15) indicates that the Boltzmann equation is a dissipative equation. More precisely, the dissipation through D is degenerate in the sense that it is only on the velocity variable. However, the interplay between this degenerate dissipation and the conservative operator for the free transport yields full dissipation in all variables in the convergence to the equilibrium. This phenomena is now called hypercoercivity which appears also in other kinetic equations such as Fokker-Planck equation, oscillator chains and Oseen's vortices, cf. [98] and references therein. Moreover, this fact is essentially used in the L^1 theory of the Boltzmann equation, [31]. On the other hand, its linearized version results in the non-positivity of the linearized operator of Q at the uniform Maxwellian \mathbf{M} , and is a key ingredient in the theory of the Boltzmann equation near the Maxwellian.

This non-positivity is formulated as follows. Introduce the bilinear symmetric operator associated with the quadratic operator Q :

$$Q(f, g) = \frac{1}{2} \int_{\mathbb{R}^n \times S^{n-1}} q(v, \theta) (f' g'_* + g' f'_* - f g_* + g' f'_*) d\xi_* d\omega.$$

Let f be a function near \mathbf{M} in the form of

$$f = \mathbf{M} + \epsilon \mathbf{M}^{1/2} u,$$

for small $\epsilon \in \mathbb{R}$ and some function $u = u(\xi)$. Then, by virtue of [P1, 3] and the bilinearity of Q , we get

$$\langle \log f, Q(f, f) \rangle = \langle \log(1 + \epsilon \mathbf{M}^{-1/2} u), \epsilon \mathbf{M}^{1/2} (\mathbf{L}u + \epsilon \Gamma(u, u)) \rangle,$$

where

$$\mathbf{L}u = 2\mathbf{M}^{-1/2} Q(\mathbf{M}, \mathbf{M}^{1/2} u), \quad \Gamma(u, v) = \mathbf{M}^{-1/2} Q(\mathbf{M}^{1/2} u, \mathbf{M}^{1/2} v). \quad (1.17)$$

Here, \mathbf{L} , the *linearized collision operator*, is a linearized operator of Q around \mathbf{M} while Γ , being its remainder, is a bilinear symmetric operator. Now, we see at least formally that

$$\frac{1}{\epsilon^2} \langle \log f, Q(f, f) \rangle \rightarrow \langle u, \mathbf{L}u \rangle \quad (\epsilon \rightarrow 0),$$

concluding, by the aid of [P2], that

$$\langle u, \mathbf{L}u \rangle \leq 0. \quad (1.18)$$

1.2.3 Assumptions on cross-sections

Needless to say that all the properties of Q in the preceding subsection are substantiated only when the relevant integrals are convergent.

The collision kernel q in Q has no singularity for the hard sphere gas (1.6). However, (1.7) for the inverse power law potential has a strong singularity at $\theta = \pi/2$ which corresponds to the grazing collision, and as a consequence, the integral over S^{n-1} in (1.4) does not converge under a mild assumption on f, g such that they are only bounded. Actually, it is well-known that $Q(f, g)$ is well-defined only for sufficiently smooth f, g as a nonlinear pseudo-differential operator, see e.g. [1, 28, 90]. However, this is a too strong restriction to solve the Boltzmann equation in the full generality. To be more precise, for the non-cutoff cross section of inverse power laws, the collision operator behaves like a fraction of Laplacian, that is,

$$Q(f, f) = -C_f(-\Delta)^{\frac{\gamma}{2}} f + \text{more regular terms},$$

for some constant $\gamma = \frac{2}{s}$ in three dimensional space with $s > 1$. In other words, the non-cutoff cross section has smoothing effect on the solution which is an active and difficult topic on Boltzmann equation and we will not cover it in this notes.

In order to avoid this difficulty, Grad [40] introduced an idea to cut off the singularity at $\theta = \pi/2$ so that $q_0(\theta) \in L^1(S^{n-1})$. This assumption has been widely used and is now called Grad's angular cutoff assumption.

In the following discussion, we assume that $q(v, \theta)$ is a non-negative measurable function satisfying

$$\int_{S^{n-1}} q(v, \theta) d\omega \geq q_0 v^\gamma, \quad q(v, \theta) \leq q_1 (1+v)^\gamma |\cos \theta|, \quad (1.19)$$

for some constants $q_0, q_1 > 0$ and $\gamma \in [0, 1]$. Clearly, this is satisfied by the hard sphere gas (1.6) with $\gamma = 1$ and by the inverse power law potential case (1.7) under the Grad's cutoff with $\gamma = 1 - 2(n-1)/s$ for $s \geq 2(n-1)$. Thus, (1.19) is a slightly generalized version of Grad's cutoff hard potentials.

1.2.4 Basic properties of the collision operators

The Grad's cutoff assumption ensures well-definedness of the nonlinear and linearized collision operators Q and \mathbf{L} . Here, we shall summarize their properties which will be used in essential ways throughout this notes. Most of them go back to Grad [40]. All the results below hold for any choice of the Maxwellian \mathbf{M} in (1.17), but they are equivalent to each other by a simple scaling and translation of the velocity variables. This can be seen from the definition of Q in (1.4). In particular, Q is

translation invariant: With the translation $\tau_c u = u(\xi - c)$, $c \in \mathbb{R}^n$, it holds that

$$\tau_c Q(f, g) = Q(\tau_c f, \tau_c g).$$

Thus, in the sequel, we fix \mathbf{M} to be the standard Maxwellian $\mathbf{M}_{[1,0,1]}$, without loss of generality.

Proposition 1.1. ([40]). *Under the assumption (1.19), the linearized collision operator \mathbf{L} defined by (1.17) has the decomposition*

$$\mathbf{L}u = -\nu(\xi)u + Ku, \quad (1.20)$$

where $\nu(\xi)$ is a non-negative measurable function of ξ satisfying

$$\nu_0(1 + |\xi|)^\gamma \leq \nu(\xi) \leq \nu_1(1 + |\xi|)^\gamma, \quad \xi \in \mathbb{R}^3, \quad (1.21)$$

with $\gamma \in [0, 1]$ as in (1.19) and for some constants $\nu_0, \nu_1 > 0$, while K is a linear integral operator in ξ ;

$$Ku = \int_{\mathbb{R}^n} K(\xi, \xi_*)u(\xi_*)d\xi_*,$$

whose kernel is real symmetric and enjoys the estimates

$$\int_{\mathbb{R}^n} |K(\xi, \xi_*)|^2 d\xi_* \leq C, \quad (1.22)$$

$$\int_{\mathbb{R}^n} |K(\xi, \xi_*)|(1 + |\xi_*|)^{-\beta} d\xi_* \leq C_\beta(1 + |\xi|)^{-\beta-1}, \quad \beta \geq 0, \quad (1.23)$$

for some constants C, C_β independent of ξ .

Actually, the function $\nu(\xi)$ is defined by

$$\nu(\xi) = \int_{\mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \theta) \mathbf{M}^{1/2}(\xi_*) d\xi_* d\omega, \quad (1.24)$$

so (1.21) follows easily from (1.19). The proof of (1.22), (1.23) can be found in [40] for the case $n = 3$ and $\gamma = 1$. (1.22) should be modified for the other cases but is assumed here for sake of simplicity. For the hard sphere gas, the explicit expressions are available:

$$\nu(\xi) = (2\pi)^{1/2} \left\{ (|\xi| + |\xi|^{-1}) \int_0^{|\xi|} \exp(-u^2/2) du + \exp(-|\xi|^2/2) \right\}.$$

$$\begin{aligned} K(\xi, \xi_*) = & (2\pi)^{1/2} |\xi - \xi_*|^{-1} \exp\left(-\frac{1}{8} \frac{(|\xi|^2 - |\xi_*|^2)^2}{|\xi - \xi_*|^2} - \frac{1}{8} |\xi - \xi_*|^2\right) \\ & - \frac{1}{2} |\xi - \xi_*| \exp\left(-(|\xi|^2 + |\xi_*|^2)/4\right). \end{aligned}$$

The operator K is to be considered in various function spaces such as

$$\begin{aligned} L^p &= L^p(\mathbb{R}_\xi^n), \quad p \in [1, \infty], \\ L_\beta^\infty &= L^\infty(\mathbb{R}_\xi^n; (1 + |\xi|)^\beta d\xi), \quad \beta \in \mathbb{R}. \end{aligned}$$

Note that $L_0^\infty = L^\infty$. The following lemma is partly from [20, 40, 91].

Lemma 1.2. (a) Let $2 \leq p \leq r \leq \infty$. Then, the operator

$$K : L^p \rightarrow L^r \tag{1.25}$$

is bounded.

(b) Let $\beta \geq 0$. Then, the operator

$$K : L_\beta^\infty \rightarrow L_{\beta+1}^\infty \tag{1.26}$$

is bounded.

Proof. Consider first the part (a) when $p = r \in [2, \infty]$. The case $p = r = \infty$ comes from (1.23):

$$|Ku(\xi)| \leq \int_{\mathbb{R}^n} |K(\xi, \xi_*)| d\xi_* \|u\|_{L^\infty}.$$

When $p < \infty$, by (1.23) again and the Hölder inequality,

$$|Ku(\xi)|^p \leq \left(\int_{\mathbb{R}^n} |K(\xi, \xi_*)| d\xi_* \right)^{p-1} \int_{\mathbb{R}^n} |K(\xi, \xi_*)| |u(\xi_*)|^p d\xi_*,$$

whence follows

$$\|Ku\|_{L^p}^p \leq C_0^{p-1} C_{p-1} \int_{\mathbb{R}^n} (1 + |\xi_*|)^{-p} |u(\xi_*)|^p d\xi_*. \tag{1.27}$$

Thus,

$$K : L^p \rightarrow L^p \tag{1.28}$$

is bounded for any $p \in [2, \infty]$.

Next, we prove the lemma for the case with $p = 2$ and $r = \infty$ (1.22) and the Schwarz inequality give

$$|Ku(\xi)|^2 \leq \int_{\mathbb{R}^n} |K(\xi, \xi_*)|^2 d\xi_* \int_{\mathbb{R}^n} |u(\xi_*)|^2 d\xi_* \leq C \|u\|_{L^2}^2,$$

which indicates that

$$K : L^2 \rightarrow L^\infty \tag{1.29}$$

is bounded.

Finally, the usual interpolation (see e.g. [4]) between (1.28) for $p = \infty$ and (1.29) proves (1.25) for the case $p \in [2, \infty]$ and $r = \infty$, and then the interpolation between this result and (1.28) for $p \in [2, \infty]$ proves (1.25) for $p \leq r$ for any $r \in [p, \infty]$, proving (a).

For the proof of (1.26), use again (1.23) to compute

$$|Ku(\xi)| \leq \int_{\mathbb{R}^n} |K(\xi, \xi_*)|(1 + |\xi|)^{-\beta} d\xi_* \|u\|_{L_\beta^\infty} \leq C_\beta (1 + |\xi|)^{-\beta-1} \|u\|_{L_\beta^\infty}.$$

Thus the proof of the lemma is complete. \square

Remark 1.3. (a) For easy reference, we include the Riesz-Thorin interpolation theorem here. Let T be a bounded linear operator from L^{p_1} to L^{p_2} and at the same time from L^{q_1} to L^{q_2} . Then it is a bounded operator from L^{r_1} to L^{r_2} with

$$\frac{1}{r_1} = \frac{\theta}{p_1} + \frac{1-\theta}{q_1}, \quad \frac{1}{r_2} = \frac{\theta}{p_2} + \frac{1-\theta}{q_2},$$

for any $\theta \in [0, 1]$.

(b) More exact regularity estimate of K can be obtained for cutoff potentials in the sense that

$$K : L_\beta^\infty \rightarrow L_{\beta+2-\gamma}^\infty,$$

for $\gamma > -3$ defined in (1.19), cf. [74, 100].

The following classical lemma about the compact property of the operator K was proved in [40].

Lemma 1.4. *K is a self-adjoint compact operator on L^2 .*

Proof. K is a bounded operator on L^2 according to (a) of the previous lemma and it is self-adjoint since the integral kernel is real symmetric according to Proposition 1.1. To prove that it is compact, for $R > 0$, let $\chi_R(\xi)$ be a characteristic function for $|\xi| < R$ and put $\bar{\chi}_R = 1 - \chi_R$. By (1.27) for $p = 2$, we get

$$\begin{aligned} \|K\bar{\chi}_R u\|_{L^2}^2 &\leq C_0 C_1 \int_{\mathbb{R}^n} (1 + |\xi_*|)^{-2} \bar{\chi}_R(\xi_*) |u(\xi_*)|^2 d\xi_* \\ &\leq C_0 C_1 (1 + R)^{-2} \|u\|_{L^2}^2, \end{aligned}$$

which indicates that

$$\|K\bar{\chi}_R\| \rightarrow 0 \quad (R \rightarrow \infty),$$

in the operator norm of L^2 . The same is true for $\bar{\chi}_R K$ by the direct calculation or because it is the adjoint to $K\bar{\chi}_R$. On the other hand, the

operator $\chi_R K \chi_R$ is a compact operator on L^2 , or more precisely, it is of Hilbert-Schmidt type as (1.22) implies that its integral kernel is in $L^2(\mathbb{R}_\xi^n \times \mathbb{R}_{\xi_*}^n)$. Then the proof is completed by [49, Theorem III. 4.7]. \square

Fundamental properties of the operator \mathbf{L} are now summarized in

Proposition 1.5. *Assume (1.19) with $\gamma \in [0, 1]$ and consider the operator \mathbf{L} defined by (1.17) with the domain of definition*

$$D(\mathbf{L}) = \{u \in L^2 \mid \nu(\xi)u \in L^2\}.$$

Then, the following holds:

- (1) \mathbf{L} is a linear densely defined closed operator in L^2 .
- (2) \mathbf{L} is self-adjoint and non-positive in L^2 .
- (3) Its spectrum $\sigma(\mathbf{L})$ satisfies the following:
 - (a) $\sigma(\mathbf{L}) \subset (-\infty, 0]$.
 - (b) The part $\sigma(\mathbf{L}) \cap (-\nu_*, 0]$ consists only of discrete semi-simple eigenvalues and its only possible accumulation point is $-\nu_*$ where

$$\nu_* \equiv \inf_{\xi \in \mathbb{R}^n} \nu(\xi). \quad (1.30)$$

Note from (1.21) that $\nu_* \geq \nu_0$.

- (c) 0 is a semi-simple eigenvalue of \mathbf{L} . Its eigenspace, which is the null space of \mathbf{L} , denoted by \mathcal{N} , is $(n+2)$ -dimensional and spanned by collision invariants weighted by $\mathbf{M}^{1/2}$,

$$\mathcal{N} = \text{span}\{\mathbf{M}^{1/2}, \xi_i \mathbf{M}^{1/2} \ (i = 1, 2, \dots, n), \frac{1}{2}|\xi|^2 \mathbf{M}^{1/2}\}.$$

- (4) Let $\{\psi_i\}_{i=0}^{n+1}$ be an orthonormal basis of \mathcal{N} in L^2 and put

$$\mathbf{P}u = \sum_{i=0}^{n+1} \langle u, \psi_i \rangle \psi_i, \quad (1.31)$$

which gives the orthogonal eigenprojection for the eigenvalue 0 :

$$\mathbf{P} : L^2 \rightarrow \mathcal{N}.$$

It holds that

$$\mathbf{P}\mathbf{L}u = 0 \quad (\forall u \in D(\mathbf{L})).$$

Moreover, $\mathbf{P} : L^2 \rightarrow L_\beta^\infty$ is a bounded operator for any $\beta \geq 0$.

Proof. (i) is true for the multiplication operator ν defined by

$$\nu u = \nu(\xi)u(\xi), \quad D(\nu) = \{u \in L^2 \mid \nu(\xi)u \in L^2\}, \quad (1.32)$$

where $\nu(\xi)$ is the function in (1.24). Since K is bounded owing to Lemma 1.2 (a), \mathbf{L} is a bounded perturbation of $-\nu$, and (i) follows from [49].

It is easy to see that ν is self-adjoint, and consequently, so is \mathbf{L} since K is self-adjoint and bounded, owing to Lemma 1.2(a). Now, it is clear that (1.18) is valid for any $u \in D(\mathbf{L})$, hence the non-positivity. This proves (2) and (3)(a) is a simple consequence of (2).

On the other hand, (3)(b) follows from Weyl's theorem on the compact perturbation, [49, Theorem IV. 5.35], since $\sigma(\nu) \subset (-\infty, -\nu_*]$ due to (1.21) and K is compact. And, (4) is a direct consequence of [P1], see [20, 40]. Thus, the proposition is proved. \square

Note that (3)(b-c) in the above proposition imply the spectral gap: There exists a constant $\mu_1 > 0$ such that

$$\langle u, \mathbf{L}u \rangle \leq -\mu_1 \| (I - \mathbf{P})u \|_{L^2}^2, \quad \forall u \in D(\ell) \quad (1.33)$$

holds. Actually, $-\mu_1$ is the first largest non-zero eigenvalue of \mathbf{L} in $(-\nu_*, 0)$ or if it does not exist, $\mu_1 = \nu_*$. This can be strengthened as

Lemma 1.6. *There is a constant $\mu_* > 0$ such that*

$$\langle u, \mathbf{L}u \rangle \leq -\mu_* \|\nu^{1/2}(I - \mathbf{P})u\|_{L^2}^2, \quad \forall u \in D(\mathbf{L}),$$

where $\nu^{1/2}$ is the multiplication operator by the function $\nu^{1/2}(\xi)$.

This lemma was proved in [38] coming from the decomposition (1.20) and the spectral gap (1.33).

Moreover, note that the operator K can absorb the collision frequency as a bounded operator.

Lemma 1.7. *The operator $K\nu$ is a bounded operator on L^2 .*

Proof. Use again (1.27) for $p = 2$ and recall (1.21). Then,

$$\|K\nu u\|_{L^2}^2 \leq C_0 C_1 \int_{\mathbb{R}^n} (1 + |\xi_*|)^{-2} \nu(\xi_*)^2 |u(\xi_*)|^2 d\xi_* \leq C \|u\|_{L^2}^2,$$

whence the lemma follows. \square

The main purpose of Sections 4 and 5 is to obtain the well-posedness of the Boltzmann equation in a new function space which is weaker than

the function space used previously for the perturbative solutions. That is, we consider the following function spaces. Set, for $\beta \geq 0$,

$$X_\beta = L^2 \cap L_\beta^\infty, \quad (1.34)$$

$$L^2 = L^2(\mathbb{R}_x^n \times \mathbb{R}_\xi^n), \quad L_\beta^\infty = \{u \mid (1 + |\xi|)^\beta u \in L^\infty(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)\},$$

$$\|u\|_{X_\beta} = \|u\|_{L^2(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)} + \|(1 + |\xi|)^\beta u\|_{L^\infty(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)},$$

and the supplementary function space mainly for initial data

$$Z_q = L^2(\mathbb{R}_\xi^n; L^q(\mathbb{R}_x^n)), \quad (1.35)$$

$$\|u\|_{Z_q} = \left(\int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} |u(x, \xi)|^q dx \right)^{2/q} d\xi \right)^{1/2},$$

when $q \in [1, 2]$.

Notice that the operators ν , K , \mathbf{L} , and \mathbf{P} can be defined also in various spaces of function in v variable while they are viewed as constant operators in x . For example, ν^{-1} and K are bounded operators on X_β . We now give some estimates for the nonlinear operator Γ in the space X_β and other related spaces. Set

$$Y_\beta = L_\beta^\infty = \{u(x, \xi) \mid (1 + |\xi|)^\beta u \in L^\infty(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)\}. \quad (1.36)$$

Lemma 1.8. *Assume (1.19) with $\gamma \in [0, 1]$. Then, for any $\delta \in [0, 1]$, there is a constant $C > 0$ such that the following holds:*

$$\|\nu^{-\delta}\Gamma(u, v)\|_{Z_1} \leq C(\|\nu^{1-\delta}u\|_{Z_2}\|v\|_{Z_2} + \|u\|_{Z_2}\|\nu^{1-\delta}v\|_{Z_2}). \quad (1.37)$$

$$\|\nu^{-\delta}\Gamma(u, v)\|_{Z_2} \leq C(\|u\|_{Y_\beta}\|v\|_{Z_2} + \|u\|_{Z_2}\|v\|_{Y_\beta}), \quad \beta > \frac{n}{2} + \gamma(1 - \delta). \quad (1.38)$$

$$\|\nu^{-\delta}\Gamma(u, v)\|_{Y_\beta} \leq C(\|\nu^{1-\delta}u\|_{Y_\beta}\|v\|_{Y_\beta} + \|u\|_{Y_\beta}\|\nu^{1-\delta}v\|_{Y_\beta}), \quad \beta \geq 0. \quad (1.39)$$

Moreover,

$$\mathbf{P}\Gamma(u, v) = 0 \quad (1.40)$$

holds for any $u, v \in X_\beta$.

Proof. (1.40) is a property of the collision invariants, see [40]. Write

$$\Gamma(u, v) = \frac{1}{2}\{\Gamma_1(u, v) + \Gamma_1(v, u) - \Gamma_2(u, v) - \Gamma_2(v, u)\},$$

with

$$\Gamma_1(u, v) = \int_{\mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \theta) \mathbf{M}(\xi_*)^{1/2} u(\xi') v(\xi'_*) d\xi_* d\omega,$$

$$\Gamma_2(u, v) = \int_{\mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \theta) \mathbf{M}(\xi_*)^{1/2} u(\xi) v(\xi_*) d\xi_* d\omega.$$

Clearly, the estimates in the lemma follow from those for each Γ_j , $j = 1, 2$.

(i) Proof of (1.37) for $j = 1$: We compute

$$\begin{aligned} & \|\Gamma_1(u, v)(\cdot, \xi)\|_{L^1(\mathbb{R}_x^n)} \\ & \leq \int_{\mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \theta) \mathbf{M}(\xi_*)^{1/2} \|u(\cdot, \xi') v(\cdot, \xi'_*)\|_{L^1(\mathbb{R}_x^n)} d\xi_* d\omega, \end{aligned}$$

which is bounded by the Schwartz inequality by

$$\begin{aligned} & \left(\int_{\mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \theta)^2 \mathbf{M}(\xi_*) d\xi_* d\omega \right)^{1/2} \\ & \quad \times \left(\int_{\mathbb{R}^n \times S^{n-1}} \|u(\cdot, \xi') v(\cdot, \xi'_*)\|_{L^1(\mathbb{R}_x^n)}^2 d\xi_* d\omega \right)^{1/2} \\ & \leq C\nu(\xi) \left(\int_{\mathbb{R}^n \times S^{n-1}} \|u(\cdot, \xi') v(\cdot, \xi'_*)\|_{L^1(\mathbb{R}_x^n)}^2 d\xi_* d\omega \right)^{1/2}. \end{aligned}$$

The last line comes from

$$\int_{\mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \theta)^2 \mathbf{M}(\xi_*) d\xi_* d\omega \leq C(1 + |\xi|)^{2\gamma} \leq C\nu(\xi)^2,$$

which holds by virtue of (1.19) and (1.21). Consequently,

$$\begin{aligned} & \int_{\mathbb{R}^n} \left(\nu(\xi)^{-\delta} \|\Gamma_1(u, v)(\cdot, \xi)\|_{L^1(\mathbb{R}_x^n)} \right)^2 d\xi \\ & \leq C \int_{\mathbb{R}^n \times \mathbb{R}^n \times S^{n-1}} \nu(\xi)^{2(1-\delta)} \|u(\cdot, \xi') v(\cdot, \xi'_*)\|_{L^1(\mathbb{R}_x^n)}^2 d\xi d\xi_* d\omega. \end{aligned}$$

We shall evaluate the last integral. To this end, note that the Schwartz inequality implies

$$\|u(\cdot, \xi') v(\cdot, \xi'_*)\|_{L^1(\mathbb{R}_x^n)} \leq \|u(\cdot, \xi')\|_{L^2(\mathbb{R}_x^n)} \|v(\cdot, \xi'_*)\|_{L^2(\mathbb{R}_x^n)},$$

and that (1.4) and (1.21) yield

$$\begin{aligned} \nu(\xi) & \leq C(1 + |\xi|)^\gamma = C(1 + |\xi' - \{(\xi' - \xi'_*) \cdot \omega\}\omega|)^\gamma \\ & \leq C(2 + |\xi'| + |\xi'_*|)^\gamma \leq C(\nu(\xi') + \nu(\xi'_*)). \end{aligned}$$

Finally, the Jacobian of the change of variable $(\xi, \xi_*, \omega) \leftrightarrow (\xi', \xi'_*, -\omega)$ is unity. Therefore, we have

$$\begin{aligned} & \int_{\mathbb{R}^n} \left(\nu(\xi)^{-\delta} \|\Gamma_1(u, v)(\cdot, \xi)\|_{L^1(\mathbb{R}_x^n)} \right)^2 d\xi \\ & \leq C \int_{\mathbb{R}^n \times \mathbb{R}^n \times S^{n-1}} \left(\nu(\xi')^{2(1-\delta)} + \nu(\xi'_*)^{2(1-\delta)} \right) \\ & \quad \times \|u(\cdot, \xi')\|_{L^2(\mathbb{R}_x^n)}^2 \|v(\cdot, \xi'_*)\|_{L^2(\mathbb{R}_x^n)}^2 d\xi' d\xi_* d\omega, \end{aligned}$$

which proves (1.37) for $j = 1$.

(ii) Proof of (1.38) for $j = 1$: In the same way as in (i), we have

$$\begin{aligned} \|\nu^{-\delta}\Gamma_1(u, v)\|_{Z_2}^2 &= \int_{\mathbb{R}^n} \left(\nu(\xi)^{-\delta} \|\Gamma_1(u, v)(\cdot, \xi)\|_{L^2(\mathbb{R}_x^n)} \right)^2 d\xi \\ &\leq C \int_{\mathbb{R}^n \times \mathbb{R}^n \times S^{n-1}} \left(\nu(\xi')^{2(1-\delta)} + \nu(\xi'_*)^{2(1-\delta)} \right) \\ &\quad \times \|u(\cdot, \xi')v(\cdot, \xi'_*)\|_{L^2(\mathbb{R}_x^n)}^2 d\xi' d\xi'_* d\omega. \end{aligned}$$

Now, assume $\beta - \gamma(1 - \delta) > n/2$. Then,

$$\begin{aligned} &\int_{\mathbb{R}^n \times \mathbb{R}^n} \nu(\xi')^{2(1-\delta)} \|u(\cdot, \xi')v(\cdot, \xi'_*)\|_{L^2(\mathbb{R}_x^n)}^2 d\xi' d\xi'_* \\ &= \int_{\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n} \nu(\xi')^{2(1-\delta)} |u(x, \xi')|^2 |v(x, \xi'_*)|^2 dx d\xi' d\xi'_* \\ &\leq \int_{\mathbb{R}^n \times \mathbb{R}^n} (1 + |\xi'|)^{-2(\beta - \gamma(1 - \delta))} \|u\|_{Y_\beta}^2 \|v(\cdot, \xi'_*)\|_{L^2(\mathbb{R}_x^n)}^2 d\xi' d\xi'_* \\ &\leq C \|u\|_{Y_\beta}^2 \|v\|_{Z_2}^2, \end{aligned}$$

and similarly,

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \nu(\xi')^{2(1-\delta)} \|u(\cdot, \xi')v(\cdot, \xi'_*)\|_{L^2(\mathbb{R}_x^n)}^2 d\xi' d\xi'_* \leq C \|u\|_{Z_2}^2 \|v\|_{Y_\beta}^2,$$

which prove (1.38) for $j = 1$.

(iii) Proof of (1.39) for $j = 1$. It follows from (1.4), (1.19), and (1.21) that

$$\begin{aligned} q(|\xi - \xi_*|, \theta)^{1-\delta} &\leq C\nu_1(1 + |\xi - \xi_*|)^{\gamma(1-\delta)} \\ &\leq C\nu_1(1 + |\xi'| + |\xi'_*|)^{\gamma(1-\delta)} \\ &\leq C(\nu(\xi')^{1-\delta} + \nu(\xi'_*)^{1-\delta}) \end{aligned}$$

and

$$\int_{\mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \theta)^\delta \mathbf{M}(\xi'_*)^{1/2} d\xi'_* d\omega \leq C(1 + |\xi|)^{\delta\gamma} \leq C\nu(\xi)^\delta.$$

On the other hand,

$$\begin{aligned} (1 + |\xi'|^2)^{-\beta/2} (1 + |\xi'_*|^2)^{-\beta/2} &\leq (1 + |\xi'|^2 + |\xi'_*|^2)^{-\beta/2} \\ &= (1 + |\xi|^2 + |\xi_*|^2)^{-\beta/2} \leq C(1 + |\xi|)^{-\beta}, \end{aligned}$$

where we used the conservation law $|\xi'|^2 + |\xi'_*|^2 = |\xi|^2 + |\xi_*|^2$. Combining these estimates yields

$$\begin{aligned} & |\Gamma_1(u, v)(x, \xi)| \\ & \leq C \int_{\mathbb{R}^n \times S^{n-1}} q(|\xi - \xi_*|, \theta)^\delta M(\xi_*)^{1/2} (1 + |\xi'|)^{-\beta} (1 + |\xi'_*|)^{-\beta} d\xi_* d\omega \\ & \quad \times (\|\nu^{1-\delta} u\|_{Y_\beta} \|v\|_{Y_\beta} + \|u\|_{Y_\beta} \|\nu^{1-\delta} v\|_{Y_\beta}) \\ & \leq C\nu(\xi)^\delta (1 + |\xi|)^{-\beta} (\|\nu^{1-\delta} u\|_{Y_\beta} \|v\|_{Y_\beta} + \|u\|_{Y_\beta} \|\nu^{1-\delta} v\|_{Y_\beta}), \end{aligned}$$

which is what was desired.

The estimates for $\Gamma_2(u, v)$ can be carried out in a similar way but are more straightforward. The details are omitted. Now, the proof of the lemma is completed. \square

The function $\nu(\xi)$ is bounded if $\gamma = 0$ and unbounded of order $O(|\xi|^\gamma)$ if $\gamma > 0$. Thus, Lemma 1.8 asserts that Γ is a bounded operator if $\gamma = 0$ whereas it is well-defined but unbounded with the weight loss of order $O(|\xi|^\gamma)$ if $\gamma > 0$.

In the space of function in ξ only, the same computation as above proves the

Lemma 1.9. *Let $L^p = L^p(\mathbb{R}_\xi^n)$. For any $p \in [1, \infty]$ and $\delta \in [0, 1]$, there is a constant $C > 0$ such that*

$$\|\nu^{-\delta} \Gamma(u, v)\|_{L^p} \leq C (\|\nu^{1-\delta} u\|_{L^p} \|v\|_{L^p} + \|u\|_{L^p} \|\nu^{1-\delta} v\|_{L^p}).$$

The case ($p = \infty, \delta = 1$) was first proved in [40] and the case ($p = 2, \delta = 1/2$) is due to [38]. Lemma 1.8 is an extension of Lemma 1.9.

2 Expansions and their unification

It is well known that the Boltzmann equation is closely related to the systems of fluid dynamics when the Knudsen number is small. In fact, in different physical scales, the Boltzmann equation reveals a mathematical hierarchy of fluid dynamical systems which include compressible and incompressible Euler equations and Navier-Stokes equations in both linear and nonlinear settings. Moreover, some non-classical fluid dynamical systems can also be derived from the Boltzmann equation for example when the bulk velocity is of the order of the Knudsen number to some power in a very rarefied gas. One of the typical examples is the thermal creep flow, [84].

In this section, we will use the decomposition of the solution into the macroscopic and microscopic components to reformulate the Boltzmann

equation as a system of conservation laws for the macroscopic components coupled with an equation for the microscopic component. This kind of thinking is similar to the Hilbert and Chapman-Enskog expansions where the leading term is a local Maxwellian with its macroscopic components governed by the conservation laws, either the compressible Euler equations or Navier-Stokes equations. The main difference between the reformulation introduced later and the classical expansions is that there is no approximation or truncation in the reformulation so that it is equivalent to the Boltzmann equation. Moreover, the system of conservation laws for the local Maxwellian has the framework of compressible Navier-Stokes equations with a source term determined by the microscopic component. In other words, this decomposition can be viewed as one kind of unification of the Hilbert and Chapman-Enskog expansions in some sense.

For the completeness and the convenience of the readers, the Hilbert and Chapman-Enskog expansions are reviewed in the following subsection before the decomposition and reformulation are given. Notice that in this section, we set the space dimension to be three.

2.1 Classical expansions

In this subsection, we will introduce two classical expansions to the Boltzmann equation when the Knudsen number is small. Consider the Boltzmann equation,

$$f_t + \xi \cdot \nabla_x f = \frac{1}{\kappa} Q(f, f), \quad (2.1)$$

where κ is the Knudsen number which is proportional to the mean free path. Here, we assume κ is a small constant and use it as the parameter for the expansion. In 1912, Hilbert [45] introduced the following classical expansion of the solution to the Boltzmann equation:

$$f = \sum_{n=0}^{\infty} \kappa^n f_n. \quad (2.2)$$

Putting this expansion into the Boltzmann equation (2.1) and comparing the terms by the order of κ yield the following equations for f_n :

$$\begin{aligned} Q_0 &= 0, \\ (f_{n-1})_t + \xi \cdot \nabla_x f_{n-1} &= Q_n, \quad n \geq 1, \end{aligned} \quad (2.3)$$

where

$$\begin{aligned} Q_0 &= Q(f_0, f_0), \\ Q_n &= 2Q(f_0, f_n) + \sum_{k=1}^{n-1} Q(f_k, f_{n-k}), \quad n \geq 1. \end{aligned} \quad (2.4)$$

Hence, by using the property [P2] in Section 1.2.2, the first equation in (2.4) implies that f_0 is a local Maxwellian, i.e.,

$$f_0 = \mathbf{M}_0 \equiv \mathbf{M}_{[\rho^0, u^0, \theta^0]} = \frac{\rho^0}{(2\pi R\theta^0)^{\frac{3}{2}}} \exp \left\{ -\frac{|\xi - u^0|^2}{2R\theta^0} \right\},$$

where ρ^0 , u^0 and θ^0 are functions of (t, x) . And f_0 satisfies

$$f_{0t} + \xi \cdot \nabla_x f_0 = Q_1. \quad (2.5)$$

Here, Q_1 is a microscopic component which is orthogonal to the five collision invariants $\psi_\alpha(\xi)$, $\alpha = 0, 1, \dots, 4$, given by [P1] in Section 1.2.2:

$$\begin{cases} \psi_0(\xi) \equiv 1, \\ \psi_i(\xi) \equiv \xi_i, \quad i = 1, 2, 3, \text{ or } \psi(\xi) = \xi, \\ \psi_4(\xi) \equiv \frac{1}{2}|\xi|^2. \end{cases}$$

The solvability condition for (2.5) gives the system of conservation laws

$$\int_{\mathbb{R}^3} \psi_\alpha(f_{0t} + \xi \cdot \nabla_x f_0) d\xi = 0,$$

which are exactly the compressible Euler equations

$$\begin{cases} \rho_t^0 + \nabla_x \cdot (\rho^0 u^0) = 0, \\ (\rho^0 u^0)_t + \nabla_x \cdot (\rho^0 u^0 \otimes u^0) + \nabla_x p^0 = 0, \\ [\rho^0 (\frac{|u^0|^2}{2} + \mathcal{E}^0)]_t + \nabla_x \cdot \{[\rho^0 (\frac{|u^0|^2}{2} + \mathcal{E}^0) + p^0] u\} = 0, \end{cases}$$

where the pressure function is given by $p^0 = R\rho^0\theta^0$ and the internal energy is $\mathcal{E}^0 = \frac{3}{2}R\theta^0$.

For $n \geq 1$, if we denote

$$S_n = \sum_{i=1}^{n-1} Q(f_k, f_{n-k}),$$

and

$$\mathbf{L}_{\mathbf{M}_0} h = 2Q(h, f_0),$$

which is the linearized collision operator with respect to the local Maxwellian \mathbf{M}_0 , then under the solvability condition for (2.3)₂

$$\int_{\mathbb{R}^3} \psi_\alpha((f_{n-1})_t + \xi \cdot \nabla_x f_{n-1}) d\xi = 0,$$

f_n can be represented in terms of f_i for $i = 0, 1, \dots, n-1$ by

$$f_n = \sum_{\alpha=0}^4 c_{\alpha}^{(n)} \psi_{\alpha} \mathbf{M}_0 + \mathbf{L}_{\mathbf{M}_0}^{-1} \{(f_{n-1})_t + \xi \cdot \nabla_x f_{n-1} - S_n\}.$$

Thus, the conservation laws

$$\int_{\mathbb{R}^3} \psi_{\alpha} (f_{nt} + \xi \cdot \nabla_x f_n) d\xi = 0, \quad \alpha = 0, 1, \dots, 4,$$

are the system of linear non-homogeneous first order hyperbolic linear partial differential equations for $c_{\alpha}^n, \alpha = 0, \dots, 4$ for the macroscopic components in f_n with source term determined by the f_k with $0 \leq k \leq n-1$.

Since to determine the value of f_n in the Hilbert expansion involves the differentiation of f_{n-1} , by induction, the convergence of this expansion can only be expected when the solution is infinitely differentiable and bounded with respect to the Knudsen number κ . Therefore, usually, the Hilbert expansion does not converge, especially in the present of initial layer, shock layer and boundary layer where the value of the differentiation grows when κ decreases.

Another classical expansion called Chapman-Enskog expansion was introduced by Chapman and Enskog in 1916 and 1917 independently. The main idea in this expansion is to expand both the equation and the solution, but to keep the conservative quantities unexpanded. The advantage of this expansion is that the first order correction yields the compressible Navier-Stokes equations for the macroscopic components so that the viscosity and heat conductivity are correctly represented. However, the drawback of the Chapman-Enskog expansion is that the higher order approximations give differential equations of higher order, such as the Burnett and super-Burnett equations for which there is no satisfactory mathematical theory. Basically, there is no established mathematical theory on this expansion beyond the compressible Navier-Stokes level.

Formally, we can write

$$\frac{\partial f}{\partial t} = \sum_{n=0}^{\infty} \kappa^n \frac{\partial^{(n)} f}{\partial t^n}. \quad (2.6)$$

The assumption on the unexpanded conserved quantities requires that for $n \geq 1$,

$$\int_{\mathbb{R}^3} \psi_{\alpha} f_n d\xi = 0, \quad \alpha = 0, 1, 2, 3, 4,$$

which implies that all the functions f_n for $n \geq 1$ are microscopic. By substituting (2.2) and (2.6) into the Boltzmann equation, we have

$$Q_0 = Q(f_0, f_0) = 0, \quad (2.7)$$

and for $n \geq 1$,

$$\sum_{i=0}^{n-1} \frac{\partial^{(i)} f_{n-i-1}}{\partial t} + \xi \cdot \nabla_x f_{n-1} = 2Q(f_0, f_n) + S_n, \quad (2.8)$$

where the notation has the same meaning as for the Hilbert expansion. However, one should notice that here each f_n is a functional of the conserved quantities which are not expanded. Again, the equation (2.7) implies that f_0 must be a local Maxwellian, i.e.,

$$f_0 = \mathbf{M} \equiv \mathbf{M}_{[\rho, u, \theta]} = \frac{\rho}{(2\pi R\theta)^{\frac{3}{2}}} \exp\left\{-\frac{|\xi - u|^2}{2R\theta}\right\}.$$

Therefore, the equation (2.8) for $n = 1$ can be written as

$$\frac{\partial^{(0)} f_0}{\partial t} + \xi \cdot \nabla_x f_0 = \mathbf{L}_M f_1. \quad (2.9)$$

The solvability condition for (2.9) immediately gives the following Euler equations

$$\begin{cases} \frac{\partial^{(0)} \rho}{\partial t} = -\frac{\partial}{\partial x_i}(\rho u_i), \\ \frac{\partial^{(0)} u_i}{\partial t} = -u_j \frac{\partial u_i}{\partial x_j} - \frac{1}{\rho} \frac{\partial p}{\partial x_i}, \quad i = 1, 2, 3, \\ \frac{\partial^{(0)} \theta}{\partial t} = -u_i \frac{\partial \theta}{\partial x_i} - \frac{2}{3} \theta \frac{\partial u_i}{\partial x_i}, \end{cases} \quad (2.10)$$

where $p = R\rho\theta$, here and in what follows, the summation is over any repeated indices. By plugging the expression of the local Maxwellian of f_0 into the equation (2.9), we have

$$\frac{1}{\rho} \mathbf{MB}^0 \rho + \frac{1}{\theta} \left(\frac{c^2}{2R\theta} - \frac{3}{2} \right) \mathbf{MB}^0 \theta + \frac{1}{R\theta} c_j \mathbf{B}^0 u_j = \mathbf{L}_M f_1, \quad (2.11)$$

where $c = \xi - u$ is the random velocity and \mathbf{B}^0 is the following linear operator

$$\mathbf{B}^0 = \frac{\partial^{(0)}}{\partial t} + \xi \cdot \nabla_x.$$

By substituting the time derivative $\frac{\partial^{(0)}}{\partial t}$ of (2.10) into the equation (2.11), we have

$$\left(\frac{c^2}{2R\theta} - \frac{5}{2} \right) \mathbf{M} \frac{c_i}{\theta} \frac{\partial \theta}{\partial x_i} + \frac{1}{R\theta} \left(c_i c_j - \frac{1}{3} c^2 \delta_{ij} \right) \mathbf{M} \frac{\partial u_i}{\partial x_j} = \mathbf{L}_M f_1.$$

In terms of the Burnett functions defined by

$$\begin{cases} A_j(\xi) = \frac{|\xi|^2 - 5}{2} \xi^j, & j = 1, 2, 3, \\ B_{ij}(\xi) = \xi^i \xi^j - \frac{1}{3} \delta_{ij} |\xi|^2, & i, j = 1, 2, 3, \end{cases}$$

$$f_1 = \mathbf{L}_M^{-1} \left(\sqrt{R} A_i \left(\frac{c}{\sqrt{R\theta}} \right) M \frac{\partial \theta}{\partial x_i} + B_{ij} \left(\frac{c}{\sqrt{R\theta}} \right) M \frac{\partial u_i}{\partial x_j} \right).$$

Notice that we have used the fact that for Maxwell molecules or cutoff hard potentials, the range of the operator \mathbf{L}_M is \mathcal{N}^\perp and the operator \mathbf{L}_M is invertible and bounded from \mathcal{N}^\perp to \mathcal{N}^\perp .

The properties of the Burnett functions which are also macroscopic are given in the following proposition.

Proposition 2.1. *Denote*

$$A' = \mathbf{L}_M^{-1} A, \quad B' = \mathbf{L}_M B.$$

Then there exist positive functions $a(r)$ and $b(r)$ defined on $[0, \infty)$ such that

$$A'(\xi) = -a(|\xi|) A(\xi), \quad B'(\xi) = -b(|\xi|) B(\xi).$$

And the following properties hold,

- $\langle -A_i, A'_i \rangle$ is positive and independent of i .
- $\langle A_i, A'_j \rangle = 0$ for any $i \neq j$.
- $\langle A_i, B'_{jk} \rangle = 0$ for any i, j, k .
- $\langle B_{ij}, B'_{kl} \rangle = \langle B_{kl}, B'_{ij} \rangle = \langle B_{ji}, B'_{kl} \rangle$ holds and is independent of i, j for any fixed k, l .
- $-\langle B_{ij}, B'_{ij} \rangle$ is positive and independent of i, j when $i \neq j$.
- $-\langle B_{ii}, B'_{jj} \rangle$ is positive and independent of i, j when $i \neq j$.
- $-\langle B_{ii}, B'_{jj} \rangle$ is positive and independent of i .
- $\langle B_{ij}, B'_{kl} \rangle = 0$ unless either $(i, j) = (k, l)$ or (l, k) , or $i = j$ and $k = l$.
- $\langle B_{ii}, B'_{ii} \rangle - \langle B_{ii}, B'_{jj} \rangle = 2 \langle B_{ij}, B'_{ij} \rangle$ holds for any $i \neq j$.

The proof of this proposition can be found in [3] and we omit the details here.

With the Burnett functions, the viscosity $\mu(\theta)$ and heat conductivity coefficient $\kappa(\theta)$ can be represented by

$$\begin{cases} \mu(\theta) = -\kappa R \theta \int_{\mathbb{R}^3} B_{ij} \left(\frac{c}{\sqrt{R\theta}} \right) \mathbf{L}_M^{-1} \left(B_{ij} \left(\frac{c}{\sqrt{R\theta}} \right) M \right) d\xi > 0, & i \neq j, \\ \kappa(\theta) = -\kappa R^2 \theta \int_{\mathbb{R}^3} A_l \left(\frac{c}{\sqrt{R\theta}} \right) \mathbf{L}_M^{-1} \left(A_l \left(\frac{c}{\sqrt{R\theta}} \right) M \right) d\xi > 0. \end{cases}$$

Note that these coefficients are independent of the density function ρ for the Boltzmann equation where the collision operator is bi-linear.

Now if we put f_1 into the conservation laws to include the first order approximation, then the conservation laws take the form

$$\int_{\mathbb{R}^3} \psi_\alpha((f_0)_t + \xi \cdot \nabla_x(f_0 + \kappa f_1))d\xi = 0.$$

Since f_1 is microscopic, its contribution to the conservation of mass is zero. And its contribution to the equations of conservation of momentum and energy is represented by the stress tensor and heat flux:

$$p_{ij}^{(1)} = \kappa \int_{\mathbb{R}^3} c_i c_j f_1 d\xi, \quad q_i^{(1)} = \frac{\kappa}{2} \int_{\mathbb{R}^3} c_i c^2 f_1 d\xi.$$

With Proposition 2.1, it is straightforward to calculate the stress tensor and heat flux in terms of the fluid variables:

$$\begin{cases} p_{ij}^{(1)} = -\mu(\theta) \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \frac{2}{3}\mu(\theta) \frac{\partial u_k}{\partial x_k} \delta_{ij}, \\ q_i^{(1)} = -\varkappa(\theta) \frac{\partial \theta}{\partial x_i}. \end{cases}$$

In summary, the first order approximation in the Chapman-Enskog expansion is the compressible Navier-Stokes equations:

$$\left\{ \begin{array}{l} \rho_t + \nabla_x \cdot (\rho u) = 0, \\ (\rho u^i)_t + \nabla_x \cdot (\rho u^i u) + p_{x_i} = [\mu(\theta)(u_{x_j}^i + u_{x_i}^j - \frac{2}{3}\delta_{ij}\nabla_x \cdot u)]_{x_j}, \quad i = 1, 2, 3, \\ \left[\rho \left(\frac{1}{2}|u|^2 + \mathcal{E} \right) \right]_t + \nabla_x \cdot \left(\left[\rho \left(\frac{1}{2}|u|^2 + \mathcal{E} \right) + p \right] u \right) \\ \quad = \mu(\theta)u^i \left(u_{x_j}^i + u_{x_i}^j - \frac{2}{3}\delta_{ij}\nabla_x \cdot u \right)_{x_j} + (\varkappa(\theta)\theta_{x_j})_{x_j}. \end{array} \right.$$

Again, by similar but tedious calculation, we can find the next terms, f_2, f_3, \dots , in the Chapman-Enskog expansion, however, without good mathematical theory. In the next subsection, we will give a decomposition and reformulation of the Boltzmann equation without any expansion so that the structure of the systems for fluid dynamics together with the effects from the microscopic component become clear and exact. In fact, one can compare it with the Hilbert and Chapman-Enskog expansions so that some similarities and subtle difference can be found.

2.2 Unification by decomposition

Based on the decomposition of the solution into its macroscopic (fluid dynamic) and microscopic (kinetic) component, we will reformulate the Boltzmann equation into a system of conservation laws for the time evolution of the macroscopic variables and an equation for the time evolution of the microscopic variable. The main idea is not to have any approximation, but a complete description of the solutions to the Boltzmann equation so that the analytic techniques such as traditional energy method from the theory of conservation laws can be applied in the study of the Boltzmann equation. The stability of wave patterns for the Boltzmann equation by using energy method was initiated in [70] by the study of the shock profile constructed in [16], through a rigorous analysis of the Chapman-Enskog expansion where the macro-micro decomposition is defined around the local Maxwellian given by the Chapman-Enskog expansion. For the stability of the rarefaction waves and contact discontinuity, see [47, 69].

The reformulation of the Boltzmann equation using the macro-micro decomposition with respect to the local Maxwellian defined by the solution itself was introduced in [68] as explained in the following. Since this reformulation gives the Euler and Navier-Stokes equations directly, it can be viewed as a unification of the above two classical expansions.

To be precise, let $f(t, x, \xi)$ be the solution to the Boltzmann equation. We decompose it into the macroscopic component in the form of the local Maxwellian $\mathbf{M} = \mathbf{M}(x, t, \xi) = \mathbf{M}_{[\rho, u, \theta]}(\xi)$, and the microscopic component $\mathbf{G} = \mathbf{G}(x, t, \xi)$:

$$f(t, x, \xi) = \mathbf{M}(t, x, \xi) + \mathbf{G}(t, x, \xi).$$

Here, $\mathbf{M}(t, x, \xi)$ is the local Maxwellian with its five fluid parameters (ρ, u, θ) defined by the five conserved quantities of f , the mass density $\rho(t, x)$, momentum $m(t, x) = \rho(t, x)u(t, x)$ and energy density $\mathcal{E}(t, x) + |u(t, x)|^2/2$:

$$\left\{ \begin{array}{l} \rho(t, x) \equiv \int_{\mathbb{R}^3} f(t, x, \xi) d\xi, \\ m^i(t, x) = \rho u \equiv \int_{\mathbb{R}^3} \psi_i(\xi) f(t, x, \xi) d\xi, \quad i = 1, 2, 3, \\ [\rho (\mathcal{E} + \frac{1}{2}|u|^2)](t, x) \equiv \int_{\mathbb{R}^3} \psi_4(\xi) f(t, x, \xi) d\xi. \end{array} \right.$$

To have an orthogonal basis for the subspace of the macroscopic components, we first define an inner product in the space \mathbf{L}_ξ^2 with a

weight. For this, let $\tilde{\mathbf{M}} = \tilde{\mathbf{M}}_{[\tilde{\rho}, \tilde{u}, \tilde{\theta}]}$ be any given Maxwellian. Define

$$\langle h, g \rangle_{\tilde{\mathbf{M}}} \equiv \int_{\mathbb{R}^3} \frac{h(\xi)g(\xi)}{\tilde{\mathbf{M}}} d\xi$$

for functions h and g of ξ such that the integral is well defined. Using this inner product, the subspace spanned by the collision invariants has the following set of orthogonal basis:

$$\left\{ \begin{array}{l} \chi_0^{\tilde{\mathbf{M}}} = \chi_0(\xi; \tilde{\rho}, \tilde{u}, \tilde{\theta}) \equiv \frac{1}{\sqrt{\tilde{\rho}}} \tilde{\mathbf{M}}, \\ \chi_i^{\tilde{\mathbf{M}}} = \chi_i(\xi; \tilde{\rho}, \tilde{u}, \tilde{\theta}) \equiv \frac{\xi^i - \tilde{u}^i}{\sqrt{R\tilde{\theta}\tilde{\rho}}} \tilde{\mathbf{M}}, \quad i = 1, 2, 3, \\ \chi_4^{\tilde{\mathbf{M}}} = \chi_4(\xi; \tilde{\rho}, \tilde{u}, \tilde{\theta}) \equiv \frac{1}{\sqrt{6\tilde{\rho}}} \left(\frac{|\xi - \tilde{u}|^2}{R\tilde{\theta}} - 3 \right) \tilde{\mathbf{M}}, \\ \langle \chi_\alpha^{\tilde{\mathbf{M}}}, \chi_\beta^{\tilde{\mathbf{M}}} \rangle_{\tilde{\mathbf{M}}} = \delta_{\alpha\beta}, \quad \text{for } \alpha, \beta = 0, 1, 2, 3, 4. \end{array} \right.$$

With this basis, define the macroscopic projection $\mathbf{P}_0^{\tilde{\mathbf{M}}}$ and microscopic projection $\mathbf{P}_1^{\tilde{\mathbf{M}}}$ by:

$$\left\{ \begin{array}{l} \mathbf{P}_0^{\tilde{\mathbf{M}}} h \equiv \sum_{\alpha=0}^4 \langle h, \chi_\alpha^{\tilde{\mathbf{M}}} \rangle_{\tilde{\mathbf{M}}} \chi_\alpha^{\tilde{\mathbf{M}}}, \\ \mathbf{P}_1^{\tilde{\mathbf{M}}} h \equiv h - \mathbf{P}_0^{\tilde{\mathbf{M}}} h. \end{array} \right.$$

Notice that the operators $\mathbf{P}_0^{\tilde{\mathbf{M}}}$ and $\mathbf{P}_1^{\tilde{\mathbf{M}}}$ are orthogonal projections, that is,

$$\mathbf{P}_0^{\tilde{\mathbf{M}}} \mathbf{P}_0^{\tilde{\mathbf{M}}} = \mathbf{P}_0^{\tilde{\mathbf{M}}}, \quad \mathbf{P}_1^{\tilde{\mathbf{M}}} \mathbf{P}_1^{\tilde{\mathbf{M}}} = \mathbf{P}_1^{\tilde{\mathbf{M}}}, \quad \mathbf{P}_0^{\tilde{\mathbf{M}}} \mathbf{P}_1^{\tilde{\mathbf{M}}} = \mathbf{P}_1^{\tilde{\mathbf{M}}} \mathbf{P}_0^{\tilde{\mathbf{M}}} = 0.$$

Now, the system of conservation laws

$$\int_{\mathbb{R}^3} \psi_\alpha(f_t + \xi \cdot \nabla_x f) d\xi = 0, \quad \alpha = 0, 1, \dots, 4,$$

takes the following form

$$\left\{ \begin{array}{l} \rho_t + \operatorname{div}_x m = 0, \\ m_t^i + \left(\sum_{j=1}^3 u^j m^i \right)_{x^j} + p_{x^i} + \int_{\mathbb{R}^3} \psi_i(\xi) (\xi \cdot \nabla_x \mathbf{G}) d\xi = 0, \quad i = 1, 2, 3, \\ \left[\rho \left(\frac{|u|^2}{2} + \mathcal{E} \right) \right]_t + \sum_{j=1}^3 \left\{ u^j \left[\rho \left(\frac{|u|^2}{2} + \mathcal{E} \right) + p \right]_{x^j} \right\}_{x^j} + \int_{\mathbb{R}^3} \psi_4(\xi) (\xi \cdot \nabla_x \mathbf{G}) d\xi = 0. \end{array} \right. \quad (2.12)$$

The equation of the state is for the monatomic gas, with the gas constant R chosen to be $\frac{2}{3}$ without loss of generality, given by

$$p = \frac{2}{3} \rho \mathcal{E}.$$

And the macroscopic entropy S can be defined as:

$$S = -\frac{2}{3} \ln \rho + \ln \left(\frac{4}{3} \pi \theta \right) + 1.$$

The microscopic equation is obtained by applying the microscopic projection $\mathbf{P}_1^{\mathbf{M}}$ to the Boltzmann equation (2.1):

$$\mathbf{G}_t + \mathbf{P}_1^{\mathbf{M}} \left(\xi \cdot \nabla_x \mathbf{G} + \xi \cdot \nabla_x \mathbf{M} \right) = \frac{1}{\kappa} \mathbf{L}_{\mathbf{M}} \mathbf{G} + \frac{1}{\kappa} Q(\mathbf{G}, \mathbf{G}), \quad (2.13)$$

where $\mathbf{L}_{\mathbf{M}}$ is the linearized collision operator around the local Maxwellian \mathbf{M} .

From (2.13), we have

$$\begin{aligned} \mathbf{G} &= \kappa \mathbf{L}_{\mathbf{M}}^{-1} (\mathbf{P}_1^{\mathbf{M}} (\xi \cdot \nabla_x \mathbf{M})) \\ &\quad + \mathbf{L}_{\mathbf{M}}^{-1} (\kappa (\partial_t \mathbf{G} + \mathbf{P}_1^{\mathbf{M}} \xi \cdot (\nabla_x \mathbf{G})) - Q(\mathbf{G}, \mathbf{G})) \\ &= \kappa \mathbf{L}_{\mathbf{M}}^{-1} (\mathbf{P}_1^{\mathbf{M}} (\xi \cdot \nabla_x \mathbf{M})) + \Theta. \end{aligned} \quad (2.14)$$

Substituting (2.14) into (2.12) yields the following fluid-type system for the macroscopic components:

$$\left\{ \begin{array}{l} \rho_t + \operatorname{div}_x m = 0, \\ \\ m_t^i + \left(\sum_{j=1}^3 u^j m^i \right)_{x^j} + p_{x^i} + \kappa \int_{\mathbb{R}^3} \psi_i(\xi) (\xi \cdot \nabla_x \mathbf{L}_{\mathbf{M}}^{-1} (\mathbf{P}_1^{\mathbf{M}} (\xi \cdot \nabla_x \mathbf{M}))) d\xi \\ \quad + \int_{\mathbb{R}^3} \psi_i(\xi) (\xi \cdot \nabla_x \Theta) d\xi = 0, \quad i = 1, 2, 3, \\ \\ \left[\rho \left(\frac{|u|^2}{2} + \mathcal{E} \right) \right]_t + \sum_{j=1}^3 \left\{ u^j \left[\rho \left(\frac{|u|^2}{2} + \mathcal{E} \right) + p \right] \right\}_{x^j} \\ \quad + \kappa \int_{\mathbb{R}^3} \psi_4(\xi) (\xi \cdot \nabla_x \mathbf{L}_{\mathbf{M}}^{-1} (\mathbf{P}_1^{\mathbf{M}} (\xi \cdot \nabla_x \mathbf{M}))) d\xi \\ \quad + \int_{\mathbb{R}^3} \psi_4(\xi) (\xi \cdot \nabla_x \Theta) d\xi = 0. \end{array} \right. \quad (2.15)$$

A straightforward calculation from (2.15) gives the following fluid-type

system

$$\left\{ \begin{array}{l} \rho_t + \operatorname{div}_x m = 0, \\ \\ m_t^i + \sum_{j=1}^3 (u^j m^i)_{x^j} + p_{x^i} = \sum_{j=1}^3 \left[\mu(\theta) \left(u_{x^j}^i + u_{x^i}^j - \frac{2}{3} \delta_{ij} \operatorname{div}_x u \right) \right]_{x^j} \\ \quad - \int_{\mathbb{R}^3} \psi_i(\xi) (\xi \cdot \nabla_x \Theta) d\xi, \quad i = 1, 2, 3, \\ \\ [\rho(\frac{1}{2}|u|^2 + \mathcal{E})]_t + \sum_{j=1}^3 \left(u^j \left(\rho \left(\frac{1}{2}|u|^2 + \mathcal{E} \right) + p \right) \right)_{x^j} \\ = \sum_{i,j=1}^3 \left\{ \mu(\theta) u^i \left(u_{x^j}^i + u_{x^i}^j - \frac{2}{3} \delta_{ij} \operatorname{div}_x u \right) \right\}_{x^j} \\ \quad + \sum_{j=1}^3 (\kappa(\theta) \theta_{x^j})_{x^j} - \int_{\mathbb{R}^3} \psi_4(\xi) (\xi \cdot \nabla_x \Theta) d\xi. \end{array} \right. \quad (2.16)$$

From this fluid-type system, one can easily see the structure of the compressible Euler and the compressible Navier-Stokes equations. For instance, when the Knudsen number κ and Θ are set zero, the system (2.16) becomes the compressible Euler equations. On the other hand, when Θ is set to be zero in (2.16), it becomes the compressible Navier-Stokes equations. These fluid equations as derived through the Hilbert and Chapman-Enskog expansions are approximations to the Boltzmann equation. However, the above system is part of the Boltzmann equation. Nevertheless, this reformulation is consistent in spirit with the Chapman-Enskog expansion in that the higher order terms beyond zeroth order in the expansions must be microscopic. Therefore, it is interesting to notice that the first order approximation in Chapman-Enskog expansion f_1 is just the leading term in the microscopic component expression (2.14), that is,

$$f_1 = \mathbf{L}_M^{-1}(\mathbf{P}_1^M(\xi \cdot \nabla_x M)).$$

This re-formulation of the Boltzmann equation is useful to investigate the solution behavior of the Boltzmann equation which has counterpart in fluid dynamics, especially in the Navier-Stokes equations. And it is applied to study the wave patterns, non-trivial solution profiles, optimal convergence rates, time-periodic solutions etc. However, there is another kind of solution behavior of the Boltzmann equation which can not be described by the classical fluid dynamical systems which is called the ghost effect such as the thermal creep flow and thermal edge flow. Even though we will not touch the topic of ghost effect in this notes, it should be noted that it is an important subject for rarefied gas and there is no nonlinear mathematical theory for this phenomena so far.

Moreover, we should mention that the energy method based on another form of decomposition was also introduced by Guo in [41] where the solution is decomposed with respect to the global Maxwellian. This kind of decomposition is closely related to Grad's 13 moments' estimation. In addition, it is proved to be useful for the study of the large time behavior of solutions with global Maxwellian as the time asymptotic state even with external force.

3 Detour to hyperbolic conservation laws

In the last section, the relation between the Boltzmann equation and the typical examples of conservation laws becomes clear when the Knudsen number is small. Thus, it is now appropriate to have a detour to the mathematical theories for conservation laws. In this section, we will give a brief presentation of the well-posedness theory of hyperbolic conservation laws in one dimensional space in order to help the readers to have a better picture of the solution behavior in the fluid dynamic level, that is, at local equilibrium.

The study of conservation laws has a long history and the earliest mathematical work can be traced back to Euler in 1755 on the study of acoustic waves. And the pioneer nonlinear formulation on fluid dynamics was done by Riemann through the consideration of two stationary gases separated by a membrane when the membrane is suddenly removed. This fundamental work bearing the name of "Riemann problem" is so fundamental that it is still a hot topic in more general mathematical and physical settings. The Riemann problem is used to obtain global existence of entropy solution, a unique picture of large time behavior of the solution with the typical non-interacting wave patterns.

With the efforts of many prestigious mathematicians, like von Neumann, Courant, Friedrichs, Lax, Glimm, etc., we now have a satisfactory well-posedness theory for hyperbolic conservation laws when the total variation of the solution is small. In fact, the theory for scalar conservation laws was established in late 1960s through the classical papers by Oleinik, Volpert and Kruzhkov. And then the fundamental paper by Glimm in 1965 on the global existence of weak solution is based on the Glimm scheme and the solution to Riemann problem by Lax. Since then, the stability and uniqueness remained open until the breakthrough made by Bressan in 1995. Later, a new functional approach was introduced by Liu-Yang and then gives a satisfactory theory for the system. Besides the fundamental existence theory of Glimm, the recent L^1 stability theory leads to the mature stability and uniqueness theory for hyperbolic conservation laws with small total variation. Indeed, the first breakthrough on L^1 stability was made by Bressan, et. al. By using

the generalized entropy functional, called Liu-Yang functional, now the functional approach gives an elegant way to show the intrinsic mechanism in the L^1 topology. However, for large data, it is almost open even though there is an elegant method called compensated compactness, for the scalar equations and the 2×2 systems by Murat, Tartar, DiPerna and Chen-Ding-Luo, and some other works on piecewise constant solutions separated by non-interacting large waves.

There are many books and monographs on the theory of hyperbolic conservation laws, such as the classical books by Courant-Friedrichs [22] and Courant-Hilbert [23], the one used in this field for many years by Smoller [83], the recent ones by Bressan [8], Dafermos [24], LeFloch [55] and Serre [80]. Moreover, the multi-dimensional problems are discussed by Madja [72], numerical computation by LeVeque [57] and physical models with asymptotic analysis by Witham [101]. Therefore, this chapter can only serve as a very brief introduction to this rich area.

To be concentrated, we will present the well-posedness theory for solutions with small total variation. We start from the definition of weak solution through the Burgers equation. Then the construction of wave curves, the solution to the Riemann problem, the solution to the Cauchy problem, stability in L^1 norm and uniqueness will be given accordingly. Furthermore, we will state the result on the vanishing viscosity by Bianchini-Bressan and the convergence rate estimate by Bressan-Yang.

3.1 Scalar conservation laws

By using the scalar equation, we will introduce the concepts of shock waves, rarefaction waves and entropy for hyperbolic conservation laws. For the complete study of the well-posedness theory for scalar conservation laws see [52].

Consider the scalar equation

$$u_t + f(u)_x = 0, \quad (3.1)$$

where $u = u(x, t)$ is the unknown function and $f(u)$ is a smooth nonlinear function. When $f(u) = \frac{u^2}{2}$, it is the classical inviscid Burgers equation. The Riemann problem is to study the (3.1) when the initial data is given by

$$u(x, 0) = \begin{cases} u_l, & x < 0, \\ u_r, & x > 0. \end{cases}$$

Since both the equation and the initial data are invariant under the scaling $x \rightarrow \kappa x, t \rightarrow \kappa t$ for any positive constant κ , the solution to the Riemann problem depends only on $\frac{x}{t}$. Hence, we can look for the self-similar solution $u(x, t) = \phi(\frac{x}{t}) = \phi(\xi)$. Plug this into the equation, we

have

$$f'(\phi)\phi' = \xi\phi',$$

where ' denotes the differentiation with respect to the corresponding variable. Therefore, we have either $\phi'(\xi) = 0$ or $f'(\phi) = \xi$. When two end states are not the same, $\phi'(\xi)$ may be defined in the weak sense. In fact, it is straightforward to show that even though the initial data can be arbitrarily smooth, the solution to (3.1) may blowup in finite time. For example, let $f''(u) > 0$, differentiating (3.1) gives

$$v_t + f'(u)v = -f''(u)v^2,$$

where $v = u_x$. Therefore, along the characteristic $\frac{dx}{dt} = f'(u(x,t))$, we have

$$v(x(t), t) = \frac{v(x(0), 0)}{1 + v(x(0), 0) \int_0^t f''(u(x(s), s)) ds}.$$

It is then obvious that if $v(x(0), 0) < 0$ at some point $x(0)$, there exists $0 < T < \infty$ such that $\lim_{t \rightarrow T^-} v(x(t), t) = -\infty$.

For this, we need to consider the weak solution. That is, we will consider weak solution defined as follows.

Definition 3.1. A bounded measurable function $u(x, t)$ is a weak solution of (3.1) with initial data $u(x, 0) = u_0(x)$ if and only if

$$\int_0^\infty \int_{-\infty}^\infty [\phi_t u + \phi_x f(u)](x, t) dx dt + \int_{-\infty}^\infty \phi(x, 0) u_0(x) dx = 0$$

for any smooth function $\phi(x, t)$ of compact support in

$$\{(x, t) \mid (x, t) \in \mathbf{R}^2, t \geq 0\}.$$

As a consequence of the weak formulation, a discontinuity (u_-, u_+) in the weak solution with speed s satisfies the Rankine-Hugoniot (jump) condition

$$s(u_+ - u_-) = f(u_+) - f(u_-), \quad (3.2)$$

where u_- and u_+ are the left and right states of the discontinuity respectively.

This prompts the introduction of the Hugoniot curves $H(u_0)$ passing through a given state u_0 as follows:

$$H(u_0) \equiv \{u : \sigma(u_0 - u) = f(u_0) - f(u)\},$$

for some scalar $\sigma = \sigma(u_0, u)$. The Rankine-Hugoniot condition says that if $u_+ \in H(u_-)$, then the function consisting of two states u_\pm separated by a straight line in $x - t$ plane with slope $s = \sigma(u_-, u_+)$ is a weak solution.

Back to the solution to the Riemann problem for the scalar conservation law, by construction, we can have the following two cases when the flux function $f(u)$ is convex, i.e., $f''(u) > 0$.

- Case 1. When $u_r > u_l$, the solution is given by a rarefaction wave defined by

$$u(x, t) = \begin{cases} u_l, & \frac{x}{t} < f'(u_l), \\ (f')^{-1}\left(\frac{x}{t}\right), & f'(u_l) < \frac{x}{t} < f'(u_r), \\ u_r, & \frac{x}{t} > f'(u_r), \end{cases}$$

where $(f')^{-1}$ represents the inverse function of f' which exists because of $f'' > 0$.

- Case 2. When $u_r < u_l$, the solution is given by a shock wave defined by

$$u(x, t) = \begin{cases} u_l, & \frac{x}{t} < s, \\ u_r, & \frac{x}{t} > s, \end{cases}$$

where s is the shock speed given by (3.2).

However, the weak solution for Case 1 is not unique. In fact, we can construct infinitely many weak solutions for Case 1. For example,

$$u(x, t) = \begin{cases} u_l, & \frac{x}{t} < f'(u_l), \\ (f')^{-1}\left(\frac{x}{t}\right), & f'(u_l) < \frac{x}{t} < f'(u_1), \\ u_1, & f'(u_1) < \frac{x}{t} < \tilde{s}, \\ u_2, & \tilde{s} < \frac{x}{t} < f'(u_2), \\ (f')^{-1}\left(\frac{x}{t}\right), & f'(u_2) < \frac{x}{t} < f'(u_r), \\ u_r, & \frac{x}{t} > f'(u_r), \end{cases}$$

where u_1 and u_2 are any two constants satisfying $u_l < u_1 < u_2 < u_r$, and $f(u_2) - f(u_1) = \tilde{s}(u_2 - u_1)$.

In order to obtain a unique physical solution, we need the entropy condition for discontinuity. For the case with convex flux, this is given by the Lax' entropy condition saying that a discontinuity (u_l, u_r) satisfying (3.2) is admissible if $f'(u_r) < s < f'(u_l)$. This entropy condition is consistent with the Oleinik entropic condition and the vanishing viscosity limit of

$$u_t + f_x(u) = \epsilon u_{xx},$$

when $\epsilon > 0$ tends to 0.

There are other versions of entropy condition, especially for general flux functions. For example, when the flux function is neither convex nor concave, we have the Oleinik entropy condition saying that a discontinuity (u_l, u_r) is admissible if the curve of $y = f(u)$ is above the line segment

connecting the two points $(u_l, f(u_l))$ and $(u_r, f(u_r))$ if $u_r > u_l$, otherwise, it is below the line segment, [77]. And there are other entropy conditions, such as the maximizing entropy production and vanishing viscosity criterion, [25]. For the relation between different versions of these entropy conditions, please refer to the paper [63]. In the later presentation for system, we will use the condition proposed in [62] which is a generalized version of the Lax's entropy condition for genuinely nonlinear characteristic fields.

For general flux function, the solution to the Riemann problem consists of composite waves, that is, a composition of rarefaction waves and shocks with increasing propagation speeds.

There are many special properties of scalar conservation laws. Firstly, the solution operator of a scalar conservation law is an L_1 contraction semigroup as stated in the following lemma, cf. [51].

Lemma 3.1. *Let u_i , $i = 1, 2$ be two solutions of (3.1) satisfying the entropy condition, then*

$$\|u_1(x, t) - u_2(x, t)\|_{L_1} \leq \|u_1(x, s) - u_2(x, s)\|_{L_1}, \quad \text{for } s \leq t.$$

Furthermore, the inequality sign holds at time t only when there exists a shock wave of one solution which crosses the solution curve of the other solution and the decrease is guaranteed by the entropy condition.

For scalar conservation laws, any convex function of u can be an entropy. By choosing the particular convex entropy $\eta(u) = \frac{u^2}{2}$ with entropy flux $q(u) = \int^u s f'(s) ds$, we have the following entropy estimate.

Lemma 3.2. *Let $u(x, t)$ be a weak solution to the scalar conservation law (3.1) consisting of countably many admissible shocks, denoted by $\{\alpha_i\}$. Then we have*

$$\frac{d}{dt} \int_{\mathbb{R}} u^2(x, t) dx = -2 \sum_{\alpha_i} A(\alpha_i).$$

Here, for any admissible shock $\alpha = (u^-, u^+)$, $A(\alpha)$ denotes the area bounded by the curve $y = f(u)$ and the straight line segment connecting the end points $(u^-, f(u^-))$ and $(u^+, f(u^+))$ in the $u - y$ plane.

Proof. If the solution is smooth, then it is obvious that

$$\frac{d}{dt} \int_{\mathbb{R}} u^2(x, t) dx = 0$$

if $u_+ = u_-$. Without loss of generality, we consider the contribution of a single shock to this derivative. Let $\alpha_i = (u^-, u^+)$ be an admissible shock

located at $x = x(t)$. We have

$$\begin{aligned} \frac{d}{dt} \int_{\mathbb{R}} \frac{u^2}{2}(x, t) dx &= \frac{1}{2} \dot{x}(t)(u^{-2} - u^{+2}) - q(u^-) + q(u^+) \\ &\quad + \text{other terms not related to } (u^-, u^+), \end{aligned}$$

where $q' = uf'$ is the corresponding entropy flux. The term on the right hand side of the above equality can be calculated as follows.

$$\begin{aligned} &\frac{1}{2} \dot{x}(t)(u^{-2} - u^{+2}) - q(u^-) + q(u^+) \\ &= \frac{1}{2} (f(u^-) - f(u^+))(u^- + u^+) - f(u^-)u^- + f(u^+)u^+ + \int_{u^+}^{u^-} f(t) dt \\ &= \frac{1}{2} (u^+ - u^-)(f(u^+) + f(u^-)) - \int_{u^-}^{u^+} f(t) dt = -A(\alpha_i). \end{aligned}$$

Summing the terms for all shocks gives the proof of this lemma. \square

This entropy estimate is closely related to the bifurcation of the Hugoniot curve from the rarefaction wave curve in a general system.

Finally in this subsection, we mention that for both scalar equation and system, the solution to the Riemann problem gives the time asymptotic behavior of the solution to the Cauchy problem. And for initial data of compact support, the time asymptotic behavior is given by the N-wave, [37]. However, we will not touch these topics in this section.

Since the discontinuity in the solution satisfies the Rankine-Hugoniot condition, any nonlinear transformation of the unknown function leads to different propagation speeds so that nonlinear transformation of the equation is not allowed when we discuss weak solutions.

3.2 Riemann problem for systems

We now turn to the Cauchy problem for a general system of hyperbolic conservation laws

$$u_t + f(u)_x = 0, \tag{3.3}$$

$$u(x, 0) = u_0(x), \tag{3.4}$$

where $u = u(x, t) = (u^1(x, t), \dots, u^n(x, t))$ and $f(u)$ are n -vectors.

The system is assumed to be strictly hyperbolic, that is, the eigenvalues of the $n \times n$ matrix $f'(u)$ are real and distinct:

$$f'(u)r_i(u) = \lambda_i(u)r_i(u),$$

$$l_i(u)f'(u) = \lambda_i(u)l_i(u),$$

$$l_i(u) \cdot r_j(u) = \delta_{ij}, \quad i, j = 1, 2, \dots, n,$$

$$\lambda_1(u) < \lambda_2(u) < \dots < \lambda_n(u).$$

By a linear transformation, if necessary, we may assume that the i -th component u^i of the vector u is strictly increasing in the direction of r_i . This can be done at least for a small neighborhood of a given state. In the following we will use u^i to measure the wave strength of an i -wave.

The same as for the scalar equation, because of the dependence of the characteristics $\lambda_i(u)$ on the dependent variables u , waves may compress and smooth solutions in general do not exist globally in time.

It follows easily from the strict hyperbolicity of the system that in a small neighborhood of a given state u_0 , the set $H(u_0)$ consists of n smooth curves $H_i(u_0)$, $i = 1, 2, \dots, n$ through u_0 , such that $\sigma_i(u_0, u)$ tends to $\lambda_i(u_0)$ as u moves along $H_i(u_0)$ toward u_0 . Here we use the notation $\sigma_i(u_0, u)$ to denote the scalar $\sigma(u_0, u)$ in $H_i(u_0)$. A discontinuity (u_-, u_+) , $u_+, u_- \in H_i(u_-)$, is called an i -discontinuity.

As shown for the scalar equation, weak solution to the initial value problem (3.3) and (3.4) is not unique. Certain admissibility condition, the entropy condition, needs to be imposed on the weak solution to rule out non-physical discontinuities and we will adopt the one introduced in [62].

Definition 3.2. A discontinuity (u_-, u_+) is admissible if

$$\sigma(u_-, u_+) \leq \sigma(u_-, u),$$

for any state u on the Hugoniot curve $H(u_-)$ between u_- and u_+ .

If a characteristic field of the system (3.3) is genuinely nonlinear, [54], in the sense that

$$\nabla \lambda_i(u) \cdot r_i(u) \neq 0 \quad (g.nl.),$$

then the entropy condition is reduced to the Lax's entropy condition

$$\lambda_i(u_+) < \sigma_i(u_-, u_+) < \lambda_i(u_-).$$

If a characteristic field of the system (3.3) is linearly degenerate, i.e.,

$$\nabla \lambda_i(u) \cdot r_i(u) \equiv 0 \quad (l.dg.),$$

then the entropy condition is reduced to the one for linear waves

$$\lambda_i(u_+) = \sigma_i(u_-, u_+) = \lambda_i(u_-).$$

When each characteristic field is either genuinely nonlinear or linearly degenerate, there is a classical existence theory of Glimm, [36]. An important physical example of such a system is the Euler equations

in gas dynamics. Other physical systems, such as those in elasticity and magneto-hydrodynamics, for instance, are not necessarily genuinely nonlinear or linearly degenerate.

In the following, we will present the existence theory for a general system. Thus for a given characteristic field $\lambda_i(u)$, we allow the linearly degenerate manifold $LG_i \equiv \{u : \nabla \lambda_i(u) \cdot r_i(u) = 0\}$ to be neither the empty space, as in the case of genuine nonlinearity, nor the whole space, as in the case of linear degeneracy.

The solution to the Riemann problem for the general system (3.3) was solved in [60], [61]. We enclose the following lemmas on the properties on wave curves proved in [61] for the self-containedness.

The i -rarefaction wave curve from a state u_0 , denoted by $R_i(u_0)$, is the integral curve of the right eigenvector r_i passing through u_0 , $i = 1, 2, \dots, n$. In general, the Hugoniot curve $H_i(u_0)$ and the rarefaction wave curve $R_i(u_0)$ have second order contact at the initial state u_0 , [54], while no higher-order contact is expected when the characteristic field is genuinely nonlinear. However, as we will see in the following lemmas, the situation is more interesting for non-genuinely nonlinear characteristic fields. The following lemmas are needed for the construction of the wave curve $W_i(u_0)$ through the state u_0 . As mentioned before, the strength of the i -wave is measured by the difference of the parameter u^i between the right and left states.

Lemma 3.3. *For any $u \in H_i(u_0)$ in a small neighborhood of u_0 , we have*

(i) $\lambda_i(u) > \sigma(u_0, u)$ (or $\lambda_i(u) < \sigma(u_0, u)$) if and only if

$$\frac{d}{du^i} \sigma(u_0, u) > 0 \quad (\text{or} \quad \frac{d}{du^i} \sigma(u_0, u) < 0);$$

(ii) $H_i(u_0)$ is tangent to $R_i(u)$ at u on $H_i(u_0)$ if $\sigma(u_0, u) = \lambda_i(u)$.

Proof. Let

$$\begin{aligned} u - u_0 &= \sum_{j=1}^n \alpha_j r_j(u), \\ \frac{du}{du^i} &= \sum_{j=1}^n \beta_j r_j(u). \end{aligned}$$

It is known that $H_i(u_0)$ and $R_i(u_0)$ have second contact at $u = u_0$. Then it implies

$$\begin{aligned} \alpha_i \beta_i &> 0 \quad \text{for } u \neq u_0, \\ \frac{|\alpha_j|}{|u - u_0|^2} &\quad \text{is bounded for } j \neq i. \end{aligned} \tag{3.5}$$

By differentiating

$$\sigma(u_0, u)(u_0 - u) = f(u_0) - f(u),$$

with respect to u^i , we have

$$\alpha_j \frac{d}{du^i} \sigma(u_0, u) = (\lambda_j(u) - \sigma(u_0, u))\beta_j, \quad j = 1, 2, \dots, n. \quad (3.6)$$

Thus (i) follows from (3.5) and (3.6)_j. Since $\sigma(u_0, u)$ is close to λ_i , by strict hyperbolicity and (3.6)_j, we have (ii). \square

The following lemma gives an estimate on the interaction of two shock waves in the same direction and shows that the interaction of two admissible shock waves yields an admissible shock plus a cubic order error term. We will not prove this lemma here.

Lemma 3.4. *Suppose that (u_0, u_1) and (u_1, u_2) with $u_2^i > u_1^i > u_0^i$ are two admissible i -shocks with strengths α_1 and α_2 and speeds σ_1 and σ_2 respectively, cf. Definition 3.2. Let $u_* \in H_i(u_0)$ be the state with $u_2^i = u_*^i$, then*

- (i) (u_0, u_*) is admissible;
- (ii) $|u_2 - u_*| = O(1)\alpha_1\alpha_2(\sigma_1 - \sigma_2)$;
- (iii) $\sigma\alpha = \sigma_1\alpha_1 + \sigma_2\alpha_2 + O(1)\alpha_1\alpha_2(\sigma_1 - \sigma_2)$, where α and σ are the strength and speed of the admissible shock (u_0, u_*) respectively. The same estimate holds for the case when $u_0^i > u_1^i > u_2^i$.

We next construct the i -wave curve from a state u_l , $i = 1, 2, \dots, n$, with the property that any state $u \in W_i(u_l)$ can be connected to u_l on the left by i -waves. That is, we will construct a curve $W_i(u_l)$ through u_l such that it passes through a single state u on each hyperplane with fixed u^i in a small neighborhood of u_l . For definiteness, we consider the case $u_l^i < u^i$. The case when $u_l^i > u^i$ can be discussed similarly. First we find a unique state u_1 with the following properties:

- (i) $u_l^i \leq u_1^i \leq u^i$;
- (ii) (u_l, u_1) is an admissible discontinuity such that $u_1^i - u_l^i$ is maximum.

If $u_1^i = u^i$, then we are done with $u = u_1$. If not, by Lemma 3.4, there is no admissible discontinuity with left state u_1 and the u^i component of the right state lies in $(u_1^i, u^i]$. Therefore, according to Lemma 3.3, we have $\nabla\lambda_i \cdot r_i(u_1) \geq 0$, and $\nabla\lambda_i \cdot r_i(u) > 0$ for states $u \in R_i(u_1)$ near u_1 with i -th component larger than u_1^i . Thus, there exists a unique state $u_2 \in R_i(u_1)$ with the following properties:

- (i) u_1 and u_2 are connected by i -rarefaction wave and $u_1^i < u_2^i \leq u^i$.
- (ii) u_2^i is the maximum in the sense that there is no state $u_* \in R_i(u_1)$ with the property that there exists admissible discontinuity (u_*, u_{**}) with $u_1^i < u_*^i < u_2^i$ and $u_*^i < u_{**}^i \leq u^i$.

If $u_2^i = u^i$, then $u = u_2$ and we are done. If not, the above procedure can be continued until we finally reach the state u on the curve $W_i(u_i)$ with the given u^i . Thus (u_i, u) forms an elementary i -wave described above when $u \in W_i(u_i)$. The wave curves are Lipschitz continuous, but have the following basic stability property:

Lemma 3.5. *Wave curves $W_i(\bar{u}_0)$ and $W_i(\tilde{u}_0)$ through different initial states have the following C^2 -like property: Given a state \bar{u} on $W_i(\bar{u}_0)$, there exists a state \tilde{u} on $W_i(\tilde{u}_0)$ such that*

$$\bar{u} - \tilde{u} = \bar{u}_0 - \tilde{u}_0 + O(1)|\bar{u}_0 - \tilde{u}_0| |\bar{u} - \tilde{u}|.$$

We will not prove this lemma here. Note that a wave curve is in general only Lipschitz continuous when two mixed curves meet. This corresponds to the vanishing of the rarefaction wave between two discontinuities to form a single discontinuity, cf. [5]. With the above preparation, we can now prove the existence of the solution to the Riemann problem, [60].

Theorem 3.6. *Suppose that system (3.3) is strictly hyperbolic with flux function $f(u) \in C^3$, and that for each characteristic field $\lambda_i(u)$ the linear degeneracy manifold LD_i either is the whole space or consists of a finite number of smooth manifolds of codimension one, each transversal to the characteristic vector $r_i(u)$. Then the Riemann problem (3.3) and (3.4) has a unique solution in the class of elementary waves satisfying the entropy condition, cf. Definition 3.2, provided that the states are in a small neighborhood of a given state.*

Proof. The i -waves, $i = 1, 2, \dots, n$, are the building blocks for the solution of the Riemann problem. The i -waves take values along the wave curves W_i . Since the wave curves W_i have tangent r_i at the initial state, it follows from the independency of the vectors r_i , $i = 1, 2, \dots, n$, and the implicit function theorem that the Riemann problem can be solved uniquely in the class of elementary waves. \square

3.3 Well-posedness theory for systems

3.3.1 Existence

Now the well-posedness theory for the hyperbolic conservation laws for solutions with small total variation can be stated in a satisfactory way. The pioneering work on this theory is the classical paper of Glimm in 1965 when he introduced the famous Glimm scheme and proved the first global existence result under the assumption that each characteristic is either genuinely nonlinear or linearly degenerate. The existence theory has now been generalized to many mathematical and physical settings,

such as the inhomogeneous systems and isolated large shocks, etc. In the following, we just state the result on general hyperbolic conservation laws based on the solution to the Riemann problem given in the previous subsection.

Theorem 3.7. *Suppose that system (3.3) is strictly hyperbolic with flux function $f(u) \in C^3$, and that for each characteristic field $\lambda_i(u)$ the linear degeneracy manifold LD_i either is the whole space or consists of a finite number of smooth manifolds of codimension one, each transversal to the characteristic vector $r_i(u)$. Then for the initial data (3.4) with sufficiently small total variation $T.V.$, there exists a global weak admissible solution $u(x, t)$ to the Cauchy problem (3.3) and (3.4) satisfying the total variation of $u(\cdot, t) = O(1)T.V.$*

The proof of this theorem depends on the construction of the Glimm functional. However, the traditional Glimm functional needs to be modified in order to take care of the possible degeneracy in the characteristic fields. The main idea for this improvement was introduced by Liu to include the effective interaction angle in the potential of waves of the same family. To be precise, for an i -wave α to the left of an i -wave β , we define $\Theta(\alpha, \beta)$ to represent the effective angle between them:

$$\Theta(\alpha, \beta) \equiv \theta_\alpha^+ + \theta_\beta^- + \sum \theta_\gamma.$$

Here θ_α^+ represents the value of λ_i at the right state of α minus its wave speed. It is negative if α is a shock and is set zero if it is an i -rarefaction wave. Similarly the term θ_β^- denotes the difference between the speed of β and the value of λ_i at its left end state. θ_γ is the value of λ_i at the right state of the wave γ minus that at the left state. It is positive if γ is a rarefaction wave and is negative if it is a shock. The sum $\sum \theta_\gamma$ is over the i -waves γ between α and β . Subject to wave interactions of distinct families, $-\Theta(\alpha, \beta)$ represents the angle between α and β when waves of other characteristic families between them propagate away. When $\Theta(\alpha, \beta)$ is positive, the two waves will not likely to meet and should not be included in the potential wave interaction functional. When $\Theta(\alpha, \beta)$ is negative, the two waves may eventually meet and interact. In this case $|\alpha||\beta||\Theta(\alpha, \beta)|$ reflects accurately the potential interactions of waves of the same characteristic family.

We now recall the Glimm scheme. The Glimm scheme is a finite difference scheme involving a random sequence a_i , $i = 0, 1, \dots$, $0 < a_i < 1$. Let $r = \Delta x$, $s = \Delta t$ be the mesh sizes satisfying the (C-F-L) condition

$$\frac{r}{s} > 2|\lambda_i(u)|, \quad 1 \leq i \leq n, \quad (3.7)$$

for all states u under consideration. The approximate solutions $u(x, t) = u_r(x, t)$ depends on the random sequence $\{a_k\}$ and is defined inductively

in time as follows:

$$u(x, 0) = u_0((h + a_0)r), \quad hr < x < (h + 1)r,$$

$$u(x, ks) = u((h + a_i)r - 0, ks - 0), \quad hr < x < (h + 1)r, \\ k = 0, \pm 1, \pm 2, \dots$$

Thus the approximate solution is a step function for each layer $t = ks$, $k = 1, 2, \dots$. Between the layers it consists of elementary waves by solving the Riemann problems at grid points $x = hr$, $h = 0, \pm 1, \dots$. Due to (C-F-L) condition (3.7) these elementary waves do not interact within the layer. Thus the approximate solution is an exact solution except at the interfaces $t = ks$, $k = 1, 2, \dots$. The numerical error depends on the random sequence.

The functional $F(J)$ is defined on any spacelike curve J . It consists of a linear part $L(J)$, measuring the total variation, a quadratic part $Q_h(J)$ and a cubic part $Q_s(J)$, measuring the potential wave interaction. The curve J incorporates the scheme and consists of line segments connecting points $((h \pm a_k)r, ks)$ and $(hr, (k \pm 1/2)s)$. The elementary waves issued from the grid points (hr, ks) will cross the line segments. These functionals are defined as follows:

$$L(J) \equiv \sum \{|\alpha| : \alpha \text{ any wave crossing } J\},$$

$$Q_h(J) \equiv \sum \{|\alpha||\beta| : \alpha \text{ and } \beta$$

interacting waves of distinct characteristic families crossing $J\}$,

$$Q_s(J) \equiv \sum_{i=1}^n Q_s^i,$$

$$Q_s^i \equiv \sum \{|\alpha||\beta| \max\{-\Theta(\alpha, \beta), 0\} : \alpha \text{ and } \beta i - \text{waves crossing } J, \\ \alpha \text{ to the left of } \beta\},$$

$$Q(J) \equiv Q_h(J) + Q_s(J),$$

$$F(J) \equiv L(J) + MQ(J).$$

Here M is a sufficiently large constant to be chosen later. In the above definition of $Q_h(J)$, an i -wave to the left of a j -wave is interacting if $i > j$.

The main estimate is that, for any curves J_1 and J_2 , J_2 lies toward larger time than J_1 ,

$$F(J_2) \leq F(J_1),$$

provided that the total variation TV of the initial data is small and that M is chosen sufficiently large. The proof of this being long, interested readers can refer to [64] for details.

3.3.2 Stability and uniqueness

For the discussion on the stability and uniqueness of entropy solutions in this subsection, we will restrict ourselves to the case when each characteristic is either genuinely nonlinear or linearly degenerate.

With the existence, the next main question is whether the solution is stable and whether it is unique. It is known that the natural and suitable norm for hyperbolic conservation laws is L_1 . One can consider the following two simple solutions, one as a simple compressible wave forming a shock in finite time and another is just a shift of the former one in space to a small distance. By straightforward calculation, one can show that the L_p distance between these two solutions is not stable when $p > 1$ even though it is stable when $p = 1$. On the other hand, it is shown in [87] that there is no L_1 metric so that the system of conservation laws is stable even though the scalar conservation law is contractive in L_1 norm.

The study of the L_1 stability has been attempted by many researchers, the first breakthrough was made by Bressan and his group. Under the assumption that each characteristic is either genuinely nonlinear or linearly degenerate, by using a homotopy idea on measuring the distance between two solutions, Bressan-Colombo [2] and Bressan-Piccoli [3] are able to establish the well-posedness for systems. The approach requires careful studies of the topology of wave interactions through painstaking front-tracking construction of approximate solutions.

The new approach based on the construction of the nonlinear functional was introduced by Liu-Yang so that the time evolution of the L_1 topology of two solutions becomes clear. The functional is sufficiently robust that the effects of wave interactions on the $L_1(x)$ distance can be separately dealt with by the generalized Glimm functional and the idea of wave tracing. The main effort is then to study the $L_1(x)$ distance between two wave patterns of *linear superposition of nonlinear waves*. Thus this analysis is orthogonal to and complements of Glimm's.

As a corollary of the continuous dependence of the solution on its initial value in the $L_1(x)$ topology, one obtains the uniqueness theory for solutions of the random choice method. This can easily be generalized to solutions obtained by other related characteristic methods, c.f. [9, 11, 25, 29]. For the study of the uniqueness problem based on the $L_2(x)$ topology, see [26, 30, 56, 59, 76].

By applying the above hyperbolic methods, the L_1 stability theorem can be stated as follows.

Theorem 3.8. *Under the assumption that each characteristic is either genuinely nonlinear or linearly degenerate, suppose that the total variations of the initial data $u_0(x)$ and $v_0(x)$ are sufficiently small and that $u_0(x) - v_0(x) \in L_1(\mathbb{R})$. Then, for the corresponding weak solutions $u(x, t)$ and $v(x, t)$ of (3.3) constructed by Glimm scheme or wave front tracking method, there exists a constant L independent of time such that*

$$\|u(x, t) - v(x, t)\|_{L_1} \leq L \|u(x, s) - v(x, s)\|_{L_1},$$

for any s, t , $0 \leq s \leq t < \infty$.

Remark 3.9. By applying the vanishing viscosity method, the L_1 stability of weak solutions obtained by vanishing viscosity holds for general systems without the assumption that each characteristic is either genuinely nonlinear or linearly degenerate. However, how to apply the above hyperbolic method, especially, how to construct the Liu-Yang functional for general scalar conservation laws remains open.

For the brevity of the presentation of this subsection, we will not give the detailed definition of the nonlinear functional here. We would like to emphasize that this functional $H[u, v](t)$ is constructed explicitly by the two solutions to the system and it satisfies the following three conditions:

- $H[u, v](t)$ is equivalent to $\|u(\cdot, t) - v(\cdot, t)\|_{L_1}$;
- $H[u, v](t)$ is non-increasing in time;
- $H[u, v](t)$ is defined solely by the waves in the solutions of $u(x, t)$ and $v(x, t)$ and the L_1 distance of the components in different families with respect to the location of the waves and their wave speeds.

Even though such nonlinear functional can now be well-defined when each characteristic is either genuinely nonlinear or linearly degenerate, the construction of this nonlinear functional is not clear for general system. The main difficulty is to construct the so called Liu-Yang functional for the scalar conservation laws. In fact, for scalar conservation law, we know that the L_1 norm is insensitive to the time evolution between two solutions except when there are some shocks of one solution crossed by the other one. The Liu-Yang functional is a generalized functional which is decreasing except the case when two solutions evolve linearly with respect to each other. More precisely, let $u(x, t)$ and $v(x, t)$ be two solution to the scalar conservation law (3.1). The main purpose for introducing this generalized entropy functional is to show that

$$\begin{aligned} & \int_0^\infty \int_{-\infty}^\infty (|u_x(x, t)| + |v_x(x, t)|) |u(x+, t) - v(x+, t)| \\ & \quad \times (|\sigma(\delta u(x, t)) - \sigma(u(x+, t), v(x+, t))| \\ & \quad + |\sigma(\delta v(x, t)) - \sigma(v(x+, t), u(x+, t))|) dx dt \\ & \leq O(1)(T.V.(u_0) + T.V.(v_0)) \|u_0 - v_0\|_{L^1}, \end{aligned}$$

where $\delta u(x, t) = (u(x-, t), u(x+, t))$ is viewed as a wave with speed $\sigma(\delta u(x, t))$, $T.V.$ denotes the total variation in x direction and $\|\cdot\|_{L^1}$ is the L^1 -norm in x . Notice that when $u(x+, t) = u(x-, t)$, the speed of the wave is the characteristic speed $f'(u(x\pm, t))$, and $u_x(x, t)$ is viewed as a wave in $u(x, t)$ located at x and time t with strength $u(x+, t) - u(x-, t)$.

For scalar conservation law with convex flux function, i.e., $f''(u) > 0$, assume that the initial data satisfy $u(x, 0) - v(x, 0) = u_0(x) - v_0(x) \in L^1(\mathbb{R})$. We can first define the L_1 distance between $u(x, t)$ and $v(x, t)$ on the left or right with respect to the location x depending on the relative propagation speed between the wave located at x and the virtual wave $(u(x+, t), v(x+, t))$.

Set

$$\begin{aligned} L(u, v)(x, t) = & \int_x^\infty (u - v)_{sign(u-v)(x+)}(y, t) dy \\ & + \int_\infty^x (u - v)_{-sign(u-v)(x+)}(y, t) dy, \end{aligned}$$

where $f_\pm = |f|$ if $\pm f \geq 0$, otherwise it is zero.

Now the Liu-Yang functional can be defined by

$$\begin{aligned} E(u, v)(t) = & \int_{-\infty}^\infty |u_x(x, t)| L(u, v)(x, t) dx \\ & + \int_{-\infty}^\infty |v_x(x, t)| L(v, u)(x, t) dx. \end{aligned}$$

The following theorem gives the time decay estimate of the nonlinear functional $E(u, v)(t)$.

Theorem 3.10. *Let $u(x, t)$, $v(x, t)$ and $E(u, v)(t)$ be defined above. The nonlinear functional $E(u, v)(t)$ is decreasing in time, and except at the time of wave interaction it satisfies*

$$\begin{aligned} & \frac{d}{dt} E(u, v)(t) \\ & \leq -c \int_{-\infty}^\infty (|u_x(x, t)| |u(x+, t) - v(x+, t)| \\ & \quad \times |\sigma(u_x(x, t)) - \sigma(u(x+, t), v(x+, t))| \\ & \quad + |v_x(x, t)| |v(x+, t) - u(x+, t)| \\ & \quad \times |\sigma(v_x(x, t)) - \sigma(v(x+, t), u(x+, t))|) dx. \end{aligned}$$

And at the time of wave interaction, the functional $E(u, v)(t)$ is not increasing in time.

We also omit the proof of this theorem. Interested readers can refer to [67]. Note that the above functional can be defined for each characteristic field in a system and its time decay estimate is crucial to control the time evolution of the L_1 distance between two weak solutions caused by the effect from waves of the same characteristic family. The effect from

waves of different characteristic families can be defined in a less subtle way because they have different speeds of propagation.

We now come to the uniqueness of the solution. Before the establishment of L^1 stability, the uniqueness in various function spaces was investigated by many researchers including Deafemros-Geng, DiPerna, LeFloch-Xin, Liu, Oleinik etc. Based on the L^1 stability, the uniqueness was proved by using the semigroup approach introduced by Bressan. And then the uniqueness of the entropy solutions constructed by either front tracking method or Glimm scheme follows from the work of Bressan-LeFloch [11], Bressan-Goatin [10] and Bressan-Lewicka [12]. For more information on the development and the detailed description, please refer to [8]. Here, we just state the main result in the following. For this, we first recall the definition of standard Riemann semigroup introduced by Bressan.

Definition 3.3. Let $\mathcal{D} \subset L^1(\mathbb{R}; \mathbb{R}^n)$ be a closed domain. A map $S : \mathcal{D} \times [0, \infty[\rightarrow \mathcal{D}$ is called a standard Riemann semigroup generated by the system of conservation laws if the following three conditions hold:

- For every $\bar{u} \in \mathcal{D}$, $t, s \geq 0$, we have

$$S_0\bar{u} = \bar{u}, \quad S_t S_s \bar{u} = S_{s+t} \bar{u}. \quad (3.8)$$

- There exists positive constants L and \bar{L} such that for all $\bar{u}, \bar{v} \in \mathcal{D}$, $t, s \geq 0$, we have

$$\|S_t \bar{u} - S_s \bar{v}\|_{L^1} \leq L \|\bar{u} - \bar{v}\|_{L^1} + \bar{L} |t - s|. \quad (3.9)$$

- If the initial data \bar{u} is piecewise constant, then there exists small $\delta > 0$ such that for $t \in [0, \delta]$, the trajectory $u(t, \cdot) = S_t \bar{u}$ coincides with the solution by piecing all the local Riemann solutions together before the interaction.

Note that the first assumption is the standard property of semigroup, the second one on Lipschitz continuity is crucial while the third one is needed for the consistency with the Riemann problem as the building block locally in space and time. Notice also that even though it is not required explicitly that each trajectory of the standard Riemann semigroup to be a solution to the Cauchy problem of the conservation laws, it turns out to be a consequence of the above conditions. In fact, it was proved by Bressan that the semigroup generated by a system of conservation laws is unique in the space

$$\mathcal{D} = cl\{u \in L^1(\mathbb{R}, \mathbb{R}^n); u \text{ is piecewise constant, } F(u) < \delta\},$$

for some positive constant δ , where $F(u)$ is the Glimm functional for the given system of conservation laws whose characteristic fields are either

genuinely nonlinear or linearly degenerate. For this case, the part Q_s^i in the Glimm's functional is defined as

$$Q_s^i \equiv \sum |\alpha||\beta|,$$

where the summation is over all waves in the i -th family at time t and at least one of the waves in the product is a shock.

Moreover, the uniqueness of the entropy solution can be stated as follows, cf. [8].

Theorem 3.11. *Let $S : \mathcal{D} \times [0, \infty[\rightarrow \mathcal{D}$ be the semigroup generated by (3.3). Let $u : [0, T] \rightarrow \mathcal{D}$ be a continuous map in L^1 . Suppose that the following three conditions hold.*

- $u(t, x)$ is a weak solution of the Cauchy problem (3.3) and (3.4).
- $u(t, x)$ satisfies the Lax entropy condition in the following sense. Let $u(t, x)$ have an approximate discontinuity at some point $(\tau, \xi) \in]0, T[\times R$, and there exist two states $u^\pm \in \Omega$ and a speed $\sigma \in \mathbb{R}$. Denote

$$U(t, x) = \begin{cases} u^-, & x < \sigma t, \\ u^+, & x > \sigma t. \end{cases}$$

If

$$\lim_{\rho \rightarrow 0+} \frac{1}{\rho^2} \int_{\tau-\rho}^{\tau+\rho} \int_{\xi-\rho}^{\xi+\rho} |u(t, x) - U(t - \tau, x - \xi)| dx dt = 0,$$

then for some $i \in \{1, \dots, n\}$, the entropy inequality

$$\lambda_i(u^+) \leq \sigma \leq \lambda_i(u^-),$$

holds.

- There exists $\delta > 0$ such that for every bounded space-like curve $\{t = \gamma(x); x \in [a, b]\}$ with

$$|\gamma(x_1) - \gamma(x_2)| \leq \delta|x_1 - x_2|,$$

for any $x_1, x_2 \in [a, b]$, the function $u(\gamma(x), x)$ has bounded variation.

Then

$$u(t, \cdot) = S_t u_0, \quad t \in [0, T].$$

That is, the weak solution of the Cauchy problem of (3.3) and (3.4) satisfying the above three conditions is unique. In particular, the weak solution constructed by either the Glimm scheme or the front-tracking approximation is unique.

Note that the third condition above can be replaced by a tame oscillation condition, cf. [8] for details.

3.4 Vanishing viscosity

Consider a strictly hyperbolic system of conservation laws (3.3) together with the viscous approximations

$$u_t^\varepsilon + A(u^\varepsilon)u_x^\varepsilon = \varepsilon u_{xx}^\varepsilon. \quad (3.10)$$

Here $A(u) \doteq Df(u)$ is the Jacobian matrix of f . Given an initial data $u(0, x) = u_0(x)$ having small total variation, the recent important result in [6] shows that the corresponding solutions u^ε of (3.10) exist for all $t \geq 0$, and have uniformly small total variation. Moreover, they converge to a unique solution of (3.3) as $\varepsilon \rightarrow 0$. The precise statement of the theorem borrowed from [6] can be stated as follows.

Theorem 3.12. *Consider the Cauchy problem for the hyperbolic system with artificial viscosity*

$$u_t + A(u)u_x = \varepsilon u_{xx}, \quad u(0, x) = u_0(x). \quad (3.11)$$

Assume that the matrix $A(u)$ is strictly hyperbolic, smoothly depending on u in a neighborhood of a compact set $K \subset \mathbb{R}^n$. Then there exist constants c, L, L' and $\delta > 0$ such that the following holds. If

$$T.V.\{u_0\} < \delta, \quad \lim_{x \rightarrow -\infty} u_0(x) \in K,$$

then for each $\varepsilon > 0$ the Cauchy problem (3.11) has a unique solution u^ε , defined for all $t \geq 0$, denoted by $u^\varepsilon(t, \cdot) = S_t^\varepsilon(u_0)$. In addition,

$$\begin{aligned} T.V.\{S_t^\varepsilon u_0\} &\leq CT.V.\{u_0\}, \\ \|S_t^\varepsilon u_0 - S_t^\varepsilon v_0\|_{L^1} &\leq L\|u_0 - v_0\|_{L^1}, \\ \|S_t^\varepsilon u_0 - S_s^\varepsilon u_0\|_{L^1} &\leq L'(|t-s| + |\sqrt{\varepsilon t} - \sqrt{\varepsilon s}|). \end{aligned}$$

Moreover, when $\varepsilon \rightarrow 0+$, the solution u^ε converge to the trajectories of a semigroup S such that

$$\|S_t u_0 - S_s v_0\|_{L^1} \leq L\|u_0 - v_0\|_{L^1} + L'|t-s|.$$

These vanishing viscosity limits can be regarded as the unique vanishing viscosity solutions of the hyperbolic Cauchy problem

$$u_t + A(u)u_x = 0, \quad u(0, x) = u_0(x).$$

In the conservative case when $A(u) = Df(u)$, every vanishing viscosity solution is a weak solution of

$$u_t + f(u)_x = 0, \quad u(0, x) = u_0(x)$$

satisfying the entropy condition given above.

Furthermore, if each characteristic field is genuinely nonlinear or linearly degenerate, the vanishing viscosity solutions coincide with the unique limits of the Glimm and front-tracking approximations.

Based on this stability, the distance $\|u^\varepsilon(t) - u(t)\|_{L^1}$ can be analyzed and it provides a convergence rate for the vanishing viscosity approximations. For the case when each characteristic field is genuinely nonlinear, the convergence rate result can be stated as follows, [15]. Note that it is interesting to generalize this convergence estimate for general hyperbolic systems.

Theorem 3.13. *Let the system (3.3) be strictly hyperbolic and assume that each characteristic field is genuinely nonlinear. Then, given any initial data $u(0, x) = u_0(x)$ with small total variation, for every $\tau > 0$ the corresponding solutions u, u^ε of (3.3) and (3.10) satisfy the estimate*

$$\|u^\varepsilon(\tau, \cdot) - u(\tau, \cdot)\|_{L^1} = O(1) \cdot (1 + \tau) \sqrt{\varepsilon} |\ln \varepsilon| T.V.\{u_0(x)\}. \quad (3.12)$$

Remark 3.14. For a fixed time $\tau > 0$, a similar convergence rate was proved in [14] for approximate solutions generated by the Glimm scheme, namely

$$\|u^{Glimm}(\tau, \cdot) - u(\tau, \cdot)\|_{L^1} = o(1) \cdot \sqrt{\varepsilon} |\ln \varepsilon|.$$

Here $\varepsilon \approx \Delta x \approx \Delta t$ measures the mesh of the grid.

Remark 3.15. For a scalar conservation law, the method of Kuznetsov in [50] shows that the convergence rate in (3.12) is $O(1) \cdot \varepsilon^{1/2}$. As shown in [86], this rate is sharp in the general case.

In the case of hyperbolic systems, in [39] Goodman and Xin have studied the viscous approximation of piecewise smooth solutions having a finite number of non-interacting shocks. With these regularity assumptions, they obtain the convergence rate $O(1) \cdot \varepsilon^\gamma$ for any $\gamma < 1$. On the other hand, the estimate (3.12) applies to any general BV solution, possibly with a countable everywhere dense set of shocks.

To appreciate the estimate in (3.12), denote by S_t and S_t^ε the semi-groups generated by the systems (3.3) and (3.10) respectively. The previous theorems show that they are Lipschitz continuous with respect to the initial data, namely

$$\begin{aligned} \|S_t \bar{u} - S_t \bar{v}\|_{L^1} &\leq L \|\bar{u} - \bar{v}\|_{L^1}, \\ \|S_t^\varepsilon \bar{u} - S_t^\varepsilon \bar{v}\|_{L^1} &\leq L \|\bar{u} - \bar{v}\|_{L^1}. \end{aligned} \quad (3.13)$$

The Lipschitz constant L here does not depend on t, ε . By (3.13), a trivial error estimate is

$$\begin{aligned} \|u^\varepsilon(\tau) - u(\tau)\|_{L^1} &= L \cdot \int_0^\tau \left\{ \lim_{h \rightarrow 0+} \frac{\|u^\varepsilon(t+h) - S_h u^\varepsilon(t)\|_{L^1}}{h} \right\} dt \\ &= L \cdot \int_0^\tau \|\varepsilon u_{xx}^\varepsilon(t)\|_{L^1} dt. \end{aligned}$$

However, $\|u_{xx}^\varepsilon(t)\|_{L^1}$ grows like ε^{-1} in the appearance of shock wave, hence the right hand side in the above estimate does not converge to zero as $\varepsilon \rightarrow 0$.

4 Spectral analysis on the linearized Boltzmann operator

Now let us come back to the Boltzmann equation and study the solution f to (1.3) without external force or source near an absolute (global) Maxwellian \mathbf{M} in the form

$$f = \mathbf{M} + \mathbf{M}^{1/2}u. \quad (4.1)$$

By a suitable Galilean translation and scaling of the velocity variable ξ , \mathbf{M} can be taken, without loss of generality, to be the standard Maxwellian,

$$\mathbf{M} = \mathbf{M}_{[1,0,1]}(\xi) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{|\xi|^2}{2}\right).$$

Plug (4.1) into the Boltzmann equation to deduce the equation for the new unknown function $u = u(t, x, \xi)$,

$$\frac{\partial u}{\partial t} + \xi \cdot \nabla_x u = \mathbf{L}u + \Gamma(u, u), \quad (t, x, \xi) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n, \quad (4.2)$$

where

$$\begin{aligned} \mathbf{L}u &= 2\mathbf{M}^{-1/2}Q(\mathbf{M}, \mathbf{M}^{1/2}u), \\ \Gamma(u, v) &= \mathbf{M}^{-1/2}Q(\mathbf{M}^{1/2}u, \mathbf{M}^{1/2}v) \end{aligned}$$

are the same as in (1.17).

Consider the Cauchy problem

$$\begin{cases} \frac{\partial u}{\partial t} + \xi \cdot \nabla_x u = \mathbf{L}u + \Gamma(u, u), & (t, x, \xi) \in \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^n, \\ u(0, x, \xi) = u_0(x, \xi), & (x, \xi) \in \mathbb{R}^n \times \mathbb{R}^n. \end{cases} \quad (4.3)$$

We shall discuss (4.3) in the mild form defined in the function space X_β mentioned in (1.34), that is,

$$\begin{aligned} X_\beta &= L^2 \cap L_\beta^\infty, \\ L^2 &= L^2(\mathbb{R}_x^n \times \mathbb{R}_\xi^n), \quad L_\beta^\infty = \{u \mid (1 + |\xi|)^\beta u \in L^\infty(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)\}, \\ \|u\|_{X_\beta} &= \|u\|_{L^2(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)} + \|(1 + |\xi|)^\beta u\|_{L^\infty(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)}. \end{aligned}$$

The linear part of (4.2) gives rise to the *linearized Boltzmann operator*:

$$B = -\xi \cdot \nabla_x + \mathbf{L}. \quad (4.4)$$

The aim of this section is to present some basic decay estimates in time t for the semi-group e^{tB} which are useful in solving the nonlinear problems in the space X_β . The fundamental techniques used here are those developed in [88, 89] for the space defined in (5.21) while some estimates are improved in [96] for the study in function space X_β . Thus, after recalling some basic properties of the collision operators \mathbf{L} and Γ , we derive decay estimates for the semi-group e^{tB} under the presence of the unbounded operator ν , which is crucial for the construction of solutions in the space X_β .

4.1 Smoothing properties of e^{tA}

Write

$$Au = -\xi \cdot \nabla_x u - \nu(\xi)u, \quad (4.5)$$

which is part of the linear part of (4.4). It is easy to show that under the condition (1.21), (4.5) gives a well-defined operator, denoted again by A , in the space $Z_2 = L^2(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)$ with the domain of definition

$$D(A) = \{u \in Z_2 \mid -\xi \cdot \nabla_x u, \nu(\xi)u \in Z_2\}, \quad (4.6)$$

and it generates a C_0 semi-group e^{tA} , with the explicit expression

$$e^{tA}u = e^{-\nu(\xi)t}u(x - t\xi, \xi). \quad (4.7)$$

The formula (4.7) implies that the operator e^{tA} can be well-defined in various other function spaces. The main purpose of this subsection is to show that it has a smoothing effect in those spaces if coupled with the integral operator K . In the sequel, the notation L^p stands for the space

$$L^p = L^p(\mathbb{R}_x^n \times \mathbb{R}_\xi^n),$$

unless otherwise stated.

Lemma 4.1. *For any $2 \leq p \leq r \leq \infty$, there is a constant $C > 0$ such that*

$$\|Ke^{tA}Ku\|_{L^r} \leq Ct^{-\kappa_{p,r}}e^{-\nu_* t}\|u\|_{L^p}, \quad t > 0, \quad \kappa_{p,r} \left(\frac{1}{p} - \frac{1}{r} \right),$$

holds for any $u \in L^p$ where ν_ is given in (1.30).*

Proof. For simplicity, put $w(t, x, \xi) = e^{tA}u$. Also write $L_x^p = L^p(\mathbb{R}_x^n)$ etc. First, consider the case $r < \infty$. By (4.7) and the change of variable $\xi \mapsto y = x - t\xi$, we get

$$\begin{aligned}\|w(t, x, \cdot)\|_{L_\xi^p}^p &= \int_{\mathbb{R}_\xi^n} e^{-p\nu(\xi)t} |u(x - t\xi, \xi)|^p d\xi \\ &\leq e^{-p\nu_* t} \int_{\mathbb{R}_\xi^n} |u(y, (x - y)/t)|^p t^{-n} dy.\end{aligned}$$

Put $r = sp$ with $s \geq 1$ and take the L_x^s norm of the above to deduce

$$\begin{aligned}\|w(t)\|_{L_x^r(L_\xi^p)}^r &= \| \|w(t, x, \cdot)\|_{L_\xi^p}^p \|_{L_x^s}^s \\ &\leq \left(t^{-n} e^{-p\nu_* t} \left\| \int_{\mathbb{R}^n} |u(y, (\cdot - y)/t)|^p dy \right\|_{L_x^s} \right)^s \\ &\leq \left(t^{-n} e^{-p\nu_* t} \int_{\mathbb{R}^n} \| |u(y, (\cdot - y)/t)|^p \|_{L_x^s} dy \right)^s.\end{aligned}$$

By the change of variable $x \mapsto \xi = (x - y)/t$,

$$\begin{aligned}\| |u(y, (\cdot - y)/t)|^p \|_{L_x^s} &= \left(\int_{\mathbb{R}^n} |u(y, (x - y)/t)|^{sp} dx \right)^{1/s} \\ &= \left(t^n \int_{\mathbb{R}^n} |u(y, \xi)|^{sp} d\xi \right)^{1/s} = t^{n/s} \|u(y, \cdot)\|_{L_\xi^r}^p.\end{aligned}$$

Combining the above two estimates yields

$$\begin{aligned}\|w(t)\|_{L_x^r(L_\xi^p)}^r &\leq \left(t^{-n} e^{-p\nu_* t} \int_{\mathbb{R}^n} t^{n/s} \|u(y, \cdot)\|_{L_\xi^r}^p dy \right)^s \\ &= t^{-n(s-1)} e^{-r\nu_* t} \|u\|_{L_x^p(L_\xi^r)}^r\end{aligned}$$

or

$$\|e^{tA}u\|_{L_x^r(L_\xi^p)} \leq t^{-\kappa_{p,r}} e^{-\nu_* t} \|u\|_{L_x^p(L_\xi^r)} \quad (4.8)$$

with $\kappa_{p,r}(s-1)/r(1/p-1/r)$.

The case $r = \infty$ can be proved similarly but more straightforward.

Finally, note from Lemma 1.2 (a) that for any $q \in [1, \infty]$ and $2 \leq p \leq r \leq \infty$, the operator

$$K : L_x^q(L_\xi^p) \rightarrow L_x^q(L_\xi^r)$$

is bounded, which, with the choice $q = p, r$ and combined with (4.8), completes the proof of the lemma. \square

The following lemma obtained by the bootstrap argument based on the above lemma is crucial to establish the decay estimates of e^{tB} in the

space X_β . Denote the convolution in t by $*$:

$$g * h = \int_0^t g(t-s)h(s)ds. \quad (4.9)$$

For each $k \in \mathbb{N}$, define the operator $F_k(t)$ by induction,

$$F_1(t) = e^{tA}K * e^{tA}K, \quad F_k(t) = F_1(t) * F_{k-1}(t) \quad (k = 2, 3, \dots). \quad (4.10)$$

Lemma 4.2. *Let $\beta \geq 0$ and set*

$$N = [(n + \beta)/2] + 2.$$

For any $\epsilon > 0$, there is a constant $C > 0$ such that

$$\|F_N(t)u\|_{Y_\beta} \leq C e^{-(\nu_* - \epsilon)t} \|u\|_{Z_2}, \quad t \geq 0,$$

holds for any $u \in Z_2$. Here, Y_β is defined by (1.36) and $Z_2 = L^2$ by (1.35).

Proof. First, we note from (4.7) that for $r \in [1, \infty]$,

$$\|e^{tA}u\|_{L^r} \leq e^{-\nu_* t} \|u\|_{L^r}.$$

Noting that $F_1(t) = e^{tA} * K e^{tA}K$ and by virtue of Lemma 4.1, we get for $2 \leq p \leq r \leq \infty$,

$$\begin{aligned} \|F_1(t)u\|_{L^r} &\leq C \int_0^t e^{-\nu_*(t-s)} s^{-\kappa_{p,r}} e^{-\nu_* s} \|u\|_{L^p} ds \\ &= C e^{-\nu_* t} \int_0^t s^{-\kappa_{p,r}} ds \|u\|_{L^p}. \end{aligned}$$

The final integral converges if $\kappa_{p,r} < 1$ or

$$\frac{1}{p} - \frac{1}{r} < \frac{1}{n}, \quad 2 \leq p \leq r \leq \infty. \quad (4.11)$$

Then, we conclude

$$\|F_1(t)u\|_{L^r} \leq C t^{1-\kappa_{p,r}} e^{-\nu_* t} \|u\|_{L^p} \leq C e^{-(\nu_* - \epsilon)t} \|u\|_{L^p}.$$

In view of (4.11) and by an induction argument, we get

$$\|F_k(t)u\|_{L^\infty} \leq C e^{-(\nu_* - \epsilon)t} \|u\|_{L^2}, \quad (4.12)$$

for $k > n/2$. Furthermore, it follows from (4.7) that

$$\|e^{tA}u\|_{Y_\beta} \leq e^{-\nu_* t} \|u\|_{Y_\beta}.$$

This and Lemma 1.2 (c) yields

$$\|F_1(t)u\|_{Y_{\beta+2}} \leq C e^{-(\nu_* - \epsilon)t} \|u\|_{Y_\beta}$$

for $\beta \geq 0$, which, together with (4.12), completes the proof of the lemma. \square

Let h be a function of t, x, ξ and set

$$\Psi_0[h](t) = e^{tA} * (\nu h). \quad (4.13)$$

Here ν is the multiplication operator (1.32). The following lemma shows that the unboundedness of the function $\nu(\xi)$ in (1.21) can be controlled by the integration in t .

Lemma 4.3. *Let $\sigma \geq 0$.*

(a) *For any $\delta \in [0, 1]$, there is a constant $C(\sigma) > 0$ such that*

$$\|\Psi_0[h](t)\|_{L^2} \leq C(\sigma)(1+t)^{-\sigma} \sup_{t \geq 0} (1+t)^\sigma \|\nu^{(1-\delta)} h(t)\|_{L^2}.$$

(b) *For any $\beta \geq 0$, there is a constant $C(\sigma) > 0$ such that*

$$\|\Psi_0[h](t)\|_{Y_\beta} \leq C(\sigma)(1+t)^{-\sigma} \sup_{t \geq 0} (1+t)^\sigma \|h(t)\|_{Y_\beta}.$$

Here, $C(\sigma) \rightarrow \infty$ when $\sigma \rightarrow \infty$.

Proof. We start from

$$\|\Psi_0[h](t, \cdot, \xi)\|_{L_x^2} \leq \int_0^t e^{-\nu(\xi)(t-s)} \nu(\xi) \|h(s, \cdot, \xi)\|_{L_x^2} ds.$$

Since for any $\delta \geq 0$,

$$\begin{aligned} e^{-\nu(\xi)t} \nu(\xi) &= \left(e^{-\nu(\xi)t/2} \nu(\xi)^\delta \right) e^{-\nu(\xi)t/2} \nu(\xi)^{(1-\delta)} \\ &\leq C t^{-\delta} e^{-\nu_* t/2} \nu(\xi)^{(1-\delta)} \end{aligned}$$

for some constant $C > 0$, we get

$$\begin{aligned} &\|\Psi_0[h](t, \cdot, \cdot)\|_{L^2} \\ &\leq C \int_0^t (t-s)^{-\delta} e^{-\nu_*(t-s)/2} \|\nu^{(1-\delta)} h(s, \cdot, \cdot)\|_{L^2} ds \\ &\leq C \int_0^t (t-s)^{-\delta} e^{-\nu_* (t-s)/2} (1+s)^{-\sigma} ds \\ &\quad \times \sup_{t \geq 0} (1+t)^\sigma \|\nu^{(1-\delta)} h(t, \cdot, \cdot)\|_{L^2}. \end{aligned} \quad (4.14)$$

The last integral is bounded by

$$(t/2)^{-\delta} e^{-\nu_* t/4} \int_0^{t/2} (1+s)^{-\sigma} ds \\ + (1+t/2)^{-\sigma} \int_{t/2}^t (t-s)^{-\delta} e^{-\nu_*(t-s)/2} ds,$$

which is, in turn, bounded by $C(1+t)^{-\sigma}$, whence (a) follows.

The proof of (b) is more straightforward:

$$|\Psi_0[h](t, x, \xi)| \leq \int_0^t e^{-\nu(\xi)(t-s)} \nu(\xi) (1+s)^{-\sigma} ds \\ \times (1+|\xi|)^{-\beta} \sup_{t \geq 0} (1+t)^\sigma \|h(t, \cdot, \cdot)\|_{Y_\beta}.$$

The last integral is bounded by

$$e^{-\nu_* t/2} \int_0^{t/2} e^{-\nu(\xi)(t-s)/2} \nu(\xi) (1+s)^{-\sigma} ds \\ + (1+t/2)^{-\sigma} \int_{t/2}^t e^{-\nu(\xi)(t-s)} \nu(\xi) ds \\ \leq e^{-\nu_* t/2} \int_0^{t/2} e^{-\nu(\xi)(t-s)/2} \nu(\xi) ds + (1+t/2)^{-\sigma} \int_{t/2}^t e^{-\nu(\xi)(t-s)} \nu(\xi) ds \\ \leq C(1+t)^{-\sigma},$$

which proves (b). Now the proof of the lemma is complete. \square

Remark 4.4. The part (a) of the above lemma shows that the integration in t controls the weight loss in ξ up to order $|\xi|^{\gamma\delta}$ for $\delta < 1$ in the space L^2 and (b) for $\delta = 1$ in the space L^∞ . The part (b) is essentially due to [40].

4.2 Spectral properties of B

Since K is a bounded operator on $Z_2 = L^2$, the linearized Boltzmann operator $B = -\xi \cdot \nabla + \mathbf{L}$ can be taken to be a bounded perturbation of the operator A with the same domain of definition $D(A)$ in (4.6), namely,

$$B = A + K, \quad D(B) = D(A),$$

so that B generates a C_0 semi-group e^{tB} on Z_2 , see e.g. [49].

We shall study the Fourier transform of e^{tB} . Introduce

$$\hat{u}(k, \xi) = \mathcal{F}(u)(k, \xi) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-ik \cdot x} u(x, \xi) dx,$$

which is the Fourier transform of a function $u(x, \xi)$ with respect to x , where $k = (k_1, k_2, \dots, k_n) \in \mathbb{R}^n$ is the dual variable to x and \cdot is the inner product of \mathbb{R}^n .

For $u \in D(A) = D(B)$, we have

$$\mathcal{F}(Au) = (-i\xi \cdot k - \nu(\xi))\hat{u}, \quad \mathcal{F}(Bu) = (-i\xi \cdot k + \mathbf{L})\hat{u} = \mathcal{F}(Au) + K\hat{u}. \quad (4.15)$$

For $w = w(\xi)$, put

$$\hat{A}(k)w = (-i\xi \cdot k - \nu(\xi))w, \quad \hat{B}(k)w = (-i\xi \cdot k + \mathbf{L})w = \hat{A}(k)w + Kw.$$

Here, we regard $k \in \mathbb{R}^n$ as a parameter and consider $\hat{A}(k)$ and $\hat{B}(k)$ as operators in the space L_ξ^2 with the domain of definition

$$D(\hat{A}(k)) = D(\hat{B}(k)) = \{w \in L_\xi^2 \mid k \cdot \xi w, \nu(\xi)w \in L_\xi^2\}. \quad (4.16)$$

Since, clearly, $\hat{A}(k)$ generates a C_0 semi-group of the form

$$e^{t\hat{A}(k)}w = e^{(-i\xi \cdot k - \nu(\xi))t}w(\xi),$$

so does $\hat{B}(k)$ as a bounded perturbation of $\hat{A}(k)$. Set

$$\Phi(t, k) = e^{t\hat{B}(k)} = \frac{1}{2\pi i} \int_{\lambda_0-i\infty}^{\lambda_0+i\infty} e^{\lambda t} (\lambda - \hat{B}(k))^{-1} d\lambda.$$

Clearly, it follows from (4.15) that

$$e^{tB} = \mathcal{F}^{-1} \left\{ \Phi(t, k) \right\} \mathcal{F}. \quad (4.17)$$

The following theorem summarizes a central result of the spectral analysis of $\hat{B}(k)$ and gives a semi-implicit formula of $\Phi(t, k)$. The parts (1), (2) below are essentially due to [35] ($n = 3$) and (3) due to [88, 89], see also [91]. Interested readers please refer to these references for the proof.

Theorem 4.5. *Assume (1.19) with $\gamma \in [0, 1]$. Then, there exist two positive numbers κ_0 and σ_0 such that the following holds:*

(1) *For each $k \in \mathbb{R}^n, |k| \leq \kappa_0$, the spectrum of $\hat{B}(k)$ in the right half complex plane*

$$\{\lambda \in \mathbb{C} \mid \operatorname{Re}\lambda \geq -\sigma_0\}$$

consists only of $(n+2)$ discrete semi-simple eigenvalues $\lambda_j(k)$, $j = 0, \dots, n+1$. Denote the corresponding eigenprojections by $\mathbf{P}_j(k)$.

(2) *Set $\tilde{k} = |k|^{-1}k \in S^{n-1}$. For $j = 0, \dots, n+1$, the asymptotic expansions*

$$\begin{aligned} \lambda_j(k) &= i\lambda_{j,1}|k| - \lambda_{j,2}|k|^2 + O(|k|^3), \\ &O(1), \end{aligned}$$

hold for $|k| \leq \kappa_0$, with

$$\lambda_{j,1} \in \mathbb{R}, \quad \lambda_{j,2} > 0, \quad \sum_{j=0}^{n+1} \mathbf{P}_{j,0}(\tilde{k}) = \mathbf{P},$$

where \mathbf{P} is the orthogonal projection defined by (1.31) and $O(1)$ presents a bounded operator on L^2_ξ with uniform operator norm for $|k| \leq \kappa_0$.

(3) The formula

$$\Phi(t, k) = e^{t\hat{A}(k)} + \sum_{j=0}^{n+2} \Phi_j(t, k), \quad t \geq 0, \quad k \in \mathbb{R}^n$$

holds where for $j = 0, \dots, n+1$,

$$\Phi_j(t, k) = e^{\lambda_j(k)t} \mathbf{P}_j(k) \chi(|k| < \kappa_0),$$

$\chi(|k| < \kappa_0)$ being the characteristic function for $|k| < \kappa_0$, and $\Phi_{n+2}(t, k)$ is a linear bounded operator on L^2_ξ with the operator norm

$$\|\Phi_{n+2}(t, k)\| \leq c_0 e^{-\sigma_0 t}, \quad k \in \mathbb{R}^n, \quad t \geq 0,$$

with some constant $c_0 > 0$ independent of t and k .

The following lemma, which is crucial to solve (4.2) in the space X_β , is an improvement of the estimates for $\mathbf{P}_j(k)$ and $\Phi_{n+2}(t, k)$ given in the above theorem and says that these operators are still uniformly bounded if they are multiplied by the unbounded operator ν .

Lemma 4.6. *Let $\|\cdot\|$ denote the operator norm of the space L^2_ξ . There is a constant $C > 0$ such that*

$$\|\mathbf{P}_j(k)\nu\| \leq C, \quad j = 0, \dots, n+1, \quad |k| \leq \kappa_0, \quad (4.18)$$

$$\|\Phi_{n+2}(t, k)\nu\| \leq C e^{-\sigma_0 t}, \quad t \geq 0, \quad k \in \mathbb{R}^n. \quad (4.19)$$

Proof. Since the eigenvalue $\lambda_j(k)$ is semi-simple, its eigenprojection has the form

$$\mathbf{P}_j(k)w = \sum_{m=1}^{m_0} (w, \varphi_{j,m}^*(k))_{L^2_\xi} \varphi_{j,m}(k), \quad (4.20)$$

where m_0 is the geometric multiplicity of $\lambda_j(k)$ and $\{(\varphi_{j,m}(k), \varphi_{j,m}^*(k))\}$ is a bi-orthonormal set of the pairs of corresponding eigenfunctions and adjoint eigenfunctions. Since \mathbf{L} is self-adjoint, we see that the adjoint of $\hat{B}(k)$ is given by $\hat{B}^*(k) = i\xi \cdot k + \mathbf{L} = \hat{B}(-k)$ with the same domain of definition as (4.16). As a consequence, $\varphi_{j,m}^*(k) = \varphi_{j,m}(-k)$, [35]. We

shall prove that $\nu\varphi_{j,m}(-k)$ and hence $\nu\varphi_{j,m}(k)$ are uniformly bounded in $|k| \leq \kappa_0$. Note that $\varphi_{j,m}(k)$ is a normalized eigenfunction. The eigenvalue problem

$$\varphi_{j,m}(k) \in D(\hat{B}(k)), \quad \hat{B}(k)\varphi_{j,m}(k) = \lambda_j(k)\varphi_{j,m}(k)$$

yields the relation

$$\nu(\xi)\varphi_{j,m}(k) = \frac{\nu(\xi)}{-i\xi \cdot k - \nu(\xi)} \left(-K\varphi_{j,m}(k) + \lambda_j(k)\varphi_{j,m}(k) \right).$$

The first factor on the right hand side is bounded by 1 in magnitude while the L_ξ^2 norm of the second factor is bounded by $\|K\| + |\lambda_j(k)|$. This and (4.20) prove (4.18).

Finally, we shall say that the Φ_{n+2} in the previous theorem is nothing but U_{∞,σ_0} in [89], [91, p. 61]. Here, notice that

$$\begin{aligned} (\lambda - \hat{B}(k))^{-1} &= (\lambda - \hat{A}(k))^{-1} + Z(\lambda), \\ Z(\lambda) &= (\lambda - \hat{A}(k))^{-1}(I - G(\lambda))^{-1}G(\lambda), \end{aligned}$$

with

$$G(\lambda) = K(\lambda - \hat{A}(k))^{-1}.$$

The estimate of it given there is reproduced in the present notation as

$$\Phi_{n+2}(t, k) = \int_{-\sigma_0-i\infty}^{-\sigma_0+i\infty} e^{(-\sigma_0+i\tau)t} Z(-\sigma_0 + i\tau) d\tau,$$

and

$$\begin{aligned} &|(\Phi_{n+2}(t, k)u, v)_{L_\xi^2}| \\ &\leq C_1 \|K\| e^{-\sigma_0 t} \int_{-\infty}^{\infty} \|(\lambda - \hat{A}(k))^{-1}u\|_{L_\xi^2} \|(\bar{\lambda} - \hat{A}^*(k))^{-1}v\|_{L_\xi^2} d\tau, \end{aligned}$$

where $\lambda = -\sigma_0 + i\tau$. Obviously, we can have

$$\begin{aligned} &|(\Phi_{n+2}(t, k)\nu u, v)_{L_\xi^2}| \\ &\leq C_1 \|K\nu\| e^{-\sigma_0 t} \int_{-\infty}^{\infty} \|(\lambda - \hat{A}(k))^{-1}u\|_{L_\xi^2} \|(\bar{\lambda} - \hat{A}^*(k))^{-1}v\|_{L_\xi^2} d\tau, \end{aligned}$$

which, together with Lemma 1.7, yields (4.19). The proof of the lemma is complete. \square

4.3 Decay rates of e^{tB} in X_β

We shall now establish the decay estimates of the semi-group e^{tB} as $t \rightarrow \infty$. Firstly, substituting the decomposition in Theorem 4.5 (3) into the formula (4.17) yields the decomposition

$$e^{tB} = e^{tA} + E_1(t) + E_2(t), \quad (4.21)$$

$$E_1(t) = \sum_{j=0}^{n+1} \mathcal{F}^{-1} \left\{ \Phi_j(t, k) \right\} \mathcal{F}, \quad E_2(t) = \mathcal{F}^{-1} \left\{ \Phi_{n+2}(t, k) \right\} \mathcal{F}.$$

We shall show that $E_1(t)\nu$ has the algebraic decay while $E_2(t)\nu$ has the exponential decay, as $t \rightarrow \infty$. Set

$$\sigma_{q,m} = \frac{n}{2} \left(\frac{1}{q} - \frac{1}{2} \right) + \frac{m}{2}, \quad (4.22)$$

and recall the space Z_q in (1.35). The basic decay estimates are stated in the theorem below.

Theorem 4.7. (1) Let $q \in [1, 2]$ and $\alpha \in \mathbb{N}^n$. If $\partial^{\alpha'} u \in Z_q$ for some $\alpha' \leq \alpha$, then, for any $\delta \in [0, 1]$, we have

$$\partial_x^\alpha E_1(t)\nu^\delta u \in C^0([0, \infty); Z_2),$$

and the algebraic decay estimates

$$\|\partial_x^\alpha E_1(t)\nu^\delta u\|_{Z_2} \leq b_1(1+t)^{-\sigma_{q,m}} \|\partial_x^{\alpha'} u\|_{Z_q}, \quad (4.23)$$

$$\|\partial_x^\alpha E_1(t)(I - \mathbf{P})\nu^\delta u\|_{Z_2} \leq b_2(1+t)^{-\sigma_{q,m+1}} \|\partial_x^{\alpha'} u\|_{Z_q} \quad (4.24)$$

for all $t \geq 0$, where $m = |\alpha - \alpha'|$ and b_1, b_2 are positive constants depending only on q and m .

(2) Let $\alpha \in \mathbb{N}^n$ and $\partial^\alpha u \in Z_2$. Then, for any $\delta \in [0, 1]$, we have

$$\partial_x^\alpha E_2(t)\nu^\delta u \in C^0([0, \infty); Z_2),$$

and the exponential decay estimate

$$\|\partial_x^\alpha E_2(t)\nu^\delta u\|_{Z_2} \leq b_3 e^{-\sigma_0 t} \|\partial_x^\alpha u\|_{Z_2} \quad (4.25)$$

for all $t \geq 0$, where σ_0 and b_3 are positive constants independent of δ, α , u , and t .

Remark 4.8. This theorem shows that the higher order x -derivatives of e^{tB} decay faster than the lower order derivatives as $t \rightarrow \infty$. The ξ -derivatives, on the other hand, have no such property. Theorem 5.3 indicates that this feature is inherited by the nonlinear problem. The fact that the constants b_1, σ_0 and b_3 are independent of α is crucial for the proof of Theorem 5.3. Notice that the heat kernel enjoys the same decay properties.

Remark 4.9. The above theorem is well-established for the case $\delta = 0$ in the space (5.21), see [89], [91]. The improvement presented here is necessary for handling the nonlinear problems in the space X_β .

Proof of Theorem 4.7. It suffices to prove the theorem for the case $\delta = 1$. Put

$$I_j(t, k) = \|\Phi_j(t, k)\nu\hat{u}(k, \cdot)\|_{L^2_\xi}.$$

Write $k^\alpha = k_1^{\alpha_1} k_2^{\alpha_2} \cdots k_n^{\alpha_n}$ and note that

$$k^\alpha \Phi_j(t, k)\nu\hat{u}(k, \cdot) = \Phi_j(t, k)\nu(k^\alpha \hat{u}(k, \cdot)), \quad j = 0, 1, \dots, n+2,$$

holds point-wise for $k \in \mathbb{R}^n$ in the space L^2_ξ .

For $j = 0, \dots, n+1$, choosing κ_0 sufficiently small in Theorem 4.5 (2) if necessary, we have

$$\operatorname{Re}\lambda_j(k) = -\lambda_j^{(2)}|k|^2(1 + O(|k|)) \geq -a_0|k|^2 \quad (|k| \leq \kappa_0),$$

with some constants a_0 independent of k , and hence, in virtue of Theorem 4.5 (3) and Lemma 4.6,

$$\begin{aligned} & \|k^\alpha I_j(t, k)\|_{L^2(\mathbb{R}_k^n)}^2 \\ & \leq c_0 \int_{|k| \leq \kappa_0} |k^{\alpha - \alpha'}|^2 e^{2\operatorname{Re}\lambda_j(k)} \|k^{\alpha'} \hat{u}(k, \cdot)\|_{L^2_\xi}^2 dk \\ & \leq c_0 \left(\int_{|k| \leq \kappa_0} |k|^{2p'm} e^{-2p'a_0|k|^2 t} dk \right)^{1/p'} \left(\int_{\mathbb{R}^n} \|k^{\alpha'} \hat{u}(k, \cdot)\|_{L^2_\xi}^{2q'} dk \right)^{1/q'}, \end{aligned} \tag{4.26}$$

with $m = |\alpha - \alpha'|$ and $p' \in [1, \infty)$, $1/p' + 1/q' = 1$. Here, $c_0 > 0$ is a constant independent of all of $q', p', t, u, \alpha, \alpha'$. Note that

$$\int_{|k| \leq \kappa_0} |k|^{2p'm} e^{-2p'a_0|k|^2 t} dk \leq b(1+t)^{-n/2-p'm},$$

with a constant $b > 0$ depending only on m . Moreover, by the well known property of the Fourier transformation, we have

$$\left(\int_{\mathbb{R}^n} \|k^{\alpha'} \hat{u}(k, \cdot)\|_{L^2_\xi}^{2q'} dk \right)^{1/q'} \leq c_1 \|\partial_x^{\alpha'} u\|_{Z_q}^2, \quad \frac{1}{q} + \frac{1}{2q'} = 1,$$

with a constant $c_1 > 0$ independent of α' . Thus, (4.23) follows.

Similarly, if $\mathbf{P}\nu u = 0$, we have $\mathbf{P}\nu\hat{u}(k, \xi) = 0$ for all k , and hence from Theorem 4.5 (2) and Lemma 4.6,

$$\begin{aligned} \|\mathbf{P}_j(k)\nu\hat{u}(k, \cdot)\|_{L^2_\xi} &= \|(\mathbf{P}_j(k) - \mathbf{P}_{j,0}(\tilde{k}))\nu\hat{u}(k, \cdot)\|_{L^2_\xi} \\ &\leq c_0 |k| \|\hat{u}(k, \cdot)\|_{L^2_\xi} \quad (|k| \leq \kappa_0). \end{aligned}$$

Consequently, the same computation as (4.26) with m replaced by $m+1$ gives (4.24).

On the other hand, it follows from Lemma 4.6 and by the aid of Parseval's relation that

$$\|k^\alpha I_{n+2}(t, k)\|_{L^2(\mathbb{R}_k^n)} \leq C_1 e^{-\sigma_0 t} \|k^\alpha \hat{u}\|_{L^2(\mathbb{R}_k^n \times \mathbb{R}_\xi^n)} = C_1 e^{-\sigma_0 t} \|\partial_x^\alpha u\|_{Z_2},$$

which proves the part (2).

Finally, the continuity properties with respect to t can be easily seen from those of $\Phi_j(t, k)$ in Theorem 4.5. Now, the proof of the theorem is complete. \square

In (4.24), the extra decay rate $1/2$ is obtained under the assumption $\nu u \in \mathcal{N}^\perp$. This will be essential when combined with the property (1.40) of the nonlinear operator Γ .

The decay estimates in the space $Z_2 = L^2$ is not enough for solving the nonlinear problems because they are usually to be manipulated in a Banach algebra that is a property which Z_2 does not possesses. On the other hand, it is easily seen that $X_\beta = Y_\beta \cap Z_2$ defined in (1.34) is a Banach algebra if $\beta \geq 0$, and moreover, the nonlinear operator $\nu^{-\delta}\Gamma$ is bounded in this space if $\beta > n/2$ as is seen from Lemma 1.8. Thus, we shall still derive the decay estimates in the space Y_β . This can be done by means of the bootstrap argument starting from the L^2 decay estimates provided in Theorem 4.7 and based on the smoothing effect stated in Lemma 4.3.

More precisely, let $*$ denote the convolution in (4.9) and define the operators G_j , $j \in \mathbb{N}$, inductively by

$$G_0(t) = e^{tA},$$

$$G_j(t) = (G_0(t)K) * G_{j-1}(t) = G_{j-1}(t) * (KG_0(t)), \quad j = 1, 2, \dots.$$

Iterate Duhamel's formula

$$e^{tB} = e^{tA} + (e^{tA}K) * e^{tB}$$

to deduce

$$e^{tB} = \sum_{j=0}^k G_j(t) + G_k(t)K * e^{tB} \quad (4.27)$$

for $k \in \mathbb{N}$. Recall the operator F_k defined by (4.10) and note that

$$F_k(t) = G_{2k}(t)K, \quad k = 1, 2, \dots,$$

which rewrites (4.27) for $k = 2N$ as

$$e^{tB} = G_0(t) + G_1(t) + \sum_{j=1}^{N-1} F_j(t) * \{G_0(t) + G_1(t)\} + F_N(t) * e^{tB}. \quad (4.28)$$

Substituting the decomposition (4.21) into the last term of (4.28) yields

$$e^{tB} = D_0(t) + D_1(t) + D_2(t), \quad (4.29)$$

$$D_0(t) = G_0(t) = e^{tA},$$

$$D_1(t) = F_N(t) * E_1(t),$$

$$\begin{aligned} D_2(t) &= G_1(t) + \sum_{j=1}^{N-1} F_j(t) * (G_0(t) + G_1(t)) \\ &\quad + F_N(t) * (G_0(t) + E_2(t)). \end{aligned}$$

This is a desired decomposition in the space X_β if $N \geq [(n + \beta)/2] + 2$ as seen from the theorem following.

Theorem 4.10. *Recall $\sigma_{q,m}$ in (4.22) and σ_0 in (4.25). Let $\beta \geq 0$.*

(1) *Let $q \in [1, 2]$ and $\alpha \in \mathbb{N}^n$. If $\partial_x^\alpha u \in Z_q$ for some $\alpha' \leq \alpha$, then, for any $\delta \in [0, 1]$, we have*

$$\partial_x^\alpha D_1(t) \nu^\delta u \in L^\infty(0, \infty; X_\beta),$$

and the algebraic decay estimates

$$\|\partial_x^\alpha D_1(t) \nu^\delta u\|_{X_\beta} \leq b_1 (1+t)^{-\sigma_{q,m}} \|\partial_x^{\alpha'} u\|_{Z_q}, \quad (4.30)$$

$$\|\partial_x^\alpha D_1(t) (I - \mathbf{P}) \nu^\delta u\|_{X_\beta} \leq b_2 (1+t)^{-\sigma_{q,m+1}} \|\partial_x^{\alpha'} u\|_{Z_q}, \quad (4.31)$$

for a.a. $t \geq 0$, where $m = |\alpha - \alpha'|$ and b_1, b_2 are positive constants depending only on m and β .

(2) *Let $\alpha \in \mathbb{N}^n$ and $\partial_x^\alpha u \in X_\beta$. Then, for any $\delta \in [0, 1]$, we have*

$$\partial_x^\alpha D_2(t) \nu^\delta u \in L^\infty(0, \infty; X_\beta),$$

and the exponential decay estimate

$$\|\partial_x^\alpha D_2(t) \nu^\delta u\|_{X_\beta} \leq b_3 e^{-\sigma_0 t} \|\partial_x^\alpha u\|_{X_\beta}. \quad (4.32)$$

for a.a. $t \geq 0$, where b_3, σ_0 are positive constants independent of α, δ, u , and t .

Proof. Notice that for any numbers $\kappa_1 > 0$ and $\kappa_2 \geq 0$,

$$e^{-\kappa_1 t} * (1+t)^{-\kappa_2} \leq C(1+t)^{-\kappa_2}, \quad e^{-\kappa_1 t} * e^{-\kappa_2 t} \leq C e^{-\kappa_2 t} \quad (\kappa_1 > \kappa_2) \quad (4.33)$$

hold for $t \geq 0$ with some positive constant $C > 0$. The first inequality can be concluded by the computation,

$$\begin{aligned} \int_0^t e^{-\kappa_1(t-s)}(1+s)^{-\kappa_2}ds &= \int_0^{t/2} + \int_{t/2}^t \\ &\leq e^{-\kappa_1 t/2} \int_0^{t/2} (1+s)^{-\kappa_2} ds + (1+t/2)^{-\kappa_2} \int_{t/2}^t e^{-\kappa_1(t-s)} ds \\ &\leq C(1+t)^{-\kappa_2}, \end{aligned}$$

whereas the second inequality comes by a direct computation.

Now, since $\partial_x^\alpha F_N(t) = F_N(t)\partial_x^\alpha$,

$$\partial_x^\alpha D_1(t) = F_N(t) * \left(\partial_x^\alpha E_1(t) \right)$$

holds. Therefore, it follows from the first inequality in (4.33), Lemma 4.2, and Theorem 4.7 that (4.30) holds in Y_β in place of X_β ;

$$\begin{aligned} \|\partial_x^\alpha D_1(t)\nu^\delta u\|_{Y_\beta} &\leq Ce^{-(\nu_*-\epsilon)t} * (1+t)^{-\sigma_{q,m}} \|\partial_x^{\alpha'} u\|_{Z_q}, \\ &\leq C(1+t)^{-\sigma_{q,m}} \|\partial_x^{\alpha'} u\|_{Z_q}, \end{aligned}$$

while, since $F_N(t)$ can be also taken to be a bounded operator on $Z_2 = L^2$ with the estimate

$$\|F_N(t)u\|_{Z_2} \leq Ce^{-(\nu_*-\epsilon)t} \|u\|_{Z_2},$$

(4.30) holds also in Z_2 ;

$$\|\partial_x^\alpha D_1(t)\nu^\delta u\|_{Z_2} \leq b_1(1+t)^{-\sigma_{q,m}} \|\partial_x^{\alpha'} u\|_{Z_q}.$$

Combining these two estimate proves (4.30). Also, (4.31) can be proved in the same way by using the estimate (4.24).

On the other hand, it follows from (1.21), Lemma 1.2 (b) and (4.7) that

$$\|G_1(t)\nu u\|_{Y_\beta}, \|F_k(t) * G_j(t)\nu u\|_{Y_\beta} \leq Ce^{-(\nu_*-\epsilon)t} \|u\|_{Y_\beta},$$

and from Lemma 1.7 that

$$\|G_1(t)\nu u\|_{Z_2}, \|F_k(t) * G_j(t)\nu u\|_{Z_2} \leq Ce^{-(\nu_*-\epsilon)t} \|u\|_{Z_2}.$$

Combining these estimates and (4.25) in Theorem 4.7 proves (4.32). Now the proof of the theorem is complete. \square

Remark 4.11. A point of the above theorem is the fact that the operators $D_1(t), D_2(t)$ can recover the weight loss of order $O(|\xi|^\gamma \delta)$ coming from the unbounded operator ν^δ ($\delta > 0$). This is crucial for controlling the unboundedness of the collision operator Γ stated in Lemmas 1.8 and 1.9. On the other hand, the operator $G_0(t) = e^{tA}$, the first term of the decomposition (4.29), does not enjoy this property. However, the weight loss can be again controlled, though by the smoothing effect of the integration in t as shown in Lemma 4.3.

We now summarize the decay estimates of e^{tB} which holds when the weight ν is removed.

Theorem 4.12. *Let $q \in [1, 2]$ and $\beta \geq 0$. Then, there is a positive constant c_0 such that for any $u \in X_\beta \cap Z_q$, it holds that*

$$\|e^{tB}u\|_{X_\beta} \leq c_0(1+t)^{-\sigma_{q,0}} \left\{ \|u\|_{X_\beta} + \|u\|_{Z_q} \right\}, \quad (4.34)$$

$$\|e^{tB}(I - \mathbf{P})u\|_{X_\beta} \leq c_0(1+t)^{-\sigma_{q,1}} \left\{ \|u\|_{X_\beta} + \|u\|_{Z_q} \right\}. \quad (4.35)$$

Proof. These two estimates follow immediately from the above theorem with $\delta = 0$ since the operator $G_0(t)$ enjoys the exponential decay in X_β as seen from the explicit expression (4.7). \square

Since $\sigma_{2,0} = 0$, the estimate (4.34) does not provide any decay rate for the case $q = 2$. Indeed, there is some reason to believe that there is no decay for some u in the space Y_β if $q = 2$. However, the decay property shows up in some other spaces. For example,

Theorem 4.13. *Any $u \in Z_2 = L^2$ enjoys the decay property*

$$\|e^{tB}u\|_{Z_2} \rightarrow 0 \quad (t \rightarrow \infty).$$

Proof. Recall the decomposition (4.21). For $q = 2$, (4.23) for $\alpha = \alpha' = 0$ asserts that the operator $E_1(t)$ is uniformly bounded for $t \geq 0$ on the space Z_2 . On the other hand, it is easy to see that the intersection $Z_2 \cap Z_q$ for $q \in [1, 2)$ ($q \neq 2$) is a dense subset of Z_2 . This and (4.23) for $q < 2$ imply that

$$\|E_1(t)u\|_{Z_2} \rightarrow 0 \quad (t \rightarrow \infty).$$

Since the remaining two terms in the decomposition (4.21) enjoy the exponential decay, this concludes the theorem. \square

This theorem is also true in the spaces (5.21) and (5.22) with a slight modification. The details are omitted. See [81], [91].

4.4 Effect of external force

In this subsection, we will give a brief presentation of the decay rate estimates on the solution operator of the linearized Boltzmann equation with external force. Notice that the linearized Boltzmann operator without forcing generates a semi-group as shown in the previous subsections. However, this is no longer true for the case with a time dependent force. In fact, to our knowledge, there is no general spectral theory for the linearized Boltzmann operator with forcing. The following decay estimate on the solution operator is obtained by combining the spectral estimates on the linearized Boltzmann operator without forcing and the energy estimate on the case with forcing. This kind of combination is useful in the sense that the decay estimates of lower order derivatives can be obtained by using the spectral estimates on the corresponding force-free case while the higher order derivatives can be estimated by using the L^2 -energy method developed recently by [41] so that the decay estimate can be closed. This method was also used recently in the study of some hyperbolic-parabolic systems, such as Navier-Stokes equations [32], and some hyperbolic systems with dissipative mechanism, such as Euler equations with frictional damping [99].

Now consider the Boltzmann equation for the hard-sphere gas in n -dimensional space under the influence of an external force

$$\partial_t f + \xi \cdot \nabla_x f + F \cdot \nabla_\xi f = Q(f, f). \quad (4.36)$$

Here, the external force field $F = F(t, x)$ is a given time dependent function and Q is the usual Boltzmann collision operator.

Again, we consider the perturbative solution near an absolute Maxwellian so that the equation for the perturbation u is:

$$\partial_t u + \xi \cdot \nabla_x u + F \cdot \nabla_\xi u - \frac{1}{2} \xi \cdot F u = \mathbf{L}u + \Gamma(u) + \mathbf{M}^{1/2} \xi \cdot F, \quad (4.37)$$

where $\mathbf{L}u$ and $\Gamma(u, u)$ are the linear and nonlinear operators defined before.

Corresponding to the previous subsection about the decay estimate on e^{tB} , we now study the decay in time properties of the solution operator for the linearized Boltzmann equation corresponding to (4.37), that is,

$$\partial_t u + \xi \cdot \nabla_x u + F \cdot \nabla_\xi u - \frac{1}{2} \xi \cdot F u = \mathbf{L}u. \quad (4.38)$$

Unlike the case without external force discussed in the previous section, we can not obtain similar result in the space X_β . Instead, the decay estimates given below are obtained in some Sobolev space weighted in velocity variables. For any initial time $s \in \mathbb{R}$, consider (4.38) for $t \geq s$

with initial data

$$u(t, x, \xi)|_{t=s} = u_0(x, \xi), \quad x \in \mathbb{R}^n, \quad \xi \in \mathbb{R}^n. \quad (4.39)$$

Formally the solution to the initial value problem (4.38)-(4.39) is written as

$$u(t, x, \xi) = U(t, s)u_0, \quad -\infty < s \leq t < \infty,$$

where $U(t, s)$ is called the solution operator for the linear equation (4.38). For later presentation, we need the following notations of norms on the perturbation u . For any numbers $\ell, k \in \mathbb{N}$, define a norm $[[\cdot]]_{0,k}^{(\ell)}$ and a seminorm $[[\cdot]]_{1,k}^{(\ell)}$ over the Sobolev space $H^\ell(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)$ by

$$[[u]]_{0,k}^{(\ell)} = \sum_{0 \leq |\alpha| \leq \ell} \|\nu^k \partial_{x,\xi}^\alpha u\|, \quad (4.40)$$

$$[[u]]_{1,k}^{(\ell)} = \sum_{1 \leq |\beta| \leq \ell} \|\partial_x^\beta \mathbf{P} u\| + \sum_{0 \leq |\alpha| \leq \ell} \|\nu^k \partial_{x,\xi}^\alpha \{\mathbf{I} - \mathbf{P}\} u\|, \quad (4.41)$$

where $u = u(x, \xi)$. Notice that

$$[[u]]_{0,k}^{(\ell)} \sim [[u]]_{1,k}^{(\ell)} + \|u\|.$$

With these notations, the decay estimates on the solution operator $U(t, s)$ can be summarized in the following theorem, [34].

Theorem 4.14. *Suppose that*

- (i) *the integers $n \geq 3$, $\ell \geq 2$ and $1 \leq q < \frac{2n}{n+2}$;*
- (ii) *there is a constant $\delta > 0$ such that*

$$\sum_{0 \leq |\beta| \leq \ell} \|(1 + |x|) \partial_x^\beta F(t, x)\|_{L_{t,x}^\infty} + \sum_{0 \leq |\beta| \leq \ell-1} \|(1 + |x|) \partial_t \partial_x^\beta F(t, x)\|_{L_{t,x}^\infty} \leq \delta,$$

and

$$\left\| |x| F(t, x) \right\|_{L_t^\infty(L_x^{2q/(2-q)})} \leq \delta.$$

Then for any $k \geq 1$, there exist constants $\delta_0 > 0$ and $C_0 > 0$ such that for any $\delta \leq \delta_0$, the linear solution operator $U(t, s)$, $-\infty < s \leq t < \infty$, corresponding to the linear equation (4.38) satisfies the decay in time estimates

$$[[U(t, s)u_0]]_{m,k}^{(\ell)} \leq C_0(1 + t - s)^{-\sigma_{q,m}} ([[u_0]]_{m,k}^{(\ell)} + \|u_0\|_{Z_q}), \quad m = 0, 1,$$

for any $u_0 = u_0(x, \xi)$ such that the right hand side of the above inequality is bounded, where the constant C_0 depends only on n, ℓ, q, k and δ_0 .

Remark 4.15. The decay estimates on the solution operator given in Theorem 4.14 are optimal in the corresponding norms which are the same as the case without forcing. The draw-back is that there is no decomposition of the solution operator according to the long and short waves for the more precise time decay estimate. Therefore, the effect of the macroscopic and microscopic components on the time decay is not given. In addition, these decay estimates may not be enough to prove the global existence of the time periodic solutions to the Boltzmann equation with time periodic external force when the space dimension is less or equal to 4, see [34], [92].

5 Global existence and convergence rates

There is an extensive literature on the study of the global solutions to the Cauchy problem for the Boltzmann equation. Roughly speaking, most of the existence theorems established so far are categorized into two types. One is about the existence theorems of the unique global strong solution near the global Maxwellian. This has been achieved in various function spaces which are more or less regular with respect to the space variable, see e.g. [41, 68, 71, 81, 88, 89, 91, 95]. The other is that of the renormalized solutions in the space L^1 , which was developed in [31] under a general situation that no smallness conditions nor regularity assumptions are imposed on the initial data. However, the uniqueness problem on the renormalized solutions remains open, which is now widely understood to be one of the most important issues in the theory of the Boltzmann equation. Another issue, which is also important, is to establish the uniqueness criteria that are as general as possible, or in other words, to find the “optimal function spaces” for the well-posedness of the Cauchy problem. Based on the spectral analysis on the semigroup generated by the linearized Boltzmann operator given in the previous sections, one attempt in this direction has been given in [96] where the global existence and optimal convergence rates are obtained in the function space X_β . And this will be explained in the following subsections.

5.1 Global existence

In this subsection, the global existence of solutions is proved in X_β defined in (1.34) which does not assume any regularity properties but sets the Cauchy problem well-posed globally in a mild sense. One of the advantages of this space is that it can reveal the dependency of the decay rate on the order of spatial derivatives of solutions.

Consider the Cauchy problem

$$\begin{cases} \frac{\partial u}{\partial t} + \xi \cdot \nabla_x u = \mathbf{L}u + \Gamma(u, u), & (t, x, \xi) \in \mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^n, \\ u(0, x, \xi) = u_0(x, \xi), & (x, \xi) \in \mathbb{R}^n \times \mathbb{R}^n. \end{cases} \quad (5.1)$$

We shall discuss (5.1) in the mild form defined in the function space X_β .

The precise definition of the mild formulation of (5.1) will be given below, and the corresponding solution will be called a mild solution. The following theorem states that the mild formulation is well-posed in the space X_β globally in time.

Theorem 5.1. *Let $n \geq 3$ and $\beta > n/2$, and assume (1.19) for some $\gamma \in [0, 1]$. Then, there are two positive constants a_0, a_1 such that for any initial data u_0 satisfying*

$$u_0 \in X_\beta, \quad \|u_0\|_{X_\beta} \leq a_0, \quad (5.2)$$

the Cauchy problem (5.1) has a unique global mild solution u in the function class

$$u \in L^\infty(0, \infty; X_\beta), \quad \sup_{t \geq 0} \|u(t)\|_{X_\beta} \leq a_1 \|u_0\|_{X_\beta}. \quad (5.3)$$

Furthermore, the solution map defined by $u_0 \mapsto u$ is a continuous map from the neighborhood of 0 in X_β defined by (5.2) to the space $L^\infty(0, \infty; X_\beta)$.

Firstly, we shall make precise the definition of the mild solution to the Cauchy problem (5.1). Rewrite (5.1) into the form of the integral equation by means of Duhamel's formula as

$$u(t) = e^{tB}u_0 + \int_0^t e^{(t-s)B}\Gamma(u(s), u(s))ds \quad (5.4)$$

and call u a mild solution if it solves (5.4). Actually, this definition makes sense only when the last integral is well-defined.

It is just the decay estimates in Theorem 4.10 for the linear semigroup e^{tB} established in the previous section that enable us to justify the well-definedness of this integral. Furthermore, they also provide the uniform boundedness of the integral for all $t \geq 0$, which is, then, used in a combination with the contraction mapping principle, to establish the existence of global solution to the nonlinear integral equation (5.4).

We shall regard (5.4) as an equation in the space

$$W_{\beta, \sigma} = \{u \in L^\infty(0, \infty; X_\beta) \mid |||u|||_{\beta, \sigma} < \infty\}, \quad (5.5)$$

$$|||u|||_{\beta, \sigma} = \sup_{t \geq 0} (1+t)^\sigma \|u(t)\|_{X_\beta},$$

where $\beta, \sigma \geq 0$. Consider the integral in (5.4) in the form

$$\Psi[u, v](t) = e^{tB} * \Gamma(u, v) = \int_0^t e^{(t-s)B} \Gamma(u(s), v(s)) ds. \quad (5.6)$$

Recall $\sigma_{1,1}/4 + 1/2$ from (4.22). The following estimate on $\Phi[u, v]$ is crucially used in the proofs of the existence here and the optimal convergence rates in the next subsection.

Lemma 5.2. *Let $\beta > n/2$, $\sigma \in [0, \sigma_{1,1}]$ and $u, v \in W_{\beta, \sigma}$. Then, $\Psi[u, v]$ is in $W_{\beta, \sigma}$ and satisfies*

$$|||\Psi[u, v]|||_{\beta, \sigma} \leq c_1 |||u|||_{\beta, \sigma} |||v|||_{\beta, \sigma}, \quad (5.7)$$

for a constant $c_1 > 0$ independent of u, v .

Proof. Put $\nu h = \Gamma(u, v) = (I - \mathbf{P})\Gamma(u, v)$ and decompose Ψ according to the decomposition (4.29) as

$$\begin{aligned} \Psi[u, v](t) &= D_0(t) * (\nu h) + D_1(t) * (\nu h) + D_2(t) * (\nu h) \\ &= \Psi_0[h] + \Psi_1[h] + \Psi_2[h]. \end{aligned}$$

Evidently, it suffices to prove the lemma for each Ψ_j .

Ψ_0 is the same as (4.13). Let $\beta > n/2$ and choose $\delta < 1$ so close to 1 that $\beta > n/2 + \gamma(1 - \delta)$ can hold. Then, (1.38) in Lemma 1.8 and Lemma 4.3 (a) yield

$$\begin{aligned} \|\Psi_0[h](t)\|_{Z_2} &\leq C_1(1+t)^{-2\sigma} \sup_{t \geq 0} (1+t)^{2\sigma} \|\nu^{-\delta} \Gamma(u(t), v(t))\|_{Z_2} \\ &\leq C_2(1+t)^{-2\sigma} \sup_{t \geq 0} (1+t)^{2\sigma} \left(\|u(t)\|_{Y_\beta} \|v(t)\|_{Z_2} + \|u(t)\|_{Z_2} \|v(t)\|_{Y_\beta} \right) \\ &\leq C_2(1+t)^{-2\sigma} |||u|||_{\beta, \sigma} |||v|||_{\beta, \sigma} \end{aligned}$$

for $\sigma \geq 0$. On the other hand, (1.39) and Lemma 5.2 (b) yield

$$\begin{aligned} \|\Psi_1[h](t)\|_{Y_\beta} &\leq C_1(1+t)^{-2\sigma} \sup_{t \geq 0} (1+t)^{2\sigma} \|\nu^{-1} \Gamma(u(t), v(t))\|_{Y_\beta} \\ &\leq C_2(1+t)^{-2\sigma} \sup_{t \geq 0} (1+t)^{2\sigma} \|u(t)\|_{Y_\beta} \|v(t)\|_{Y_\beta} \\ &\leq C_2(1+t)^{-2\sigma} |||u|||_{\beta, \sigma} |||v|||_{\beta, \sigma}. \end{aligned}$$

Adding these two estimates proves the lemma for Ψ_0 .

The estimate for Ψ_1 comes from (1.37) in Lemma 1.8 and (4.31) in

Theorem 4.10 for $q = 1, \alpha = \alpha' = 0$:

$$\begin{aligned} \|\Psi_1[h](t)\|_{X_\beta} &\leq C \int_0^t (1+t-s)^{-\sigma_{1,1}} \|\nu^{-1}\Gamma[u(s), v(s)]\|_{Z_1} ds \\ &\leq C(1+t)^{-\sigma_{1,1}} * (1+t)^{-2\sigma} \sup_{t \geq 0} (1+t)^{2\sigma} \|\nu^{-1}\Gamma[u(t), v(t)]\|_{Z_1} \\ &\leq C(1+t)^{-\sigma_*} \sup_{t \geq 0} (1+t)^{2\sigma} \|u(t)\|_{Z_2} \|v(t)\|_{Z_2} \\ &\leq C(1+t)^{-\sigma_*} |||u|||_{\beta,\sigma} |||v|||_{\beta,\sigma}, \end{aligned}$$

where $\sigma_* = \min(\sigma_{1,1}, 2\sigma)$ and we have used the estimate

$$\begin{aligned} &(1+t)^{-\sigma_{1,1}} * (1+t)^{-2\sigma} \\ &\leq (1+t/2)^{-\sigma_{1,1}} \int_0^{t/2} (1+s)^{-2\sigma} ds + \int_{t/2}^t (1+t-s)^{-\sigma_{1,1}} ds (1+t/2)^{-2\sigma}. \end{aligned}$$

The estimate for Ψ_2 is derived exactly in the same way as for Ψ_1 by using (1.38) and (1.39) in Lemma 1.8, (4.32) in Theorem 4.10 for $q = 1, \alpha = \alpha' = 0$, and the estimate

$$e^{-\sigma_0 t} * (1+t)^{-2\sigma} \leq C(1+t)^{-2\sigma}.$$

This completes the proof of the lemma. \square

This lemma assures that the integral equation (5.4) can be regarded as an equation in the space X_β , and the definition of the mild solution is thus well-defined. Now, we are in the position to prove Theorems 5.1.

Proof of Theorem 5.1. Notice that (5.2) and (5.3) are respectively the special case of (5.13) and (5.14) for $q = 2$ stated below in Theorem 5.3. Indeed, we will prove the existence of the mild solutions satisfying (5.14) under the assumption (5.13) for arbitrary $q \in [1, 2]$. Put

$$\Phi[u](t) = e^{tB} u_0 + \Psi[u, u]. \quad (5.8)$$

This defines a nonlinear map and (5.4) is written as $u = \Phi[u]$, that is, the mild solution u is a fixed point of the map Φ . We now show that Lemma 5.2 ensures that Φ is a contraction map if u_0 is small.

In the sequel, we fix

$$q \in [1, 2], \quad \beta > n/2,$$

and for simplifying the notation, set

$$\begin{aligned} W &= W_{\beta, \sigma_{q,0}}, \quad |||u||| = |||u|||_{\beta, \sigma_{q,0}}, \\ U_0 &= \|u_0\|_{Y_\beta} + \|u_0\|_{Z_q}. \end{aligned} \quad (5.9)$$

Then, for $u_0 \in X_\beta \cap Z_q$, Theorem 4.12 gives

$$|||e^{tB}u_0||| \leq c_0 U_0.$$

Combine this with Lemma 5.2 for $\sigma = \sigma_{q,0}$ to deduce

$$|||\Phi[u]||| \leq c_0 U_0 + c_1 |||u|||^2, \quad u \in W. \quad (5.10)$$

Here, note from the definition (4.22),

$$\sigma_{q,k} = \frac{n}{2} \left(\frac{1}{q} - \frac{1}{2} \right) + \frac{k}{2},$$

for $q \in [1, 2]$ and $k \in \mathbb{N}$ that $\sigma_{q,0} \leq \sigma_{1,1}$ holds.

Furthermore, since the operator Γ is bilinear symmetric, that is,

$$\Gamma[u, u] - \Gamma[v, v] = \Gamma[u + v, u - v]$$

holds, we have

$$\Phi[u] - \Phi[v] = \Psi[u, u] - \Psi[v, v] = \Psi[u + v, u - v],$$

and, in virtue of Lemma 5.2,

$$|||\Phi[u] - \Phi[v]||| = |||\Psi[u + v, u - v]||| \leq c_1 |||u + v||| |||u - v|||. \quad (5.11)$$

Here, the constant c_1 is the same as the one in (5.10) coming from (5.7). This fact is essential for the construction of the mild solutions.

Now, consider the quadratic equation of a ,

$$c_1 a^2 - a + c_0 U_0 = 0,$$

which has two positive roots if U_0 is so small that

$$D \equiv 1 - 4c_0 c_1 U_0 > 0.$$

Then, its smaller positive root, denoted by a_* , is given by

$$a_* = \frac{1}{2c_1} (1 - \sqrt{D}). \quad (5.12)$$

With this choice, set

$$W_* = \{u \in W \mid |||u||| \leq a_*\}.$$

Clearly, W_* is a complete metric space with the metric induced by the norm $|||\cdot|||$. Now, it follows from (5.10) that for any $u \in W_*$,

$$|||\Phi[u]||| \leq c_0 U_0 + c_1 a_*^2 = a_*.$$

This implies that Φ maps W_* into itself. Moreover, it follows from (5.11) that for any $u, v \in W_*$,

$$\begin{aligned} |||\Phi[u] - \Phi[v]||| &\leq c_1(|||u||| + |||v|||)|||u - v||| \\ &\leq 2c_1 a_* |||u - v|||. \end{aligned}$$

Since $2c_1 a_* = 1 - \sqrt{D} < 1$, this proves that Φ is a contraction map. Now, the contraction mapping principle proves the existence part in the theorem provided that the constants a_0 is so chosen that $a_0 < 1/(4c_0 c_1)$ and then the constant a_1 is so chosen that $a_* / U_0 \leq a_1$ holds for all $U_0 \in (0, a_0]$. Clearly, the latter is possible as seen from (5.12).

It remains to prove the continuity of the map $u_0 \mapsto u$. Indeed, we will prove this continuity for arbitrary $q \in [1, 2]$. To this end, let $v \in W_*$ be another mild solution for the initial data $v_0 \in X_\beta \cap Z_q$. Then, again in virtue of the bilinear symmetry of Γ , we have

$$u - v = e^{tB}(u_0 - v_0) + \Psi[u - v, u + v],$$

and hence exactly in the same way as above and since $u, v \in W_*$,

$$|||u - v||| \leq c_0 R_0 + 2c_1 a_* |||u - v|||, \quad R_0 = \|u_0 - v_0\|_{X_\beta} + \|u_0 - v_0\|_{Z_q}.$$

Recalling that $2c_1 a_* = 1 - \sqrt{D} < 1$, we have

$$|||u - v||| \leq c_0 R_0 / \sqrt{D},$$

which proves the desired continuity. \square

5.2 Optimal convergence rates

Next, we discuss the relations between the regularity in x and the decay rate in t of the mild solution. Put

$$\begin{aligned} \alpha &= (\alpha_1, \alpha_2, \dots, \alpha_n) \in \mathbb{N}^n, \quad |\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_n, \\ \partial_x^\alpha &= \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \cdots \partial_{x_n}^{\alpha_n}. \end{aligned}$$

The optimal decay rates in t of the mild solution and its spatial derivatives are given in the following theorem.

Theorem 5.3 (Spatial regularity and Decay rate). *Under the situation of Theorem 5.1 and for $q \in [1, 2]$, there are two positive constants a_0, a_1 such that for any initial data u_0 satisfying*

$$u_0 \in X_\beta \cap Z_q, \quad \|u_0\|_{X_\beta} + \|u_0\|_{Z_q} \leq a_0, \quad (5.13)$$

the following holds. Here, the constant a_0 may be smaller and a_1 may be larger than those of Theorem 5.1.

(1) The mild solution u constructed in the previous theorem enjoys the decay estimate

$$\|u(t)\|_{\beta} \leq a_1(1+t)^{-\sigma_{q,0}}(\|u_0\|_{X_{\beta}} + \|u_0\|_{Z_q}), \quad a.a. \quad t \geq 0. \quad (5.14)$$

(2) Let $q \in [1, 2]$ and $\ell \in \mathbb{N}$. Suppose, in addition to (5.13), that

$$\partial_x^{\alpha} u_0 \in X_{\beta} \cap Z_q, \quad (\alpha \in \mathbb{N}^n, |\alpha| \leq \ell). \quad (5.15)$$

Then, the mild solution u has the spatial derivatives up to the order ℓ which satisfy

$$\partial_x^{\alpha} u \in L^{\infty}(0, \infty; X_{\beta}), \quad |\alpha| \leq \ell, \quad (5.16)$$

and the algebraic decay

$$\|\partial_x^{\alpha} u(t)\|_{X_{\beta}} \leq C_k(1+t)^{-\sigma_{q,k}}, \quad |\alpha| = k, \quad a.a. \quad t \geq 0, \quad (5.17)$$

for each $k = 0, 1, \dots, \ell$, where C_k is a positive constant depending on k and the norms of $\partial_x^{\alpha} u_0$ in $X_{\beta} \cap Z_q$ for $|\alpha| \leq k$.

Moreover, the mild solution has some regularity properties also with respect to t which depend on the initial data.

Theorem 5.4 (Temporal regularity). *The mild solution u of Theorem 5.1 is L^2 -continuous in t , that is,*

$$u \in C^0([0, \infty); L^2). \quad (5.18)$$

Moreover, if the conditions (5.13) and (5.15) are fulfilled for some $q \in [1, 2]$ and $\ell \geq 1$, then u enjoys

$$\partial_x^{\alpha} u \in C^0([0, \infty); L^2), \quad |\alpha| \leq \ell, \quad (5.19)$$

$$\partial_x^{\alpha} u \in C^0([0, \infty); X_{\beta-1}), \quad \partial_x^{\alpha} \partial_t u \in L^{\infty}(0, \infty; L_{\beta-1}^{\infty}), \quad |\alpha| \leq \ell - 1. \quad (5.20)$$

In particular, if $\ell \geq 2$, then

$$\partial_x^{\alpha} \partial_t u \in C^0([0, \infty); L_{\beta-2}^{\infty}), \quad |\alpha| \leq \ell - 2,$$

and u is a classical solution in the sense that it satisfies the equation (5.1)₁ almost everywhere in $\mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^n$ and the initial condition (5.1)₂ by the continuity (5.18)–(5.20).

We now compare these decay estimates with some previous results. First of all, the hypocoercivity theory which is closely related to, but is different from, the hypoellipticity theory has become one of the main focuses in the study of problems from mathematical physics, in particular

from kinetic theory. The main feature of this theory is that the coupling of a degenerate diffusion operator and a conservative operator may give the dissipation in all variables, and the convergence to the equilibrium state which lies in a subspace smaller than the kernel of the diffusion operator. Breakthroughs have been made and substantial results have been obtained recently, especially by Villani and his collaborators, on problems in bounded domains or a torus. However, there are still many challenging problems remained unsolved. And the Boltzmann equation is one of the key models which have this kind of phenomena in terms of the convergence to the equilibrium.

Remark 5.5. It should be mentioned that the well-posedness in Theorem 5.1 has been already shown in some other function spaces. In fact, the first well-posedness theorem has been established with the space X_β replaced by

$$L_\beta^\infty(\mathbb{R}_\xi^n; H^\ell(\mathbb{R}_x^n)), \quad \ell > \frac{n}{2}, \quad \beta > \frac{n}{2} + 1, \quad . \quad (5.21)$$

where $H^\ell(\mathbb{R}_x^n)$ is the L^2 Sobolev space of order ℓ , by using the spectral analysis of the linearized problem for (5.1). See [75, 89, 91]. The same result has been also proved in [81], though for the torus case, with the space

$$L_\beta^\infty(\mathbb{R}_\xi^n; C^m(\mathbb{T}_x^n)), \quad \beta > \frac{n}{2} + 1, \quad m = 0, 1, \dots \quad (5.22)$$

Clearly, these spaces are smaller than X_β . Notice that the weight order β is bigger by one in (5.21) and (5.22) compared with that in Theorem 5.1.

On the other hand, by means of a combination of the energy method which is familiar in the theory of PDE's and the macro-micro decomposition which was initiated in [70] and developed in [68] and also in [41] independently in a slightly different way, the well-posedness has been established also in the Sobolev space

$$\bigcap_{0 \leq \ell_0 + \ell_1 \leq N} \mathbf{H}_{t,x,\xi}^{\ell_0, \ell_1, 0}(\mathbb{R}_+ \times \mathbb{R}^n \times \mathbb{R}^n)$$

with $N \geq [(n+1)/2] + 2$. Although there is no inclusion relation between this space and X_β , the regularity properties with respect to t and x are required for the energy method.

Finally, the Green's function of the Boltzmann equation together with the existence and uniqueness of global solutions with the fine structure for initial data subject to the exponential decay in x have been established in [71].

Thus, it is a challenging problem to ask whether or not there exist any other larger spaces such that the Cauchy problem becomes well-posed around the global Maxwellian.

In this respect, the Cauchy problem near vacuum is also to be mentioned. The well-posedness is known under the smallness condition in the L^∞ -space with exponential weight both in x and ξ , [43, 48]. Thus, another challenging problem is to find function spaces larger than this space for the well-posedness, with or without smallness condition on initial data, and to elucidate the gap between such spaces and the space L^1 where the renormalized solutions have been constructed.

Remark 5.6. The decay rates in (5.17) are optimal in the sense that they coincide with those for the linearized equation, that is, those for the semigroup e^{tB} established in Theorems 4.7 and 4.10. Here, it is to be stressed that (5.15) does not require any smallness condition on the derivatives of initial data. This contrasts with the recent result by Strain-Guo [85] on the almost exponential decay of the solutions u for the Cauchy problem on the torus T^3 with the cutoff soft potential. His result is phrased as follows. Let $N \geq 4$ and let $H_{x,\xi}^N$ be the Sobolev space of order N in both x and ξ . For any k , it holds that if $a_k \equiv \|u_0\|_{H_{x,\xi}^{N+k}}$ is sufficiently small, then

$$(AED) \quad \|u(t)\|_{H_{x,\xi}^N} \leq C a_k (1 + \frac{t}{k})^{-k}.$$

Here, it is required that $a_k \rightarrow 0$ as $k \rightarrow \infty$. Notice that the convergence for the hard potential case is exponential on the torus, [81, 88].

On the other hand, Desvillettes-Villani [28] also established (AED) but in a quite different context: Consider the Cauchy problem on the torus or in a bounded domain with the specular or reverse reflection boundary condition. Assume that u is a smooth global solution satisfying

$$u(t) \in BC^0([0, \infty); H_{x,\xi}^\ell)$$

for sufficiently large $\ell > k$. Then, [28] asserts that (AED) holds. The smallness condition on u_0 is not assumed, but the existence of such smooth global solutions is a big open problem at the present.

Remark 5.7. Theorem 5.3 (2) does not cover the case $q = 2$. However, we can recover the decay rate $\sigma_{2,N}$ if we choose a_0 smaller with N . The point here is again that the derivatives of the initial data need not to be small. For the proof, see Remark 5.9 after the proof of Theorem 5.3.

Remark 5.8. In contrast to the x -derivatives $\partial_x^\alpha u$, the ξ -derivatives $\partial_\xi^\alpha u$ do not decay faster than $O(t^{-\sigma_{q,0}})$. Thus, the solution diffuses fast in the x -space but slow in the ξ -space. In other words, the linearized Boltzmann operator B has a smoothing property similar to the spatial Laplacian Δ_x , but it has not a counterpart for the ξ variables. This is one of the features of the Boltzmann equation to be compared with

other kinetic equations having some smoothing property with respect to ξ such as the Fokker-Planck-Boltzmann equation [44, 58],

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + \mu \nabla_v \cdot (\xi f) - \nu \Delta_v f = Q(f, f),$$

and the classical Landau equation [85], which is the same as the Boltzmann equation except

$$Q(f, f) = \nabla_\xi \cdot \left\{ \int_{\mathbb{R}^3} \phi(v - v') [f(v') \nabla_v f(v) - f(v) \nabla_v f(v')] dv' \right\}$$

where $\phi^{ij}(v) = |v|^{\gamma+2} (\delta_{ij} - \frac{v_i v_j}{|v|^2})$, $\gamma \in (-n, 1]$, and so on. It may be interesting to see whether the Vlasov-Poisson (Maxwell)-Boltzmann equations have the same properties.

Proof of Theorem 5.3. First of all, the Theorem 5.3 (1) follows directly from the proof of Theorem 5.1. For the Theorem 5.3 (2), the proof will be done by induction. For some $N \in \mathbb{N}$, suppose that Theorem 5.3 is true up to $\ell = N - 1$ and that the condition (5.15) is fulfilled for $\ell = N$.

In the sequel, $\alpha \in \mathbb{N}^n$ is fixed so that $|\alpha| = N$. Apply ∂_x^α to the map Φ in (5.8):

$$\partial_x^\alpha \Phi[u] = \partial_x^\alpha e^{tB} u_0 + \partial_x^\alpha \Psi[u, u](t). \quad (5.23)$$

Exactly in the same way as (4.34), the first term on the right hand side is evaluated as

$$\|\partial_x^\alpha e^{tB} u_0\|_{X_\beta} \leq c(1+t)^{-\sigma_{q,N}} (\|\partial_x^\alpha u_0\|_{X_\beta} + \|u_0\|_{Z_q}) \leq c(1+t)^{-\sigma_{q,N}}. \quad (5.24)$$

Here and hereafter, c denotes various positive constants that may depend on N and the norms $\|\partial_x^{\alpha'} u_0\|_{X_\beta} + \|u_0\|_{Z_q}$ for $|\alpha'| \leq N$.

Recall (5.6) and put $\nu h = \Gamma(u, u)$. According to the decomposition (4.21), we have the decomposition

$$\partial_x^\alpha \Psi[u, u](t) = \partial_x^\alpha D_0(t) * (\nu h) + \partial_x^\alpha D_1(t) * (\nu h) + \partial_x^\alpha D_2(t) * (\nu h). \quad (5.25)$$

We shall estimate term by term.

Estimate of $\partial_x^\alpha D_0(t) * (\nu h)$. By virtue of the bilinear symmetry of Γ , the Leibnitz rule gives

$$\partial_x^\alpha (\nu h) = \sum_{\alpha' \leq \alpha} C_{\alpha'}^\alpha \Gamma(\partial_x^{\alpha-\alpha'} u, \partial_x^{\alpha'} u) = 2\Gamma(\partial_x^\alpha u, u) + \sum_{0 < \alpha' < \alpha} = \nu h_1^\alpha + \nu h_2^\alpha. \quad (5.26)$$

Hence,

$$\partial_x^\alpha D_0(t) * (\nu h) = \Psi_0(h_1^\alpha) + \Psi_0(h_2^\alpha),$$

where Ψ_0 is as in (4.13). By virtue of Lemma 4.3 combined with Lemma 1.8, we get

$$\|\Psi_0(h_1^\alpha)(t)\|_{X_\beta} \leq 2c(\sigma)(1+t)^{-\sigma-\sigma_{q,0}} \sup_{t \geq 0} (1+t)^{\sigma+\sigma_{q,0}} \|\partial_x^\alpha u(t)\|_{X_\beta} \|u(t)\|_{X_\beta},$$

for any $\sigma \geq 0$. Here and hereafter $c(\sigma)$ denotes various positive constants which depend only on σ . In the above, $c(\sigma)$ stands for the constants C in Lemma 4.3. The estimate (5.14) then leads to

$$\|\Psi_0(h_1^\alpha)(t)\|_{X_\beta} \leq 2c(\sigma)a_1 U_0(1+t)^{-\sigma-\sigma_{q,0}} \|\partial_x^\alpha u\|_{\beta,\sigma}, \quad (5.27)$$

where U_0 and $\|\cdot\|_{\beta,\sigma}$ are as in (5.5) while a_1 is as in (5.14).

On the other hand, a similar computation leads to

$$\begin{aligned} \|\Psi_0(h_2^\alpha)(t)\|_{X_\beta} &\leq c_0 \sum_{0 < \alpha' < \alpha} C_{\alpha'}^\alpha (1+t)^{-\sigma_{q,\ell-\ell'}-\sigma_{q,\ell'}} \\ &\quad \times \sup_{t \geq 0} (1+t)^{\sigma_{q,\ell-\ell'}+\sigma_{q,\ell'}} \|\partial_x^{\alpha-\alpha'} u(t)\|_{X_\beta} \|\partial^{\alpha'} u(t)\|_{X_\beta}, \end{aligned}$$

where $\ell = |\alpha| = N$ and $\ell' = |\alpha'|$. Clearly, $\sigma_{q,\ell-\ell'} + \sigma_{q,\ell'} = \sigma_{q,N} + \sigma_{q,0} \geq \sigma_{q,N}$. Here, $c_0 = c(\sigma_{q,N})$. Hence by the induction hypothesis that (5.17) holds for $\ell \leq N-1$, we can conclude

$$\|\Psi_0(h_2^\alpha)(t)\|_{X_\beta} \leq c(1+t)^{-\sigma_{q,N}}. \quad (5.28)$$

Estimate of $\partial_x^\alpha D_1(t) * (\nu h)$. Introduce the splitting

$$\partial_x^\alpha D_1(t) * (\nu h) = \Phi_{11}^\alpha + \Phi_{12}^\alpha$$

with

$$\begin{aligned} \Phi_{11}^\alpha &= \int_0^{t/2} \partial_x^\alpha D_1(t-s) \Gamma[u(s), u(s)] ds, \\ \Phi_{12}^\alpha &= \int_{t/2}^t \partial_x^\alpha D_1(t-s) \Gamma[u(s), u(s)] ds. \end{aligned}$$

Use Theorem 4.10 (1) with $|\alpha| = N$ and $\alpha' = 0$ and Lemma 1.8 to deduce

$$\begin{aligned} \|\Phi_{11}^\alpha\|_{X_\beta} &\leq c_0 \int_0^{t/2} (1+t-s)^{-\sigma_{1,N+1}} \|\nu^{-1} \Gamma[u(s), u(s)]\|_{Z_1} ds \\ &\leq c \int_0^{t/2} (1+t-s)^{-\sigma_{1,N+1}} (1+s)^{-2\sigma_{q,0}} (1+s)^{2\sigma_{q,0}} \|u(s)\|_{Z_2}^2 ds \\ &\leq c \int_0^{t/2} (1+t-s)^{-\sigma_{1,N+1}} (1+s)^{-2\sigma_{q,0}} ds \|u\|^2 \\ &\leq c(1+t/2)^{-\sigma_{1,N+1}+\max(0,1-2\sigma_{q,0})}, \end{aligned} \quad (5.29)$$

where $\|\cdot\|$ is as in (5.9). Note that $\sigma_{1,N+1} - \max(0, 1 - 2\sigma_{q,0}) \geq \sigma_{q,N}$.

Use Theorem 4.10 (1) again, but this time with $\alpha = \alpha'$, $|\alpha| = N$, to deduce

$$\begin{aligned}\|\Phi_{12}^\alpha\|_{X_\beta} &\leq b_2 \int_{t/2}^t (1+t-s)^{-\sigma_{1,1}} (\|h_1^\alpha(s)\|_{Z_1} + \|h_2^\alpha(s)\|_{Z_1}) ds \\ &= b_2(J_1 + J_2),\end{aligned}$$

where the constant b_2 is independent of N because $m = |\alpha - \alpha'| = 0$ (see Theorem 4.10, (1)). By the aid of Lemma 1.8, since $h = \nu^{-1}\Gamma$, we have

$$\begin{aligned}J_1 &= 2 \int_{t/2}^t (1+t-s)^{-\sigma_{1,1}} \|\partial_x^\alpha u(s)\|_{Z_2} \|u(s)\|_{Z_2} ds \\ &\leq 2 \int_{t/2}^t (1+t-s)^{-\sigma_{1,1}} (1+s)^{-\sigma_{q,0}} \|\partial_x^\alpha u(s)\|_{Z_2} ds \|\cdot\|,\end{aligned}\tag{5.30}$$

where we used Theorem 5.3 (1). Then, for any $\sigma \geq 0$,

$$\begin{aligned}J_1 &\leq 2 \int_{t/2}^t (1+t-s)^{-\sigma_{1,1}} (1+s)^{-\sigma_{q,0}-\sigma} ds \|\partial_x^\alpha u\|_{\beta,\sigma} \|\cdot\| \\ &\leq 2a_1 U_0 \int_{t/2}^t (1+s/2)^{-\sigma_{q,0}-\sigma} (1+t-s)^{-\sigma_{1,1}} ds \|\partial_x^\alpha u\|_{\beta,\sigma} \\ &\leq 2a_1 U_0 c(\sigma) (1+t)^{-\sigma_{q,0}-\sigma} \|\partial_x^\alpha u\|_{\beta,\sigma},\end{aligned}\tag{5.31}$$

where the fact $\sigma_{1,1} > 1$ was used.

On the other hand, by induction hypothesis,

$$\begin{aligned}J_2 &\leq \sum_{0 < \alpha' < \alpha} C_{\alpha'}^\alpha \int_{t/2}^t (1+t-s)^{-\sigma_{1,1}} \|\partial_x^{(\alpha-\alpha')} u(s)\|_{Z_2} \|\partial_x^{\alpha'} u(s)\|_{Z_2} ds \\ &\leq c \sum_{m=1}^{N-1} \int_{t/2}^t (1+t-s)^{-\sigma_{1,1}} (1+s)^{-\sigma_{q,m}-\sigma_{q,N-m}} ds \\ &\leq c(1+t/2)^{-\sigma_{q,N}-\sigma_{q,0}} \int_{t/2}^t (1+t-s)^{-\sigma_{1,1}} ds \\ &\leq c(1+t)^{-\sigma_{q,N}}.\end{aligned}\tag{5.32}$$

In the last line, we used again $\sigma_{1,1} > 1$.

Estimate of $\partial_x^\alpha D_2(t) * (\nu h)$. Recall (5.26) and introduce the splitting

$$\partial_x^\alpha D_2(t) * (\nu h) = D_2(t) * (\nu h_1^\alpha) + D_2(t) * (\nu h_2^\alpha) = \Phi_{21}^\alpha + \Phi_{22}^\alpha,$$

By means of Theorem 4.10 (2) for $|\alpha| = N$, we get

$$\begin{aligned} \|\Phi_{21}^\alpha(t)\|_{X_\beta} &\leqslant 2b_3 \int_0^t e^{-\sigma_0(t-s)} \|h_1^\alpha(s)\|_{X_\beta} ds \\ &\leqslant 2b_5 \int_0^t e^{-\sigma_0(t-s)} \|\partial_x^\alpha u(s)\|_{X_\beta} \|u(s)\|_{X_\beta} ds \\ &\leqslant 2b_5 \int_0^t e^{-\sigma_0(t-s)} (1+s)^{-\sigma_{q,0}-\sigma} ds \|\partial_x^\alpha u\|_{\beta,\sigma} \|u\| \\ &\leqslant 2a_1 U_0 c(\sigma) (1+t)^{-\sigma_{q,0}-\sigma} \|\partial_x^\alpha u\|_{\beta,\sigma}, \end{aligned} \quad (5.33)$$

where $b_5 > 0$ is a constant independent of N and $\sigma > 0$. In the above, we have used

$$\begin{aligned} &\int_0^t e^{-\sigma_0(t-s)} (1+s)^{-\sigma_{q,0}-\sigma} ds \\ &\leqslant e^{-\sigma_0 t/2} \int_0^{t/2} (1+s)^{-\sigma_{q,0}} ds + (1+t/2)^{-\sigma_{q,0}-\sigma} \int_{t/2}^t e^{-\sigma_0(t-s)} ds \\ &\leqslant c(\sigma) (1+t)^{-\sigma_{q,0}-\sigma}. \end{aligned}$$

Finally, again by Theorem 4.10 (2) for $|\alpha| = |\alpha'| = N$, and by induction hypothesis for (5.17),

$$\begin{aligned} \|\Phi_{22}(t)\|_{X_\beta} &\leqslant \sum_{0<\alpha'<\alpha} C_{\alpha'}^\alpha \int_0^t e^{-\sigma_0(t-s)} \|\partial_x^{(\alpha-\alpha')} u(s)\|_{X_\beta} \|\partial_x^{\alpha'} u(s)\|_{X_\beta} ds \\ &\leqslant c \sum_{m=1}^{N-1} \int_0^t e^{-\sigma_0(t-s)} (1+s)^{-\sigma_{q,N-m}-\sigma_{q,m}} ds \\ &\leqslant c(1+t)^{-\sigma_{q,N}}. \end{aligned} \quad (5.34)$$

Apply all the estimates (5.27)–(5.34) to (5.25), and combine the resulting estimate with (5.24) in (5.23), to deduce

$$\|\Phi^\alpha[u]\|_{X_\beta} \leqslant c(1+t)^{-\sigma_{q,N}} + c(\sigma) a_0 a_1 (1+t)^{-(\sigma+\sigma_{q,0})} \|\partial_x^\alpha u\|_{\beta,\sigma}, \quad (5.35)$$

for $|\alpha| = N$ and $\sigma \geqslant 0$. Here, a_0, a_1 are as in Theorem 5.3 (1), and $c > 0$ depends on N . To simplify the notation, fix q, β as in Theorem 5.3 (2) and recall (5.5) to write

$$[[u]]_\sigma = \|\|u\|\|_{\beta,\sigma}.$$

Since u is a fixed point $u = \Phi[u]$, (5.35) implies

$$[[\partial_x^\alpha u]]_{\sigma+\sigma_{q,0}} \leqslant c + c(\sigma) a_0 a_1 [[\partial_x^\alpha u]]_\sigma \quad (5.36)$$

for $\sigma \geq 0$ such that $\sigma + \sigma_{q,0} \leq \sigma_{q,N}$. First, we put $\sigma = 0$ and choose the constant a_0 smaller if necessary so that $c(0)a_0a_1 < 1$ holds. Then, we get

$$[[\partial_x^\alpha u]]_0 \leq \frac{c}{1 - c(0)a_0a_1}.$$

On the other hand, since $q \in [1, 2)$ is assumed, we know $\sigma_{q,0} > 0$ and therefore, we can solve the recurrence inequality (5.36) and after a finite number of steps (actually, $[\sigma_{q,N}/\sigma_{q,0}] + 1$ steps) we get

$$[[\partial_x^\alpha u]]_{\sigma_{q,N}} \leq c_1 + c_2 [[\partial_x^\alpha u]]_0.$$

Combining these two estimates confirms the induction hypothesis for $\ell = N$, and the proof of Theorem 5.3 is now complete. \square

Remark 5.9. The proof of the statement in Remark 5.7 follows directly from (5.36). If $q = 2$, then $\sigma_{q,0} = 0$. Take $\sigma = \sigma_{2,N}$ and assume that a_0 is so small that $c(\sigma_{2,N})a_0a_1 < 1$ holds. Then, (5.36) gives

$$[[\partial_x^\alpha u]]_{\sigma_{2,N}} \leq \frac{c}{1 - c(\sigma_{2,N})a_0a_1}.$$

Proof of Theorem 5.4. First, we need the following properties of e^{tA} .

Lemma 5.10. (1) Assume

$$\partial_x^\alpha u_0 \in L^2, \quad |\alpha| \leq 1, \quad (5.37)$$

and put $v = e^{tA}u_0$. Then,

$$\partial_x^\alpha v \in C^0([0, \infty); L^2), \quad (1 + |\xi|)^{-1}v \in C^1([0, \infty); L^2),$$

and

$$\frac{\partial v}{\partial t} + \xi \cdot \nabla_x v + \nu(\xi)v = 0 \quad (5.38)$$

holds almost everywhere in $(0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n$.

(2) Assume

$$\nu^{(1-\delta)}h \in L^\infty(0, \infty; L^2), \quad (5.39)$$

for some $\delta \in [0, 1)$. Recall (4.13) and put $w = \Psi_0[h]$. Then,

$$w \in C^0([0, \infty); L^2). \quad (5.40)$$

(3) Assume further

$$\partial_x^\alpha h \in L^\infty(0, \infty; L^2), \quad |\alpha| \leq 1. \quad (5.41)$$

Then,

$$(1 + |\xi|)^{-2}\partial_t w \in L^\infty([0, \infty); L^2),$$

and

$$\frac{\partial w}{\partial t} + \xi \cdot \nabla_x w + \nu(\xi)w = \nu h \quad (5.42)$$

holds almost everywhere in $(0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n$.

Admitting this lemma for the time being, we prove the theorem. Let u be a mild solution of Theorem 5.1 and recall the integral equation (5.4):

$$u(t) = e^{tB}u_0 + \int_0^t e^{(t-s)B}\Gamma(u(s), u(s))ds = e^{tB}u_0 + \Psi[u, u]. \quad (5.43)$$

First, we prove the continuity (5.18) and (5.19). Thus, we shall consider (5.43) term by term. By the assumptions (5.13) and (5.15), $\partial_x^\alpha u_0 \in L^2$, $|\alpha| \leq \ell$ for some $\ell \geq 0$. Since e^{tB} is a C_0 semi-group on L^2 and since B commutes with ∂_x^α , we have

$$\partial_x^\alpha(e^{tB}u_0) = e^{tB}(\partial_x^\alpha u_0) \in C^0([0, \infty); L^2).$$

We shall use Duhamel's formula in the form

$$\begin{aligned} \partial_x^\alpha \Psi[u, u] &= e^{tB} * (\nu \partial_x^\alpha h) = e^{tA} * (\nu \partial_x^\alpha h) + e^{tA} * K e^{tB} * (\nu \partial_x^\alpha h) \\ &= \Psi_0[\partial_x^\alpha h] + \Psi_0[\partial_x^\alpha h_1], \end{aligned} \quad (5.44)$$

where we put $\nu h = \Gamma[u, u]$ and $\nu h_1 = K\Psi[u, u]$. In virtue of (5.14), (5.16) and (1.38), it holds that for $\delta \in [0, 1)$,

$$\nu^{1-\delta} \partial_x^\alpha h \in L^\infty(0, \infty; L^2), \quad (5.45)$$

while the proof of Theorem 5.1 given in the first part of the previous subsection shows that the estimates (5.14), (5.16), (5.17) are valid also for $\Psi[u, u]$ and thereby

$$\nu \partial_x^\alpha h_1 \in L^\infty(0, \infty; L^2). \quad (5.46)$$

In view of Lemma 5.10 (2), we have

$$\partial_x^\alpha \Psi[u, u] = \Psi_0[\partial_x^\alpha h] + \Psi_0[\partial_x^\alpha h_1] \in C^0([0, \infty); L^2),$$

and (5.18), (5.19) thus follow.

To discuss the differentiability of u with respect to t , suppose $\ell \geq 1$. Duhamel's formula gives

$$e^{tB}u_0 = e^{tA}u_0 + e^{tA} * K e^{tB}u_0 = v + v_1.$$

Evidently, Lemma 5.10 (1) applies to v while Lemma 5.10 (3) applies to v_1 since

$$\partial_x^\alpha \nu^{-1} K e^{tB}u_0 = \nu^{-1} K e^{tB} \partial_x^\alpha u_0 \in C^0([0, \infty); L^2).$$

This implies that Lemma 5.10(1) applies also to $e^{tB}u_0$ and in place of (5.38),

$$\frac{\partial(v + v_1)}{\partial t} + \xi \cdot \nabla_x(v + v_1) + \nu(\xi)(v + v_1) = Ke^{tB}u_0 = K(v + v_1)$$

holds almost everywhere in $(0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n$.

Further, the estimates (5.45) and (5.46) imply

$$\partial_x^\alpha h, \quad \partial_x^\alpha h_1 \in L^\infty([0, \infty); L^2).$$

Then, Duhamel's formula (5.44) combined with Lemma 5.10 (3) implies that $\Psi[u, u]$ also enjoys the conclusion of Lemma 5.10 (3).

Summarizing, we concluded that the mild solution u satisfies the Boltzmann equation

$$\frac{\partial u}{\partial t} + \xi \cdot \nabla_x u + \nu(\xi)u = Ku + \Gamma[u, u] \quad (5.47)$$

in L^2 and hence almost everywhere in $(0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n$. In view of (5.16), this assures for $|\alpha| \leq \ell - 1$,

$$\partial_x^\alpha u_t \in L^\infty(0, \infty; L_{\beta-1}^\infty),$$

which, in turn, implies

$$\partial_x^\alpha u \in C^0([0, \infty); L_{\beta-1}^\infty).$$

Plug this into (5.47) to see that

$$\partial_x^\alpha u_t \in C^0([0, \infty); L_{\beta-2}^\infty)$$

holds for $|\alpha| \leq \ell - 2$. Thus, u is a classical solution if $\ell \geq 2$. This completes the proof of Theorem 5.4. We close this subsection with the following proof.

Proof of Lemma 5.10. (1) Put $z = (1+|\xi|)^{-1}v$ and $z_0 = (1+|\xi|)^{-1}u_0$. Since e^{tA} commutes with $(1+|\xi|)^{-1}$, we have $z = e^{tA}z_0$. The assumption (5.37) implies $z_0 \in D(A)$ where $D(A)$ is the domain of definition (4.6) of A . Consequently, the semi-group theory (see e.g. [49]) says that

$$z = e^{tA}z_0 \in C^1([0, \infty); L^2), \quad z(t) \in D(A) \quad \text{for all } t \geq 0,$$

and

$$\frac{dz}{dt} = Az$$

holds in the space L^2 and hence almost everywhere in $(0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n$. Multiplying this equation by $(1+|\xi|)$ yields (5.38).

(2) Let $r > 0, t \geq 0$ and write

$$\begin{aligned} \Psi_0[h](t+r) - \Psi_0[h](t) &= \int_t^{t+r} e^{(t+r-s)A} \nu h(s) ds \\ &\quad + \int_0^t e^{(t-s)A} \nu^\delta (e^{rA} - I) \nu^{1-\delta} h(s) ds \equiv \psi_1 + \psi_2, \end{aligned}$$

where we used the same commutativity as mentioned above. By the same computation as for (4.14) and by the assumption (5.39), we get

$$\|\psi_1\|_{L^2} \leq \int_t^{t+r} (t+r-s)^{-\delta} e^{-\nu_* (t+r-s)/2} ds \sup_{t \geq 0} \|\nu^{(1-\delta)} h(t)\|_{L^2} \rightarrow 0,$$

and

$$\|\psi_2\|_{L^2} \leq \int_0^t (t-s)^{-\delta} e^{-\nu_* (t-s)/2} \| (e^{rA} - I) \nu^{1-\delta} h(s) \|_{L^2} ds \rightarrow 0,$$

as $r \rightarrow 0$. The last convergence is due to Lebesgue's dominated convergence theorem. This proves (5.40).

(3) Set $H = \nu H_0 = (1 + |\xi|)^{-2} \nu h$. The assumption (5.41) implies $H \in D(A)$ and

$$\|AH(s)\|_{L^2} + \left\| \frac{1}{r} (e^{rA} - I) H(s) \right\|_{L^2} \leq C(\|h\|_{L^2} + \|\nabla_x h(s)\|_{L^2}), \quad (5.48)$$

$$\left\| \frac{1}{r} (e^{rA} - I) H(s) - AH(s) \right\|_{L^2} \rightarrow 0 \quad (r \rightarrow +0), \quad (5.49)$$

both for each s . Now, we write

$$\begin{aligned} \frac{1}{r} \left(\Psi_0[H_0](t+r) - \Psi_0[H_0](t) \right) &= \frac{1}{r} \int_t^{t+r} e^{(t+r-s)A} H(s) ds \\ &\quad + \frac{1}{r} \int_0^t e^{(t-s)A} (e^{rA} - I) H(s) ds \equiv \chi_1 + \chi_2. \end{aligned}$$

Consider the decomposition

$$\chi_1 = \frac{1}{r} \int_0^r (e^{sA} - I) H(s) ds + \frac{1}{r} \int_0^r H(t+s) ds,$$

and take the limit as $r \rightarrow +0$. The first term goes to 0 in L^2 owing to (5.48), while the second term converges to $H(t)$ for almost all t . On the other hand, χ_2 converges to $e^{tA} * (AH)$ in L^2 owing to (5.48) and

(5.49), and by Lebesgue's theorem on the differentiation under integral sign. Clearly,

$$\chi_2 = \frac{1}{r}(e^{rA} - I) \int_0^t e^{(t-s)A} H(s) ds = \frac{1}{r}(e^{rA} - I)\Psi_0[H_0]$$

holds. This and the convergence of χ_1, χ_2 show that

$$\Psi_0[H_0] \in D(A), \quad \frac{d}{dt}\Psi_0[H_0] \in L^\infty(0, \infty; L^2),$$

and that

$$\frac{d}{dt}\Psi_0[H_0] = \nu H_0(t) + A\Psi_0[H_0]$$

holds in the space L^2 for almost all t and hence almost all $(0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n$. Multiplying this equation by $(1 + |\xi|)^2$ yields (5.42). Now, the proof of the lemma is complete. \square

5.3 External force, revisited

The decay estimates on the solution operator with forcing can be used to prove global existence of solutions to the Boltzmann equation with external force and source. However, since it is given in some Sobolev space with regularity requirement which is different from the case without forcing given in the previous subsections, the global existence is also given in the corresponding Sobolev space. Indeed, the theorem can be stated as follows.

Consider the global existence and decay rates of the solution to the Cauchy problem for the nonlinear Boltzmann equation (1.3) in the external force F and with the external source S . Setting as usual $f = \mathbf{M} + \mathbf{M}^{1/2}u$, we shall consider the Cauchy problem (4.37):

$$\partial_t u + \xi \cdot \nabla_x u + F \cdot \nabla_\xi u - \frac{1}{2}\xi \cdot Fu = \mathbf{L}u + \Gamma(u) + \tilde{S}, \quad (5.50)$$

$$u(t, x, \xi)|_{t=0} = u_0(x, \xi), \quad (5.51)$$

where $u = u(t, x, \xi)$, $(t, x, \xi) \in \mathbb{R}^+ \times \mathbb{R}^n \times \mathbb{R}^n$, and

$$\tilde{S} = \mathbf{M}^{-\frac{1}{2}}S + \mathbf{M}^{\frac{1}{2}}\xi \cdot F. \quad (5.52)$$

Recall the norms $[[\cdot]]_{m,k}$ defined by (4.40) and (4.41).

Theorem 5.11. *Suppose that*

- (B1) *the integers $n \geq 3$, $\ell \geq [n/2] + 2$;*
- (B2) *the functions $F = F(t, x)$, $S = S(t, x, \xi)$ and $u_0 = u_0(x, \xi)$ satisfy*

$$F \in C_b^i(\mathbb{R}_t^+; H^{\ell-i}(\mathbb{R}_x^n)), \quad i = 0, 1, \quad S \in C_b^0(\mathbb{R}_t^+; H^\ell(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)),$$

$$u_0 \in H^\ell(\mathbb{R}_x^n \times \mathbb{R}_\xi^n).$$

(B3) there are constants $\delta > 0$, $k \geq 1$ and $\kappa > 1$ such that F and u_0 satisfy

$$\begin{aligned} \sum_{0 \leq |\beta| \leq \ell} \left\| (1 + |x|) \partial_x^\beta F(t, x) \right\|_{L_{t,x}^\infty} + \sum_{0 \leq |\beta| \leq \ell-1} \left\| (1 + |x|) \partial_t \partial_x^\beta F(t, x) \right\|_{L_{t,x}^\infty} \\ + \left\| |x| F(t, x) \right\|_{L_t^\infty(L_x^2)} \leq \delta, \end{aligned} \quad (5.53)$$

$$[[u_0]]_{0,k+1/2} + \|u_0\|_{Z_1} \leq \delta. \quad (5.54)$$

Moreover, F and S decay in time like

$$\|F(t)\|_{H_x^\ell \cap L_x^1} \leq \delta(1+t)^{-\kappa}, \quad (5.55)$$

$$[[\mathbf{M}^{-1/2} S(t)]]_{0,k-1/2} + \left\| \mathbf{M}^{-1/2} S(t) \right\|_{Z_1} \leq \delta(1+t)^{-\kappa}. \quad (5.56)$$

Then there are constants $\delta_1 > 0$ and $C_1 > 0$ such that for any $\delta \leq \delta_1$, the Cauchy problem (5.50)–(5.51) corresponding to (4.36) has a unique global classical solution

$$u \in C_b^i(\mathbb{R}_t^+; H^{\ell-i}(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)), \quad i = 0, 1, \quad (5.57)$$

which satisfies

$$\sup_{t \geq 0} (1+t)^{2\kappa_0} [[u(t)]]_{0,k}^2 + \int_0^\infty [[u(s)]]_{0,k+1/2}^2 ds \leq C_1^2, \quad (5.58)$$

where C_1 can be also taken as $C_1 = C'_1 \delta$ for another constant C'_1 independent of δ , and κ_0 is given by

$$\begin{cases} \frac{1}{2} < \kappa_0 < \kappa - \frac{1}{2}, & \text{if } \sigma_{1,0} \geq \kappa - \frac{1}{2}, \\ \kappa_0 = \sigma_{1,0}, & \text{if } \sigma_{1,0} < \kappa - \frac{1}{2}. \end{cases} \quad (5.59)$$

Furthermore, it holds that

$$\sum_{0 \leq |\alpha| \leq \ell-1} \left\| \nu^{k-1} \partial_t \partial_{x,\xi}^\alpha u(t) \right\| \leq C \delta (1+t)^{-\kappa_0}, \quad (5.60)$$

for some constant C .

In order to prove the above theorem, we introduce a function space $\mathbb{S}(C_1)$ given by

$$\mathbb{S}(C_1) = \left\{ u = u(t, x, \xi) \mid u \in C_b^0(\mathbb{R}_t^+; H^\ell(\mathbb{R}_x^n \times \mathbb{R}_\xi^n)), |||u|||_{k,\kappa_0} \leq C_1 \right\},$$

where $C_1 > 0$ is some constant to be determined, and the norm $\|\cdot\|_{k,\kappa_0}$ is defined by

$$\|u\|_{k,\kappa_0}^2 = \sup_{t \geq 0} (1+t)^{2\kappa_0} [[u(t)]]_{0,k}^2 + \int_0^\infty [[u(s)]]_{0,k+1/2}^2 ds.$$

Clearly, $\mathbb{S}(C_1)$ is a complete metric space with the metric induced by the norm $\|\cdot\|_{k,\kappa_0}$. Under some conditions, the solution to (5.50)–(5.51) will be obtained by applying the contraction mapping theorem to find a fixed point in $\mathbb{S}(C_1)$ for some nonlinear mapping Ψ , where Ψ is defined by

$$\Psi(u) = U(t,0)u_0 + \int_0^t U(t,s)\{\Gamma(u(s),u(s)) + \tilde{S}(s)\}ds. \quad (5.61)$$

Thus one has to estimate the time integral in (5.61) in terms of the norm $\|\cdot\|_{k,\kappa_0}$. For this, in what follows, given a function $\phi = \phi(t, x, \xi)$, we will first consider the estimate on the general time integral

$$(\mathbf{T}\phi)(t, x, \xi) = \int_0^t U(t, s)\phi(s, x, \xi)ds.$$

This time integral will be written into two parts again by decomposing the linearized Boltzmann operator and by using the Duhamel's principle. In fact, define the solution operator $U_1(t, s)$ for any $0 \leq s \leq t$ in the sense that for any $v_0 = v_0(x, \xi)$, $v = v(t, x, \xi) = U_1(t, s)v_0$ denotes the solution to the following initial value problem:

$$\begin{aligned} \partial_t v + \nu v + \xi \cdot \nabla_x v + F \cdot \nabla_\xi v - \frac{1}{2}\xi \cdot Fv &= 0, \\ v(t, x, \xi)|_{t=s} &= v_0(x, \xi). \end{aligned}$$

Note that $\mathbf{L} = -\nu + K$. Then again by the Duhamel's formula, the solution operator $U(t, s)$ can be rewritten as

$$U(t, s) = U_1(t, s) + U_2(t, s), \quad 0 \leq s \leq t,$$

where

$$U_2(t, s) = \int_s^t U(t, \tau)KU_1(\tau, s)d\tau.$$

We further define

$$(\mathbf{T}_j\phi)(t, x, \xi) = \int_0^t U_j(t, s)\phi(s, x, \xi)ds, \quad j = 1, 2.$$

Then

$$\mathbf{T}\phi = \mathbf{T}_1\phi + \mathbf{T}_2\phi,$$

and the following estimates hold.

Lemma 5.12. *Under the assumption (5.53), if $\delta > 0$ is small enough, then we have*

$$\begin{aligned} (1+t)^{2m}[[\mathbf{T}_1\phi(t)]]_{0,k}^2 + \int_0^t (1+s)^{2m}[[\mathbf{T}_1\phi(s)]]_{0,k+1/2}^2 ds \\ \leq C \int_0^t (1+s)^{2m}[[\phi(s)]]_{0,k-1/2}^2 ds, \end{aligned} \quad (5.62)$$

for any $m \geq 0$ and any k , and

$$\begin{aligned} (1+t)^{2m}\|\mathbf{T}_1\phi(t)\|_{Z_1}^2 + \int_0^t (1+s)^{2m}\|\mathbf{T}_1\phi(s)\|_{Z_1}^2 ds \\ \leq C \int_0^t (1+s)^{2m} \left([[\phi(s)]]_{0,k-1/2}^2 + \|\phi(s)\|_{Z_1}^2 \right) ds, \end{aligned} \quad (5.63)$$

for any $m \geq 0$ and any $k \geq 1/2$.

Proof. For simplicity, write $w = \mathbf{T}_1\phi$, which by the definitions of \mathbf{T}_1 and $U_1(t, s)$, satisfies the following Cauchy problem with zero initial data:

$$\partial_t w + \nu w + \xi \cdot \nabla_x w + F \cdot \nabla_\xi w - \frac{1}{2} \xi \cdot F w = \phi, \quad (5.64)$$

$$w(t, x, \xi)|_{t=0} = 0. \quad (5.65)$$

By applying energy method, we have the following energy inequality by tedious calculations which are omitted here,

$$\frac{d}{dt} J_{0,k}[w(t)] + c J_{0,k+1/2}[w(t)] \leq C [[\phi(t)]]_{0,k-1/2}^2, \quad (5.66)$$

for any k , where the nonlinear functional $J_{0,k}[\cdot]$ satisfies

$$J_{0,k}[w(t)] \sim [[w(t)]]_{0,k}. \quad (5.67)$$

After integration, (5.66) implies

$$J_{0,k}[w(t)] + \int_0^t J_{0,k+1/2}[w(s)] ds \leq C \int_0^t [[\phi(s)]]_{0,k-1/2}^2 ds. \quad (5.68)$$

On the other hand, multiplying (5.66) by $(1+t)^{2m}$ with $m \geq 0$ and further integrating it give

$$\begin{aligned} (1+t)^{2m} J_{0,k}[w(t)] + c \int_0^t (1+s)^{2m} J_{0,k+1/2}[w(s)] ds \\ \leq 2m \int_0^t (1+s)^{2m-1} J_{0,k}[w(s)] ds + C \int_0^t (1+s)^{2m} [[\phi(s)]]_{0,k-1/2}^2 ds \\ \leq \frac{c}{2} \int_0^t (1+s)^{2m} J_{0,k+1/2}[w(s)] ds + C \int_0^t J_{0,k+1/2}[w(s)] ds \\ + C \int_0^t (1+s)^{2m} [[\phi(s)]]_{0,k-1/2}^2 ds. \end{aligned} \quad (5.69)$$

Then (5.69) together with (5.67) and (5.68) yields (5.62).

Next consider the estimate (5.63) in the norm $\|\cdot\|_{Z_1}$. It can be based on the explicit form for the solution w from (5.64)–(5.65):

$$w(t, x, \xi) = \int_0^t e^{-\nu(\xi)(t-s)} \{F \cdot \nabla_\xi w - \xi/2 \cdot Fw + \phi\} (s, x - (t-s)\xi, \xi) ds,$$

which implies

$$\begin{aligned} \|w(t, \xi)\|_{L^1(\mathbb{R}_x^n)} &\leq C \int_0^t e^{-\nu_0(t-s)} (\|\nabla_\xi \nabla_x w(s, \xi)\|_{L^2(\mathbb{R}_x^n)} \\ &\quad + \nu \|\nabla_x w(s, \xi)\|_{L^2(\mathbb{R}_x^n)} + \|\phi(s, \xi)\|_{L^1(\mathbb{R}_x^n)}) ds. \end{aligned}$$

Further taking the norm $\|\cdot\|_{L^2(\mathbb{R}_\xi^n)}$ gives

$$\|w(t)\|_{Z_1} \leq C \int_0^t e^{-\nu_0(t-s)} G(s) ds, \quad (5.70)$$

where for simplicity, we have used the notion

$$G(s) = \|\nabla_\xi \nabla_x w(s)\|_{Z_2} + \|\nu \nabla_x w(s)\|_{Z_2} + \|\phi(s)\|_{Z_1}. \quad (5.71)$$

From (5.70), we claim that for any $m \geq 0$,

$$(1+t)^{2m} \|w(t)\|_{Z_1}^2 + \int_0^t (1+s)^{2m} \|w(s)\|_{Z_1}^2 ds \leq C \int_0^t (1+s)^{2m} G(s)^2 ds. \quad (5.72)$$

In fact, by the Hölder inequality, it is straightforward to have from (5.70) that

$$\begin{aligned} \|w(t)\|_{Z_1}^2 &\leq C \int_0^t e^{-2\nu_0(t-s)} (1+s)^{-2m} ds \int_0^t (1+s)^{2m} G(s)^2 ds \\ &\leq C(1+t)^{-2m} \int_0^t (1+s)^{2m} G(s)^2 ds. \end{aligned} \quad (5.73)$$

On the other hand, again by (5.70), one has

$$\int_0^t (1+s)^{2m} \|w(s)\|_{Z_1}^2 ds \leq \int_0^t (1+s)^{2m} \left[\int_0^s e^{-\nu_0(s-\tau)} G(\tau) d\tau \right]^2 ds. \quad (5.74)$$

By the Schwarz inequality, it holds that

$$\begin{aligned} &\left[\int_0^s e^{-\nu_0(s-\tau)} G(\tau) d\tau \right]^2 \\ &\leq \int_0^s e^{-\nu_0(s-\tau)} (1+\tau)^{-2m} d\tau \int_0^s e^{-\nu_0(s-\tau)} (1+\tau)^{2m} G(\tau)^2 d\tau \\ &\leq C(1+s)^{-2m} \int_0^s e^{-\nu_0(s-\tau)} (1+\tau)^{2m} G(\tau)^2 d\tau, \end{aligned}$$

which together with (5.74) gives

$$\begin{aligned} \int_0^t (1+s)^{2m} \|w(s)\|_{Z_1}^2 ds &\leq C \int_0^t \int_0^s e^{-\nu_0(s-\tau)} (1+\tau)^{2m} G(\tau)^2 d\tau ds \\ &= C \int_0^t d\tau (1+\tau)^{2m} G(\tau)^2 \int_\tau^t e^{-\nu_0(s-\tau)} ds \\ &\leq C \int_0^t (1+\tau)^{2m} G(\tau)^2 d\tau. \end{aligned} \quad (5.75)$$

Thus (5.72) follows from (5.73) and (5.75). Furthermore, notice from (5.71) and $k \geq 1/2$ that

$$\begin{aligned} G(s)^2 &\leq C (\|\nabla_\xi \nabla_x w(s)\|^2 + \|\nu \nabla_x w(s)\|^2 + \|\phi(s)\|_{Z_1}^2) \\ &\leq C ([w(t)]_{0,k+1/2}^2 + \|\phi(s)\|_{Z_1}^2), \end{aligned}$$

by (5.62), implies

$$\begin{aligned} &\int_0^t (1+s)^{2m} G(s)^2 ds \\ &\leq C \int_0^t (1+s)^{2m} ([w(t)]_{0,k+1/2}^2 + \|\phi(s)\|_{Z_1}^2) ds \\ &\leq C \int_0^t (1+s)^{2m} ([\phi(t)]_{0,k-1/2}^2 + \|\phi(s)\|_{Z_1}^2) ds. \end{aligned} \quad (5.76)$$

With the notion $w = \mathbf{T}_1 \phi$, combining (5.72) and (5.76) leads to (5.63). This completes the proof of the lemma. \square

Lemma 5.13. *Under the assumption (5.53), if $\delta > 0$ is small enough, then one has*

$$\begin{aligned} &(1+t)^{2m} [[\mathbf{T}_2 \phi(t)]]_{0,k}^2 + \int_0^t [[\mathbf{T}_2 \phi(s)]]_{0,k+1/2}^2 ds \\ &\leq C \int_0^t (1+s)^{2m} ([\phi(s)]_{0,k-1/2}^2 + \|\phi(s)\|_{Z_1}^2) ds, \end{aligned} \quad (5.77)$$

for any $1/2 < m \leq \sigma_{1,0}$ and any $k \geq 1$.

Proof. First fix some m and k with $1/2 < m \leq \sigma_{1,0}$ and $k \geq 1$. Set $z = \mathbf{T}_2 \phi$ for simplicity. By the definitions of \mathbf{T}_i and $U_i(t, s)$, $i = 1, 2$, note that

$$z(t) = \mathbf{T}_2 \phi(t) = \int_0^t U_2(t, s) \phi(s) ds = \int_0^t U(t, s) K \mathbf{T}_1 \phi(s) ds.$$

Then by Theorem 4.14 and Lemma 5.12, it holds that

$$\begin{aligned}
& [[z(t)]]_{0,k}^2 \\
& \leq C \left| \int_0^t (1+t-s)^{-\sigma_{1,0}} ([[K\mathbf{T}_1\phi(s)]]_{0,k} + \|K\mathbf{T}_1\phi(s)\|_{Z_1}) ds \right|^2 \\
& \leq C \left| \int_0^t (1+t-s)^{-\sigma_{1,0}} ([[T_1\phi(s)]]_{0,k-1} + \|T_1\phi(s)\|_{Z_1}) ds \right|^2 \\
& \leq C \int_0^t (1+t-s)^{-2\sigma_{1,0}} (1+s)^{-2m} ds \\
& \quad \times \int_0^t (1+s)^{2m} \left([[T_1\phi(s)]]_{0,k+1/2}^2 + \|T_1\phi(s)\|_{Z_1}^2 \right)^2 ds \\
& \leq C(1+t)^{-2m} \int_0^t (1+s)^{2m} \left([[\phi(s)]]_{0,k-1/2}^2 + \|\phi(s)\|_{Z_1}^2 \right) ds. \tag{5.78}
\end{aligned}$$

On the other hand, $z = z(t, x, \xi)$ is the solution to the following initial value problem with zero initial data:

$$\begin{aligned}
& \partial_t z + \nu z + \xi \cdot \nabla_x z + F \cdot \nabla_\xi z - \frac{1}{2}\xi \cdot Fz = Kz + KT_1\phi, \\
& z(t, x, \xi)|_{t=0} = 0.
\end{aligned}$$

This means that

$$z = T_1(Kz + KT_1\phi).$$

Use (5.62) with $m = 0$ to deduce

$$\begin{aligned}
\int_0^t [[z(s)]]_{0,k+1/2}^2 ds & \leq C \int_0^t [[Kz + KT_1\phi]]_{0,k-1/2}^2 ds \\
& \leq C \int_0^t [[z(s)]]_{0,k-3/2}^2 ds + C \int_0^t [[T_1\phi(s)]]_{0,k-3/2}^2 ds,
\end{aligned}$$

where further, it holds from (5.78) that

$$\begin{aligned}
& \int_0^t [[z(s)]]_{0,k-3/2}^2 ds \leq \int_0^t [[z(s)]]_{0,k}^2 ds \\
& \leq C \int_0^t (1+s)^{-2m} ds \sup_{0 \leq s \leq t} \int_0^s (1+\tau)^{2m} \left([[\phi(\tau)]]_{0,k-1/2}^2 + \|\phi(\tau)\|_{Z_1}^2 \right) d\tau \\
& \leq C \int_0^t (1+\tau)^{2m} \left([[\phi(\tau)]]_{0,k-1/2}^2 + \|\phi(\tau)\|_{Z_1}^2 \right) d\tau,
\end{aligned}$$

and again from (5.62) with $m = 0$ that

$$\int_0^t [[T_1\phi(s)]]_{0,k-3/2}^2 ds \leq \int_0^t [[T_1\phi(s)]]_{0,k+1/2}^2 ds \leq C \int_0^t [[\phi(s)]]_{0,k-1/2}^2 ds.$$

Then,

$$\int_0^t [[z(s)]]_{0,k+1/2}^2 ds \leq C \int_0^t (1+s)^{2m} \left([[\phi(s)]]_{0,k-1/2}^2 + \|\phi(s)\|_{Z_1}^2 \right) ds. \quad (5.79)$$

Thus (5.77) follows from (5.78) and (5.79). This completes the proof of the lemma. \square

Corollary 5.14. *Under the assumption (5.53), if $\delta > 0$ is small enough, then one has*

$$(1+t)^{2m} [[\mathbf{T}\phi(t)]]_{0,k}^2 + \int_0^t [[\mathbf{T}\phi(s)]]_{0,k+1/2}^2 ds \leq C \int_0^t (1+s)^{2m} \left([[\phi(s)]]_{0,k-1/2}^2 + \|\phi(s)\|_{Z_1}^2 \right) ds,$$

for any $1/2 < m \leq \sigma_{1,0}$ and any $k \geq 1$.

Before giving the proof of the global existence, let us prove the following lemma on the nonlinear collision operator in the norm $[[\cdot]]_{0,k}$.

Lemma 5.15. *Let $k \geq 0$ and $k_0 \leq 1$. Suppose that $\ell \geq [n/2] + 2$. Then for any $u = u(x, \xi)$ and $v = v(x, \xi)$, it holds that*

$$[[\Gamma(u, v)]]_{0,k-k_0} \leq C([[u]]_{0,k+1-k_0} [[v]]_{0,k} + [[u]]_{0,k} [[v]]_{0,k+1-k_0}),$$

where C is some constant.

Proof. Write

$$\Gamma(u, v) = \frac{1}{2} \{ \Gamma_1(u, v) + \Gamma_1(v, u) - \Gamma_2(u, v) - \Gamma_2(v, u) \},$$

with

$$\begin{aligned} \Gamma_1(u, v) &= \int_{\mathbb{R}^n \times S^{n-1}} |(\xi - \xi_*) \cdot \omega| \mathbf{M}_*^{1/2} u(\xi') v(\xi'_*) d\xi_* d\omega, \\ \Gamma_2(u, v) &= \int_{\mathbb{R}^n \times S^{n-1}} |(\xi - \xi_*) \cdot \omega| \mathbf{M}_*^{1/2} u(\xi) v(\xi_*) d\xi_* d\omega. \end{aligned}$$

It is obvious that the lemma holds if it does for each Γ_j , $j = 1, 2$.

First consider Γ_1 . As in [42], after taking change of variable $z = \xi - \xi_*$, Γ_1 can be rewritten as

$$\Gamma_1(u, v)(\xi) = \int_{\mathbb{R}^n \times S^{n-1}} |z \cdot \omega| \mathbf{M}^{1/2} (\xi - z) u(\xi') v(z') dz d\omega, \quad (5.80)$$

where

$$\xi' = \xi - z_{\parallel}, \quad z' = \xi - z_{\perp},$$

with $z_{\parallel} = (z \cdot \omega)\omega$, $z_{\perp} = z - z_{\parallel}$. Applying $\partial_{x,\xi}^{\alpha} = \partial_x^{\beta}\partial_{\xi}^{\gamma}$ with $0 \leq |\alpha| \leq \ell$ and $\alpha = \beta + \gamma$ to (5.80) yields

$$\begin{aligned} & \partial_{x,\xi}^{\alpha} \Gamma_1(u, v)(\xi) \\ &= \sum_{\beta_1+\beta_2=\beta} C_{\beta_1}^{\beta} \partial_{\xi}^{\gamma} \int_{\mathbb{R}^n \times S^{n-1}} |z \cdot \omega| \mathbf{M}^{1/2}(\xi - z) (\partial_x^{\beta_1} u)(\xi') (\partial_x^{\beta_2} v)(z') dz d\omega \\ &= \sum_{\substack{\beta_1+\beta_2=\beta \\ \gamma_1+\gamma_{21}+\gamma_{22}=\gamma}} C_{\beta_1}^{\beta} C_{\gamma_1}^{\gamma} C_{\gamma_{21}}^{\gamma-1} \\ & \quad \int_{\mathbb{R}^n \times S^{n-1}} |z \cdot \omega| \partial_{\xi}^{\gamma_1} \mathbf{M}^{1/2}(\xi - z) (\partial_x^{\beta_1} \partial_{\xi}^{\gamma_{21}} u)(\xi') (\partial_x^{\beta_2} \partial_{\xi}^{\gamma_{22}} v)(z') dz d\omega. \end{aligned}$$

Notice that for any γ_1 ,

$$|\partial_{\xi}^{\gamma_1} \mathbf{M}^{1/2}(\xi - z)| \leq C \mathbf{M}^{1/4}(\xi - z).$$

Then

$$\begin{aligned} |\partial_{x,\xi}^{\alpha} \Gamma_1(u, v)(\xi)| &\leq C \sum_{\alpha_1+\alpha_2 \leq \alpha} \int_{\mathbb{R}^n \times S^{n-1}} |z \cdot \omega| \mathbf{M}^{1/4}(\xi - z) \\ & \quad \times |\partial_{x,\xi}^{\alpha_1} u(\xi')| |\partial_{x,\xi}^{\alpha_2} v(z')| dz d\omega. \end{aligned} \quad (5.81)$$

Without loss of generality, suppose $|\alpha_1| \leq |\alpha|/2$ in (5.81). Then by integrating (5.81) over \mathbb{R}_x^n with respect to the space variable and using the Sobolev inequality, one has

$$\|\partial_{x,\xi}^{\alpha} \Gamma_1(u, v)(\xi)\|_{L_x^2} \leq C \sum_{|\alpha_1| \leq |\alpha|/2} \Gamma_{\alpha_1}(\xi),$$

where

$$\begin{aligned} \Gamma_{\alpha_1}(\xi) &= \int_{\mathbb{R}^n \times S^{n-1}} |z \cdot \omega| \mathbf{M}^{1/4}(\xi - z) \\ & \quad \times \| |\nabla_x|^{[(n-1)/2]+1} \partial_{x,\xi}^{\alpha_1} u(\xi') \|_{H_x^1} \| \partial_{x,\xi}^{\alpha_2} v(z') \|_{L_x^2} dz d\omega. \end{aligned}$$

Noting that for any $k \geq 0$,

$$\nu^k(\xi') \nu^k(z') = \nu^k(\xi - z_{\parallel}) \nu^k(\xi - z_{\perp}) \geq C \nu^k(\xi), \quad (5.82)$$

where the constant $C > 0$, then for each α_1 , one has

$$\begin{aligned} & \nu^k \Gamma_{\alpha_1}(\xi) \\ & \leq C \int_{\mathbb{R}^n \times S^{n-1}} |z \cdot \omega| \mathbf{M}^{1/4}(\xi - z) \\ & \quad \times \|\nu^k |\nabla_x|^{[(n-1)/2]+1} \partial_{x,\xi}^{\alpha_1} u(\xi')\|_{H_x^1} \|\nu^k \partial_{x,\xi}^{\alpha_2} v(z')\|_{L_x^2} dz d\omega \\ & \leq C \nu(\xi) \left\{ \int_{\mathbb{R}^n \times S^{n-1}} \left[\|\nu^k |\nabla_x|^{[(n-1)/2]+1} \partial_{x,\xi}^{\alpha_1} u(\xi')\|_{H_x^1} \right. \right. \\ & \quad \times \left. \left. \|\nu^k \partial_{x,\xi}^{\alpha_2} v(z')\|_{L_x^2} \right]^2 dz d\omega \right\}^{1/2}. \end{aligned}$$

Taking further integration over \mathbb{R}_ξ^n with respect to the velocity variable gives

$$\begin{aligned} & \|\nu^{k-k_0} \Gamma_{\alpha_1}\|_{L_\xi^2}^2 \\ & \leq C \int_{\mathbb{R}^n \times S^{n-1}} \nu^{2-2k_0}(\xi) \|\nu^k \nabla_x \partial_{x,\xi}^{\alpha_1} u(\xi')\|_{H_x^1}^2 \|\nu^k \partial_{x,\xi}^{\alpha_2} v(z')\|_{L_x^2}^2 d\xi dz d\omega \\ & \leq C \int_{\mathbb{R}^n \times S^{n-1}} [\nu^{2-2k_0}(\xi') + \nu^{2-2k_0}(z')] \\ & \quad \times \|\nu^k \nabla_x \partial_{x,\xi}^{\alpha_1} u(\xi')\|_{H_x^1}^2 \|\nu^k \partial_{x,\xi}^{\alpha_2} v(z')\|_{L_x^2}^2 d\xi' dz' d\omega, \end{aligned}$$

where we have used the inequality (5.82) since $2 - 2k_0 \geq 0$ and taken change of variables $(\xi, z) \rightarrow (\xi', z')$, whose Jacobian is unity. Hence

$$\|\nu^{k-k_0} \Gamma_{\alpha_1}\|_{L_\xi^2}^2 \leq C ([u]_{0,k+1-k_0}^2 [v]_{0,k}^2 + [u]_{0,k}^2 [v]_{0,k+1-k_0}^2). \quad (5.83)$$

Thus combining (5.3) and (5.83) implies that Lemma 5.15 holds for Γ_1 .

Finally it is more straightforward to carry out the estimates on $\Gamma_2(u, v)$ in a similar way. The details are omitted. This completes the proof of the lemma. \square

Now we are ready to prove the global existence of the solution to the Cauchy problem for the nonlinear Boltzmann equation with external forcing and source.

Proof of Theorem 5.11. First we prove that there is a constant $C_1 > 0$ such that Ψ is a contraction mapping from $\mathbb{S}(C_1)$ to itself, and thus it has a fixed point in $\mathbb{S}(C_1)$ which is a unique solution to the Cauchy problem (5.50)–(5.51). For this purpose, we start with a claim that there is a constant C such that for any $u, v \in \mathbb{S}(C_1)$,

$$|||\Psi(u)|||_{k,\kappa_0} \leq C\delta + C|||u|||_{k,\kappa_0}^2, \quad (5.84)$$

$$|||\Psi(u) - \Psi(v)|||_{k,\kappa_0} \leq C|||u + v|||_{k,\kappa_0} |||u - v|||_{k,\kappa_0}. \quad (5.85)$$

In fact, recall the definition (5.61) of Ψ , and then it is straightforward to compute

$$\begin{aligned} & \|U(t, 0)u_0\|_{k, \kappa_0}^2 \\ & \leq \sup_{t \geq 0} (1+t)^{2\kappa_0} [[U(t, 0)u_0]]_{0, k}^2 + \int_0^\infty [[U(s, 0)u_0]]_{0, k+1/2}^2 ds \\ & \leq C \sup_{t \geq 0} (1+t)^{2\kappa_0 - 2\sigma_{1,0}} [[u_0]]_{0, k}^2 + C \int_0^\infty (1+s)^{-2\sigma_{1,0}} ds [[u_0]]_{0, k+1/2}^2 \\ & \leq C [[u_0]]_{0, k+1/2}^2 \leq C\delta^2, \end{aligned} \quad (5.86)$$

where we used (5.54), and the inequalities $\kappa_0 \leq \sigma_{1,0}$ and $2\sigma_{1,0} > 1$ since $n \geq 3$. Furthermore, by noticing from (5.59) and $n \geq 3$ that $1/2 < \kappa_0 \leq \sigma_{1,0}$, we can apply Corollary 5.14 with $m = \kappa_0$ to obtain

$$\begin{aligned} & \left\| \left\| \int_0^t U(t, s)\Gamma(u(s), u(s))ds \right\| \right\|_k^2 \\ & \leq C \int_0^\infty (1+s)^{2\kappa_0} \left([[\Gamma(u(s), u(s))]_{0, k-1/2}^2 + \|\Gamma(u(s), u(s))\|_{Z_1}^2 \right) ds \\ & \leq C \int_0^\infty (1+s)^{2\kappa_0} [[u(s)]]_{0, k+1/2}^2 [[u(s)]]_{0, k}^2 ds \\ & \leq C \int_0^\infty [[u(s)]]_{0, k+1/2}^2 ds \sup_{s \geq 0} (1+s)^{2\kappa_0} [[u(s)]]_{0, k}^2 \\ & \leq C \|u\|_{k, \kappa_0}^2, \end{aligned} \quad (5.87)$$

where Lemma 5.15 was used. Furthermore, (5.55) and (5.56) together with (5.52) imply

$$[[\tilde{S}(s)]]_{0, k-1/2} + \|\tilde{S}(s)\|_{Z_1} \leq C\delta(1+s)^{-\kappa}.$$

Similarly applying Corollary 5.14 with $m = \kappa_0$ yields

$$\begin{aligned} & \left\| \left\| \int_0^t U(t, s)\tilde{S}(s)ds \right\| \right\|_k^2 \\ & \leq C \int_0^\infty (1+s)^{2\kappa_0} \left([[\tilde{S}(s)]]_{0, k-1/2}^2 + \|\tilde{S}(s)\|_{Z_1}^2 \right) ds \\ & \leq C\delta^2 \int_0^\infty (1+s)^{2\kappa_0 - 2\kappa} ds \leq C\delta^2, \end{aligned} \quad (5.88)$$

where by (5.59), $\kappa_0 < \kappa - 1/2$ was used. Thus by (5.61), combining (5.86), (5.87) and (5.88) proves (5.84). For (5.85), notice that since Γ is bilinear,

$$\Gamma(u, u) - \Gamma(v, v) = \Gamma(u+v, u-v).$$

Then it holds that

$$\Psi(u) - \Psi(v) = \int_0^t U(t, s)\Gamma(u + v, u - v)(s)ds,$$

which similar to the proof of (5.87), implies (5.85).

Now suppose $u, v \in \mathbb{S}(C_1)$. Then based on (5.84) and (5.85), it is straightforward to show that

$$\Psi(u), \Psi(v) \in C_b^0(\mathbb{R}_t^+; H^\ell(\mathbb{R}_x^n)),$$

with estimates

$$\begin{aligned} |||\Psi(u)|||_{k, \kappa_0} &\leq C\delta + CC_1^2, \\ |||\Psi(u) - \Psi(v)|||_{k, \kappa_0} &\leq 2CC_1|||u - v|||_{k, \kappa_0}. \end{aligned}$$

If $\delta \leq \delta_1$ with $\delta_1 > 0$ and small enough, then there is a constant $C_1 > 0$ depending only on δ_1 and C such that

$$C\delta + CC_1^2 \leq C_1, \quad 2CC_1 < 1.$$

Thus $\Psi(u), \Psi(v) \in \mathbb{S}(C_1)$ and

$$|||\Psi(u) - \Psi(v)|||_{k, \kappa_0} \leq \mu|||u - v|||_{k, \kappa_0}, \quad \mu = 2CC_1 < 1.$$

Therefore Ψ is a contraction mapping over $\mathbb{S}(C_1)$. Thus there is a unique fixed point u in $\mathbb{S}(C_1)$ as a mild solution to the Cauchy problem (5.50)–(5.51). Then (5.57) with $i = 0$ and (5.58) are proved. In addition, it is obvious that C_1 can be also taken as $C_1 = C'_1\delta$ for another constant C'_1 independent of δ .

Finally the time-differentiability (5.57) with $i = 1$ of the solution u and the estimate (5.60) directly follow from the equation. This completes the proof of the theorem. \square

Remark 5.16. The above analysis can be applied to the study of corresponding problems for compressible Navier-Stokes equations. For stationary potential force, more information on the solutions can be obtained such as the large time behavior and the optimal convergence rates. Interested readers can refer to [32, 33] for more details.

Acknowledgement. The preparation of the notes is influenced by our recent work on the well-posedness theory of hyperbolic conservation laws, [13, 65, 66, 67]; the introduction of a new function space and the analysis by combining the spectral analysis and energy method for the Boltzmann equation, [33, 96].

The last but not least, we would like to thank the organizers of the Shanghai Summer School, Professors Gui-Qiang Chen, Ta-Tsien Li, Fanghua Lin, and Chun Liu for their kind invitation. In particular, the second author would like to thank Professor Chun Liu who attended the lectures and gave valuable comments on revising this lecture notes.

References

- [1] R. Alexandre, L. Desvillettes, C. Villani and B. Wennberg, Entropy dissipation and long-range interactions, *Arch. Ration. Mech. Anal.* **152** (2000), 327–355.
- [2] L. Arkeryd and C. Cercignani, A global existence theorem for the initial boundary value problem for the Boltzmann equation when the boundaries are not isothermal, *Arch. Rat. Mech. Anal.* **125** (1993), 271–288.
- [3] C. Bardos, F. Golse and D. Levermore, Fluid dynamic limits of kinetic equations I: Formal derivations, *J. Stat. Phys.*, **63** (1991), 323–344.
- [4] J. Bergh and J. Löfström, Interpolation spaces: An introduction, *Die Grundlehren der mathematischen Wissenschaften* 223, Springer-Verlag, Germany, U.S. - Berlin, 1976.
- [5] S. Bianchini, A note on the Riemann problem for nonconservative hyperbolic systems, *Arch. Rat. Mech. Anal.* **166** (2003), 1–26.
- [6] S. Bianchini and A. Bressan, Vanishing viscosity solutions of nonlinear hyperbolic systems, *Ann. of Math.*, **161** (2005), 223–342.
- [7] L. Boltzmann, Lectures on Gas Theory, English transl. by S.G. Brush, Dover Publ. Inc., New York, 1964.
- [8] A. Bressan, Hyperbolic Systems of Conservation Laws, the one-dimensional Cauchy problem, *Oxford Lecture Series in Mathematics and its Applications*, no. 20, Oxford University Press, 2000.
- [9] A. Bressan, G. Crasta and B. Piccoli, Well posedness of the Cauchy problem for $n \times n$ systems of conservation laws, *Memoir Amer. Math. Soc.*, 694, 2000.
- [10] A. Bressan and Goatin, Oleinik type estimates and uniqueness for $n \times n$ conservation laws, *J. Diff. Equations*, **156** (1999), 26–49.
- [11] A. Bressan and P. LeFloch, Uniqueness of weak solutions to systems of conservation laws, *Arch. Ration. Mech. Anal.* **140** (1997), 301–317.
- [12] A. Bressan and M. Lewicka, A uniqueness condition for hyperbolic systems of conservation laws, *Discrete Contin. Dyn. Syst.* **6** (2000), 673–682.
- [13] A. Bressan, T.-P. Liu and T. Yang, L^1 stability estimates for $n \times n$ conservation laws. *Archive for Rational Mechanics and Analysis*, **149** (1999), 1–22.

- [14] A. Bressan and A. Marson, Error bounds for a deterministic version of the Glimm scheme, *Arch. Rat. Mech. Anal.*, **142** (1998), 155–176.
- [15] A. Bressan and T. Yang, On the convergence rate of vanishing viscosity approximations, *Communications on Pure and Applied Mathematics*, **LVII** (2004), 1075–1109.
- [16] R.E. Caflisch and B. Nicolaenko, Shock profile solutions of the Boltzmann equation, *Commun. Math. Phys.* **86** (1982), 161–194.
- [17] T. Carleman, Sur la théorie de l'équation intégral-differentielle de Boltzmann, *Acta Math.* **60** (1932), 91–146.
- [18] ———, Probléme mathématique dans la théorie cinétique des gases, Tech. report, Almqvist and Wiksel, Uppsala, 1957.
- [19] C. Cercignani, Equilibrium states and trend to equilibrium in a gas according to the Boltzmann equation, *Rend. Math. Appl.* **10** (1990), 77–95.
- [20] C. Cercignani, R. Illner and M. Pulvirenti, *The Mathematical Theory of Dilute Gases*, Appl. Math. Sci., vol. 109, Springer-Verlag, New York-Berlin, 1994.
- [21] C. Cercignani, *Rarefied Gas Dynamics, from basic concepts to actual calculations*, Cambridge Texts In Appl. Math., 2000.
- [22] R. Courant and K.O. Friedrichs, *Supersonic Flow and Shock Waves*, Wiley Interscience, New York, 1948.
- [23] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Vol.II*, Wiley Interscience, New York, 1962.
- [24] C.M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Heidelberg, 2000.
- [25] ———, Polygonal approximations of solutions of the initial value problem for a conservation law, *J. Math. Anal. Appl.*, **38** (1972), 33–41.
- [26] ———, Entropy and the stability of classical solutions of hyperbolic systems of conservation laws, *Lecture Notes in Mathematics*, Editor: T. Ruggeri, Montecatini Terme, 1994, Springer.
- [27] L. Desvillettes, Convergence to equilibrium in large time for Boltzmann and BGK equations, *Arch. Rat. Mech. Anal.* **110** (1990), 73–91.
- [28] L. Desvillettes and C. Villani, On the trend to global equilibrium for spatially inhomogeneous kinetic systems: the Boltzmann equation, *Invent. Math.* **159** (2005), no. 2, 245–316.

- [29] R. DiPerna, Global existence of solutions to nonlinear hyperbolic systems of conservation laws, *J. Diff. Eq.* **20**(1976), 187–212.
- [30] ———, Uniqueness of solutions to hyperbolic conservation laws, *Indiana Univ. Math. J.*, **28**(1979), 137–188.
- [31] R. Diperna and P.L. Lions, On the Cauchy problem for the Boltzmann equation: Global existence and weak stability, *Ann. Maths.* **130** (1989), 321–366.
- [32] R.J. Duan, H.X. Liu, S. Ukai and T. Yang, Optimal L^p - L^q Convergence rates for the Navier-Stokes equations with potential force, *Journal of Differential Equations* **238** (2007), no. 1, 220–233.
- [33] R.J. Duan, S. Ukai, T. Yang and H.J. Zhao, Optimal convergence rates to the stationary solutions for the Boltzmann equation with potential forces, *Mathematical Models and Methods in Applied Sciences* **17** (2007), no. 5, 737–758.
- [34] ———, Optimal decay estimates on the linearized Boltzmann equation with time dependent force and their applications, *Commun. Math. Phys.* **277** (2008), no. 1, 189–236.
- [35] R.S. Ellis and M.A. Pinsky, The first and seconf fluid approximations to the linearized Boltzmann equation, *J. Math. Pures Appl.* **54** (1972), 1825–1856.
- [36] J. Glimm, Solutions in the large for nonlinear hyperbolic systems of equations, *Comm. Pure Appl. Math.*, **18** (1965), 697–715.
- [37] J. Glimm and P. D. Lax, Decay of solutions of systems of hyperbolic conservation laws, *Memoirs Amer. Math. Soc.*, **101**, 1970.
- [38] F. Golse, B. Perthame and C. Sulem, A half space problem for the nonlinear Boltzmann equation with reflection boundary condition, *Arch. Rational Mech. Anal.* **103** (1988), 81–96.
- [39] J. Goodman and Z. Xin, Viscous limits for piecewise smooth solutions to systems of conservation laws, *Arch. Rational Mech. Anal.*, **121** (1992), 235–265.
- [40] H. Grad, Asymptotic equivalence of the Navier-Stokes and nonlinear Boltzmann equations, *Proc. Symp. Appl. Math.* vol. 17 (R.Finn, ed.), AMS, Providence, 1965, 154–183.
- [41] Y. Guo, The Boltzmann equation in the whole space, *Indiana Univ. Math. J.* **53** (2004), no. 4, 1081–1094.
- [42] ———, The Vlasov-Poisson-Boltzmann system near Maxwellians, *Comm. Pure Appl. Math.*, **55** (9) (2002), 1104–1135.
- [43] K. Hamdache, Initial boundary value problems for Boltzmann equation. global existence of weak solutions, *Arch. Rat. Mech. Anal.* **119** (1992), 309–353.

- [44] F. Herau and F. Nier, Isotropic hypoellipticity and trend to equilibrium for the fokker-planck equation with a high - degree potential, *Arch. Rational Mech. Anal.* **171** (2004), 151–218.
- [45] D. Hilbert, *Grundzuge einer Allgemeinen Theorie der Linearen Integralgleichungen.* (Teubner, Leipzig), Chapter 22.
- [46] E. H. Kennard, *Knetic Theory of Gases*, McGraw-Hill, New York, 1938.
- [47] F.M. Huang and T. Yang, Stability of contact discontinuity for the Boltzmann equation, Preprint 2004.
- [48] R. Illner and M. Shinbrot, The Boltzmann equation: global existence for a rare gas in an infinite vacuum, *Comm. Math. Phys.* **95** (1984), no. 2, 217–226.
- [49] T. Kato, Perturbation Theory of Linear Operators, *Die Grundlehren der mathematischen Wissenschaften* 123, Springer-Verlag, New York, 1966.
- [50] N.N. Kuznetsov, Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation, *U.S.S.R. Comp. Math. and Math. Phys.*, **16** (1976), 105–119.
- [51] B. Keyfitz, Solutions with shocks: An example of L_1 -contractive semigroup, *Comm. Pure Appl. Math.*, **24**(1971), 125–132.
- [52] S. Kruzhkov, First-order quasilinear equations in several space variables, *Mat. Sb.* **123**(1970), 228–255; English translation in *Math. USSR Sb.*, **10** (1970), 217–273.
- [53] O. Lanford III, Time evolution of large classical system, Lec. Notes in Phys. (E.J. Moser, ed.), vol. 38, Springer-Verlag, New York, 1975, 1–111.
- [54] P.D. Lax, Hyperbolic systems of conservation laws II, *Comm. Pure Appl. Math.*, **10** (1957), 537–566.
- [55] P. LeFloch, Hyperbolic systems of conservation laws. The theory of classical and nonclassical shock waves, *Lectures in Math. ETH Zürich.* Birkhäuser Verlag, Basel, 2002.
- [56] P. LeFloch and Z.P. Xin, Uniqueness via the Adjoint Problems for Systems of Conservation laws, *Comm. Pure Appl. Math.*, **XLVI**(1993), 1499–1533.
- [57] R. LeVeque, Numerical methods for conservation laws, *Lecture Notes in Mathematics*, Birkhäuser, Basel, 1990.
- [58] H. Li and A. Matsumura, On Fokker-Planck-Boltzmann equation near maxwellian, Preprint, 2005.

- [59] T.-P. Liu, Uniqueness of weak solutions of the Cauchy problem for general 2×2 conservation laws, *J. Diff. Equa.* **20** (1976), 369–388.
- [60] ———, The Riemann problem for general system of conservation laws, *J. Diff. Eq.*, **18** (1975), 218–234.
- [61] ———, Admissible solutions of hyperbolic conservation laws, *Memoirs of Amer. Math. Soc.*, Vol. 30, 240 (1981).
- [62] ———, The entropy condition and the admissibility of shocks, *J. Math. Anal. Appl.*, **53** (1976), 78–88.
- [63] T.-P. Liu and T. Ruggeri, Entropy production and admissibility of shocks, *Acta Math. Appl. Sinica*, Vol., no. 1, 1–12, 2003.
- [64] T.-P. Liu and T. Yang, Weak solutions of general systems of hyperbolic conservation laws, *Communications in Mathematical Physics*, **230**(2002), no. 2, 289–327.
- [65] ———, L_1 stability for 2×2 systems of hyperbolic conservation laws, *J. Amer. Math. Soc.*, **12**, no. 3, 729–774, 1999.
- [66] ———, Well-posedness theory for hyperbolic conservation laws, *Comm. Pure Appl. Math.*, **52**(1999), no. 2, 1553–1586.
- [67] ———, A new entropy functional for scalar conservation law. *Communications on Pure and Applied Mathematics*, **52** (1999), 1427–1442.
- [68] T.-P. Liu, T. Yang and S.-H. Yu, Energy method for Boltzmann equation, *Physica D* **188** (2004), no. 3–4, 178–192.
- [69] T.-P. Liu, T. Yang, S.-H. Yu and H. Zhao, Nonlinear stability of rarefaction waves for the Boltzmann equation, *Arch. Ration. Mech. Anal.* **181** (2006), no. 2, 333–371.
- [70] T.-P. Liu and S.-H. Yu, Boltzmann equation: Micro-macro decompositions and positivity of shock profiles, *Commun. Math. Phys.* **246** (2004), no. 1, 133–179.
- [71] ———, The Green’s function and large-time behavior of solutions for the one-dimensional Boltzmann equation, *Comm. Pure Appl. Math.* **57** (2004), no. 12, 1543–1608.
- [72] A. Majda, Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables, Springer-Verlag, New York, 1984.
- [73] J. C. Maxwell, On the dynamical theory of gas, *Phil. Trans. Royal Soc. London* **157** (1867), 49–88.
- [74] C. Mouhot and R. Strain, Spectral gap and coercivity estimates for linearized Boltzmann collision operators without angular cutoff, preprint.

- [75] T. Nishida and K. Imai, Global solutions to the initial value problem for the nonlinear Boltzmann equation, *Publ. RIMS, Kyoto Univ.* **12** (1977), 229–239.
- [76] O. Oleinik, Uniqueness and stability of the generalized solution of the Cauchy problem for a quasilinear equation, *Usp. Mat. Nauk* **14** (1969), 165–170, *Amer. Math. Soc. Transl. Ser. 2*, **33** (1964), 285–290.
- [77] ———, Discontinuous solutions of nonlinear differential equations, *Am. Math. Soc. Transl.* **26** (1963), 95–172.
- [78] B. Perthame, Introduction to the collision models in Boltzmann’s theory, in “Modeling of Collisions”, P.-A. Raviart ed., *Research in Appl. Math.*, Masson, Paris (1997).
- [79] A. Palczewski, Stationary Boltzmann’s equation with Maxwell’s boundary conditions in a bounded domain, *Math. Methods Appl. Sci.* **15** (1992), 375–393.
- [80] D. Serre, *Systèmes de Lois de Conservation*, I, II, Diderot Editeur, Paris, 1996.
- [81] Y. Shizuta, On the classical solutions of the Boltzmann equation, *Commun. Pure Appl. Math.* **36** (1983), 705–754.
- [82] Y. Shizuta and K. Asano, Global solutions of the Boltzmann equation in a bounded convex domain, *Proc. Japan Acad. Ser. A Math. Sci.* **53** (1977), no. 1, 3–5.
- [83] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.
- [84] Y. Sone, *Kinetic Theory and Fluid Dynamics*, Birkhäuser, Berlin, 2002.
- [85] R. Strain and Y. Guo, Almost exponential decay near Maxwellian, *Commun. Partial Differential Equations*, **31** (2006), no. 1–3, 417–429.
- [86] T. Tang and Z.H. Teng, The sharpness of Kuznetsov’s $O(\sqrt{\Delta x})$ L^1 error estimate for monotone difference schemes, *Math. Comp.*, **64** (1995), 581–589.
- [87] B. Temple, No L_1 -contractive metrics for system of conservation laws, *Trans. Amer. Math. Soc.*, **288** (1985), 471–480.
- [88] S. Ukai, On the existence of global solutions of a mixed problem for the Boltzmann equation, *Proc. Japan Acad.* **50** (1974), 179–184.
- [89] ———, Les solutions globales de l’équation de Boltzmann dans l’espace tout entier et dans le demi-espace, *C. R. Acad. Sci. Paris* **282A** (1976), 317–320.

- [90] ———, Local solutions in gevrey classes to the nonlinear Boltzmann equation without cutoff, *Japan J. Appl. Math.* **1** (1984), 141–156.
- [91] ———, Solutions of the Boltzmann equation, *Pattern and Waves – Qualitative Analysis of Nonlinear Differential Equations* (eds. M.Mimura and T.Nishida), *Studies of Mathematics and Its Applications*, vol. 18, 37–96, Kinokuniya-North-Holland, Tokyo, 1986.
- [92] ———, Time-periodic solutions of the Boltzmann equation, *Discrete Continuous Dynamical Sys. - Ser. A* **14** (2006), 579–596.
- [93] S. Ukai and K. Asano, Stationary solutions of the Boltzmann equation for a gas flow past an obstacle, I. Existence, *Arch. Rational Anal. Mech.* **84** (1983), 248–291.
- [94] ———, Stationary solutions of the Boltzmann equation for a gas flow past an obstacle, II. Stability, *Publ. Res. Inst. Math. Sci., Kyoto Univ.* **22** (1986), no. 6, 1035–1062.
- [95] S. Ukai and T. Yang, Mathematical theory of Boltzmann equation, Lecture Notes Series-No. 8, Liu Bie Ju Centre for Math. Sci., City University of Hong Kong, March 2006.
- [96] ———, The Boltzmann equation in the space $L^2 \cap L_\beta^\infty$: Global and time-periodic solutions, *Analysis and Applications*, **4**(2006), no. 3, 263–310.
- [97] ———, Stationary problems of the Botlzmann equation, *Handbook of Differential Equations, Stationary Partial Differential Equations*, Edited by M. Chipot, Chapter 5, Vol. 5, 2007, 371–485.
- [98] C. Villani, Hypocoercive diffusion operators, *Proceedings Volume III, International Congress of Mathematicians Madrid 2006*.
- [99] W.K. Wang and T. Yang, Stability of planar diffusion waves for two dimensional Euler equations with damping, *Journal of Differential Equations*, to appear.
- [100] W.K. Wang, T. Yang and X.F. Yang, Existence of boundary layers to the Boltzmann equation with cutoff soft potentials, *Journal of Mathematical Physics*, to appear.
- [101] G.B. Witham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.
- [102] T. Yang and H.J. Zhao, A new energy method for the Boltzmann equation, *Journal of Mathematical Physics*, **268** (2006), 569–605.

Elementary Statistical Theories with Applications to Fluid Systems

Xiaoming Wang

Department of Mathematics

Florida State University

Tallahassee, FL32306, USA

E-mail: wxm@math.fsu.edu

1 Introduction

Consider an abstract continuous dynamical system

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}), \quad \mathbf{u} \in H \quad (1.1)$$

on the phase space H with associated solution semi-group $S(t)$, i.e., $\mathbf{u}(t) = S(t)\mathbf{u}_0$.

Discrete in time dynamical system (generated by a map T) can be considered as well.

1.1 Why statistical description

Suppose that we are interested in long time behavior of the system. Due to uncertainty in initial data and in the model, and possible sensitive dependence on initial data and/or parameters, it might be meaningless to use one single trajectory to make deterministic prediction of the far future as we can see from the following examples.

Example: binomial and normal distribution.

Example: tent map (map can be viewed as time discretization of continuous dynamical system)

$$T(x) = \begin{cases} 2x, & \text{if } 0 \leq x \leq \frac{1}{2}, \\ 2 - 2x, & \text{if } \frac{1}{2} \leq x \leq 1. \end{cases}$$

Example: logistic map $T(x) = 4x(1 - x)$ on the unit interval and its relationship (conjugacy) to the tent map.

Example: Lorenz 96 (the nonlinear part is a special case of Orszag-McLaughlin models) and Lorenz 63 model. Use numerics to demonstrate

the phenomena.

$$\frac{du_j}{dt} = (u_{j+1} - u_{j-2})u_{j-1} - u_j + F, \quad j = 0, 1, \dots, J$$

periodic in J . $J = 5, 36, 40$.

Example: Hamiltonian system of many ($k \gg 1$) particles

$$\frac{dx_j}{dt} = \frac{\partial H}{\partial p_j}, \quad \frac{dp_j}{dt} = -\frac{\partial H}{\partial x_j}$$

(x_j : position, p_j : momentum).

1.2 What characterizes statistical behavior

Uncertainty in initial data characterized by initial probability measure μ_0 on the phase space. Future uncertainty at time t characterized by (time dependent/non-stationary) statistical solution μ_t .

Characterization of statistical solutions:

- (strong formulation) pull-back

$$\mu_t(E) = \mu_0(S^{-1}(t)(E)),$$

for all measurable set E .

- (weak formulation) push-forward

$$\int_H \Phi(\mathbf{u}) d\mu_t(\mathbf{u}) = \int_H \Phi(S(t)\mathbf{u}) d\mu_0(\mathbf{u}),$$

for all suitable test functionals Φ .

Formally differentiate the push-forward formula with respect to time and utilize the push-forward formula again we arrive at the following

Liouville type equation.

$$\frac{d}{dt} \int_H \Phi(\mathbf{u}) d\mu_t(\mathbf{u}) = \int_H (\mathbf{F}(\mathbf{u}), \Phi'(\mathbf{u})) d\mu_t(\mathbf{u}). \quad (1.2)$$

If we choose special type of test functions, say $\Phi(\mathbf{u}) = e^{i(\mathbf{u}, g)}$ for some $g \in H$ we arrive at the following

Hopf's equation.

$$\frac{d}{dt} \int_H e^{i(\mathbf{u}, g)} d\mu_t(\mathbf{u}) = \int_H (\mathbf{F}(\mathbf{u}), ig) e^{i(\mathbf{u}, g)} d\mu_t(\mathbf{u}), \quad (1.3)$$

which can be viewed as an equation for the characteristic function

$$\int_H e^{i(\mathbf{u}, g)} d\mu_t(\mathbf{u}).$$

Example of statistical solutions: $\mu_0 = \delta_{\mathbf{u}_0}$, i.e. the Dirac delta measure concentrated at the point \mathbf{u}_0 , then $\{\mu_t = \delta_{\mathbf{u}(t)}, t > 0\}$ is a statistical solution. In another word, classical trajectory/solution is a special case of the statistical solution theory.

Remark. The Liouville type equations for the statistical solutions are linear although the original dynamical systems are nonlinear. The advantage then is that we can apply machinery from linear functional analysis. However, we pay the price of working in the space of probability measures.

Special case of finite dimensional dynamical systems. In this case we assume that the statistical solutions possess $p(\mathbf{u}, t)$. We may then derive from the Liouville type equation, the following classical

Liouville equation.

$$\frac{\partial p(\mathbf{u}, t)}{\partial t} + \nabla_{\mathbf{u}} \cdot (\mathbf{F}(\mathbf{u})p(\mathbf{u}, t)) = 0. \quad (1.4)$$

Again the Liouville equation is linear (PDE) although the original dynamical system (ODE) is non-linear.

Remark. It is easy to check that the Lebesgue measure is invariant under an ODE system if and only if $\nabla \cdot \mathbf{F} = 0$ (Liouville theorem). Systems satisfying the Liouville property have a very good chance of having complex behavior. Example: Hamiltonian system.

Now if the system reaches a statistical equilibrium in the sense that the statistics are time independent (stationary), the probability measure μ that describes the stationary uncertainty can be characterized via either the strong (pull-back) or weak (push-forward) formulation. In the pull-back case, we have the notion of

Invariant Measure.

$$\mu(E) = \mu(S^{-1}(t)(E)), \quad \forall t \geq 0. \quad (1.5)$$

In the case of push-forward, we have the notion of

Stationary statistical solutions.

$$\int_H (\mathbf{F}(\mathbf{u}), \Phi'(\mathbf{u})) d\mu(\mathbf{u}) = 0, \quad (1.6)$$

for all suitable test functionals Φ . In the case of PDE (infinite dimensional dynamical system) we may impose additional assumption (see the next chapter).

Example of invariant measure and stationary statistical solutions: let \mathbf{u}_0 be a steady state of the dynamical system, then $\mu = \delta_{\mathbf{u}_0}$ is an invariant measure and stationary statistical solution.

Example: Lebesgue measure for the tent map.

Example: Hamiltonian system with compact constant energy surface. The induced measure on the constant energy sub-manifold $\{(x, p) | H(x, p) = c\}$ given by $d\nu = d\sigma / |\nabla H|$ where $d\sigma$ is the measure on the constant energy surface generated by the Euclidean distance is invariant (micro-canonical distributions).

1.3 Relative stability of statistical solution over single trajectories

One obvious advantage of thinking statistically is the relative stability. Think about ensemble prediction with the statistics calculated by using ensemble averages via a group of (finite) trajectories.

$$\frac{1}{N} \sum_{j=1}^N \Phi(\mathbf{u}_j(t)),$$

where $\mathbf{u}_j(t), 1 \leq j \leq N$ are N trajectories (orbits). It is then obvious that changing one trajectory while leaving the rest unchanged would affect the statistics very little as long as N is large and the trajectories remain bounded.

Another way to see the relative stability of “genuine” statistical solutions is the following: notice that for any two statistical solutions $\{\mu_t, t \geq 0\}$ and $\{\tilde{\mu}_t, t \geq 0\}$ and any measurable set E ,

$$|\mu_t(E) - \tilde{\mu}_t(E)| = |\mu_0(S^{-1}(t)E) - \tilde{\mu}_0(S^{-1}(t)E)|$$

and therefore the total variation between two initial probability measures will not grow in time. Hence small error initially (measured in terms of total variation) will remain small for all time. Of course, this is only relevant for “genuine” statistical solutions since the singular statistical solutions (with initial data equal to a Dirac delta measure) the total variation between two delta measures is always 2 unless they are the same.

2 Stationary statistics

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}), \quad \mathbf{u} \in H. \quad (2.1)$$

2.1 Invariant measures and stationary statistical solutions

As we discussed earlier, if a dynamical system reaches a statistical equilibrium, the equilibrium statistics are characterized by the invariant measures or stationary statistical solutions of the system. We discuss these objects briefly here.

2.2 Definition, existence

As we discussed earlier, an invariant measure of the dynamical system is a probability measure that is invariant under the flow.

Definition 1 (Invariant measure). *A Borel measure μ is called an invariant measure (IM) of the dynamical system with associated solution semi-group $S(t)$ if*

$$\mu(E) = \mu(S^{-1}(t)(E)), \quad \forall t \geq 0 \quad (2.2)$$

and all Borel sets E .

The statistical invariance may also be characterized by the stationarity of the Liouville type equation. Nevertheless, there are caveats associated with dynamical systems generated by PDEs (infinite dimensional dynamical systems). In particular, in the abstract formulation, the forcing term $\mathbf{F}(\mathbf{u})$ is in general not a map from the phase space H into itself. In fact, \mathbf{F} is in general associated with certain differential operators and hence it must map from a space V with more regularity (function space with more derivatives) into a space with less regularity (say V' the dual space of V induced by the inner product on H). This implies that we should then require that the stationary statistical solution be supported on V . This also implies that we cannot take an arbitrary (smooth) test functional Φ . In fact, we should require that the Fréchet derivative be in V so that the formal duality in the Liouville type equation makes sense. Also, quadratic skew symmetric nonlinearity is a common feature of fluid dynamics due to the total derivative in the Eulerian coordinates. To summarize, we will make the following assumptions on the dynamical system:

- There exists a Hilbert space V such that

$$\mathbf{F} = \mathcal{L}\mathbf{u} + B(\mathbf{u}, \mathbf{u}) + \mathbf{f}$$

is continuous from V to V' where V' is the dual space of V w.r.t. the inner product on H , where \mathcal{L} is a linear operator while B is a bilinear operator which is skew symmetric in the sense that

$$\langle B(\mathbf{u}, \mathbf{u}), \mathbf{u} \rangle_{V', V} = 0,$$

and \mathbf{f} is independent of \mathbf{u} .

- \mathbf{F} grows at most quadratically, i.e., there exists a constant c such that

$$\|\mathbf{F}(\mathbf{u})\|_{V'} \leq c(\|\mathbf{u}\|_V^2 + 1).$$

We then introduce the following definition for stationary statistical solutions:

Definition 2 (Stationary Statistical Solutions). *A Borel/Radon probability measure μ on the phase space H is called a stationary statistical solution to the dynamical system if the following are satisfied.*

1. μ is supported on V with finite second moment, i.e.,

$$\int_H \|\mathbf{u}\|_V^2 d\mu(\mathbf{u}) < \infty. \quad (2.3)$$

2. μ is stationary in the sense that it satisfies the stationary Liouville type equation

$$\int_H \langle \mathbf{F}(\mathbf{u}), \Phi'(\mathbf{u}) \rangle_{V',V} d\mu(\mathbf{u}) = 0 \quad (2.4)$$

for all suitable test functional Φ that are bounded on bounded sets in H , Fréchet differentiable in H along V with derivative $\Phi'(\mathbf{u})$ continuous and bounded on V .

3. μ satisfies the energy type inequality in the statistical sense, i.e.

$$\int_H \langle \mathbf{F}(\mathbf{u}), \mathbf{u} \rangle_{V',V} d\mu(\mathbf{u}) = \int_H \langle \mathcal{L}\mathbf{u} + \mathbf{f}, \mathbf{u} \rangle_{V',V} d\mu(\mathbf{u}) \leq 0. \quad (2.5)$$

Remark. One advantage of the weak formulation of stationary statistical solutions is that no solution semi-group is involved explicitly. Hence such a concept/definition may be used for generalized dynamical system (no uniqueness of orbits/solutions). This is what we shall do with three dimensional Navier-Stokes equations and the Boussinesq system for Rayleigh-Bénard convection.

Example of suitable test functional: $\Phi(\mathbf{u}) = \psi((\mathbf{u}, g_1), \dots, (\mathbf{u}, g_m))$, $\psi \in C_c^1(R^m)$, $g_j \in V$. In this case $\Phi'(\mathbf{u}) = \sum_{j=1}^m \partial_j \psi((\mathbf{u}, g_1), \dots, (\mathbf{u}, g_m)) g_j$. Sometimes we simply enforce the weak invariance in the definition of the stationary statistical solutions for these cylindrical test functionals only. Of course, general test functionals can be approximated by cylindrical ones under appropriate assumptions.

Remark. It is easy to see that $\int_H \langle \mathbf{F}(\mathbf{u}), \Phi'(\mathbf{u}) \rangle d\mu(\mathbf{u})$ makes sense (converges) under the quadratic growth condition, the finite second moment assumption and the boundedness of Fréchet derivative assumption. The quadratic growth assumption is usually satisfied for fluid system with quadratic nonlinearity. The finite second moment assumption usually follows either from a priori estimates in V or simply considering invariant measure on subsets with bounded V norm.

The statistical version of energy estimates can be refined/strengthened to require statistical energy type estimates on all energy shells $e_1 \leq \| \mathbf{u} \|^2 \leq e_2$ (the current one is for $e_1 = 0, e_2 = \infty$).

Remark. The existence of invariant measures follows from either the existence of steady state or long time average (see next section). However, invariant measure is not necessarily supported on a steady state, and it could be supported on a periodic orbit or more complex object. For instance the Lebesgue measure on the unit interval is an invariant of the tent map, and $d\mu = \frac{1}{2\pi} d\theta$ on the unit circle is an invariant measure (supported on a periodic orbit) for the following ODE (in polar coordinates)

$$\frac{dr}{dt} = r(1 - r^2), \quad \frac{d\theta}{dt} = 1.$$

Also there are micro-canonical distributions on the energy level surfaces for Hamiltonian systems if energy level surfaces are compact.

In general we may have multiple invariant measures. This raises the question of which invariant measures are “physically” relevant and if there is a preferred one. One way to define physical is to say, in the finite dimensional case at least, it has a basin of attraction with positive measure. This will not eliminate the non-uniqueness (say for instance for the ODE $\frac{du}{dt} = u - u^3$). This definition also has no obvious generalization to the infinite dimension. Now if the system is “mixing” in the sense that information can be transferred from almost any part of the phase space to any other part of the phase space (there are other measures of mixing such as the decay of time correlation functions), we then should expect a unique “physical” invariant measure. How to find such a measure is a challenge. In the case of inviscid unforced finite dimensional case, we may postulate that the system tends to mix the phase space so that information is lost as quickly as possible. Therefore, it may be plausible to propose that the physically relevant IM is the one that maximizes the information loss (or maxima of the Shannon entropy) with given constraints. This is the information theoretical approach that we shall discuss later. Another approach is to acknowledge that the world is intrinsically noisy and therefore we should add noise to the system.

Generic noise actually may induce stability in the statistical sense, i.e., the noisy system may have only one (unique) IM. This is another topic that we shall investigate later. We naturally inquire what kind of noises should be added to the system (application dependent) and if the zero noise limit leads us to something special. For a special type of dynamical systems (system with axiom-A attractor), zero noise limit leads to the physical one, the so-called SRB measure. Roughly speaking, SRB measures are measures that are most compatible with volume when volume is not preserved; and they provide mechanism of how local instability on the attractor produce coherent statistics. However, for most systems that we are interested in, we do not know if they are axiom A.

In general, the set of invariant measures is not continuous with respect to the parameters in the system (let alone differentiability) as can be seen from simple ODEs with bifurcations. Even the “physical” IMs (in the sense that they have a basin of attraction with positive measure in finite dim) are not continuous as is clear from the following example

$$\frac{du}{dt} = -u(u-1)^2 + \alpha.$$

Again, for many dissipative dynamical systems, taking into consideration of noises can in fact induce stability in the statistical sense since the unique invariant measure will be continuous in physical parameters in general. We shall discuss noise induced statistical stability later.

2.3 Ergodicity

For practical purposes we need to compute statistics with respect to certain invariant measure or stationary statistical properties. Although the possible non-uniqueness issue bothers mathematician quite a bit, people have used long time average to compute/approximate statistics in practice regardless of whether the IM is unique. The rational behind the customary approach is the assumption of ergodicity which roughly means that generic trajectory will traverse almost all interesting parts of the phase space and hence time average is equivalent to spatial (in phase space) average. The mathematical question then is whether such an approach can be justified. Time average also has some obvious advantages over the ensemble approach especially in the case when ensemble is prohibitively expensive or not available(say for instance the climate system).

We first notice that for a general test functional Φ and an arbitrary orbit $\{S(t)\mathbf{u}_0, t \geq 0\}$, the long time average may not exist. This difficulty can be circumvented if we consider the so-called Banach limit LIM which is a generalized limit on the space of bounded functions on the positive real line. The Banach limits have the following properties:

1.

$$\text{LIM } g(t) \geq 0, \forall g \in \mathcal{B}([0, \infty)), \text{ s.t. } g(t) \geq 0 \forall t \geq 0; \quad (2.6)$$

2.

$$\text{LIM } g(t) = \lim g(t), \text{ if } \lim g(t) \text{ exists}; \quad (2.7)$$

3.

$$\liminf g(t) \leq \text{LIM } g(t) \leq \limsup g(t); \quad (2.8)$$

4.

$$|\text{LIM } g(t)| \leq \limsup |g(t)| \leq \sup_{t \geq 0} |g(t)|; \quad (2.9)$$

5. For any $f \in \mathcal{B}([0, \infty))$ and $g(t) = \frac{1}{t} \int_0^t f(s) ds$, we then have

$$\text{LIM } g(t + \tau) = \text{LIM } g(t); \quad (2.10)$$

6. Moreover, for a fixed $g_0 \in \mathcal{B}([0, \infty))$, there exists a special Banach limit LIM_0 such that

$$\text{LIM}_0 g_0(t) = \limsup g_0(t). \quad (2.11)$$

Remark. The existence of Banach/generalized limit is a simple application of the Hahn-Banach theorem to the subspace

$$Y = \{g \in C([0, \infty)) \mid \lim_{t \rightarrow \infty} g(t) \text{ exists}\}$$

of all bounded functions on $[0, \infty)$, i.e. $L^\infty([0, \infty))$. The special property of having a particular generalized limit that agrees with \limsup on a particular given function g_0 is also included in the proof of the Hahn-Banach theorem. Indeed, we consider the subspace $\tilde{Y} = \text{span}\{Y, g_0\}$ so that the extension to \tilde{Y} agrees with \limsup on g_0 . This is possible since for the semi-norm

$$p(g) = \limsup_{t \rightarrow \infty} |g(t)|$$

satisfies the extensible condition

$$\lim_{t \rightarrow \infty} g(t) - p(g - g_0) \leq \limsup_{t \rightarrow \infty} g_0(t) \leq \lim_{t \rightarrow \infty} g(t) + p(g - g_0), \quad \forall g \in Y.$$

With the help of Banach limit, together with the Kakutani-Riesz theorem, we may show that time average (defined through Banach limit) is in fact equivalent to spatial average with respect to appropriate stationary statistical solutions.

For convenience we first recall that a system is called dissipative if it possesses a global attractor which is defined as below.

Definition 3 (Global attractor). *The global attractor \mathcal{A} of a semigroup $\{S(t)\}$ on a phase space H is a compact set in H such that*

- It is positive invariant in the sense that

$$S(t)\mathcal{A} = \mathcal{A}, \forall t \geq 0. \quad (2.12)$$

- It attracts arbitrary bounded set B in H in the sense that

$$\lim_{t \rightarrow \infty} \text{dist}_H(S(t)B, \mathcal{A}) = 0, \quad (2.13)$$

where the Hausdorff semi-distance is defined through

$$\text{dist}_H(B, A) = \sup_{b \in B} \inf_{a \in A} \|b - a\|_H. \quad (2.14)$$

Theorem 1 (\mathcal{IM} generated by time averages). *Let $\{S(t), t \geq 0\}$ be a continuous dissipative dynamical system on a reflexive Banach space H . Then for any given initial data \mathbf{u}_0 and any choice of generalized limit LIM , there exists a unique invariant measure μ of the dynamical system so that the time average defined via LIM is equivalent to spatial average with respect to μ , i.e.,*

$$LIM_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(S(t)\mathbf{u}_0) dt = \int_H \varphi(\mathbf{u}) d\mu(\mathbf{u}), \quad \forall \varphi \in C(H). \quad (2.15)$$

Proof. Since the system possesses a compact global attractor \mathcal{A} , there exists a closed absorbing ball \mathcal{B}_a in H for each fixed initial data \mathbf{u}_0 so that

$$S(t)\mathbf{u}_0 \in \mathcal{B}_a, \quad \forall t \geq 0.$$

\mathcal{B}_a is weakly compact since H is reflexive thanks to Banach-Alaoglu theorem [15] (generalized Heine-Borel theorem). Therefore, thanks to the Kakutani-Riesz representation theorem [15], for a fixed initial data \mathbf{u}_0 and generalized limit LIM , there exists a Borel probability measure μ on \mathcal{B}_a such that

$$LIM_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(S(t)\mathbf{u}_0) dt = \int_{\mathcal{B}_a} \varphi(\mathbf{u}) d\mu(\mathbf{u}), \quad \forall \varphi \in C_w(\mathcal{B}_a)$$

since the long time limit defined through the generalized limit on the left hand side defines a continuous linear functional on $C_w(\mathcal{B}_a)$, the space of all weakly continuous functionals on \mathcal{B}_a .

Thanks to the Tietze extension theorem [15] and the fact that \mathcal{B}_a is weakly closed, any weakly continuous functional on \mathcal{B}_a can be extended to a weakly continuous functional on H . The Borel probability measure μ on \mathcal{B}_a can be extended to a Borel probability measure on the whole space in a trivial manner (assigning zero probability to $H \setminus \mathcal{B}_a$) and thus the same equivalence of spatial and temporal average relation holds. Moreover, we see that if two weakly continuous functionals agree on \mathcal{B}_a ,

then the long time averages defined through the generalized limit are the same since the whole trajectory belongs to \mathcal{B}_a . Also, the restriction of any weakly continuous functional on H onto \mathcal{B}_a is weakly continuous. Therefore, we have the desired equivalence of the spatial and temporal averages for all $\varphi \in C_w(H)$, i.e.,

$$\text{LIM}_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(S(t)\mathbf{u}_0) dt = \int_H \varphi(\mathbf{u}) d\mu(\mathbf{u}), \quad \forall \varphi \in C_w(H).$$

We need to show that μ is invariant under the flow and that the invariance is in fact valid for all continuous functionals on H (weakly continuous functionals are automatically continuous functionals but not vice versa).

We now verify that μ is invariant under the flow. For this purpose we fix $\tau > 0$ and we have

$$\begin{aligned} \int_H \varphi(S(\tau)\mathbf{u}) d\mu(\mathbf{u}) &= \text{LIM}_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(S(\tau)S(t)\mathbf{u}_0) dt \\ &= \text{LIM}_{T \rightarrow \infty} \frac{1}{T} \int_{\tau}^{T+\tau} \varphi(S(t)\mathbf{u}_0) dt \\ &= \text{LIM}_{T \rightarrow \infty} \frac{1}{T} \left\{ \int_0^T + \int_T^{T+\tau} - \int_0^{\tau} \right\} \varphi(S(t)\mathbf{u}_0) dt \\ &= \text{LIM}_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(S(t)\mathbf{u}_0) dt \\ &\quad + \text{LIM}_{T \rightarrow \infty} \frac{1}{T} \left\{ \int_T^{T+\tau} - \int_0^{\tau} \right\} \varphi(S(t)\mathbf{u}_0) dt \\ &= \int_H \varphi(\mathbf{u}) d\mu(\mathbf{u}), \end{aligned}$$

where we have used the boundedness of the weakly continuous functional φ on the weakly compact set \mathcal{B}_a which contains the whole trajectory.

Next, we need to show that the equivalence between spatial and temporal averages is in fact valid for any continuous functionals on H .

Since the weak Borel sets and strong Borel sets are the same, μ is also strongly invariant and hence its support contained in the global attractor \mathcal{A} by a result from the next section which dictates that all invariant measures are supported on the global attractor.

Now let $\varphi \in C(H)$ (continuous but not necessarily weakly continuous), the restriction of φ on \mathcal{A} is weakly continuous due to the compactness of \mathcal{A} . We also notice that \mathcal{A} is weakly closed since any weak limit would also be a strong limit thanks to the compactness. Now let $\tilde{\varphi}$ be a weakly continuous extension of $\varphi|_{\mathcal{A}}$ to H . Such an extension exists due to Tietze theorem. We also notice that the two functionals must be

asymptotically the same along the trajectory in the sense that

$$\varphi(S(t)\mathbf{u}_0) - \tilde{\varphi}(S(t)\mathbf{u}_0) \rightarrow 0, \quad t \rightarrow \infty.$$

Indeed, if it were not true, we must have a $\delta > 0$ and a time sequence $\{t_n, n = 1, 2, \dots\}$ with $t_n \rightarrow \infty, n \rightarrow \infty$ so that

$$|\varphi(S(t_n)\mathbf{u}_0) - \tilde{\varphi}(S(t_n)\mathbf{u}_0)| \geq \delta.$$

Due to the attracting nature of the global attractor \mathcal{A} , there exists a sub time sequence, still denoted $\{t_n, n = 1, 2, \dots\}$, and $\mathbf{u}_\infty \in \mathcal{A}$, so that

$$S(t_n)\mathbf{u}_0 \rightarrow \mathbf{u}_\infty \in \mathcal{A}, \quad n \rightarrow \infty.$$

Therefore, thanks to the continuity of φ and weak continuity of $\tilde{\varphi}$, and the fact that $\varphi(\mathbf{u}_\infty)$ and $\tilde{\varphi}(\mathbf{u}_\infty)$ are the same since $\mathbf{u}_\infty \in \mathcal{A}$, we have

$$\lim_{n \rightarrow \infty} \varphi(S(t_n)\mathbf{u}_0) = \varphi(\mathbf{u}_\infty) = \tilde{\varphi}(\mathbf{u}_\infty) = \lim_{n \rightarrow \infty} \tilde{\varphi}(S(t_n)\mathbf{u}_0),$$

which contradicts the choice of the time sequence.

With the asymptotic equivalence between $\varphi(S(t)\mathbf{u}_0)$ and $\tilde{\varphi}(S(t)\mathbf{u}_0)$, we have, for any $\varphi \in C(H)$,

$$\begin{aligned} \int_H \varphi(\mathbf{u}) d\mu(\mathbf{u}) &= \int_{\mathcal{A}} \varphi(\mathbf{u}) d\mu(\mathbf{u}) \quad (\text{support of } \mu) \\ &= \int_{\mathcal{A}} \tilde{\varphi}(\mathbf{u}) d\mu(\mathbf{u}) \quad (\text{equivalence of } \varphi \text{ and } \tilde{\varphi} \text{ on } \mathcal{A}) \\ &= \int_H \tilde{\varphi}(\mathbf{u}) d\mu(\mathbf{u}) \quad (\text{support of } \mu) \\ &= \text{LIM}_{T \rightarrow \infty} \frac{1}{T} \int_0^T \tilde{\varphi}(S(t)\mathbf{u}_0) dt \\ &\quad (\tilde{\varphi} \in C_w(H) \text{ and the weak equivalence}) \\ &= \text{LIM}_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(S(t)\mathbf{u}_0) dt. \\ &\quad (\text{asymptotic equivalence of } \varphi \text{ and } \tilde{\varphi}) \end{aligned}$$

This ends the proof of the theorem. \square

A proof that relies on the existence of a compact absorbing set, or a weaker result that the equivalence is valid for only weakly continuous functionals on H , i.e. $C_w(H)$ is essentially included in [11, 30].

Although the introduction of Banach limit enabled us to establish connection between generalized time and spatial average, we still do not know if these stationary statistical solutions are ergodic (different initial data, different generalized limit). We first introduce the notion of ergodicity.

Definition 4 (Ergodicity). An invariant measure μ of the dynamical system is called ergodic if all invariant sets are trivial, i.e., $S^{-1}(t)(E) = E, \forall t \geq 0$ implies either $\mu(E) = 0$ or $\mu(H \setminus E) = 0$.

In order to show that this ergodicity definition leads to the equivalence of spatial and temporal averages, we need the following results.

It is easy to verify the following result.

Lemma 1. Let μ be an invariant probability measure of the dynamical system $\{S(t), t \geq 0\}$. Let U_t be the Koopman operators defined as $U_t\varphi(\mathbf{u}) = \varphi(S(t)\mathbf{u})$. Then μ is ergodic if and only if the only fixed points of the Koopman operators are constants.

Proof. We first show the sufficiency through BWOC.

Suppose μ is not ergodic. Then there exists a measurable set $E \subset H$ such that $S^{-1}(t)(E) = E, \forall t \geq 0$ and $0 < \mu(E) < 1$.

Let $\varphi(\mathbf{u}) = \chi_E(\mathbf{u})$. Then

$$\begin{aligned} U_t\varphi(\mathbf{u}) &= \varphi(S(t)\mathbf{u}) \\ &= \chi_E(S(t)\mathbf{u}) \\ &= \chi_{S^{-1}(t)(E)}(\mathbf{u}) \\ &= \chi_E(\mathbf{u}) \\ &= \varphi(\mathbf{u}), \end{aligned}$$

which is a contradiction since φ is a non-constant fixed point of the Koopman operators. This proves the necessity.

Next we show the necessity via BWOC again.

Suppose φ is a non-constant fixed point of the Koopman operators U_t . Then there exists r such that

$$E_r = \{\mathbf{u}, \varphi(\mathbf{u}) < r\}$$

is nontrivial in the sense that $0 < \mu(E_r) < 1$.

Hence

$$\begin{aligned} S^{-1}(t)(E_r) &= \{\mathbf{u} | S(t)\mathbf{u} \in E_r\} \\ &= \{\mathbf{u} | \varphi(S(t)\mathbf{u}) < r\} \\ &= \{\mathbf{u} | (U_t\varphi)(\mathbf{u}) < r\} \\ &= \{\mathbf{u} | \varphi(\mathbf{u}) < r\} \\ &= E_r, \end{aligned}$$

which leads to a contradiction.

This ends the proof of the Koopman Lemma. \square

We are now in the position to prove the following well-known important result.

Theorem 2 (Birkhoff's ergodic theorem, discrete version). *Let μ be an invariant probability measure of the dynamical system generated by the map T . Let $\varphi \in L^\infty \cap L^1$. Then there exists a $\varphi^* \in L^\infty \cap L^1$ such that*

$$\varphi^*(\mathbf{u}) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=0}^{j=N-1} \varphi(T^j \mathbf{u}) \quad a.e. \quad (2.16)$$

Moreover, if μ is ergodic, then

$$\int_H \varphi(\mathbf{u}) d\mu = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=0}^{j=N-1} \varphi(T^j \mathbf{u}) \quad a.e. \quad (2.17)$$

Proof. Without loss of generality we assume

$$0 \leq \varphi \leq 1. \quad (\text{why?})$$

Define

$$\begin{aligned} \Phi_n(\mathbf{u}) &= \frac{1}{n+1} \sum_{j=0}^n \varphi(T^j \mathbf{u}), \\ \varphi^*(\mathbf{u}) &= \limsup_{n \rightarrow \infty} \Phi_n(\mathbf{u}), \\ \varphi_*(\mathbf{u}) &= \liminf_{n \rightarrow \infty} \Phi_n(\mathbf{u}). \end{aligned}$$

It is easy to see that it is sufficient to show that

$$\int_H \varphi^* d\mu \leq \int_H \varphi d\mu,$$

since if the above inequality is valid for all φ , we may apply the same inequality to $1 - \varphi$ and we have

$$\int_H \varphi_* d\mu \geq \int_H \varphi d\mu.$$

Combining the two inequalities we deduce

$$\int_H (\varphi^* - \varphi_*) d\mu = 0,$$

which implies that the desired limit exists.

Now for fixed $\varepsilon > 0$, we define a map from the phase space H into the set of positive integers N ,

$$\tau(\mathbf{u}) = \min\{n | \Phi_n(\mathbf{u}) \geq \varphi^*(\mathbf{u}) - \varepsilon\}.$$

We see that τ is finite almost everywhere.

We first make the *assumption* that there exists M such that

$$\tau(\mathbf{u}) \leq M, \quad a.s.$$

This implies that

$$\tau(T^j \mathbf{u}) \leq M, \quad \forall j, a.s.$$

Let $\mathbf{u} \in H$ be such a point and we consider its orbit $\mathbf{u}, T\mathbf{u}, \dots, T^{j_1}\mathbf{u}, \dots, T^n\mathbf{u}$. We may group the orbit in the following fashion:

$$\begin{aligned} & \mathbf{u}, \quad T\mathbf{u} \dots, \quad T^{j_1}\mathbf{u} \quad (j_1 = \tau(\mathbf{u})) \\ & T^{j_1+1}\mathbf{u}, \quad T^{j_1+2}\mathbf{u} \dots, \quad T^{j_2}\mathbf{u} \quad (j_2 - j_1 - 1 = \tau(T^{j_1+1}\mathbf{u})) \\ & T^{j_{r-1}+1}\mathbf{u}, \quad T^{j_{r-1}+2}\mathbf{u} \dots, \quad T^{j_r}\mathbf{u} \quad (j_r - j_{r-1} - 1 = \tau(T^{j_{r-1}+1}\mathbf{u})) \\ & T^{j_r+1}\mathbf{u}, \quad T^{j_r+2}\mathbf{u} \dots, \quad T^n\mathbf{u}, \end{aligned}$$

with $n - j_r \leq M$.

Now consider the sum

$$S_n(\mathbf{u}) = (n+1)\Phi_n(x) = \sum_{j=0}^n \varphi(T^j \mathbf{u}).$$

We have

$$\begin{aligned} S_n(\mathbf{u}) &= S_{j_1}(\mathbf{u}) + S_{j_2-j_1-1}(T^{j_1+1}\mathbf{u}) + \dots \\ &\quad + S_{j_{r-1}-j_{r-1}-1}(T^{j_{r-1}+1}\mathbf{u}) + S_{n-j_r-1}(T^{j_r+1}\mathbf{u}) \\ &\geq S_{j_1}(\mathbf{u}) + S_{j_2-j_1-1}(T^{j_1+1}\mathbf{u}) + \dots \\ &\quad + S_{j_{r-1}-j_{r-1}-1}(T^{j_{r-1}+1}\mathbf{u}) \quad (\text{positivity}) \\ &\geq j_1(\varphi^*(\mathbf{u}) - \varepsilon) + (j_2 - j_1)(\varphi^*(T^{j_1+1}\mathbf{u}) - \varepsilon) + \dots \\ &\quad + (j_r - j_{r-1})(\varphi^*(T^{j_{r-1}+1}\mathbf{u}) - \varepsilon) \quad (\text{choice of } j) \\ &= j_1(\varphi^*(\mathbf{u}) - \varepsilon) + (j_2 - j_1)(\varphi^*(\mathbf{u}) - \varepsilon) + \dots \\ &\quad + (j_r - j_{r-1})(\varphi^*(\mathbf{u}) - \varepsilon) \quad (\text{invariance}) \\ &= j_r(\varphi^*(\mathbf{u}) - \varepsilon) \\ &\geq (n - M)(\varphi^*(\mathbf{u}) - \varepsilon). \end{aligned}$$

This implies that

$$\Phi_n(\mathbf{u}) = \frac{1}{n+1} S_n(\mathbf{u}) \geq \frac{n-M}{n+1} (\varphi^*(\mathbf{u}) - \varepsilon).$$

Integrating over H with respect to μ and utilize the invariance of μ with respect to T we have

$$\int_H \varphi d\mu \geq \frac{n-M}{n+1} \left(\int_H \varphi^* d\mu - \varepsilon \right).$$

Letting $n \rightarrow \infty$ we deduce

$$\int_H \varphi d\mu \geq \int_H \varphi^* d\mu - \varepsilon,$$

which further implies the desired result since ε is arbitrary.

Now we consider the general case of essentially unbounded τ . For a given ε , we may choose an M such that

$$\mu(\{\mathbf{u} | \tau(\mathbf{u}) > M\}) < \varepsilon,$$

since μ is finite and τ is finite a.e..

Now consider

$$\varphi' = \varphi + \chi_{\{\mathbf{u} | \tau(\mathbf{u}) > M\}}$$

and define

$$S'_n(\mathbf{u}) = \sum_{j=0}^n \varphi'(T^j \mathbf{u}).$$

Observe that

$$\tau'(\mathbf{u}) = \min\{n | \frac{1}{n+1} S'_n(\mathbf{u}) \geq \varphi^*(\mathbf{u}) - \varepsilon\} \leq M.$$

Indeed, if $\tau(\mathbf{u}) \leq M$, then $\tau'(\mathbf{u}) \leq \tau(\mathbf{u})$ due to the non-negativity of the characteristic function. If $\tau(\mathbf{u}) > M$, we then have $\tau'(\mathbf{u}) = 0$ since $1 \geq \varphi^*$.

Therefore, we may apply the same grouping technique to the partial sum and we may deduce

$$S'_n(\mathbf{u}) \geq (n - M)(\varphi^*(\mathbf{u}) - \varepsilon).$$

Integrating over the phase space with respect to the invariant measure μ , utilizing the invariance of μ under T (the push forward version in particular), we have

$$(n+1) \int_H \varphi d\mu + (n+1)\varepsilon \geq (n-M)(\int_H \varphi^* d\mu - \varepsilon).$$

Dividing by $n+1$ and letting $n \rightarrow \infty$ followed by sending ε to zero we arrive at the desired result.

We are left to prove that φ^* must be a constant in the case of ergodic μ . This follows from the Koopman theorem and the simple observation that φ^* is invariant under T and hence fixed point of the Koopman operator. The constant must be the right one as one can easily verify.

It is easy to see that $\varphi^* \in L^\infty$ if $\varphi \in L^\infty$. Likewise $\varphi^* \in L^1$ if $\varphi \in L^1$.

This ends the proof of the discrete version of Birkhoff's ergodic theorem. \square

We may generalize the discrete version to the continuous version.

Theorem 3 (Birkhoff's ergodic theorem, continuous version). *Let μ be an invariant probability measure of the dynamical system $\{S(t), t \geq 0\}$. Let $\varphi \in L^\infty \cap L^1$. Then there exists a $\varphi^* \in L^\infty \cap L^1$ such that*

$$\varphi^*(\mathbf{u}) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(S(t)\mathbf{u}) dt, \quad a.e. \quad (2.18)$$

Moreover, we have

$$\int_H \varphi d\mu = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(S(t)\mathbf{u}) dt, \quad a.e. \quad (2.19)$$

if μ is ergodic.

Proof. The idea is to approximate the continuous case with the discrete case and utilize the discrete result.

Define

$$\psi(\mathbf{u}) = \int_0^1 \varphi(S(t)\mathbf{u}) dt.$$

For $T = n + \alpha$ where n is a positive integer and $\alpha \in [0, 1)$, we have

$$\begin{aligned} \frac{1}{T} \int_0^T \varphi(S(t)\mathbf{u}) dt &= \frac{1}{n} \int_0^n \varphi(S(t)\mathbf{u}) dt + \left(\frac{1}{T} - \frac{1}{n} \right) \int_0^n \varphi(S(t)\mathbf{u}) dt \\ &\quad + \frac{1}{T} \int_n^T \varphi(S(t)\mathbf{u}) dt. \end{aligned}$$

The last two terms approach zero as $T \rightarrow \infty$ by the boundedness of φ . As for the first term we have

$$\begin{aligned} \frac{1}{n} \int_0^n \varphi(S(t)\mathbf{u}) dt &= \frac{1}{n} \sum_{j=0}^{n-1} \int_j^{j+1} \varphi(S(t)\mathbf{u}) dt \\ &= \frac{1}{n} \sum_{j=0}^{n-1} \int_j^{j+1} \varphi(S(t-j)S(j)\mathbf{u}) dt \\ &= \frac{1}{n} \sum_{j=0}^{n-1} \int_0^1 \varphi(S(t')S(j)\mathbf{u}) dt' \\ &= \frac{1}{n} \sum_{j=0}^{n-1} \psi(S(j)\mathbf{u}) \\ &= \frac{1}{n} \sum_{j=0}^{n-1} \psi(S^j(1)\mathbf{u}) \\ &\rightarrow \psi^*(\mathbf{u}) \quad a.e.\mu, \end{aligned}$$

where we have applied the discrete version of Birkhoff's ergodic theorem with $S(1)$ being the map and ψ being the function.

This ends the proof of the first part of the theorem.

We observe that ψ^* is invariant under the flow and hence Koopman's theorem implies that ψ^* must be a constant function when μ is ergodic. It is then easy to see that the only constant that works is the one dictated by the theorem.

This ends the proof of the theorem. \square

Not every invariant measure is ergodic. However, ergodic invariant measures form the building blocks of the set of invariant measures of a given dynamical system in many cases.

Ergodicity may be established if we look at extremals in the set of invariant measures. This is not surprising as nature favors various extremals in many situations. Better property of extremals also fits with our information theoretical approach with maximum Shannon entropy.

Indeed, we have the following result.

Theorem 4 (Ergodicity and extremal points). *Let IM be the set of all invariant probability measures of the dynamical system $\{S(t), t \geq 0\}$. Then $\mu \in IM$ is ergodic if μ is an extreme point of IM . Moreover, if the dynamical system possesses a global attractor \mathcal{A} and the dynamics is injective on the global attractor, every ergodic invariant measure must be an extremal point of IM .*

Proof. We prove the sufficiency first.

Assume that μ is an extremal point of the set of invariant measures of the dynamical system.

Suppose that μ is not ergodic. Then there must exist an invariant set E_0 such that $0 < \mu(E_0) < 1$. Define two measures on the phase space as

$$\begin{aligned}\mu_1(E) &= \frac{\mu(E \cap E_0)}{\mu(E_0)}, \\ \mu_2(E) &= \frac{\mu(E \cap (H \setminus E_0))}{\mu(E_0)}.\end{aligned}$$

It is then easy to see that $\mu_1, \mu_2 \in IM$. Indeed,

$$\mu_1(S^{-1}(t)E) = \frac{\mu((S^{-1}(t)E) \cap E_0)}{\mu(E_0)} = \frac{\mu(E \cap E_0)}{\mu(E_0)} = \mu_1(E).$$

On the other hand

$$\mu = \mu(E_0)\mu_1 + (1 - \mu(E_0))\mu_2, \quad \mu(E_0) \in (0, 1)$$

and

$$\mu \neq \mu_1.$$

This contradicts the assumption that μ is an extremal point of IM .

Next we prove the necessity using BWOC.

We first borrow the fact from the next subsection that the support of all invariant measures are included in the global attractor. Since the dynamical system is injective onto the global attractor, and that the global attractor is compact by definition, $S(t)$ is continuously invertible on \mathcal{A} .

Suppose that μ is ergodic and μ is not an extremal point of IM . Then there exists $\mu_1, \mu_2 \in IM$ and $\lambda \in (0, 1)$ such that

$$\mu = \lambda\mu_1 + (1 - \lambda)\mu_2.$$

Let

$$f(\mathbf{u}) = \frac{d\mu_1}{d\mu} \in L^1(H, \mu)$$

be the Radon-Nikodym derivative of μ_1 with respect to μ . Then for any “good” test functional $\varphi(\mathbf{u})$ we have

$$\begin{aligned} \int_H \varphi(S(t)\mathbf{u}) f(S(t)\mathbf{u}) d\mu(\mathbf{u}) &= \int_H \varphi(\mathbf{u}) f(\mathbf{u}) d\mu(\mathbf{u}) && \text{invariance} \\ &= \int_H \varphi(\mathbf{u}) d\mu_1(\mathbf{u}) && \text{definition} \\ &= \int_H \varphi(S(t)\mathbf{u}) d\mu_1(\mathbf{u}) && \text{invariance} \\ &= \int_H \varphi(S(t)\mathbf{u}) f(\mathbf{u}) d\mu(\mathbf{u}) && \text{definition.} \end{aligned}$$

In particular, for the special choice of test functional $\psi(\mathbf{u}) = \varphi(S^{-1}(t)\mathbf{u})$ (this is allowed by the fact that we only restrict to the global attractor and $S^{-1}(t)$ exists and is continuous on \mathcal{A}), we have

$$f(S(t)\mathbf{u}) = f(\mathbf{u}).$$

Hence f must be a constant by Koopman’s theorem. The constant must be one since both μ_1 and μ are probability measures. This is a contradiction.

This ends the proof of the necessity and hence we end the proof of the theorem. \square

On the other hand, extremal points of the set IM exist for reasonable dynamical system. In fact, if IM is compact (either due to dissipativity or due to other constraints), then the Krein-Milman theorem guarantees the existence of extremal points and therefore the existence of

ergodic invariant measures. Moreover, these ergodic invariant measures are building blocks of the set of invariant measures in the sense that the convex hull of these ergodic ones give us all \mathcal{IM} .

The injectivity of the solution semi-group (on the global attractor) is usually related to backward uniqueness result.

In the case of three dimensional incompressible Navier-Stokes equations or Boussinesq equations, we do not have a dynamical system and we do not have an absorbing ball in a finer/smaller space. In this case we use weak topology on the absorbing ball in the phase space to get the required compactness and therefore Kakutani-Riesz theorem can still be applied.

2.4 Invariant measure, stationary statistical solution and attractor

For dissipative dynamical system which possesses global attractors, we also ask the relationship between invariant measure/stationary statistical properties and the global attractor since all of them are associated with long time behavior of the underlying dynamical system. We will show that the support of the invariant measure/stationary statistical solution is always included in the global attractor. However, the union of the support of all \mathcal{IM} s may not be the whole global attractor since there may be no \mathcal{IM} supported on a hetero-clinic orbit.

Theorem 5 (\mathcal{IM} and the global attractors). *The support of any invariant measure μ of a given continuous dissipative dynamical system is included in the global attractor. Moreover, the set of all invariant measures, \mathcal{IM} , is a convex compact set (with respect to the weak topology) in the space of Borel measures on the phase space.*

Proof. Since the invariant measure μ is finite, for any $\varepsilon > 0$, there exists $R > 0$ such that

$$\mu(B_R) \geq 1 - \varepsilon,$$

where B_R is a ball of radius R .

Since $S^{-1}(t)S(t)B_R \supset B_R, \forall t \geq 0$, we have together with the invariance of μ ,

$$\mu(S^{-1}(t)S(t)B_R) = \mu(S(t)B_R) \geq \mu(B_R) \geq 1 - \varepsilon.$$

Since \mathcal{A} is the global attractor and hence it attracts B_R , we see that for any $\delta > 0$, there exists $T_\delta > 0$ such that

$$dist_H(S(t)B_R, \mathcal{A}) < \delta,$$

i.e., $S(t)B_R$ is within the open delta neighborhood of \mathcal{A} , denoted \mathcal{A}_δ .

On the other hand, μ is regular (since it is a Borel measure) and hence for the given $\varepsilon > 0$, there exists a $\delta_0 > 0$ such that

$$\mu(\mathcal{A}) \geq \mu(\mathcal{A}_{\delta_0}) - \varepsilon.$$

Combining the above inequalities we have

$$\begin{aligned} \mu(\mathcal{A}) &\geq \mu(\mathcal{A}_{\delta_0}) - \varepsilon \\ &\geq \mu(S(t)B_R) - \varepsilon \quad \forall t \geq T_{\delta_0} \\ &\geq 1 - 2\varepsilon. \end{aligned}$$

This finishes the proof that any invariant measure is supported on the global attractor.

It is easy to see that the set of all invariant measures, \mathcal{IM} , is a convex and closed (under the weak convergence topology) set utilizing the push-forward (weak) formulation for instance. The compactness follows from Prokharov's tightness theorem, the compactness of the global attractor, and the fact that all invariant measures are supported on the global attractor.

This ends the proof of the theorem. \square

For generalized dynamical systems the concept of invariant measure does not apply while the concept of stationary statistical solutions is still applicable. It is easy to see that for smooth dynamical systems invariance of a Borel probability measure under the flow is equivalent to the measure being a solution to the stationary Liouville type equation. We will show next that the two concepts are the same for reasonable dynamical system under suitable physical assumptions. Indeed we have the following theorem.

Theorem 6 (*IM* and stationary statistical solutions). *Let $\{S(t), t \geq 0\}$ be a continuous dynamical system defined on a separable Hilbert space H and generated by $\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u})$. Suppose there exists another Hilbert space V which is compactly imbedded in H and that $S(t)$ is continuous on V . Denote V' the dual of V with respect to the inner product on H and assume \mathbf{F} is continuous from V to V' and bounded on bounded sets. Let $\{\mathbf{v}_j, j = 1, 2, \dots\}$ be an orthonormal basis for H which also form a basis for V . Now for any Borel probability measure μ on H which is supported on a ball in V , μ is invariant under the dynamics is equivalent to μ being a solution to the stationary Liouville type equation with any smooth cylindrical test functionals determined by the $\{\mathbf{v}_j, j = 1, 2, \dots\}$.*

Proof. Fix a positive integer m and let $\psi(y_1, \dots, y_m) \in C_c^1(R^m)$. We define the following m -dimensional cylindrical test functional

$$\varphi(\mathbf{u}) = \psi((\mathbf{u}, \mathbf{v}_1), \dots, (\mathbf{u}, \mathbf{v}_m)).$$

It is easy to check that $\varphi(S(t)\mathbf{u})$ is differentiable in t . Indeed, we formally have

$$\begin{aligned} \frac{d}{dt} \varphi(S(t)\mathbf{u}) &= \sum_{j=1}^m \frac{\partial \psi}{\partial y_j} ((S(t)\mathbf{u}, \mathbf{v}_1), \dots, \\ &\quad (S(t)\mathbf{u}, \mathbf{v}_m)) < \mathbf{F}(S(t)\mathbf{u}), \mathbf{v}_j >_{V', V} \in C^0([0, \infty)). \end{aligned}$$

Moreover, this derivative is uniformly bounded on any finite time interval $[0, T]$ and $\mathbf{u} \in B_V$ (the bounded ball in V that contains the support of μ). Hence by Lebesgue dominated convergence theorem together with a finite difference approximation and mean value theorem we have

$$\begin{aligned} \frac{d}{dt} \int_H \varphi(S(t)\mathbf{u}) d\mu(\mathbf{u}) &= \int_H \sum_{j=1}^m \frac{\partial \psi}{\partial y_j} ((S(t)\mathbf{u}, \mathbf{v}_1), \dots, \\ &\quad (S(t)\mathbf{u}, \mathbf{v}_m)) < \mathbf{F}(S(t)\mathbf{u}), \mathbf{v}_j >_{V', V} d\mu(\mathbf{u}). \end{aligned}$$

Therefore if μ is invariant implies μ is a solution to the stationary Liouville type equation with any smooth cylindrical test functionals determined by the $\{\mathbf{v}_j, j = 1, 2, \dots\}$.

Conversely, if μ is a solution to the stationary Liouville type equation, the previous argument proves the weak invariance (push-forward formulation) of μ under the flow for the smooth cylindrical test functionals determined by the $\{\mathbf{v}_j, j = 1, 2, \dots\}$.

Now that every continuous m -dimensional cylindrical test functional on $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ can be approximated by smooth m -dimensional cylindrical test functional on $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ with the usual smoothing (mollifier) technique in R^m without increasing the L^∞ norm. Hence the weak invariance is also valid for any finite dimensional continuous test functional.

Finally, for any given bounded continuous functional φ on H , it can be approximated by $\varphi \circ P_m$ where P_m is the orthogonal projection from H onto the subspace spanned by $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$. Notice that $\varphi \circ P_m$ is a continuous cylindrical test functional depending on $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$. Hence the integral of it over H with respect to μ is invariant under the flow. A simple application of the Lebesgue dominated convergence theorem implies that $\int_H \varphi(\mathbf{u}) d\mu(\mathbf{u})$ is invariant under the flow, i.e., $\int_H \varphi(S(t)\mathbf{u}) d\mu(\mathbf{u})$ is independent of time.

This ends the proof of the theorem. □

2.5 Dependence on parameters

For an abstract dynamical system with parameters, there is no obvious reason that the statistical properties should depend in some nice way

on the parameters, even if trajectories converge on any given finite time interval. For instance, consider the 2D Navier-Stokes system in vorticity-stream function formulation (doubly periodic boundary condition)

$$\frac{\partial \omega}{\partial t} + \nabla^\perp \psi \cdot \nabla \omega = \epsilon \Delta \omega + \sqrt{\epsilon} F,$$

$$\omega = \Delta \psi.$$

It is easy to check (for smooth enough F and initial data), solutions of the NSE converge to the solution of the Euler system on any given finite time interval (thus we have trajectory convergence on finite time interval). However, for a special choice of saying F being one of the eigenfunctions corresponding to the 1st eigenvalue, the set of invariant measure of the NSE does not converge to that of the Euler system as ϵ approaches 0.

2.6 Regular perturbation

Although the general question of perturbation of stationary statistical properties is not easy, the situation is relatively simple if we consider uniformly dissipative systems only. What we can show, for uniformly dissipative system, is that the set of invariant measures is upper semi-continuous with respect to regular perturbation of parameters. The general argument is that if we have enough a priori estimates for a given parameter α close to α_0 , we have tightness/weak compactness of the set of invariant measures for all interested parameter regions thanks to Prokhorov's theorem. The limit of invariant measures of the perturbed system must be an invariant measure of the limit system by taking the limit in the Liouville type equation or the weak invariance formulation.

More precisely, we have the following theorem:

Theorem 7 (Upper semi-continuity of \mathcal{IM} , regular version). *Suppose the family of continuous dynamical systems $\{S(t, \epsilon), t \geq 0\}$ on a Hilbert space H is uniformly dissipative in the sense that $K = \bigcup_{0 < |\epsilon| \leq \epsilon_0} \mathcal{A}_\epsilon$ is pre-compact where \mathcal{A}_ϵ denotes the global attractor for the system with parameter ϵ . Moreover, we assume that the trajectories converge on any finite time interval uniformly on the attractors, i.e.,*

$$\lim_{\epsilon \rightarrow 0} \sup_{\mathbf{u} \in \mathcal{A}_\epsilon} \|S(t, \epsilon)\mathbf{u} - S(t, 0)\mathbf{u}\|_H = 0, \quad \forall t \geq 0.$$

Then the set of invariant measures are upper semi-continuous in the sense that for any $\{\mu_\epsilon \in \mathcal{IM}_\epsilon, 0 < |\epsilon| \leq \epsilon_0\}$, there exists $\mu_0 \in \mathcal{IM}_0$ and a subsequence (still denoted μ_ϵ) such that

$$\lim_{\epsilon \rightarrow 0} \mu_\epsilon = \mu_0$$

(the convergence is in the weak sense).

Proof. Since the support of any invariant measure is contained in the global attractor \mathcal{A}_ϵ which is contained in the same pre-compact set K , we see that $\{\mu_\epsilon, 0 < |\epsilon| \leq \epsilon_0\}$ is tight in $\mathcal{PM}(H)$, the space of Borel probability measures on H , thanks to Prokhorov's theorem. Without loss of generality we assume that μ_ϵ weakly converges to μ_0 .

Our goal now is to show that $\mu_0 \in \mathcal{IM}_0$.

Now for any $t > 0$, since $\mu_\epsilon \in \mathcal{IM}_\epsilon$, for any continuous test functional φ we have

$$\int_H \varphi(S(t, \epsilon)\mathbf{u}) d\mu_\epsilon(\mathbf{u}) = \int_H \varphi(\mathbf{u}) d\mu_\epsilon(\mathbf{u}).$$

Thanks to the weak convergence we also have

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \int_H \varphi(\mathbf{u}) d\mu_\epsilon(\mathbf{u}) &= \int_H \varphi(\mathbf{u}) d\mu_0(\mathbf{u}), \\ \lim_{\epsilon \rightarrow 0} \int_H \varphi(S(t, 0)\mathbf{u}) d\mu_\epsilon(\mathbf{u}) &= \int_H \varphi(S(t, 0)\mathbf{u}) d\mu_0(\mathbf{u}). \end{aligned}$$

Hence, for any smooth cylindrical test functional φ ,

$$\begin{aligned} &|\int_H \varphi(S(t, 0))\mathbf{u} d\mu_0(\mathbf{u}) - \int_H \varphi(\mathbf{u}) d\mu_0(\mathbf{u})| \\ &= \lim_{\epsilon \rightarrow 0} |\int_H \varphi(S(t, 0))\mathbf{u} d\mu_\epsilon(\mathbf{u}) - \int_H \varphi(\mathbf{u}) d\mu_\epsilon(\mathbf{u})| \\ &\leq \lim_{\epsilon \rightarrow 0} |\int_H (\varphi(S(t, \epsilon)\mathbf{u}) - \varphi(\mathbf{u})) d\mu_\epsilon(\mathbf{u})| + \lim_{\epsilon \rightarrow 0} |\int_H \varphi(S(t, \epsilon)\mathbf{u}) d\mu_\epsilon(\mathbf{u}) \\ &\quad - \int_H \varphi(S(t, 0)\mathbf{u}) d\mu_\epsilon(\mathbf{u})| \\ &\leq \lim_{\epsilon \rightarrow 0} \int_H \sup \|\varphi'(\mathbf{u})\| \|S(t, \epsilon)\mathbf{u} - S(t, 0)\mathbf{u}\| d\mu_\epsilon(\mathbf{u}) \\ &= 0, \end{aligned}$$

where we have used the push-forward invariance of μ_ϵ under $S(t, \epsilon)$, mean value theorem, and the uniform convergence of trajectories starting on \mathcal{A}_ϵ .

Now for an arbitrary continuous test functional $\varphi(\mathbf{u})$, it can be approximated by cylindrical continuous test functional $\varphi(P_m \mathbf{u})$ where P_m is the orthogonal projection onto the subspace spanned by the first m elements of a given orthonormal basis of H . And we have

$$\int_H \varphi(\mathbf{u}) d\mu_0(\mathbf{u}) = \lim_{m \rightarrow \infty} \int_H \varphi(P_m \mathbf{u}) d\mu_0(\mathbf{u})$$

by the Lebesgue dominated convergence theorem and the fact that μ_0 is supported on the compact global attractor \mathcal{A} and hence φ is bounded

on \mathcal{A} . The finite dimensional cylindrical test functionals $\varphi(P_m \mathbf{u})$ can be further approximated by smooth cylindrical test functionals utilizing standard finite dimensional smoothing (mollifier) techniques, i.e., there exists smooth cylindrical test functionals $\varphi_{m,k}$ such that

$$\begin{aligned}\varphi_{m,k}(\mathbf{u}) &\rightarrow \varphi(P_m \mathbf{u}), \quad k \rightarrow \infty, \\ \|\varphi_{m,k}\|_{L^\infty} &\leq \|\varphi \circ P_m\|_{L^\infty}.\end{aligned}$$

Therefore,

$$\int_H \varphi(P_m \mathbf{u}) d\mu_0(\mathbf{u}) = \lim_{k \rightarrow \infty} \int_H \varphi_{m,k}(\mathbf{u}) d\mu_0(\mathbf{u}).$$

Combining the above we end the proof of the theorem. \square

A corollary of this upper semi-continuity is that the extremal statistics (defined through long time averages) are upper semi-continuous in the following sense.

Theorem 8 (Upper semi-continuity of extremal time averaged statistics). *Under the same assumption as in the previous theorem, i.e., uniform dissipativity plus finite time convergence, we have for any fixed continuous test functional φ_0 , the extremal statistics are saturated by ergodic invariant measures, i.e., there exist ergodic invariant measures $\nu_\epsilon \in \mathcal{IM}_\epsilon, \nu_0 \in \mathcal{IM}_0$ such that*

$$\begin{aligned}\sup_{\mathbf{u} \in H} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, \epsilon) \mathbf{u}) dt &= \int_H \varphi_0(\mathbf{u}) d\nu_\epsilon(\mathbf{u}), \\ \sup_{\mathbf{u} \in H} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, 0) \mathbf{u}) dt &= \int_H \varphi_0(\mathbf{u}) d\nu_0(\mathbf{u}).\end{aligned}$$

Moreover, the extremal statistics are upper semi-continuous in the parameter ϵ , i.e.,

$$\begin{aligned}&\limsup_{\epsilon \rightarrow 0} \sup_{\mathbf{u} \in H} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, \epsilon) \mathbf{u}) dt \\ &\leq \sup_{\mathbf{u} \in H} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, 0) \mathbf{u}) dt.\end{aligned}$$

Proof. Recall that for a fixed initial data \mathbf{u}_0 and a fixed continuous test functional φ_0 , there exists a special Banach/generalized limit that agrees with the \limsup for long time average on φ_0 . This implies, thanks to the equivalence between spatial and temporal averages, there exist

$\mu_{\epsilon, \mathbf{u}_0} \in \mathcal{IM}_{\epsilon}$ and $\mu_{0, \mathbf{u}_0} \in \mathcal{IM}_0$ such that

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, \epsilon) \mathbf{u}_0) dt = \int_H \varphi_0(\mathbf{u}) d\mu_{\epsilon, \mathbf{u}_0},$$

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, 0) \mathbf{u}_0) dt = \int_H \varphi_0(\mathbf{u}) d\mu_{0, \mathbf{u}_0}.$$

Now for fixed ϵ , the set \mathcal{IM}_{ϵ} is tight and hence the set $\{\mu_{\epsilon, \mathbf{u}_0}, \mathbf{u}_0 \in H\}$ must contain a subsequence that converges to some μ_{ϵ} so that

$$\sup_{\mathbf{u}_0 \in H} \int_H \varphi_0(\mathbf{u}) d\mu_{\epsilon, \mathbf{u}_0} = \int_H \varphi_0(\mathbf{u}) d\mu_{\epsilon},$$

$$\sup_{\mathbf{u}_0 \in H} \int_H \varphi_0(\mathbf{u}) d\mu_{0, \mathbf{u}_0} = \int_H \varphi_0(\mathbf{u}) d\mu_0.$$

It is easy to see that $\mu_{\epsilon} \in \mathcal{IM}_{\epsilon}, \mu_0 \in \mathcal{IM}_0$ since the sets of invariant measures are closed.

Next, we define

$$EIM_0 = \{\tilde{\mu}_0 \in \mathcal{IM}_0, \sup_{\mu \in \mathcal{IM}_0} \int_H \varphi_0(\mathbf{u}) d\mu = \int_H \varphi_0(\mathbf{u}) d\tilde{\mu}_0\},$$

$$EIM_{\epsilon} = \{\tilde{\mu}_{\epsilon} \in \mathcal{IM}_{\epsilon}, \sup_{\mu \in \mathcal{IM}_{\epsilon}} \int_H \varphi_0(\mathbf{u}) d\mu = \int_H \varphi_0(\mathbf{u}) d\tilde{\mu}_{\epsilon}\}.$$

It is easy to see that both EIM_0 and EIM_{ϵ} are non-empty compact sets by the compactness of \mathcal{IM}_0 and \mathcal{IM}_{ϵ} . The set of extremals is non-empty by the Krein-Milman theorem and our assumption of uniform dissipativity. Let ν_0 be an extremal point of EIM_0 and ν_{ϵ} be an extremal point of EIM_{ϵ} .

It is easy to see that ν_0 must be an extremal point of \mathcal{IM}_0 and ν_{ϵ} is an extremal point of \mathcal{IM}_{ϵ} . Indeed, if for some $\lambda \in (0, 1)$ and $\mu_1, \mu_2 \in \mathcal{IM}_0$ we have

$$\nu_0 = \lambda \mu_1 + (1 - \lambda) \mu_2,$$

then

$$\int_H \varphi_0(\mathbf{u}) d\nu_0 = \lambda \int_H \varphi_0(\mathbf{u}) d\mu_1 + (1 - \lambda) \int_H \varphi_0(\mathbf{u}) d\mu_2.$$

Since $\int_H \varphi_0(\mathbf{u}) d\nu_0 = \sup_{\mu \in \mathcal{IM}_0} \int_H \varphi_0(\mathbf{u}) d\mu$, we see that

$$\int_H \varphi_0(\mathbf{u}) d\mu_1 = \int_H \varphi_0(\mathbf{u}) d\mu_2 = \sup_{\mu \in \mathcal{IM}_{\epsilon}} \int_H \varphi_0(\mathbf{u}) d\mu,$$

and therefore, both μ_1 and μ_2 are elements of EIM_0 which contradicts the assumption that ν_0 is an extremal of EIM_0 . The same argument works for ν_{ϵ} .

Thanks to the previous theorem, the $\limsup_{\epsilon \rightarrow 0}$ on the right hand side of the last inequality in the theorem is attained and the limit is satisfied by an element $\tilde{\nu}_0$ of \mathcal{IM}_0 . Hence

$$\begin{aligned}
& \limsup_{\epsilon \rightarrow 0} \sup_{\mathbf{u}_0 \in H} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, \epsilon) \mathbf{u}_0) dt \\
&= \limsup_{\epsilon \rightarrow 0} \sup_{\mathbf{u}_0 \in H} \int_H \varphi_0(\mathbf{u}) d\mu_{\epsilon, \mathbf{u}_0} \\
&= \limsup_{\epsilon \rightarrow 0} \int_H \varphi_0(\mathbf{u}) d\mu_\epsilon \\
&\leq \limsup_{\epsilon \rightarrow 0} \int_H \varphi_0(\mathbf{u}) d\nu_\epsilon \\
&= \int_H \varphi_0(\mathbf{u}) d\tilde{\nu}_0 \\
&\leq \int_H \varphi_0(\mathbf{u}) d\nu_0 \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, 0) \mathbf{u}) dt \quad (\text{a.s. w.r.t. } \nu_0) \\
&\leq \sup_{\mathbf{u}_0 \in H} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, 0) \mathbf{u}_0) dt,
\end{aligned}$$

where in the second to the last step we have used the fact that extremals of \mathcal{IM}_0 are necessarily ergodic, and hence spatial and temporal averages are equivalent.

This ends the proof of the theorem. \square

A corollary of this is the upper semi-continuity of the Nusselt number for the Boussinesq system on parameters (Rayleigh number etc) if they are defined as the supremum over all trajectories.

2.7 Singular perturbation

The case of singular perturbation of parameter is much more difficult in general. However, for a singular perturbation problem of two time scale of relaxation type, upper semi-continuity of statistical properties are still valid in some appropriate sense. The singularity usually involves an initial layer in time and hence renders the problem not that singular if one considers long time behavior (such as stationary statistical properties). In terms of long time statistics, we can imagine that the fast variable quickly relaxes and hence is essentially slaved by the slow variable at large time. Therefore we have that the long time statistics are essentially given by the slow dynamics with the fast dynamics slaved by the

slow variable, i.e., the limit dynamics. To be more precise, we consider the following type of two time scale problem of relaxation type

$$\epsilon \left(\frac{dx_1}{dt} + g(x_1, x_2) \right) = f_1(x_1, x_2), \quad x_1(0) = x_{10}, \quad (2.20)$$

$$\frac{dx_2}{dt} = f_2(x_1, x_2), \quad x_2(0) = x_{20}, \quad (2.21)$$

where X_1, X_2 are two separable Hilbert spaces. The limit problem for $\epsilon = 0$ is given by

$$0 = f_1(x_1^0, x_2^0), \quad (2.22)$$

$$\frac{dx_2^0}{dt} = f_2(x_1^0, x_2^0), \quad x_2(0) = x_{20}. \quad (2.23)$$

This is a two time scale problem with x_1 being the fast variable and x_2 being the slow variable.

Theorem 9 (Upper semi-continuity of \mathcal{IM} , singular version). *Consider a generalized dynamical system on $X_1 \times X_2$ with two explicitly separated time scales given by (2.20, 2.21) with the limit system given by (2.22, 2.23).*

We postulate the following assumptions:

H1 (uniform dissipativity of the perturbed system) *The two-time-scale system (2.20, 2.21) possesses a global attractor \mathcal{A}_ϵ for all small positive ϵ such that $K = \bigcup_{0 < \epsilon < \epsilon_0} \mathcal{A}_\epsilon$ is pre-compact in $X_1 \times X_2$.*

H2 (dissipativity of the limit system) *The limit system is wellposed and possesses a global attractor \mathcal{A}_0 in X_2 .*

H3 (convergence of the slow variable) *The slow variable of the solutions of the two time scale system converges uniformly on \mathcal{A}_ϵ , i.e.*

$$\lim_{\epsilon \rightarrow 0} \sup_{x_2 \in \mathcal{P}_2 \mathcal{A}_\epsilon} \|\mathcal{P}_2 S(t, \epsilon)(F_1(x_2, 0), x_2) - S(t, 0)x_2\|_{X_2} = 0, \quad \forall t \geq 0,$$

where \mathcal{P}_2 is the projection from $X_1 \times X_2$ to X_2 defined as $\mathcal{P}_2(x_1, x_2) = x_2$, and $x_1 = F_1(x_2, 0)$ is the unique solution to the first part of the limit system, i.e. $0 = f_1(F_1(x_2, 0), x_2)$.

H4 (smallness of the perturbation) *The two-time-scale problem (2.20, 2.21) is a uniformly small perturbation of the limit problem (2.22, 2.23) when confined to the global attractors, i.e.,*

$$\lim_{\epsilon \rightarrow 0} \sup_{(x_1, x_2) \in \mathcal{A}_\epsilon} \|\epsilon \left(\frac{dx_1}{dt} + g(x_1, x_2) \right)\|_{X_1} = 0.$$

H5 (continuity of the slave relation) The first equation in the limit system (2.22) can be solved continuously for x_1^0 with given x_2^0 and a nontrivial left hand side, i.e., there exists a continuous function $F_1 : X_2 \times X_1 \rightarrow X_1$ such that

$$y = f_1(F_1(x_2, y), x_2).$$

Moreover, we assume F_1 is uniformly continuous for $y = 0$ and $x_2 \in \mathcal{P}_2 K$.

Then the stationary statistical properties are upper semi-continuous after lifting in this singular limit, i.e., for $\{\mu_\epsilon \in \mathcal{IM}_\epsilon, 0 < \epsilon \leq \epsilon_0\}$, there exists a weakly convergent subsequence, still denoted $\{\mu_\epsilon\}$, and $\mu_0 \in \mathcal{IM}_0$ such that

$$\mu_\epsilon \rightharpoonup \mathcal{L}\mu_0,$$

where \mathcal{L} is the lift from X_2 to $X_1 \times X_2$ defined by

$$\int_{X_1 \times X_2} \varphi(x_1, x_2) d(\mathcal{L}\mu)(x_1, x_2) = \int_{X_2} \varphi(F_1(x_2, 0), x_2) d\mu(x_2).$$

Proof. We first show that the statistical properties converge in the projected sense, i.e., the statistical properties converge when restricted to the slow manifold (equivalent of taking marginal distribution in appropriate sense).

We first perform the following change of variables

$$x_1 = y_1 + F_1(y_2, 0), \quad x_2 = y_2,$$

where F_1 is the one defined through the slave relation.

Let μ_ϵ be an invariant measure of the perturbed system on $X_1 \times X_2$, the change of variable induces another Borel probability measure $\tilde{\mu}_\epsilon$ on $X_1 \times X_2$ which is defined by

$$\int \varphi(x_1, x_2) d\mu_\epsilon(x_1, x_2) = \int \varphi(y_1 + F_1(y_2, 0), y_2) d\tilde{\mu}_\epsilon(y_1, y_2).$$

Thanks to the uniform dissipativity assumption, the set $\{\mu_\epsilon\}$ is tight in the space of Borel probability measures. This also implies that the set $\{\tilde{\mu}_\epsilon\}$ is tight in the space of Borel probability measures on $X_1 \times X_2$ since F_1 is continuous and K is pre-compact. Therefore the marginal distribution of $\{\tilde{\mu}_\epsilon\}$ in X_2 , denoted $\{M\tilde{\mu}_\epsilon\}$, is also tight in the space of all Borel probability measures on X_2 . Hence there must exist a weakly convergent subsequence so that

$$M\tilde{\mu}_\epsilon \rightharpoonup \mu_0, \quad \epsilon \rightarrow 0.$$

Our first goal is to show that $\mu_0 \in \mathcal{IM}_0$, i.e., the upper semi-continuity of the stationary statistical properties in the projected sense.

For this purpose we take a smooth cylindrical test functional Φ_0 on X_2 and show the invariance of the average of Φ_0 under the flow. Indeed, for any $t > 0$,

$$\begin{aligned}
& \int_{X_2} \Phi_0(S^0(t)y_2) d\mu_0(y_2) \\
&= \lim_{\epsilon \rightarrow 0} \int_{X_2} \Phi_0(S^0(t)y_2) d(M\tilde{\mu}_\epsilon)(y_2) \\
&= \lim_{\epsilon \rightarrow 0} \int_{X_1 \times X_2} \Phi_0(S^0(t)y_2) d\tilde{\mu}_\epsilon(y_1, y_2) \\
&= \lim_{\epsilon \rightarrow 0} \int_{X_1 \times X_2} \Phi_0(S^0(t)x_2) d\mu_\epsilon(x_1, x_2) \\
&= \lim_{\epsilon \rightarrow 0} \int_{X_1 \times X_2} \Phi_0(\mathcal{P}_2 S^\epsilon(t)(F_1(x_2, 0), x_2)) d\mu_\epsilon(x_1, x_2) \\
&\quad + \lim_{\epsilon \rightarrow 0} \int_{X_1 \times X_2} (\Phi_0(S^0(t)x_2) - \Phi_0(\mathcal{P}_2 S^\epsilon(t)(F_1(x_2, 0), x_2))) d\mu_\epsilon(x_1, x_2) \\
&= \lim_{\epsilon \rightarrow 0} \int_{X_1 \times X_2} \Phi_0(\mathcal{P}_2(F_1(x_2, 0), x_2)) d\mu_\epsilon(x_1, x_2) \\
&\quad + \lim_{\epsilon \rightarrow 0} \int_{X_1 \times X_2} (\Phi_0(S^0(t)x_2) - \Phi_0(\mathcal{P}_2 S^\epsilon(t)(F_1(x_2, 0), x_2))) d\mu_\epsilon(x_1, x_2) \\
&= \lim_{\epsilon \rightarrow 0} \int_{X_1 \times X_2} \Phi_0(x_2) d\mu_\epsilon(x_1, x_2) \\
&= \int_{X_2} \Phi_0(y_2) d\mu_0(y_2),
\end{aligned}$$

where we have used the weak convergence of $M\tilde{\mu}_\epsilon$, the change of variables/measures (definition of $\tilde{\mu}_\epsilon$), the invariance of μ_ϵ under $S^\epsilon(t)$, and the following straightforward estimate

$$\begin{aligned}
& |\Phi_0(S^0(t)x_2) - \Phi_0(\mathcal{P}_2 S^\epsilon(t)(F_1(x_2, 0), x_2))| \\
&\leq \|\Phi'_0\| \|S^0(t)x_2 - \mathcal{P}_2 S^\epsilon(t)(F_1(x_2, 0), x_2)\| \rightarrow 0, \quad \epsilon \rightarrow 0
\end{aligned}$$

uniformly for $x_2 \in \mathcal{P}_2 \mathcal{A}_\epsilon$ by the uniform convergence of the slow variable assumption.

A general continuous test functional can be approximated by smooth finite dimensional cylindrical functional just as the case of regular perturbation.

This ends the proof of the convergence in the projected sense.

For the convergence in the lifted sense, we have, for any smooth

cylindrical test functional Φ on $X_1 \times X_2$

$$\begin{aligned}
& \left| \int_{X_1 \times X_2} \Phi(x_1, x_2) d\mu_\epsilon(x_1, x_2) - \int_{X_1 \times X_2} \Phi(x_1, x_2) d(\mathcal{L}\mu_0)(x_1, x_2) \right| \\
&= \left| \int_{X_1 \times X_2} \Phi(x_1, x_2) d\mu_\epsilon(x_1, x_2) - \int_{X_2} \Phi(F_1(x_2, 0), x_2) d\mu_0(x_2) \right| \\
&\leq \left| \int_{X_1 \times X_2} \Phi(F_1(x_2, 0), x_2) d\mu_\epsilon(x_1, x_2) - \int_{X_2} \Phi(F_1(x_2, 0), x_2) d\mu_0(x_2) \right| \\
&\quad + \left| \int_{X_1 \times X_2} (\Phi(x_1, x_2) - \Phi(F_1(x_2, 0), x_2)) d\mu_\epsilon(x_1, x_2) \right| \\
&\leq \left| \int_{X_2} \Phi(F_1(x_2, 0), x_2) dM\tilde{\mu}_\epsilon(x_2) - \int_{X_2} \Phi(F_1(x_2, 0), x_2) d\mu_0(x_2) \right| \\
&\quad + \int_{X_1 \times X_2} \|\Phi'\| \|x_1 - F_1(x_2, 0)\| d\mu_\epsilon(x_1, x_2) \\
&\leq \left| \int_{X_2} \Phi(F_1(x_2, 0), x_2) dM\tilde{\mu}_\epsilon(x_2) - \int_{X_2} \Phi(F_1(x_2, 0), x_2) d\mu_0(x_2) \right| \\
&\quad + \int_{X_1 \times X_2} \|\Phi'\| \left\| F_2(x_2, \epsilon \left(\frac{dx_1}{dt} + g(x_1, x_2) \right)) - F_1(x_2, 0) \right\| d\mu_\epsilon(x_1, x_2) \\
&\rightarrow 0, \quad \epsilon \rightarrow 0,
\end{aligned}$$

where we have utilized the definition of lift, marginal distribution, mean value theorem, the slave property and the smallness of the perturbation and the uniform continuity of the slave property.

Again, a general continuous functional can be approximated by smooth finite dimensional cylindrical functional just as the case of regular perturbation.

This ends the proof of the theorem. \square

A corollary of this result together with the compactness of \mathcal{IM}_0 and the fact that extremal points of \mathcal{IM} are ergodic leads us to the upper semi-continuity of extreme time averaged statistics.

Theorem 10 (Upper semi-continuity of extremal time averaged statistics, singular version). *Under the same assumption as in the previous theorem, we have for any fixed continuous test functionals $\varphi_0(x_1, x_2)$ and $\varphi_{02}(x_2)$, the extremal statistics are saturated by ergodic invariant measures, i.e., there exist ergodic invariant measures $\nu_\epsilon \in \mathcal{IM}_\epsilon, \nu_0 \in \mathcal{IM}_0$*

such that

$$\begin{aligned} & \sup_{(x_1, x_2) \in X_1 \times X_2} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, \epsilon)(x_1, x_2)) dt \\ &= \int_{X_1 \times X_2} \varphi_0(x_1, x_2) d\nu_\epsilon(x_1, x_2), \\ & \sup_{x_2 \in X_2} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_{02}(S(t, 0)x_2) dt = \int_{X_2} \varphi_{02}(x_2) d\nu_0(x_2). \end{aligned}$$

Moreover, the extremal statistics are upper semi-continuous in parameter, i.e.,

$$\begin{aligned} & \limsup_{\epsilon \rightarrow 0} \sup_{(x_1, x_2) \in X_1 \times X_2} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, \epsilon)(x_1, x_2)) dt \\ & \leq \sup_{x_2 \in X_2} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(F_1(S(t, 0)x_2, 0), S(t, 0)x_2) dt. \end{aligned}$$

Proof: The first half of the proof, saturation by ergodic measures, is exactly the same as the regular perturbation part.

For the second part, the upper semi-continuity, we have, assuming $\mu_\epsilon \rightharpoonup \mathcal{L}\mu_0$,

$$\begin{aligned} & \limsup_{\epsilon \rightarrow 0} \sup_{(x_1, x_2) \in X_1 \times X_2} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(S(t, \epsilon)(x_1, x_2)) dt \\ &= \limsup_{\epsilon \rightarrow 0} \int_{X_1 \times X_2} \varphi_0(x_1, x_2) d\nu_\epsilon \quad (\text{saturation by ergodic measure}) \\ &= \int_{X_1 \times X_2} \varphi_0(x_1, x_2) d\mathcal{L}\mu_0 \quad (\text{weak convergence after lift}) \\ &= \int_{X_2} \varphi_0(F_1(x_2, 0), x_2) d\mu_0 \quad (\text{definition of lift}) \\ &\leq \int_{X_2} \varphi_0(F_1(x_2, 0), x_2) d\nu_0 \quad (\text{definition of } \nu_0) \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(F_1(S(t, 0)x_2, 0), S(t, 0)x_2) dt \\ & \quad (\text{a.s. w.r.t. } \nu_0) \quad (\text{ergodicity of } \nu_0) \\ &= \sup_{x_2 \in X_2} \limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi_0(F_1(S(t, 0)x_2, 0), S(t, 0)x_2) dt. \\ & \quad (\text{definition of } \nu_0) \end{aligned}$$

This ends the proof of the theorem. \square

A consequence of the last result is the upper semi-continuity of the

Nusselt number in Rayleigh-Bénard convection in the singular limit of infinite Prandtl number. More details can be found in [34, 35].

2.8 Remarks on direct applications

2.8.1 An application to NSE: energy dissipation rate per unit mass

For a given dynamical system associated with a physical process (such as incompressible fluid flows governed by the incompressible Navier-Stokes equations, convection governed by the Boussinesq equations etc), there are usually a few statistical quantities that are of most interest either because they occupy a center point in the theory or because these quantities can be easily measured/computed in the experiment/observation etc. Therefore it is of great importance for theorist to investigate these special statistical quantities.

In the case of incompressible fluid flows governed by the Navier-Stokes equations, energy dissipation rate per unit mass is defined as

$$\varepsilon = \frac{\nu}{|\Omega|} < \int_{\Omega} |\nabla \mathbf{u}|^2 d\mathbf{x} >, \quad (2.24)$$

where ν is the kinematic viscosity of the fluid, \mathbf{u} represents the fluid velocity in the domain Ω , $<, >$ represents statistical average, is of great importance in the conventional theory of turbulence a la Kolmogorov.

In the Kolmogorov picture of turbulence, energy is injected at large scale (wave number) and is cascaded (essentially without dissipation) to small scales in a range of wave numbers (the inertial range) until it hits the so-called Kolmogorov dissipation wave length where dissipation dominates. Let

$$e = \frac{1}{|\Omega|} < \int_{\Omega} |\mathbf{u}|^2 d\mathbf{x} >$$

represent average energy per unit mass, then

$$t_{\varepsilon} = \frac{e}{\varepsilon}$$

represents a characteristic time for the dissipation of energy. The characteristic mean velocity is

$$U = \sqrt{2e}$$

and therefore

$$l = Ut_{\varepsilon}$$

can be viewed as the averaged distance travelled by the turbulent eddies until they are dissipated. This also means that the average rate of energy

dissipation should be of the order

$$\varepsilon \sim \frac{U^2}{t_\varepsilon} = \frac{U^3}{l}.$$

Therefore, the energy dissipation rate per unit mass should be independent of the kinematic viscosity ν and scale like U^3/l .

Kolmogorov also argued that the dissipation length l_K (the smallest effective scale) should be a function of the energy dissipation rate per unit volume and the kinematic viscosity only. We then recover via a dimensional argument, the only possible combination is

$$l_d = \left(\frac{\nu^3}{\varepsilon}\right)^{\frac{1}{4}}.$$

Therefore small scales in 3D turbulence is proportional to $\nu^{\frac{3}{4}}$ since ε is independent of ν . This small scale is one of the main obstacles in studying turbulence since this implies that the degrees of freedom for flows at large Reynolds number (defined as $Re = \frac{lU}{\nu}$) (or small viscosity) is proportional to $Re^{9/4}$.

We may also derive the Kolmogorov's energy spectrum scaling ($k^{-5/3}$ law) via a similar dimensional argument. For the inertial range (where energy flows from large to small scales essentially without dissipation) we denote $e_{k,2k}$ the energy between wave numbers k and $2k$ (or eddies of lengths between $l/2$ and l , $l = 1/k$). Following Kolmogorov, we postulate that $e_{k,2k}$ should be a function of the energy dissipation rate per unit volume, ε , and the wave number k only. A dimensional argument then leads to

$$e_{k,2k} = c \left(\frac{\varepsilon}{k}\right)^{\frac{2}{3}}. \quad (2.25)$$

Now let $\mathcal{S}(k)$ be the energy spectrum of the turbulent flow so that

$$e_{k,2k} = \int_k^{2k} \mathcal{S}(\lambda) d\lambda.$$

We see that the energy spectrum must have the $-5/3$ scaling, i.e.,

$$\mathcal{S}(k) \sim \frac{\varepsilon^{\frac{2}{3}}}{k^{\frac{5}{3}}}. \quad (2.26)$$

An important consequence of this $5/3$ scaling is that there may be singularities for the 3D incompressible Euler system. To see this, we recall that the enstrophy spectrum is related to the energy spectrum as

$$\mathcal{S}_1(k) = k^2 \mathcal{S}(k) \sim k^{\frac{1}{3}}$$

since differentiation leads to multiplication by the wave number in the Fourier space. Such kind of energy spectrum clearly indicates the possible singularity although the scaling is only valid for the inertial range for the Navier-Stokes but would be valid for all wave numbers for the Euler system (pushing the Kolmogorov dissipation length scale to zero, or pushing the dissipation wave number to infinity).

The Kolmogorov $\frac{U^3}{l}$ scaling is partially verified in the sense that there are rigorous upper bounds on energy dissipation rate per unit mass [11, 9] that agree with the Kolmogorov scaling for shear driven flow. The advantage of boundary shear driven flow is that there is a natural choice of typical velocity and typical length. Similar result holds for body force driven flow. However, typical velocity in this case is no longer an independent variable of the system or unique and is customarily defined through the viscosity, body force and the geometry of the domain etc.

Since the boundary condition is not zero for shear driven flow, we naturally try to use a “background flow” to homogenize the boundary condition, $\mathbf{u} = \mathbf{v} + \phi$ (see the section on Reynolds equation). This may also be viewed as a mean plus fluctuation decomposition. The problem becomes a variational problem with spectral constraint.

There is no non-trivial lower bound on the energy dissipation rate since all known exact analytical solutions are laminar and the Kolmogorov scaling is expected to be saturated by turbulent flows only. However, for a different physical setup of having injection and suction at the boundary, we may have an exact solution that saturates the Kolmogorov scaling.

2.8.2 An application to RBC: heat transfer in the vertical direction (Nusselt number)

For the case of Rayleigh-Bénard convection, one of the most important physical quantity is the enhancement of heat transport in the vertical direction due to convection versus pure conduction which is characterized by the Nuseelt number:

$$\begin{aligned} Nu &= \frac{J_{total}}{J_{conduct}} \\ &= \frac{\frac{1}{vol} \int (u_3 cT - c\kappa \frac{\partial T}{\partial z}) d\mathbf{x}}{c\kappa(T_{hot} - T_{cold})/h} \quad \text{dimensional c:specific heat} \\ &= 1 + \frac{1}{|\Omega|} \langle \int_{\Omega} u_3 \theta \rangle, \end{aligned} \tag{2.27}$$

where u_3 is the vertical component of the velocity field, $\theta = T - (1 - z)$ is the (perturbative, away from the conduction state $1 - z$) temperature field, and \langle , \rangle represents some statistical averages (usually time

average).

We consider the simplified model of infinite Prandtl number which is relevant for fluids such as silicone oil and the earth's mantle. At large Rayleigh number, the velocity field is expected to be large and hence the problem of infinite Prandtl number convection is a problem of large Peclét number and therefore we anticipate a thermal boundary layer at the top and the bottom. Numerical and laboratory experiments indicate that the temperature in the interior of the system is almost a constant (equal to $\frac{1}{2}$ in the non-dimensional setting). Assume that the boundary layer thickness is δ (δh in the dimensional form) and we have turbulent convection within the boundary layer but not in the interior of the domain (outside the boundary layer). It is then reasonable to postulate that the effective Rayleigh number in the boundary layer should be the critical Rayleigh number $Ra_c \sim 1708$ (this is the so-called marginal stability theory of Malkus (1954) and Howard (1964)). Therefore we have

$$\begin{aligned} Ra_c &= Ra_\delta \\ &= \frac{g\alpha(T_{hot} - T_{cold})(\delta h)^3}{2\nu\kappa} \\ &= \frac{1}{2}Ra\delta^3. \end{aligned}$$

Hence

$$\delta = (2Ra_c)^{\frac{1}{3}}Ra^{-\frac{1}{3}}, \quad (2.28)$$

which further implies that

$$Nu \sim \frac{1}{\delta} \sim Ra^{\frac{1}{3}}. \quad (2.29)$$

The governing equation under Boussinesq approximation is the following *Boussinesq system for Rayleigh-Bénard convection (non-dimensional)*:

$$\begin{aligned} \frac{1}{Pr}\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u}\right) + \nabla p &= \Delta \mathbf{u} + Ra \mathbf{k}\theta, \quad \nabla \cdot \mathbf{u} = 0, \quad \mathbf{u}|_{z=0,1} = 0, \\ \frac{\partial \theta}{\partial t} + \mathbf{u} \cdot \nabla \theta - u_3 &= \Delta \theta, \quad \theta|_{z=0,1} = 0, \end{aligned}$$

where \mathbf{u} is the fluid velocity field, p is the kinematic pressure, θ is the deviation of the temperature field from the pure conduction state $1 - z$, \mathbf{k} is the unit upward vector, Ra is the *Rayleigh number*, Pr is the *Prandtl number*, and the fluids occupy the (non-dimensionalized) region $\Omega = [0, L_x] \times [0, L_y] \times [0, 1]$ with periodicity in the horizontal directions assumed for simplicity.

Although the Boussinesq system does not define a dynamical system due to the well-known difficulty associated with the three dimensional

incompressible fluid Navier-Stokes system, we have eventual regularity at large Prandtl number [33]. Hence the generalized dynamical system can be treated as a usual dynamical system in terms of long time behavior (global attractors, stationary statistical properties) at large Prandtl number. The phase space in this case is given by $H \times L^2$ where H is the divergence free subspace of $(L^2)^3$ with zero normal trace at $z = 0, 1$. The results presented in the previous sections (except the singular perturbation one) apply to the Boussinesq system at large Prandtl number. In particular, for the continuous test functional $\varphi_0(\mathbf{u}, \theta) = 1 + \frac{1}{|\Omega|} \int_{\Omega} u_3(\mathbf{x})\theta(\mathbf{x}) d\mathbf{x}$ defined on $H \times L^2$, we have upper semi-continuity in the Rayleigh number and Prandtl number. We leave the details to the interested reader. For the singular perturbation case, we have $X_1 = H$ and $X_2 = L^2$, and the limit problem is given by the *infinite Prandtl number model*

$$\begin{aligned}\nabla p^0 &= \Delta \mathbf{u}^0 + Ra \mathbf{k} \theta^0, \quad \nabla \cdot \mathbf{u}^0 = 0, \quad \mathbf{u}^0|_{z=0,1} = 0, \\ \frac{\partial \theta^0}{\partial t} + \mathbf{u}^0 \cdot \nabla \theta^0 - u_3^0 &= \Delta \theta^0, \quad \theta^0|_{z=0,1} = 0.\end{aligned}$$

The results regarding singular perturbation also apply with $\epsilon = \frac{1}{Pr}$ and $F_1(\theta, y) = Ra A^{-1}(\mathbf{k}\theta) - A^{-1}(y)$ where A is the Stokes operator with the associated boundary conditions. In particular, we have upper semi-continuity on Prandtl number even in the singular limit of infinite Prandtl number. The interested reader is referred to [33] for more details.

The Malkus-Howard $Ra^{\frac{1}{3}}$ scaling has been partially verified in the sense that there are rigorous upper bound on the Nusselt number that agrees with the $Ra^{\frac{1}{3}}$ scaling (modulus logarithmic correction) for both the infinite Prandtl number model (due to Constantin and Doering), and the case of large but finite Prandtl number [34].

A trivial lower bound for heat transport in the vertical direction is 1 which is satisfied by the pure conduction state. It is still a challenge to derive non-trivial (in the sense of higher than constant scaling in terms of the Rayleigh number) lower bound. Just as in the case of 3D NSE, the difficulty with lower bound is that we usually derive lower bound utilizing special type of analytic solutions. On the other hand, those physical scalings are usually saturated by turbulent solutions and there is no known exact analytic solution that are “turbulent”. An idea here is to consider steady state solutions for the system. This may produce some non-trivial lower bounds since for infinite (or large Prandtl number model), the flows are sluggish and close to steady state. Tools such as boundary layer theory (asymptotic expansion in terms of the Rayleigh number) and bifurcation theory (away from the conduction state) may be useful.

The theory also applies to convection in fluid saturated porous media. The governing equation is the following *Darcy-Oberbeck-Boussinesq system (non-dimensional)* :

$$\begin{aligned} \gamma_a \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} + \nabla p &= Ra_D \mathbf{k}T, \quad \nabla \cdot \mathbf{v} = 0, \quad v_3|_{z=0,1} = 0, \\ \frac{\partial \theta}{\partial t} + \mathbf{v} \cdot \nabla \theta - v_3 &= \Delta \theta, \quad \theta|_{z=0,1} = 0, \end{aligned}$$

where \mathbf{v} is the non-dimensional seepage velocity, p is the non-dimensional kinematic pressure, $T = 1 - z + \theta$ is the non-dimensional temperature. The parameters in the system are given by the *Prandtl-Darcy number* γ_a^{-1} , and the *Rayleigh-Darcy number* Ra_D . Again we assume the fluids occupy the (non-dimensionalized) region $\Omega = [0, L_x] \times [0, L_y] \times [0, 1]$ with periodicity in the horizontal directions assumed for simplicity.

The system is partially/weakly dissipative since there is no dissipation in the velocity equation. In particular, there is no compact absorbing ball in the phase space $H \times L^2$. However, the results presented in the previous sections (except the singular perturbation one) apply to the Darcy-Boussinesq system.

2.9 Maximum entropy principle

We have focused on dissipative systems mostly in the previous sections. Similar issues such as the non-uniqueness of invariant measures and their ergodicity exist for conservative systems. Facing the possibility of non-unique invariant measures, we naturally inquire which IM is the physically relevant one and what is the typical behavior on average. If the system is strongly mixing so that the main tendency is to reduce information, it would be reasonable to speculate that the most probable IM should be the one that maximizes the lack of information among all possible probability measures. The quantity that measures the lack of information is the Shannon information entropy. This is the classical statistical mechanics approach which requires that the (ODE) system satisfies the so-called Liouville property (so that the flow is volume preserving in the phase space) and it possesses good conserved quantities (hence the system should be undamped and unforced).

Motivation via information theory

Definition 5 (Shannon entropy). *Let p be a probability measure on the finite sample space $\mathcal{A} = \{a_1, \dots, a_n\}$,*

$$p = \sum_{i=1}^n p_i \delta_{a_i}, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \tag{2.30}$$

where δ_{a_i} denotes the delta function at a point a_i .

The Shannon entropy $\mathcal{S}(p)$ of the probability p is defined as

$$\mathcal{S}(p) = \mathcal{S}(p_1, \dots, p_n) = -\sum_{i=1}^n p_i \ln p_i. \quad (2.31)$$

The functional \mathcal{S} is called the information-theoretic entropy because of the formal resemblance of the formula for $\mathcal{S}(p)$ defined above to the expression for the entropy of the canonical ensemble in statistical mechanics and because Shannon utilized $\mathcal{S}(p)$ to measure information. It is worthwhile here to briefly recall Shannon's intuition from the theory of communication for the reason that $\mathcal{S}(p)$ measures the lack of information.

The simplest such problem involves representing a *word* in a message as a sequence of binary digits with length n , i.e., one need n -digits with length n to characterize it. The set \mathcal{A}_{2^n} of all words of length n has $2^n = N$ elements and clearly, the amount of information needed to characterize one element is $n = \log_2 N$. Continuing this type of reasoning, it follows that the amount of information needed to characterize an element of any set, \mathcal{A}_N , is $\log_2 N$ for general N (think about binary search). Now consider the situation where a set $\mathcal{A} = \mathcal{A}_{N_1} \cup \dots \cup \mathcal{A}_{N_k}$ where the sets \mathcal{A}_{N_i} are pairwise disjoint from each other with \mathcal{A}_{N_i} having N_i total elements. Set p_i to be given by $p_i = N_i/N$ where $N = \sum N_i$. If we know that an element of \mathcal{A} belongs to some \mathcal{A}_{N_i} , then we need $\log_2 N_i$ additional information to determine it completely. Thus, the average amount of information we need to determine an element provided that we already know the \mathcal{A}_{N_i} to which it belongs is given by

$$\sum_i \frac{N_i}{N} \log_2 N_i = \sum p_i \log_2 p_i + \log_2 N.$$

Recall from our discussion above that $\log_2 N$ is the information that we need to determine an element given the set \mathcal{A} if we do not know to which \mathcal{A}_{N_i} a given element belongs. Thus, the corresponding average lack of information is

$$-\sum p_i \log_2 p_i$$

and we arrive at $\mathcal{S}(p)$ as a measure of lack of information.

Next we show that the Shannon entropy is essentially unique. For this purpose we first introduce the space $\mathcal{PM}_n(\mathcal{A})$ of discrete probability measures on the sample space \mathcal{A} and we recall the following result due to Jaynes.

Lemma 2. *Let H_n be a function defined on the space of discrete probability measures \mathcal{CP}_n , and satisfying the following properties:*

1. $H_n(p_1, \dots, p_n)$ is a continuous function,
2. $A(n) = H_n(1/n, \dots, 1/n)$ is monotonic increasing in n , i.e., H_n increases with increasing uncertainty.
3. Composition law: If the sample space $\mathcal{A} = \{a_1, \dots, a_n\}$ is divided into two subsets $\mathcal{A}_1 = \{a_1, \dots, a_k\}$ and $\mathcal{A}_2 = \{a_{k+1}, \dots, a_n\}$ with probabilities $w_1 = p_1 + \dots + p_k$, $w_2 = p_{k+1} + \dots + p_n$, and conditional probabilities $(p_1/w_1, \dots, p_k/w_1)$, and $(p_{k+1}/w_2, \dots, p_n/w_2)$, then the amount of uncertainty with the information split in this way is the same as it was originally

$$\begin{aligned} H_n(p_1, \dots, p_n) &= H_2(w_1, w_2) + w_1 H_k(p_1/w_1, \dots, p_k/w_1) \\ &\quad + w_2 H_{n-k}(p_{k+1}/w_2, \dots, p_n/w_2). \end{aligned}$$

Then H_n is a positive multiple of the Shannon entropy, i.e.,

$$H_n(p_1, \dots, p_n) = K\mathcal{S}(p_1, \dots, p_n) = -K \sum_{i=1}^n p_i \ln p_i$$

with $K > 0$.

Proof. Since $H_n(q_1, \dots, q_n)$ is a continuous function, it is enough to prove that equation above holds for rational values of q_1, \dots, q_n . Clearly, any set of rational numbers q_1, \dots, q_n , with $0 < q_i \leq 1$, $i = 1, \dots, n$, and $\sum_{i=1}^n q_i = 1$ can be written as $q_i = \nu_i/N$, where $N = \sum_{i=1}^n \nu_i$. Consider now the sample space $\mathcal{A} = \{a_1, \dots, a_N\}$ with probabilities $\{p_1, \dots, p_N\}$. We partition \mathcal{A} into n subsets $\mathcal{A}_i = \{a_{k_{i-1}+1}, \dots, a_{k_i}\}$, $i = 1, \dots, n$, where $k_0 = 0$, and $k_i = k_{i-1} + \nu_i$, $i = 1, \dots, n$. We also denote by w_i the probability associated with \mathcal{A}_i , $w_i = p_{k_{i-1}+1} + \dots + p_{k_i}$. From property 3 it follows that

$$H_N(p_1, \dots, p_N) = H_n(w_1, \dots, w_n) + \sum_{i=1}^n w_i H_{\nu_i}(p_{k_{i-1}+1}/w_i, \dots, p_{k_i}/w_i).$$

In particular, if we let $p_i = 1/N$, $i = 1, \dots, N$, then $w_i = n_i/n = q_i$, and equation above yields

$$H_N(1/N, \dots, 1/N) = H_n(q_1, \dots, q_n) + \sum_{i=1}^n q_i H_{\nu_i}(1/\nu_i, \dots, 1/\nu_i).$$

Utilizing property 2, and recalling that $A(n) = H_n(1/n, \dots, 1/n)$, the equation above reduces to

$$A(N) = H_n(q_1, \dots, q_n) + \sum_{i=1}^n q_i A(\nu_i).$$

In particular, let us set all $\nu_i = \nu$, then $N = n\nu$ and $q_i = \nu_i/N = 1/N$ for all i . In this case the equation above reduces to

$$A(n\nu) = A(n) + A(\nu).$$

However, it is a well known fact that the only continuous function $A(n)$ satisfying this condition is

$$A(n) = K \ln n,$$

where the constant K is chosen to be positive so that $A(n)$ is monotonically increasing, as required by property 2. Combining the results we have

$$H_n(q_1, \dots, q_n) = K \ln n - K \sum_{i=1}^n q_i \ln n_i = -K \sum_{i=1}^n q_i \ln q_i.$$

This concludes the proof of the proposition. \square

For a given probability measure $p \in \mathcal{PM}(\mathcal{A})$, various statistical measurements can be made with respect to this probability measure p . We recall that the expected value, or statistical measurement, of f with respect to p is given by

$$\langle f \rangle_p = \sum_{i=1}^n f(a_i)p_i.$$

The main principle of the subjective probability theory of Jaynes is to seek the least biased probability distribution which is consistent with the constraints imposed by the information as given, for example, by the statistical measurements of certain functions f_j , $j = 1, \dots, r$, as in the equation above. In terms of the Shannon entropy this principle is known as the maximum entropy principle.

Definition 6 (Empirical maximum entropy principle). *Given a set of constraints \mathcal{C} defined by*

$$\mathcal{C} = \{p \in \mathcal{PM}(\mathcal{A}) \mid \langle f_j \rangle_p = F_j, 1 \leq j \leq r\}, \quad (2.32)$$

the least biased probability distribution $p^ \in \mathcal{C}$ is given by maximizing the Shannon entropy $S(p)$ subject to the constraints imposed by the statistical measurements of f_j , $j = 1, \dots, r$ given by \mathcal{C} ,*

$$\max_{p \in \mathcal{C}} S(p) = S(p^*), \quad p^* \in \mathcal{C}. \quad (2.33)$$

Next we consider several examples to illustrate the maximum entropy principle. Here and elsewhere, we will proceed formally in seeking the maximum for $\mathcal{S}(p)$ rather than rigorously prove that such a maximum exists.

Example. Find the least biased probability distribution p on $\mathcal{A} = \{a_1, \dots, a_n\}$ with no additional constraints.

Since in this example there is no additional information available besides the fact that p is a probability measure, we expect that the least biased probability measure is going to be the uniform measure which assigns the same probability to every point in the sample space \mathcal{A} . We will verify this expectation by maximizing the Shannon entropy $\mathcal{S}(p_1, \dots, p_n) = -\sum_{i=1}^n p_i \ln p_i$ subject only to the constraints that $p_i \geq 0$ and that $\sum_{i=1}^n p_i = 1$. By the Lagrange multiplier rule there is a multiplier λ so that the minimum $p = p^*$ satisfies

$$-\nabla_p \mathcal{S} + \lambda \nabla_p \left(\sum_{i=1}^n p_i \right) \Big|_{p=p^*} = 0$$

and subject to the constraint of being a probability measure. Componentwise, the equation above yields n equations that must be satisfied by $p = p^*$,

$$\ln p_i^* + 1 + \lambda = 0, \quad i = 1, \dots, n,$$

and this implies that all the probabilities p_i^* , $i = 1, \dots, n$, are equal. Since the sum of the probabilities p_i^* is one, we conclude that

$$p_i^* = 1/n,$$

and conclude that the least biased measure is the uniform measure.

Example. Find the least biased probability measure that is consistent with a finite number $r \leq n - 1$ of statistical measurements F_j of given functions f_j , $j = 1, \dots, r$,

$$F_j = \langle f_j \rangle_p = \sum_{i=1}^n f_j(a_i)p_i, \quad j = 1, \dots, r.$$

In this example we want to maximize the Shannon entropy

$$\mathcal{S}(p_1, \dots, p_n) = -\sum_{i=1}^n p_i \ln p_i$$

subject to the $r + 1$ constraints

$$F_j = \langle f_j \rangle_p = \sum_{i=1}^n f_j(a_i)p_i, \quad j = 1, \dots, r,$$

$$\sum_{i=1}^n p_i = 1.$$

Notice that we restricted the number of additional constraints to be $r \leq n - 1$ so that the system of algebraic equations above is not over-determined. Once more the Lagrange multiplier rule asserts the existence of $r + 1$ multipliers λ_0 and λ_j , $j = 1, \dots, r$ such that

$$-\nabla_p \mathcal{S} + \sum_{j=1}^r \lambda_j \nabla_p < f_j >_p + \lambda_0 \nabla_p \left(\sum_{i=1}^n p_i \right) \Big|_{p^*} = 0.$$

Componentwise equation above yields a system of n equations for the unknowns p_i^* ,

$$\ln p_i^* = - \sum_{j=1}^r \lambda_j f_j(a_i) - (\lambda_0 + 1), \quad i = 1, \dots, n,$$

and solving for p_i^* we obtain

$$p_i^* = \exp \left(- \sum_{j=1}^r \lambda_j f_j(a_i) - (\lambda_0 + 1) \right).$$

To eliminate the multiplier λ_0 we utilize the constraint that the sum of all the probabilities p_i^* is one. This simplifies the formula for p_i^* in equation above to

$$p_i^* = \frac{\exp \left(- \sum_{j=1}^r \lambda_j f_j(a_i) \right)}{\sum_{i=1}^n \exp \left(- \sum_{j=1}^r \lambda_j f_j(a_i) \right)}.$$

The other constraints λ_i , $i = 1, \dots, r$ are obtained by solving for the remaining constraint equations for the measurements $< f_j >_p$. Interestingly, if we define the partition function $\mathcal{Z}(\vec{\lambda})$ by

$$\mathcal{Z}(\vec{\lambda}) = \sum_{i=1}^n \exp \left(- \sum_{j=1}^r \lambda_j f_j(a_i) \right),$$

then $\mathcal{Z}(\vec{\lambda})$ satisfies

$$-\frac{\partial}{\partial \lambda_j} \ln \mathcal{Z}(\vec{\lambda}) = < f_j >_{p^*}.$$

In this fashion we have recovered the partition function $\mathcal{Z}(\vec{\lambda})$ of statistical mechanics utilizing only the maximum entropy principle from information theory.

The definition given here has a natural generalization when the phase space is a finite dimensional Euclidean space.

$$\mathcal{S}(p) = - \int_{\mathcal{R}^N} p(\vec{X}) \ln(p(\vec{X})) d\vec{X},$$

where p is a probability density function on R^N . This is called differential entropy sometimes since it is not exactly the limit of the discrete entropy.

Given some set of constraints, \mathcal{C} , on the space of probability densities, we define the probability density with the least bias (least information) for further measurements given the constraints in \mathcal{C} through the maximum entropy principle.

Definition 7 (Maximum entropy principle). *Find the probability density $p^* \in \mathcal{C}$, so that*

$$\mathcal{S}(p^*) = \max_{p \in \mathcal{C}} \mathcal{S}(p). \quad (2.34)$$

To illustrate the maximum entropy principle, we show that the Gaussian distributions are the probability densities with the least bias given constraints \mathcal{C} defined by the first and second moments. This fact supports the idea well-known from the central limit theorem that Gaussian densities are the most universal distributions with given first and second moments. In other words, let the constraint set \mathcal{C} be defined by

$$\rho(\lambda) \geq 0, \quad \int_{\mathcal{R}^1} \rho(\lambda) d\lambda = 1,$$

$$\bar{\lambda} = \int_{\mathcal{R}^1} \lambda \rho(\lambda) d\lambda, \quad \sigma^2 = \int_{\mathcal{R}^1} (\lambda - \bar{\lambda})^2 \rho(\lambda) d\lambda.$$

We want to find the probability measure $\rho^*(\lambda)$ that maximizes the entropy under the constraints. The variational derivative of the entropy is given by

$$\frac{\delta \mathcal{S}}{\delta \rho} = -(1 + \ln \rho).$$

The first and second moment constraints are linear functionals of the density ρ so that it is easy to calculate

$$\frac{\delta \bar{\lambda}}{\delta \rho} = \lambda, \quad \frac{\delta \sigma^2}{\delta \rho} = (\lambda - \bar{\lambda})^2.$$

From the Lagrange multiplier principle we have, at the entropy maximum

$$-\frac{\delta \mathcal{S}}{\delta \rho}|_{\rho=\rho^*} = -\mu_0 - \mu_1 \frac{\delta \bar{\lambda}}{\delta \rho}|_{\rho=\rho^*} - \mu_2 \frac{\delta \sigma^2}{\delta \rho}|_{\rho=\rho^*},$$

where μ_0, μ_1, μ_2 are the Lagrange multipliers for the constraints. When combined with the previous three equations, the equation above becomes

$$\ln \rho^* = (-\mu_0 + 1) - \mu_1 \lambda - \mu_2 (\lambda - \bar{\lambda})^2.$$

This defines ρ^* as a Gaussian probability density with mean $\bar{\lambda}$ and variance σ^2 so that $\rho^* = \rho_{\bar{\lambda}, \sigma}(\lambda)$, the Gaussian density defined.

In the case there are prior information encoded in a prior distribution p_0 , the measurement of the information should take into account this prior and the appropriate object is then the *relative entropy* given by

$$S(p, p_0) = - \int p(\vec{X}) \ln \left(\frac{p(\vec{X})}{p_0(\vec{X})} \right).$$

Some times the relative entropy is written as

$$\mathcal{P}(p, p_0) = \int p \ln \left(\frac{p}{p_0} \right)$$

which is commonly utilized in studying predictability [20].

2.10 Application to ODEs

Here we briefly introduce the statistical theory for large ODE systems based on the maximum entropy principle. Application to basic geophysical flows will appear in the next section.

Consider a large nonlinear systems of ordinary differential equations (ODEs)

$$\begin{aligned} \frac{d\vec{X}}{dt} &= \vec{F}(\vec{X}), \quad \vec{X} \in \mathcal{R}^N, \quad \vec{F} = (F_1, \dots, F_N), \quad N \gg 1 \quad (2.35) \\ \vec{X}|_{t=0} &= \vec{X}_0. \end{aligned}$$

There are two main ingredients in doing statistical mechanics: Liouville property and conserved quantities.

Assume that the right hand side of the ODE satisfies the following.

Definition 8 (Liouville Property). *A vector field $\vec{F}(\vec{X})$ is said to satisfy the Liouville property if it is divergence free, i.e.,*

$$\nabla_{\vec{X}} \vec{F} = \sum_{j=1}^N \frac{\partial F_j}{\partial X_j} = 0.$$

An important consequence of the Liouville property is that it implies the flow map associated with ODE is volume preserving or measure preserving on the phase space.

It is then easily checked that the Liouville equation reduces to

$$\frac{\partial p}{\partial t} + \vec{F} \cdot \nabla_{\vec{X}} p = 0, \quad (2.36)$$

and therefore

$$p(\vec{X}, t) = p_0(S^{-1}(t)(\vec{X})), \quad (2.37)$$

where p_0 is the initial pdf.

As a consequence, any smooth function of the pdf p is transported by the flow, i.e.,

$$\frac{\partial G(p)}{\partial t} + \vec{F} \cdot \nabla_{\vec{X}} G(p) = 0.$$

This implies that

$$\frac{d}{dt} \int_{\mathcal{R}^N} G(p(\vec{X}, t)) d\vec{X} = 0,$$

which further implies the invariance of the entropy under the flow.

We now introduce the second ingredient for doing equilibrium statistical mechanics for ODEs. We assume that the ODE system possesses L conserved quantities $E_l(\vec{X}(t))$, i.e.,

$$E_l(\vec{X}(t)) = E_l(\vec{X}_0), \quad 1 \leq l \leq L.$$

These conserved quantities could be the truncated energy, enstrophy, or higher moments for the truncated quasi-geostrophic equations, or the truncated energy and linear momentum for the truncated Burgers-Hopf equation, or the Hamiltonian and angular momentum for the point vortex system etc. The ensemble average of these conserved quantities with respect to a probability density function p is defined as

$$\bar{E}_l = \langle E_l \rangle_p \equiv \int_{\mathcal{R}^N} E_l(\vec{X}) p(\vec{X}) d\vec{X}, \quad 1 \leq l \leq L.$$

We naturally expect that these ensemble averages are conserved in time. Indeed, we have the following result.

Proposition 1.

$$\langle E_l \rangle_{p(t)} = \langle E_l \rangle_{p_0}, \text{ for all } t.$$

Proof. The proof is a simple application of the Liouville property and a change of variable.

$$\begin{aligned}
\langle E_l \rangle_{p(t)} &= \int_{\mathcal{R}^N} E_l(\vec{X}) p(\vec{X}, t) d\vec{X} \\
&= \int_{\mathcal{R}^N} E_l(\vec{X}) p_0((\Phi^t)^{-1}(\vec{X})) d\vec{X} \\
&= \int_{\mathcal{R}^N} E_l(\Phi^t(\vec{Y})) p_0(\vec{Y}) d\vec{Y} \\
&= \int_{\mathcal{R}^N} E_l(\vec{Y}) p_0(\vec{Y}) d\vec{Y} \\
&= \langle E_l \rangle_{p_0},
\end{aligned}$$

where we have performed the change of variable $\vec{Y} = (\Phi^t)^{-1}(\vec{X})$, utilized the Liouville property which implies that this change of variable is volume preserving, and the assumption that E_l is conserved in time.

We recall that *the Shannon entropy* \mathcal{S} for the probability density function $p(\vec{X})$ on \mathcal{R}^N that is absolutely continuous with respect to the Lebesgue measure by

$$\mathcal{S}(p) = - \int_{\mathcal{R}^N} p(\vec{X}) \ln p(\vec{X}) d\vec{X}. \quad (2.38)$$

Note that the Shannon entropy is exactly identical with the Boltzmann entropy in the statistical mechanics of gas particles.

The question now is how to pick our probability measure p with the least bias for doing future measurements. We invoke the maximum entropy principle and claim that the probability density function p with the least bias for conducting further measurements should be the one that maximizes the Shannon entropy subject to the constraint set of measurements (conserved quantities) as defined above. More precisely, let

$$\mathcal{C} = \left\{ p(\vec{X}) \geq 0, \int_{\mathcal{R}^N} p(\vec{X}) d\vec{X} = 1, \langle E_l \rangle_p = \bar{E}_l, 1 \leq l \leq L \right\}.$$

The maximum entropy principle predicts that the most probable probability density function $p^* \in \mathcal{C}$ is the one that satisfies

$$\mathcal{S}(p^*) = \max_{p \in \mathcal{C}} \mathcal{S}(p). \quad (2.39)$$

We will see that the entropy maximizer is in fact an invariant measure of the ODE system and hence the maximizer process can be viewed

as a process over the set of invariant measures. Therefore, the maximum entropy principle may be viewed as a methodology in picking the physically most relevant (or least biased) invariant measures among all possible ones.

In order to compute the most probable probability density function, we invoke the Lagrange multiplier method. The variational derivative of the entropy with respect to the probability density function is computed as

$$\frac{\delta S(p)}{\delta p} = -(1 + \ln p),$$

and the variational derivatives of the constraints are computed easily as

$$\frac{\delta \bar{E}_l}{\delta p} = E_l(\vec{X}),$$

since the constraints are linear in p . Thus the Lagrange multiplier method dictates that the most probable state must satisfy

$$-(1 + \ln p^*) = \theta_0 + \sum_{l=1}^L \theta_l E_l(\vec{X}),$$

where θ_0 is the Lagrange multiplier for the constraint that p must be a probability density function, and θ_l is the Lagrange multiplier for \bar{E}_l for each l respectively. Equation above can be rewritten as

$$p^*(\vec{X}) = c \exp\left(-\sum_{l=1}^L \theta_l E_l(\vec{X})\right),$$

or equivalently, we may write the most probable probability density function in the form analogous to the *Gibbs measure* in statistical mechanics for gas particle systems

$$p^*(\vec{X}) = \mathcal{G}_{\vec{\theta}}(\vec{X}) = C^{-1} \exp\left(-\sum_{l=1}^L \theta_l E_l(\vec{X})\right),$$

provided that the constraints are normalizable, i.e.,

$$C = \int_{\mathcal{R}^N} \exp\left(-\sum_{l=1}^L \theta_l E_l(\vec{X})\right) d\vec{X} < \infty$$

and the θ_l 's are the Lagrange multipliers so that $\mathcal{G}_{\vec{\theta}}$ satisfies the constraints. \square

Recall that each probability density function is transported by the vector field \vec{F} thus satisfying the Liouville equation. Also, the entropy

is conserved in time. Thus we expect the Gibbs measure to solve the steady state Liouville equation since the Gibbs measure maximizes the entropy within the constraint set. Indeed, this is a special case of the following general fact.

Proposition 2. *Let E_j , $1 \leq j \leq J$ be conserved quantities of the ODE system. For any smooth function $G(E_1, \dots, E_J)$, $G(E_1, \dots, E_J)$ is a steady state solution to the Liouville equation. In particular, the most probable probability density function given above is a steady state solution to the Liouville equation, i.e.,*

$$\vec{F} \cdot \nabla_{\vec{X}} \mathcal{G}_{\vec{\theta}} = 0.$$

Hence G is an invariant measure of the ODE system.

Proof. Since $E_j(\vec{X}(t))$ is conserved in time, we have

$$0 = \frac{d}{dt} E_j(\vec{X}(t)) = \frac{\partial E_j}{\partial t} + \vec{F} \cdot \nabla_{\vec{X}} E_j = \vec{F} \cdot \nabla_{\vec{X}} E_j, \text{ for all } j.$$

Thus

$$\vec{F} \cdot \nabla_{\vec{X}} G(E_1, \dots, E_J) = \sum_{j=1}^J \vec{F} \cdot \nabla_{\vec{X}} E_j \frac{\partial G}{\partial E_j} = 0.$$

Finally setting $G = \mathcal{G}_{\vec{\theta}}$ we deduce the conclusion of the proposition.

An important consequence of a probability density function satisfying the steady state Liouville equation is that the associated probability measure is an *invariant probability measure* on the phase space.

This completes the proof of the proposition. \square

These Gibbs measures which maximize the Shannon information theoretical entropy have a very good chance of being ergodic since they are very likely to be extreme points of the set of invariant measures with the given constraints.

2.11 Application to basic geophysical systems

In this section we set-up statistical mechanics and apply the theory from the previous section to suitable truncations of the barotropic quasi-geostrophic equations without damping and forcing but with large scale mean velocity

$$\frac{\partial q}{\partial t} + J(\psi, q) = 0,$$

$$\frac{dV}{dt}(t) = -\frac{1}{4\pi^2} \int \frac{\partial h}{\partial x} \psi',$$

where

$$q = \Delta\psi' + h + \beta y,$$

$$\omega = \Delta\psi',$$

$$\psi = -V(t)y + \psi',$$

$$q' = \Delta\psi' + h,$$

$$\vec{v} = \nabla^\perp\psi = \begin{pmatrix} V(t) \\ 0 \end{pmatrix} + \nabla^\perp\psi',$$

and $\Omega = [0, 2\pi] \times [0, 2\pi]$ is the region occupied by the fluid. The derivation of the equation of the large scale mean flow is based on energy conservation [20].

Here we present the statistical theory for truncated quasi-geostrophic dynamics which is carried out in several steps:

- I. We make a finite dimensional truncation of the barotropic quasi-geostrophic equations. This amounts to making a Galerkin approximation, where the equation is projected onto a finite dimensional subspace. Since we are working with two-dimensional periodic boundary conditions, the truncation is readily accomplished with the standard Fourier basis.
- II. We verify that the finite dimensional truncated equations have the following special properties:
 - 1. The truncated equations conserve the truncated energy and enstrophy.
 - 2. The truncated equations satisfy the Liouville property. This implies that the truncated equations define an incompressible flow in phase space. Therefore the resulting flow map is measure preserving, and we can “transport” measures with the flow.
- III. We apply the equilibrium statistical mechanics theory for ODEs set up in the previous section to the truncated system, i.e., find the probability measure, the Gibbs measure, that maximizes the Shannon entropy subject to the constraints of fixed energy and enstrophy.
- IV. We compute the mean state of the Gibbs measure of the finite dimensional truncated equations and study the limit behavior of this mean state in the limit when the dimension $N \rightarrow \infty$. This is the continuum limit of the system.

We now introduce the spectrally truncated systems which approximate the barotropic system. The truncated dynamic equations are a finite dimensional approximation of the barotropic quasi-geostrophic equations which are obtained by projecting the barotropic quasi-geostrophic equations onto a subspace involving only a finite number of Fourier modes. In another words, the truncated system is a Galerkin approximation of the barotropic quasi-geostrophic equations with the use of standard Fourier basis. To derive the truncated dynamics equations we proceed as follows. We first introduce the Fourier series expansions of the truncated small scale stream function ψ'_Λ , the truncated vorticity ω_Λ , and the truncated topography h_Λ in terms of the basis

$$\begin{aligned} B_\Lambda &= \left\{ \exp\left(i\vec{k} \cdot \vec{x}\right) \mid 1 \leq |\vec{k}|^2 \leq \Lambda \right\}, \\ \psi'_\Lambda &\equiv \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \hat{\psi}_{\vec{k}}(t) e^{i\vec{x} \cdot \vec{k}} = - \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{|\vec{k}|^2} \hat{\omega}_{\vec{k}}(t) e^{i\vec{x} \cdot \vec{k}}, \\ h_\Lambda &\equiv \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \hat{h}_{\vec{k}}(t) e^{i\vec{x} \cdot \vec{k}}, \\ \omega_\Lambda &\equiv \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \hat{\omega}_{\vec{k}}(t) e^{i\vec{x} \cdot \vec{k}} = \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} (-|\vec{k}|^2 \hat{\psi}_{\vec{k}}(t)) e^{i\vec{x} \cdot \vec{k}}, \end{aligned} \quad (2.40)$$

where the amplitudes $\hat{\psi}_{\vec{k}}$, $\hat{\omega}_{\vec{k}}$ and $\hat{h}_{\vec{k}}$ satisfy the reality conditions $\hat{\psi}_{-\vec{k}} = \hat{\psi}_{\vec{k}}^*$, $\hat{\omega}_{-\vec{k}} = \hat{\omega}_{\vec{k}}^*$, and $\hat{h}_{-\vec{k}} = \hat{h}_{\vec{k}}^*$ with $*$ denoting complex conjugation here. We also assume that the mean value h_0 of the topography is zero so that the solvability condition for the associated steady state equation is automatically satisfied.

We denote by P_Λ the orthogonal projection onto the finite dimensional space $V_\Lambda = \text{span}\{B_\Lambda\}$. The truncated dynamical equations are obtained by projecting the barotropic quasi-geostrophic equations onto V_Λ ,

$$\frac{\partial q'_\Lambda}{\partial t} + \beta \frac{\partial \psi'_\Lambda}{\partial x} + V \frac{\partial q'_\Lambda}{\partial x} + P_\Lambda (\nabla^\perp \psi'_\Lambda \cdot \nabla q'_\Lambda) = 0, \quad q'_\Lambda = \omega_\Lambda + h_\Lambda, \quad (2.41)$$

$$\frac{dV}{dt} - \frac{1}{4\pi^2} \int h_\Lambda \frac{\partial \psi'_\Lambda}{\partial x} = 0. \quad (2.42)$$

Then, we get the finite dimensional system of ordinary differential equation (ODE), *the truncated dynamical equations*, for the Fourier coeffi-

cients with $1 \leq |\vec{k}|^2 \leq \Lambda$,

$$\begin{aligned} \frac{d\hat{\omega}_{\vec{k}}}{dt} - \frac{i\beta k_1}{|\vec{k}|^2} \hat{\omega}_{\vec{k}} + iV k_1 (\hat{\omega}_{\vec{k}} + \hat{h}_{\vec{k}}) \\ - \sum_{\substack{\vec{l}+\vec{m}=\vec{k}, \\ |\vec{l}|^2 \leq \Lambda, |\vec{m}|^2 \leq \Lambda}} \frac{\vec{l}^\perp \cdot \vec{m}}{|\vec{l}|^2} \hat{\omega}_{\vec{l}} (\hat{\omega}_{\vec{m}} + \hat{h}_{\vec{m}}) = 0, \end{aligned} \quad (2.43)$$

$$\frac{dV(t)}{dt} - i \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} k_1 \frac{\hat{h}_{-\vec{k}} \hat{\omega}_{\vec{k}}}{|\vec{k}|^2} = 0. \quad (2.44)$$

It is easy to see that the inviscid unforced quasi-geostrophic equations possess two or more robust conserved quantities. An advantage of utilizing Fourier truncation is that the linear and quadratic conserved quantities survive the truncation. More precisely we have the following.

Proposition 3. *The truncated energy E_Λ and enstrophy \mathcal{E}_Λ are conserved in the finite dimensionally truncated dynamics, where*

$$E_\Lambda = \frac{1}{2} V^2 + \frac{1}{2} \frac{1}{4\pi^2} \int |\nabla^\perp \psi'_\Lambda|^2 d\vec{x} = \frac{1}{2} V^2 + \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} |\vec{k}|^2 |\hat{\psi}_{\vec{k}}|^2,$$

$$\mathcal{E}_\Lambda = \beta V + \frac{1}{2} \frac{1}{4\pi^2} \int q'^2_\Lambda d\vec{x} = \beta V + \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} | - |\vec{k}|^2 \hat{\psi}_{\vec{k}} + \hat{h}_{\vec{k}} |^2.$$

Proof. In order to show that the truncated total energy E_Λ is conserved in time it is equivalent to proving that the time derivative of it is identically zero for all time. For this purpose we take the time derivative of the truncated total energy E_Λ and utilize the truncated dynamic equations (2.41) and (2.42).

$$\begin{aligned} \frac{d}{dt} E_\Lambda &= V \frac{dV}{dt} - \frac{1}{4\pi^2} \int \psi'_\Lambda \frac{\partial \omega_\Lambda}{\partial t} d\vec{x} \\ &= V \frac{dV}{dt} - \frac{1}{4\pi^2} \int \psi'_\Lambda \frac{\partial q'_\Lambda}{\partial t} d\vec{x} \\ &= V \frac{1}{4\pi^2} \int h_\Lambda \frac{\partial \psi'_\Lambda}{\partial x} + \beta \frac{1}{4\pi^2} \int \psi'_\Lambda \frac{\partial \psi'_\Lambda}{\partial x} d\vec{x} + V \frac{1}{4\pi^2} \int \psi'_\Lambda \frac{\partial q'_\Lambda}{\partial x} d\vec{x} \\ &\quad + \frac{1}{4\pi^2} \int P_\Lambda (\nabla^\perp \psi'_\Lambda \cdot \nabla q'_\Lambda) \psi'_\Lambda \\ &= V \frac{1}{4\pi^2} \int h_\Lambda \frac{\partial \psi'_\Lambda}{\partial x} + V \frac{1}{4\pi^2} \int \psi'_\Lambda \frac{\partial h_\Lambda}{\partial x} d\vec{x} \\ &\quad + \frac{1}{4\pi^2} \int (\nabla^\perp \psi'_\Lambda \cdot \nabla q'_\Lambda) \psi'_\Lambda \\ &= 0, \end{aligned}$$

where we have used the truncated dynamic equations (2.41) and (2.42) and repeatedly carried out integration by parts.

The conservation of the truncated total enstrophy can be shown in a similar fashion. We end the proof of the proposition. \square

It is easy to see that the truncated system possesses exact solutions having linear $q_\Lambda - \psi_\Lambda$ relation

$$\bar{q}_\Lambda = \mu \bar{\psi}_\Lambda.$$

This condition is the same as

$$\Delta \bar{\psi}'_\Lambda + h_\Lambda = \mu \bar{\psi}'_\Lambda, \quad \bar{V} = -\frac{\beta}{\mu}.$$

The nonlinear stability of these type of steady states can be studied using the Arnold-Kruskal techniques. More precisely, we consider a linear combination of the truncated energy and enstrophy to form a positive quadratic form for perturbations. Let $(\delta q_\Lambda, \delta V)$ be the perturbations, we then have

$$\mu E_\Lambda(q_\Lambda, V) + \mathcal{E}_\Lambda(q_\Lambda, V) = \mu E_\Lambda(\bar{q}_\Lambda, \bar{V}) + \mathcal{E}_\Lambda(\bar{q}_\Lambda, \bar{V}) + \mathcal{W}_\mu(\delta q_\Lambda, \delta V),$$

where

$$\begin{aligned} \mathcal{W}_\mu(\delta q_\Lambda, \delta V) &= \frac{\mu}{2} (\delta V)^2 + \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \left(1 + \frac{\mu}{|\vec{k}|^2}\right) (\delta q_\Lambda)^2 \\ &= \frac{\mu}{2} (V - \bar{V})^2 + \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} |\vec{k}|^2 (\mu + |\vec{k}|^2) (\hat{\psi}_{\vec{k}} - \bar{\psi}_{\vec{k}})^2. \end{aligned} \quad (2.45)$$

It then follows that the steady state solution $(\bar{q}_\Lambda, \bar{V})$ is nonlinearly stable for $\mu > 0$ in general, or for $\mu > -1$ when $V \equiv 0$.

Next we verify the Liouville property for the truncated equations (2.41) and (2.42). These equations can be written as a system of ODE's with real coefficients. Indeed, let

$$S = \left\{ \vec{k}_1, \dots, \vec{k}_M \right\}$$

be a defining set of modes for $\{1 \leq |\vec{k}|^2 \leq \Lambda\}$ satisfying

$$\vec{k} \in S \Rightarrow -\vec{k} \notin S, \quad S \cup (-S) = \{1 \leq |\vec{k}|^2 \leq \Lambda\}.$$

Let $N = 2M + 1$ and define $\vec{X} \in \mathcal{R}^N$,

$$\begin{aligned} \vec{X} &\equiv (V, \operatorname{Re} \hat{\psi}_{\vec{k}_1}, \operatorname{Im} \hat{\psi}_{\vec{k}_1}, \dots, \operatorname{Re} \hat{\psi}_{\vec{k}_M}, \operatorname{Im} \hat{\psi}_{\vec{k}_M}), \\ \vec{X} &\in R^{2M+1} = \mathcal{R}^N, \quad N \gg 1. \end{aligned}$$

We then notice that each point \vec{X} in a big space \mathcal{R}^N represents the entire state of the finite dimensional system. With these notations, the truncated dynamic equations (2.43) and (2.44) can be written in a more compact form

$$\frac{d\vec{X}}{dt} = \vec{F}(\vec{X}), \quad \vec{X}|_{t=0} = \vec{X}_0, \quad (2.46)$$

with the vector field \vec{F} satisfying the property

$$F_j(\vec{X}) = F_j(X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_N),$$

i.e., F_j does not depend on X_j . This immediately implies the Liouville property. In fact, \vec{F} satisfies the so-called detailed Liouville property in the sense that it is satisfied locally in terms of the Fourier coefficients.

It is easy to see why the relation above is true. It is obvious that F_1 is independent of $X_1 = V$. Observe that F_{2j} and F_{2j+1} correspond to $\hat{\psi}_{\vec{k}_j}$ in the truncated equation. It is obvious that the contributions from the linear terms in the truncated equation either are independent of $\hat{\psi}_{\vec{k}_j}$ or cause a rotation of $X_{2j} = \text{Re } \hat{\psi}_{\vec{k}_j}$ and $X_{2j+1} = \text{Im } \hat{\psi}_{\vec{k}_j}$. As for the nonlinear term in the truncated equation, the contribution from $\hat{\psi}_{\vec{k}_j}$ and $\hat{\psi}_{-\vec{k}_j}$ (the same as from X_{2j} and X_{2j+1}) is zero since the restriction on the summation indices requires either $\vec{l} = 0, \vec{m} = \vec{k}_j$, or $\vec{l} = \vec{k}_j, \vec{m} = 0$, or $\vec{l} = -\vec{k}_j, \vec{m} = 2\vec{k}_j$, or $\vec{l} = 2\vec{k}_j, \vec{m} = -\vec{k}_j$. In either case we have $\vec{l}^\perp \cdot \vec{m} = 0$.

We now have all the ingredients for the application of the equilibrium statistical theory introduced in the previous section. We base the theory on the truncated energy and enstrophy from which are the two conserved quantities in this truncated system.

Let α and θ be the Lagrange multipliers for the enstrophy \mathcal{E}_Λ and energy E_Λ respectively. Let

$$\mu = \frac{\theta}{\alpha}, \quad \text{if } \alpha \neq 0.$$

Then the Gibbs measure is given by

$$\begin{aligned} \mathcal{G}_{\alpha, \theta} = c \exp(-\alpha(\beta V + \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} | -|\vec{k}|^2 \hat{\psi}_{\vec{k}} + \hat{h}_{\vec{k}}|^2) \\ - \theta(\frac{1}{2} V^2 + \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} |\vec{k}|^2 |\hat{\psi}_{\vec{k}}|^2)), \end{aligned}$$

where α and θ are determined by the average enstrophy and energy constraints. Due to the resemblance of the most probable distribution

above to the thermal equilibrium ensemble, θ is sometimes referred to as the inverse temperature and α as the thermodynamic potential.

In order to guarantee that the Gibbs measure is a probability measure we need to ensure that the coefficients of the quadratic terms are negative, i.e.,

$$\begin{aligned} \alpha|\vec{k}|^4 + \theta|\vec{k}|^2 &> 0, \text{ for all } \vec{k} \text{ satisfying } |\vec{k}|^2 \leq \Lambda, \text{ and} \\ \theta &> 0, \text{ if } V \neq 0. \end{aligned}$$

This implies either

(A)

$$\alpha > 0, \quad \mu > 0,$$

or

(B)

$$V \equiv 0, \quad \alpha > 0, \quad \mu > -1,$$

or

(C)

$$\alpha < 0, \quad \mu < -\Lambda, \quad \theta > 0.$$

Obviously, case (C) is a spurious condition due to the truncation only and hence is not physically relevant.

Under the realizability condition, we may introduce the mean state

$$\bar{V} = -\frac{\beta}{\mu}, \quad \bar{\psi}_{\vec{k}} = \frac{\hat{h}_{\vec{k}}}{\mu + |\vec{k}|^2},$$

and

$$\bar{\psi}'_{\Lambda}(\vec{x}, t) = \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \bar{\psi}_{\vec{k}} e^{i\vec{x} \cdot \vec{k}}.$$

We then observe that $(\bar{V}, \bar{\psi}'_{\Lambda})$ satisfies the linear relation and hence it is nonlinearly stable under the realizability condition [20]. If $V \equiv 0$, we can also verify the nonlinearly stability of $\bar{\psi}'_{\Lambda}$ if the realizability condition is satisfied.

We may rewrite the Gibbs measure as

$$\begin{aligned} \mathcal{G}_{\alpha, \mu} &= c \exp(-(\alpha \mathcal{E}_{\Lambda} + \alpha \mu E_{\Lambda})) \\ &= c \exp(-\alpha(\mathcal{E}_{\Lambda} + \mu E_{\Lambda})) \\ &= c_{\alpha, \mu} \exp(-\alpha \mathcal{W}_{\mu}(\delta q, \delta V)) \\ &= c_{\alpha, \mu} \exp(-\alpha(\frac{\mu}{2}(V - \bar{V})^2 \\ &\quad + \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} |\vec{k}|^2 (|\vec{k}|^2 + \mu)(\hat{\psi}_{\vec{k}} - \bar{\psi}_{\vec{k}})^2)), \end{aligned} \tag{2.47}$$

or equivalently

$$\mathcal{G}_{\alpha,\mu}(\vec{X}) = \prod_{j=1}^N \mathcal{G}_{\alpha,\mu}^j(X_j).$$

Thus, the invariant Gibbs measure for the dynamics is a product of Gaussian measures with a non-zero mean. Moreover, $(\bar{V}, \bar{\psi}'_\Lambda)$ is exactly the ensemble average, or mean state, of (V, ψ'_Λ) with respect to the Gibbs measure

$$\langle \vec{X} \rangle = \int_{\mathcal{R}^N} \vec{X} \mathcal{G}_{\alpha,\mu}(\vec{X}) d\vec{X} = (\bar{V}, \bar{\hat{\psi}}_{k_1}, \dots, \bar{\hat{\psi}}_{k_M}).$$

This implies, assuming ergodicity for the Gibbs measure, the following remarkable prediction

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_{T_0}^{T_0+T} \psi'_\Lambda(\vec{x}, t) dt = \bar{\psi}'_\Lambda,$$

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_{T_0}^{T_0+T} V(t) dt = \bar{V} = -\frac{\beta}{\mu}.$$

Hence the statistical theory predicts that the time average of the solutions to the truncated equations converge to the nonlinearly stable exact steady state solutions to the truncated equations. In other words, a specific coherent large scale mean flow will emerge from the dynamics of the truncated system with generic initial data after computing a long time average.

Recall that we are interested in the dynamics of the quasi-geostrophic equations. It is only in the limit of $\Lambda \rightarrow \infty$ that we recover the solution to the barotropic quasi-geostrophic equations from solutions to the truncated equations. Thus a natural question to ask is the asymptotic behavior of the invariant Gibbs measures as well as their mean states at large Λ . We are also interested in taking the continuum limit as Λ approaches infinity.

It is apparent that all predictions of the truncated system must depend on the truncation eigenvalue Λ , thus we occasionally append a suffix Λ in order to distinguish the dependence on Λ and avoid confusion.

There is a subtle issue when we take the limit $\Lambda \rightarrow \infty$, i.e., how do we pick $\alpha = \alpha_\Lambda$, $\mu = \mu_\Lambda$ satisfying the energy and the enstrophy constraints as $\Lambda \rightarrow \infty$.

We start with two observations. First, the ensemble average of the truncated energy and enstrophy need not satisfy the exact energy/enstrophy constraint. It is in the limit of cut-off wave number approaching infinity, i.e., $\Lambda \rightarrow \infty$, that these two constraints have to be satisfied, i.e.,

$$\lim_{\Lambda \rightarrow \infty} \langle E_\Lambda \rangle = E_0,$$

$$\lim_{\Lambda \rightarrow \infty} \langle \mathcal{E}_\Lambda \rangle = \mathcal{E}_0.$$

Second, there is a constraint on the energy E_0 and enstrophy \mathcal{E}_0 imposed. Apparently \mathcal{E}_0 must be at least the minimum enstrophy associated to the given energy level, i.e.,

$$\mathcal{E}_0 \geq \min_{E(\psi)=E_0} \mathcal{E}(\psi) = \mathcal{E}_*(E_0).$$

In the case of equality, any state satisfying the energy/enstrophy constraint must be an enstrophy minimizing (selective decay) state. These enstrophy minimizing states are completely characterized [20] and there is not much randomness present. Thus, we assume that the given enstrophy level \mathcal{E}_0 is strictly more than the minimum enstrophy associated with the given energy level E_0 , i.e.,

$$\mathcal{E}_0 > \mathcal{E}_*(E_0) = \min_{E(\psi)=E_0} \mathcal{E}(\psi).$$

Since the constraints are given in terms of the ensemble average of the enstrophy and energy with respect to the Gibbs measure which is a Gaussian, we explicitly calculate that

$$\int V^2 \mathcal{G}_{\alpha,\mu} = (\alpha\mu)^{-1} + \frac{\beta^2}{\mu^2},$$

$$\int |\hat{\psi}_{\vec{k}}|^2 \mathcal{G}_{\alpha,\mu} = (\alpha|\vec{k}|^2(\mu + |\vec{k}|^2))^{-1} + \overline{|\hat{\psi}_{\vec{k}}|^2}.$$

We then observe that the ensemble average of the energy and the enstrophy naturally separate into two parts, a mean part that corresponds to the mean state and a fluctuation part

$$\langle E_\Lambda \rangle = \langle E_\Lambda \rangle_{\mathcal{G}} = \bar{E}_\Lambda + E'_\Lambda,$$

$$\begin{aligned} \bar{E}_\Lambda &= \frac{1}{2} \left(\frac{\beta^2}{\mu^2} + \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{|\vec{k}|^2 |\hat{h}_{\vec{k}}|^2}{(\mu + |\vec{k}|^2)^2} \right), \\ E'_\Lambda &= \frac{\alpha^{-1}}{2} \left(\mu^{-1} + \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{\mu + |\vec{k}|^2} \right), \end{aligned}$$

while for the enstrophy

$$\langle \mathcal{E}_\Lambda \rangle = \langle \mathcal{E}_\Lambda \rangle_{\mathcal{G}} = \bar{\mathcal{E}}_\Lambda + \mathcal{E}'_\Lambda,$$

$$\bar{\mathcal{E}}_\Lambda = -\frac{\beta^2}{\mu} + \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{\mu^2 |\hat{h}_{\vec{k}}|^2}{(\mu + |\vec{k}|^2)^2},$$

$$\mathcal{E}'_\Lambda = \frac{\alpha^{-1}}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{|\vec{k}|^2}{\mu + |\vec{k}|^2}.$$

For the case without large scale mean flow, i.e., $V \equiv 0, \beta = 0$, we have

$$\bar{E}_\Lambda = \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{|\vec{k}|^2 |\hat{h}_{\vec{k}}|^2}{(\mu + |\vec{k}|^2)^2},$$

$$E'_\Lambda = \frac{\alpha^{-1}}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{\mu + |\vec{k}|^2},$$

and

$$\bar{\mathcal{E}}_\Lambda = \frac{1}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{\mu^2 |\hat{h}_{\vec{k}}|^2}{(\mu + |\vec{k}|^2)^2},$$

$$\mathcal{E}'_\Lambda = \frac{\alpha^{-1}}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{|\vec{k}|^2}{\mu + |\vec{k}|^2}.$$

Observe that the fluctuation part of the energy is isotropic and independent of the mean state. We may estimate this part of the energy by replacing the summation by integration for large Λ ,

$$\begin{aligned} E'_\Lambda &= \frac{\alpha^{-1}}{2} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{\mu + |\vec{k}|^2} + \frac{(\alpha\mu)^{-1}}{2} \\ &\cong \frac{\alpha^{-1}}{2} 2\pi \int_1^{\sqrt{\Lambda}} \frac{|\vec{k}|}{\mu + |\vec{k}|^2} d|\vec{k}| + \frac{(\alpha\mu)^{-1}}{2} \\ &= \frac{\alpha^{-1}}{2} \pi \ln(\mu + |\vec{k}|^2) \Big|_1^{\sqrt{\Lambda}} + \frac{(\alpha\mu)^{-1}}{2} \\ &= \frac{\alpha^{-1}}{2} \pi \ln\left(\frac{\mu + \Lambda}{\mu + 1}\right) + \frac{(\alpha\mu)^{-1}}{2}. \end{aligned}$$

Likewise, the fluctuation part of the total enstrophy can be estimated as

$$\begin{aligned} \mathcal{E}'_\Lambda &= \frac{1}{2\alpha} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{|\vec{k}|^2}{\mu + |\vec{k}|^2} \\ &\cong \frac{1}{2\alpha} 2\pi \int_1^{\sqrt{\Lambda}} \frac{|\vec{k}|^3}{\mu + |\vec{k}|^2} d|\vec{k}| \\ &= \frac{\pi}{\alpha} \int_1^{\sqrt{\Lambda}} \frac{|\vec{k}|(\mu + |\vec{k}|^2) - \mu|\vec{k}|}{\mu + |\vec{k}|^2} d|\vec{k}| \\ &= \frac{\pi}{2\alpha} \left(\Lambda - 1 - \mu \ln\left(\frac{\mu + \Lambda}{\mu + 1}\right) \right). \end{aligned}$$

Immediately, we can see that if the parameters α and μ are bounded independent of Λ , the fluctuation part of the energy diverges to infinite as Λ approaches infinity which contradicts the energy constraint. Thus, one of the parameters α and μ must approach infinity. In the Appendix

(also see the arguments below), we will show that the parameter μ must remain bounded in the continuum limit. This is of course expected since large μ corresponds to small geophysical influence which contradicts physical reality. In what follows, we consider the case that μ is bounded and hence α approaches infinity. This means that the variance of the Gaussian measures for individual wave numbers $|\vec{k}|$ tends to zero, i.e., the fluctuations are suppressed at all wave numbers. For the case without large scale mean flow, the fluctuations are suppressed away from the ground energy shell since μ could approach -1 in this case (see the realizability condition). We will give more details in the sequel.

As we shall demonstrate below, the asymptotic behavior of the Gibbs measure and its mean state depend heavily on the geophysical effects. We will elaborate on the easy case with large scale mean flow only.

By the realizability condition and the explicit formula for the ensemble energy given above, we observe that $E'_\Lambda > 0$ which further implies that

$$\frac{\beta^2}{2\mu_\Lambda^2} \leq \bar{E}_\Lambda \leq E_0$$

asymptotically for large Λ . This leads to the following asymptotic lower bound on μ_Λ ,

$$\mu_\Lambda \geq \frac{\beta}{\sqrt{2E_0}}.$$

Substituting this into the second equation in the enstrophy formula, we see that

$$\bar{\mathcal{E}}_\Lambda \geq -\frac{\beta^2}{\mu_\Lambda} \geq -\beta\sqrt{2E_0}$$

asymptotically for large Λ . Thus, we arrive at an asymptotic bound on the fluctuation part of the ensemble enstrophy,

$$\mathcal{E}'_\Lambda \leq \mathcal{E}_0 - \bar{\mathcal{E}}_\Lambda \leq \mathcal{E}_0 + \beta\sqrt{2E_0}.$$

Now that μ_Λ is bounded above and is positive by the realizability condition, hence we deduce

$$\lim_{\Lambda \rightarrow \infty} \frac{\ln \frac{\mu_\Lambda + \Lambda}{\mu_\Lambda + 1}}{\Lambda} = \lim_{\Lambda \rightarrow \infty} \frac{\ln \Lambda}{\Lambda} = 0.$$

Therefore, we have

$$\alpha_\Lambda \geq \frac{\pi\Lambda}{2(\mathcal{E}_0 + \beta\sqrt{2E_0})}$$

asymptotically for large Λ . Utilizing these bounds and the boundedness of μ we deduce

$$E'_\Lambda \rightarrow 0, \quad \text{as } \Lambda \rightarrow \infty.$$

Thus, we conclude that *all energy must reside in the mean state asymptotically for large Λ* . This leads to the following choice of parameters and asymptotic behavior of the parameters as well as the mean states.

Step 1. (The choice of $\mu = \mu_\Lambda$) For the given energy E_0 and Λ , we can find a unique $\mu = \mu_\Lambda > 0$ such that

$$\bar{E}_\Lambda = E(\bar{\psi}'_{\mu_\Lambda}, \bar{V}_{\mu_\Lambda}) = E_0.$$

Then $\bar{\psi}'_{\mu_\Lambda}, \bar{V}_{\mu_\Lambda}$ satisfy

$$\Delta \bar{\psi}'_{\mu_\Lambda} + h_\Lambda = \mu_\Lambda \bar{\psi}'_{\mu_\Lambda},$$

$$\bar{V}_{\mu_\Lambda} = -\frac{\beta}{\mu_\Lambda}.$$

It is easy to see that μ_Λ is a non-decreasing function of Λ since for fixed μ , the energy associated with the mean state \bar{E}_Λ calculated earlier is a monotonic increasing function in Λ because more terms are added. Moreover, it is easy to check that there exists $\mu (= \mu_\infty) > 0$ such that

$$\lim_{\Lambda \rightarrow \infty} \mu_\Lambda = \mu$$

in a monotonic decreasing fashion. We observe that we have positive temperature for all energy level in this case.

Step 2. (The limit of the mean states) The mean states also converge. It is easy to check that

$$\bar{\psi}'_{\mu_\Lambda} \rightarrow \bar{\psi}'_\mu = \sum \frac{\hat{h}_{\vec{k}}}{\mu + |\vec{k}|^2} e^{i\vec{k} \cdot \vec{x}},$$

$$\bar{V}_{\mu_\Lambda} \rightarrow \bar{V}_\mu = -\frac{\beta}{\mu},$$

where $\mu = \mu_\infty > 0$ is the unique μ in $(0, \infty)$ such that the energy constraint is met by the limit mean state $(\bar{\psi}'_\mu, \bar{V}_\mu)$, i.e.

$$E(\bar{\psi}'_\mu, \bar{V}_\mu) = E_0.$$

We also observe that the limit mean state satisfies the limit mean field equation

$$\Delta \bar{\psi}'_\mu + h = \mu \bar{\psi}'_\mu,$$

$$\bar{V}_\mu = -\frac{\beta}{\mu},$$

and therefore nonlinearly stable according to an Arnold-Kruskal type argument.

Step 3. (The choice of $\alpha = \alpha_\Lambda$ and the enstrophy constraint) We know that for the given energy E_0 there exists the minimal enstrophy, $\mathcal{E}_*(E_0)$ which is equal to $\mathcal{E}(\bar{\psi}'_\mu, \bar{V}_\mu)$ since $(\bar{\psi}'_\mu, \bar{V}_\mu)$ is the enstrophy minimizing state (or selective decay state) with the topography h and β . On the other hand,

$$\lim_{\Lambda \rightarrow \infty} \bar{\mathcal{E}}_\Lambda = \mathcal{E}(\bar{\psi}'_\mu, \bar{V}_\mu) = \mathcal{E}_*(E_0).$$

Hence we have, for large Λ ,

$$\mathcal{E}_0 > \bar{\mathcal{E}}_\Lambda = \mathcal{E}(\bar{\psi}'_{\mu_\Lambda}, \bar{V}_{\mu_\Lambda}).$$

We now pick α_Λ so that the enstrophy constraint is satisfied for all truncations, i.e.,

$$\mathcal{E}_0 = \langle \mathcal{E}_\Lambda \rangle = \bar{\mathcal{E}}_\Lambda + \mathcal{E}'_\Lambda.$$

This amounts to requiring

$$\alpha_\Lambda \cong \frac{\frac{\pi}{2}\Lambda}{\mathcal{E}_0 - \bar{\mathcal{E}}_\Lambda}$$

for $\Lambda \gg 1$. Note that $\alpha \rightarrow \infty$ as $\Lambda_\Lambda \rightarrow \infty$. This limit is a *non-extensive thermodynamic limit* because the energy of individual fluctuation is not constant.

Step 4. (The energy constraint) It is easy to see that the energy fluctuation goes to zero as $\Lambda \rightarrow \infty$. More precisely,

$$E'_\Lambda \cong \frac{\pi \ln \Lambda}{2\Lambda} \rightarrow 0.$$

Therefore we conclude that the energy constraint is satisfied asymptotically.

The case of without large scale mean flow but with generic topography, as well as the case without geophysical effects can be studied similarly.

We may then conclude that the complete statistical mechanics theory predicts that the most probable mean state is the unique selective decay state with given energy E_0 in the case with large scale mean flow [20].

Now we see why the geophysical flows are more suitable for applying statistical theory than the ordinary 2-D Euler flows, which have no beta plane effect, no topography. In the absence of those geophysical effects, the energy is distributed only over the fluctuating part, and this makes the limiting procedure a very different one where fluctuations in the large scale dominate.

As an exercise, one can work out the statistical theory for the simplified one layer models.

We now give a sketch of the proof that the quotient of the Lagrange multiplier θ for the truncated energy and the Lagrange multiplier α for the truncated enstrophy, i.e., $\frac{\theta}{\alpha} = \mu = \mu_\Lambda$, in the Gibbs measure must be bounded from above independent of Λ if the energy and enstrophy constraints hold asymptotically for the Gibbs measure.

In what follows we assume all limits exist. Otherwise we just go through a subsequence argument.

Indeed, according to the energy and enstrophy formula, if $\mu_\Lambda \rightarrow \infty$, then all energy will be concentrated in fluctuation in high modes which shift to infinite wave number. This is intuitively not consistent with the emergence of large scale coherent structures. We may justify our intuition through the following by way of contradiction argument (BWOC).

Now suppose that

$$\mu_\Lambda \rightarrow \infty,$$

we then have, according to the energy and enstrophy formula,

$$\begin{aligned}\lim_{\Lambda \rightarrow \infty} \overline{E}_\Lambda &= 0, \\ \lim_{\Lambda \rightarrow \infty} \overline{\mathcal{E}}_\Lambda &= \frac{1}{2} \|h\|_{L^2}^2 \geq 0.\end{aligned}$$

Thus,

$$\frac{\mathcal{E}_0}{E_0} \cong \frac{\langle \mathcal{E}_\Lambda \rangle}{\langle E_\Lambda \rangle} \cong \frac{\frac{1}{2} \|h\|_{L^2}^2 + \mathcal{E}'_\Lambda}{E'_\Lambda} \geq \frac{\sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{|\vec{k}|^2}{\mu_\Lambda + |\vec{k}|^2}}{\sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{\mu_\Lambda + |\vec{k}|^2} + \frac{1}{\mu_\Lambda}}.$$

We thus observe that the generalized Dirichlet quotient (enstrophy over energy) depends only on μ_Λ asymptotically under the assumption that μ is unbounded.

It is easy to see that the fluctuation part of the energy and enstrophy in each fixed mode approaches zero as μ_Λ approaches infinity. Therefore, only the contribution from the high modes are relatively important in the last expression above. This means we may heuristically neglect the lower modes below a wave number $\sqrt{\Lambda_j}$ for any fixed Λ_j . This leads to a lower bound of Λ_j the generalized Dirichlet quotient which further leads to a contradiction since j is arbitrary. We will formalize this heuristic argument below.

On the other hand, the contribution from the high modes in the generalized Dirichlet quotient could be small as well. What we need to show is that the low modes are relatively smaller than the high modes.

For this purpose, let us fix a j and consider the following decomposition of the denominator of the last expression in the generalized Dirichlet quotient into the low modes ($|\vec{k}|^2 < \Lambda_j$) and the high modes ($|\vec{k}|^2 \geq \Lambda_j$) parts, namely,

$$\begin{aligned} \sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{\mu_\Lambda + |\vec{k}|^2} + \frac{1}{\mu_\Lambda} &= \left\{ \frac{1}{\mu_\Lambda} + \sum_{1 \leq |\vec{k}|^2 < \Lambda_j} \frac{1}{\mu_\Lambda + |\vec{k}|^2} \right\} \\ &\quad + \left\{ \sum_{\Lambda_j \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{\mu_\Lambda + |\vec{k}|^2} \right\} = I + II. \end{aligned}$$

Observe that for each fixed eigenvalue Λ_k , there are at most $2\pi\sqrt{\Lambda_k}$ number of modes corresponding to this eigenvalue. Consequently we have the following crude upper bound on the contribution from the low modes,

$$I \leq \frac{1}{\mu_\Lambda} (2\pi\sqrt{\Lambda_j} \times j) = \frac{2j\pi\sqrt{\Lambda_j}}{\mu_\Lambda}.$$

On the other hand, an approximation using integration leads to the following estimate on the contribution from the high modes,

$$II \cong \frac{\pi}{2} \ln\left(\frac{\mu_\Lambda + \Lambda}{\mu_\Lambda + \Lambda_j}\right).$$

We shall demonstrate below that I is of lower order to II in all scenarios. Without loss of generality (by going through a subsequence if necessary) we assume that the limit $\frac{\mu_\Lambda}{\Lambda}$ exists.

Case 1. First we assume that the μ_Λ grows much faster than Λ , i.e.,

$$\frac{\mu_\Lambda}{\Lambda} \rightarrow \infty.$$

In this case

$$\ln\left(\frac{\mu_\Lambda + \Lambda}{\mu_\Lambda + \Lambda_j}\right) = \ln\left(1 + \frac{\Lambda - \Lambda_j}{\mu_\Lambda + \Lambda_j}\right) \cong \frac{\Lambda - \Lambda_j}{\mu_\Lambda + \Lambda_j}.$$

Therefore, combining the last four equations, we have asymptotically

$$\frac{I}{II} \leq \frac{4j\sqrt{\Lambda_j}}{\Lambda - \Lambda_j} \frac{\mu_\Lambda + \Lambda_j}{\mu_\Lambda} \cong \frac{4j\sqrt{\Lambda_j}}{\Lambda - \Lambda_j} \rightarrow 0.$$

Case 2. Suppose now that μ_Λ and Λ grow at the same rate, i.e.,

$$\frac{\mu_\Lambda}{\Lambda} \rightarrow c \neq 0.$$

In this case

$$\ln\left(\frac{\mu_\Lambda + \Lambda}{\mu_\Lambda + \Lambda_j}\right) \cong \ln\left(\frac{c+1}{c}\right).$$

Therefore, we have

$$\frac{I}{II} \leq \frac{4j\sqrt{\Lambda_j}}{\mu_\Lambda \ln\left(\frac{c+1}{c}\right)} \rightarrow 0.$$

Case 3. Lastly we consider the case when μ_Λ grows slower than Λ , i.e.,

$$\frac{\mu_\Lambda}{\Lambda} \rightarrow 0.$$

In this case

$$\ln\left(\frac{\mu_\Lambda + \Lambda}{\mu_\Lambda + \Lambda_j}\right) \rightarrow \infty.$$

Hence we have

$$\frac{I}{II} \rightarrow 0.$$

In any case, we proved that the last relationship is true. Utilizing this relationship and the generalized Dirichelet quotient we deduce

$$\frac{\mathcal{E}_0}{E_0} \geq \frac{\sum_{1 \leq |\vec{k}|^2 \leq \Lambda} \frac{|\vec{k}|^2}{\mu_\Lambda + |\vec{k}|^2}}{\sum_{\Lambda_j \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{\mu_\Lambda + |\vec{k}|^2}} \geq \frac{\sum_{\Lambda_j \leq |\vec{k}|^2 \leq \Lambda} \frac{|\vec{k}|^2}{\mu_\Lambda + |\vec{k}|^2}}{\sum_{\Lambda_j \leq |\vec{k}|^2 \leq \Lambda} \frac{1}{\mu_\Lambda + |\vec{k}|^2}} \geq \Lambda_j.$$

This is a contradiction since j is arbitrary. Of course this implies that μ cannot be unbounded, i.e., μ_Λ must be bounded independent of Λ .

A last comment is that the approximation of the summation by integration is a valid one in the sense that the error is of lower order.

3 Remarks on time dependent statistics

Here we consider the more difficult problem of time dependent statistics which is characterized by the time dependent statistical solutions. This is necessary since a given system may not approach a statistical equilibrium or it takes an extremely long time (physically unrealistic) to reach stationarity and hence what we observe must be transient.

3.1 Definition, existence

As we mentioned earlier, the key idea in the definition of time dependent statistical solutions μ_t is the statistical formulation of the underlying dynamical system, namely the Liouville type equation that we introduced in Chapter 1.

However, we occasionally require further properties of the time dependent statistical solutions (mostly energy type), especially if the dynamical system is generated by a PDE.

To fix ideas, we will work on the Navier-Stokes system for incompressible homogeneous Newtonian fluids although the setting can be easily generalized to other systems. For the case of no-slip boundary condition with body force \mathbf{f} , we require an additional energy inequality.

Definition 9. A family $\{\mu_t, t \in [0, T]\}$ of probability measures on H is called a time dependent statistical solution of the Navier-Stokes equations on H with initial data μ_0 and forcing term \mathbf{f} if μ_t satisfies the weak Liouville type equation

$$\int_H \varphi(\mathbf{u}) d\mu_t(\mathbf{u}) = \int_H \varphi(\mathbf{u}) d\mu_0(\mathbf{u}) + \int_0^t \int_H \langle F(s, \mathbf{u}), \varphi'(\mathbf{u}) \rangle d\mu_s(\mathbf{u}) ds \quad (3.1)$$

for all cylindrical test functional φ and the energy type inequality

$$\begin{aligned} & \int_H \|\mathbf{u}\|^2 d\mu_t(\mathbf{u}) + 2\nu \int_0^t \int_H \|\nabla \mathbf{u}\|^2 d\mu_s(\mathbf{u}) ds \\ & \leq \int_H \|\mathbf{u}\|^2 d\mu_0(\mathbf{u}) + \int_0^t \int_H (\mathbf{f}(s), \mathbf{u}) d\mu_s(\mathbf{u}) ds, \quad \forall t \in [0, T] \end{aligned} \quad (3.2)$$

with the additional assumption that $\int_H \varphi(\mathbf{u}) d\mu_t(\mathbf{u})$ is measurable on $[0, T]$, $\int_H \|\mathbf{u}\|^2 d\mu_t(\mathbf{u}) \in L^\infty(0, T)$, $\int_H \|\nabla \mathbf{u}\|^2 d\mu_s(\mathbf{u}) \in L^1(0, T)$.

The existence of time dependent statistical solutions to the Navier-Stokes system is classical under the assumption that the initial probability measure has finite kinetic energy [11]. It is also known that the solution is unique in the 2D case just as in the single trajectory scenario.

Example of slow convergence to stationary solutions (algebraic and logarithmic):

$$\begin{aligned} \frac{du}{dt} &= -u^{2k+1}, \\ \frac{du}{dt} &= -\text{sign}(u)u^2 \exp(-\text{sign}(u)\frac{1}{u}). \end{aligned}$$

3.2 Applications to NSE

3.2.1 Reynolds equation for the average flow

For complex systems with turbulent/chaotic behavior such as the Navier-Stokes system at large Reynolds number, it is practically impossible to predict the detailed behavior of the system due to abundant instability. However, certain bulk quantities characterized by statistical averages are

much more robust and predictable. Here we consider the mean velocity $\bar{\mathbf{u}} = \langle \mathbf{u} \rangle$ which is the mean of an ensemble (spatial average with respect to a time dependent statistical solution). We now decompose the velocity field into the mean part $\bar{\mathbf{u}}$ and the fluctuation part $\mathbf{u}' = \mathbf{u} - \bar{\mathbf{u}}$, taking the average of the Navier-Stokes equation with respect to the time-dependent statistical solution (or simply choose a test functional of the form (\mathbf{u}, \mathbf{w}) in the Liouville type equation), we arrive at the following *Reynolds equation in functional form*

$$\bar{\mathbf{u}}(t) + \int_0^t \{ \nu A \bar{\mathbf{u}}(s) + B(\bar{\mathbf{u}}(s)) + \overline{B(\mathbf{u}'(s))} \} ds = \bar{\mathbf{u}}(0) + \int_0^t \mathbf{f}(s) ds, \quad (3.3)$$

where A is the Stokes operator and B is the usual nonlinear term in the incompressible Navier-Stokes equation.

Many of the large eddy simulation (LES) theory is related to modelling the $\overline{B(\mathbf{u}')}$ term in terms of the mean velocity so that the Reynolds equation becomes an equation for the mean velocity only.

3.2.2 Moment closure problem

Another approach to studying the statistics is by looking at the evolution of various moments. It is easy to see that the first moment will depend on the second moment, the second moment will rely on the third moment and so on. The moment equations never close automatically. One may think about truncating the moment equations artificially at level N and hope that the solution to the truncated system converge to the original moments as $N \rightarrow \infty$. Unfortunately we do not know if the convergence is true (we only know the convergence at small Reynolds number [30]). How to get good (accurate and efficient) moment approximation is a great challenge.

3.2.3 Fluctuation dissipation theory (FDT)

In many situations, we are interested in the response of a system which is at statistical equilibrium under perturbation. The most straightforward approach is a linear response theory. How to find accurate and efficient ways to estimate the leading order linear change of various physically important statistical quantities utilizing statistics of the un-perturbed is of great interest in applications. The interested reader is referred to [17] for more details and further readings.

Appendix: some useful theorems

Theorem 11 (Kakutani-Riesz representation theorem). *Let X be a locally compact Hausdorff space. Let Λ be a positive continuous linear*

functional on $C_c(X)$, the space of compactly supported, continuous real-valued functions on X . Then there exists a σ -algebra \mathcal{M} in X that contains all Borel sets in X , and there also exists a unique positive regular and complete measure μ on \mathcal{M} which represents Λ in the sense that

1.

$$\Lambda(f) = \int_X f d\mu, \forall f \in C_c(X)$$

2.

$$\mu(K) < \infty, \forall \text{compact set } K \subset X.$$

(See for instance Lax [15].)

Theorem 12 (Prokhorov's Theorem). Let $\{\mu_j\}$ be a sequence of Borel probability measures on a complete separable metric space X . Then $\{\mu_j\}$ has a weakly convergent subsequence if and only if, for each $\epsilon > 0$, there exists a compact set K_ϵ in X such that $\mu_j(K_\epsilon) \geq 1 - \epsilon, \forall j$. (See for instance Billingsley [5].)

Theorem 13 (Krein-Milman theorem). Let K be a compact subset of a locally convex topological vector space X and let E be the set of extremal points of K . Then the closed convex hull of E is the same as the closed convex hull of K . If K is convex, then K itself equals to the closed convex hull of E . (See for instance Lax [15].)

Theorem 14 (Arzela-Ascoli theorem). Let (X, d_1) be a compact metric space and let (Y, d_2) be a complete metric space. Let $C(X, Y)$ be the space of continuous functions from X to Y equipped with the metric

$$d(f_1, f_2) = \sup_{x \in X} d_2(f_1(x), f_2(x)).$$

Let $S \subset C(X, Y)$ be pointwise compact (i.e., the set $\{f(x); f \in S\}$ is pre-compact in Y for all $x \in X$) and equi-continuous. Then S is pre-compact in $C(X, Y)$. (See for instance Bourbaki [6].)

Theorem 15 (Aubin compactness theorem). Let X_0, X, X_1 be three Banach spaces and suppose that X_0 and X_1 are reflexive, $X_0 \subset X$ with compact injection, and $X \subset X_1$ with continuous injection. Let $T > 0$ and $p_0, p_1 > 1$. Consider the space

$$\mathcal{Y} = \{u \in L^{p_0}(0, T; X_0); u' = \frac{du}{dt} \in L^{p_1}(0, T; X_1)\}$$

endowed with the norm

$$\|u\|_{\mathcal{Y}} = \|u\|_{L^{p_0}(0, T; X_0)} + \|u'\|_{L^{p_1}(0, T; X_1)}.$$

Then the injection of \mathcal{Y} into $L^{p_0}(0, T; X_0)$ is compact. (See for instance Temam [29].)

Theorem 16 (Pre-compactness in L^p). Let $\Omega \subset R^n$ and $p \in [1, \infty)$. A set $S \subset L^p(\Omega)$ is pre-compact if and only if for each $\varepsilon > 0$, there exists a $\delta > 0$ and a compact subset $K \subset \Omega$ such that for all $u \in S$ and $h \in R^n$ with $|h| < \delta$ we have

$$\int_{\Omega} |\tilde{u}(x+h) - \tilde{u}(x)|^p dx < \varepsilon^p,$$

$$\int_{\Omega-K} |u(x)|^p \leq \varepsilon^p,$$

where \tilde{u} is the trivial (zero outside Ω) extension of u . (See for instance Adams [1].)

Theorem 17 (pre-compactness in Banach space valued L^p). Let X, Y be Banach spaces such that $Y \subset X$ with the imbedding being compact. Let $\mathcal{G} \subset L^p(R^1; X) \cap L^1(R^1; Y), p > 1$ be bounded in $L^p(R^1; X)$ and $L^1(R^1; Y)$. Suppose that

$$\int_{R^1} \|g(t+s) - g(s)\|_X^p ds \rightarrow 0, \quad \text{as } t \rightarrow 0$$

uniformly for $g \in \mathcal{G}$ and there exists $L > 0$ such that

$$\text{supp}(g) \subset [-L, L], \quad \forall g \in \mathcal{G}.$$

Then \mathcal{G} is pre-compact in $L^p(R^1; X)$. (See for instance Temam [27].)

Theorem 18 (Hahn-Banach theorem). Let p be a real-valued function on a real vector/linear space X . Suppose p satisfies

$$p(x+y) \leq p(x) + p(y), \quad \forall x, y \in X, \quad p(\alpha x) = \alpha p(x), \quad \forall \alpha \geq 0, x \in X.$$

Let f be a real valued functional defined on a subspace Y of X with

$$f(x) \leq p(x), \quad \forall x \in Y.$$

Then there exists a real valued linear functional F on X for which

$$F(x) = f(x), \quad \forall x \in Y; \quad F(x) \leq p(x), \quad \forall x \in X.$$

(See for instance Lax [15].)

Acknowledgement. The author acknowledges financial support from NSF and FSU, and the hospitality of Fudan University in Shanghai, and The Chinese University of Hong Kong.

References

- [1] R.A. Adams, Sobolev spaces, New York, Academic Press, 1975.
- [2] L. Arnold, Random Dynamical Systems, Springer-Verlag, Berlin, 1998.
- [3] G.K. Batchelor, An Introduction to Fluid Dynamics, Cambridge Univ. Press, 1967.
- [4] M.B. Bekka, M. Mayer, Ergodic Theory and Topological Dynamics of Group Actions on Homogeneous Spaces, London Math. Soc. Lecture Notes Ser. 269, Cambridge University Press, 2000.
- [5] P. Billingsley, Weak convergence of measures: applications in probability. SIAM, Philadelphia, 1971.
- [6] N. Bourbaki, General topology. Reading, Mass.: Addison-Wesley Pub. Co., 1966.
- [7] B. Cushman-Roisin, Introduction to Geophysical Fluid Dynamics, Prentice Hall, 1994.
- [8] G. Da Prato, J. Zabczyk, Ergodicity for infinite dimensional systems. Cambridge; New York: Cambridge University Press, 1996.
- [9] C.R. Doering, J.D. Gibbon, Applied Analysis of the Navier-Stokes Equations, Cambridge University Press, Cambridge, UK, 1995.
- [10] E. Weinan, Stochastic hydrodynamics. Current developments in mathematics, 2000, 109–147, Int. Press, Somerville, MA, 2001.
- [11] C. Foias, O. Manley, R. Rosa and R. Temam, Navier-Stokes Equations and Turbulence, Cambridge University Press, Cambridge, UK, 2001.
- [12] A.E. Gill, Atmosphere-Ocean Dynamics, New York: Academic Press, 1982.
- [13] J. Kaipio, E. Somersalo, Statistical and Computational Inverse Problems, Applied Mathematical Sciences 160, Springer-Verlag, New York, 2005.
- [14] A. Lasota, M.C. Mackey, Chaos, Fractals, and Noise, stochastic aspects of dynamics, 2nd ed., Springer-Verlag, New York, 1994.
- [15] P.D. Lax, Functional Analysis, New York: Wiley, 2002.
- [16] A. Majda, Introduction to PDEs and Waves for the Atmosphere and Ocean, AMS, 2002.
- [17] A.J. Majda, R. Abramov, M. Grote, Information theory and stochastics for multiscale nonlinear systems, CRM monograph series 25, American Mathematical Society, 2005.

- [18] A. Majda and A. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge Univ. Press, 2001.
- [19] A. Majda and X. Wang, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*, Cambridge Univ. Press, 2006.
- [20] A. Majda and X. Wang, The emergence of large scale coherent structure under small scale random bombardments, CPAM, Vol. 59, Issue 4 (2006), 467–500, 2006.
- [21] A.S. Monin, A.M. Yaglom, *Statistical fluid mechanics; mechanics of turbulence*, English ed. updated, augmented and rev. by the authors. MIT Press, Cambridge, Mass., 1975.
- [22] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1987.
- [23] D. Ruelle, Nonequilibrium statistical mechanics near equilibrium: computing higher order terms, *Nonlinearity*, 11, 1998, 5–18.
- [24] D. Ruelle, General linear response formula in statistical mechanics, and the fluctuation-dissipation theorem far from equilibrium, *Phys. Lett. A*, 245, 1998, 220–224.
- [25] D. Ruelle, Smooth dynamics and new theoretical ideas in non-equilibrium statistical mechanics, *J. Stat. Phys.* 95, 1999, 393–468.
- [26] Ya.G. Sinai, *Topics in Ergodic Theory*, Princeton University Press, 1994.
- [27] R.M. Temam, *Navier-Stokes equations and nonlinear functional analysis*, 2nd ed., Philadelphia, SIAM, 1995.
- [28] R.M. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, 2nd ed., Springer-Verlag, New York, 1997.
- [29] R.M. Temam, *Navier-Stokes Equations*, AMS Chelsea, Providence, Rhode Island, 2000.
- [30] M.I. Vishik, A.V. Fursikov, *Mathematical Problems of Statistical Hydromechanics*, Kluwer Acad. Publishers, Dordrecht/Boston/London, 1988.
- [31] P. Walters, *An introduction to ergodic theory*. Springer-Verlag, New York, 2000.
- [32] X. Wang, Lecture Notes on Introduction to Geophysical Fluid Dynamics, Shanghai Summer School on Mathematics Lecture Notes, 2004.
- [33] X. Wang, Stationary statistical properties of Rayleigh-Bénard convection at large Prandtl number, CPAM, 2007.
- [34] X. Wang, Bound on the vertical heat transport at large Prandtl number, *Physica D*, 2007.

- [35] X. Wang, Upper semi-continuity of stationary statistical properties of dissipative systems, preprint, submitted to Discrete and Continuous Dynamical Systems, special issue dedicated to Prof. Ta-Tsien Li on the occasion of his 70th birthday, 2007.
- [36] L.S. Young, What are SRB measures and which dynamical systems have them? J. Stat. Phy., vol.108, no. 5/6, 733–754, 2002.

The Compressible Euler System in Two Space Dimensions*

Yuxi Zheng

Department of Mathematics

The Pennsylvania State University, USA

E-mail: yzheng@math.psu.edu

Abstract

We analyze the self-similar isentropic irrotational Euler system via the hodograph transformation. We diagonalize the system of equations in the phase space. We use these equations to analyze the binary interaction of arbitrary planar rarefaction waves, which includes the classical problem of a wedge of gas expanding into vacuum.

Introduction

Multi-dimensional systems of conservation laws have been a major research area for many decades and substantial progress has been made in recent years. We expect that there will be great progress in the near future. This chapter is intended to introduce the topic at the first-year graduate student level and end at the research height. The primary system is the Euler system for ideal compressible gases. The issues are structures of solutions as well as the existence and uniqueness of solutions to various initial and boundary value problems. Our focus will be on special initial data that yield solutions with distinctive physical features. These types of initial data are collectively called Riemann data. Typical types of physical features are continuous expansive and compressive waves as well as regular and Mach reflections. In another word, we shall study Riemann problems in two space dimensions and explain features of shock reflections. The mathematical tools are theory of characteristics, elliptic estimates, boundary and corner regularity, fixed point theorems, numerical simulations, and asymptotic analysis. The problems are mathematically challenging and aerodynamically significant.

This chapter is organized as follows:

1. The compressible Euler system.
2. The characteristics decomposition of the pseudo-steady case.

*The author is partly supported by NSF DMS-0603859.

3. The hodograph transformation and the interaction of rarefactions.
4. Local solutions for quasi-linear systems.
5. Invariant regions for systems.
6. The pressure gradient system.
7. Open problems.

1 Physical phenomena and mathematical problems

1.1 Euler system in n dimensions

We recall that the full (or adiabatic) Euler system for an ideal fluid takes the form:

$$\begin{cases} \rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \\ (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + pI) = 0, \\ (\rho E)_t + \nabla \cdot (\rho E \mathbf{u} + p\mathbf{u}) = 0, \end{cases} \quad (1.1)$$

where $E := \frac{1}{2}|\mathbf{u}|^2 + e$, and e is the internal energy. For a polytropic gas, there holds $e = \frac{1}{\gamma-1} \frac{p}{\rho}$, where $\gamma > 1$ is the adiabatic gas constant. See Courant and Friedrichs [7]. The notations are that $\nabla \cdot$ is divergence in \mathbb{R}^n ($n = 1, 2, 3$), $\mathbf{u} \otimes \mathbf{u} = (u_i u_j)$ (an $n \times n$ matrix), and I is the identity matrix. Subscript t denotes partial derivative in t .

1.2 Phenomena

Shock waves can form in compressible gases. We see shock waves forming ahead of a flying bullet, or multiple shocks around an airplane moving at supersonic speed, see separate graphics from Van Dyke [38].

1.3 Mathematical treatment

There are a couple of problems designed as “space probes” to peek into the mysteries of the one-dimensional and two-dimensional gas dynamics.

Riemann’s shock tube problem (~ 1860). Assume that system (1.1) is only one-dimensional, say, the x direction. Assume that the initial data $(\rho, \mathbf{u}, p) =: U$ consist of two constant states separated by the position $x = 0$; i.e.,

$$U(0, x) = \begin{cases} U_-, & x < 0, \\ U_+, & x > 0, \end{cases} \quad (1.2)$$

where U_- and U_+ are two constant vectors in \mathbb{R}^n . Find the solutions to (1.1), (1.2) in $t > 0$.

This mathematical problem can be “implemented” physically as follows, see Figure 1.1. Imagine an infinitely long cylindrical tube filled

with a gas in two different states, separated by an extremely thin membrane. The membrane is broken at $t = 0$, and you are there to watch the subsequent motion of the gas.



Figure 1.1 The setup of Riemann's shock tube experiment.

Mach's oblique-shock-reflection problem (~ 1878). A shock wave, called incident shock I, of system (1.1) traveling down a flat ramp hits the ground at time $t = 0$, see Figure 1.2, where the shock is drawn at time $t = -1$. Find the subsequent reflection/diffraction patterns of the shock wave.

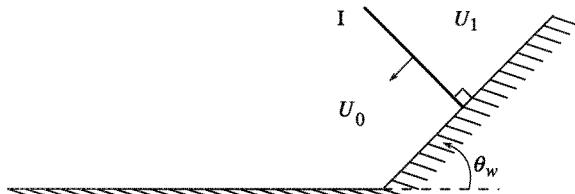
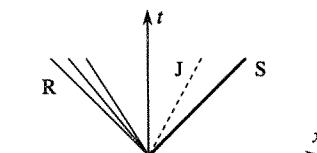


Figure 1.2 Mach's oblique-shock-reflection problem.

Riemann's shock tube problem has explicit solutions, which are illustrated in Figure 1.3. A typical solution consists of a shock, rarefaction wave, and a contact discontinuity. The shock and rarefaction waves can travel in either the positive or negative x directions. A shock wave is represented by a single bar, while a rarefaction wave is represented by a group of parallel bars in part (a) of Figure 1.3. In the evolutionary co-



(a) A snapshot in the tube (diagram).



(b) The solution in the (x, t) plane.

Figure 1.3 A solution of Riemann's shock tube experiment.

ordinates (x, t) of part (b) of Figure 1.3, a (centered) rarefaction wave is represented by a group of rays from the origin. A shock wave is a curve or surface across which the variables U has a discontinuity and the pressure behind it is greater than the pressure ahead of it. A contact discontinuity is a curve or surface across which the pressure is continuous but other variables are discontinuous. A rarefaction wave is a continuously changing wave whose pressure is decreasing in the direction of positive time.

Mach's experiment produced many patterns of solutions, illustrated in Figure 1.4, which shows the solution at $t = 1$. In Figure 1.4, the incident shock is marked with "I", the reflected shock is noted by "R". When the wedge angle θ_w is large, the experiment produces a reflection

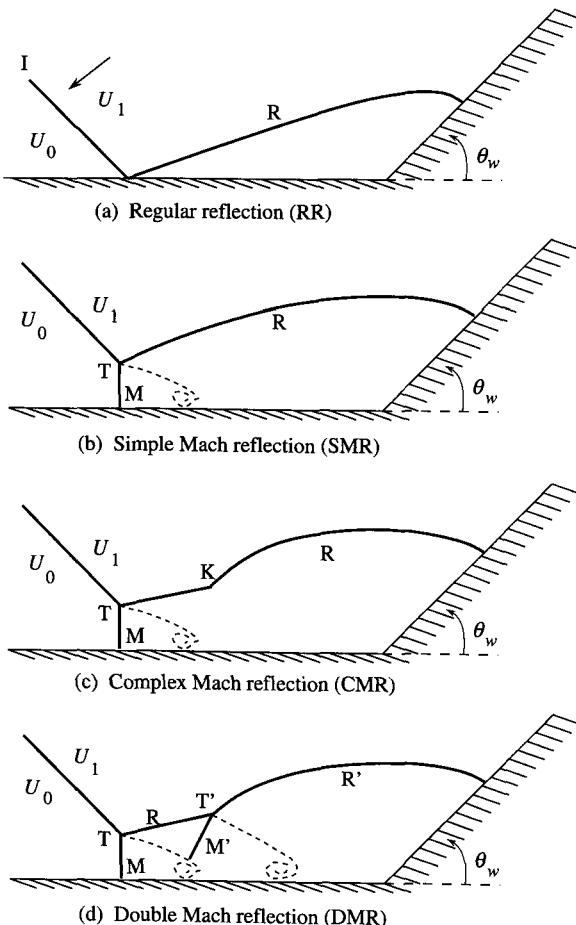


Figure 1.4 Regular and Mach reflections.

pattern as shown in part (a) of Figure 1.4, which is called *regular reflection*. When the wedge angle is small, the reflection pattern is like in part (b) of Figure 1.4, and it is called *simple Mach reflection*. In part (b), besides the incident and reflected shocks I and R, there is a new shock wave, called *Mach stem*, denoted by M, and a contact discontinuity marked by the dashed curve. The intersection point of the three shocks is denoted by “T” which is called a *triple point*. If the wedge angle is neither large nor small, then it happens that the reflected shock R may develop a kink, the point “K”, as shown in part (c) of Figure 1.4, where the segment TK is straight. The pattern in part (c) is called a *complex Mach reflection*. It is also possible, when the wedge lies in the middle, that a fourth pattern occurs, as shown in part (d) of Figure 1.4, and it is called a *Double Mach reflection*, in which a new shock wave and contact discontinuity emerge from the point K. The term *Mach reflection* refers to the three cases parts (b)–(d). These illustrations are based on physical experiments and numerical simulations, see [12]. There is little theoretical justification available so far. In addition, there may be other types of reflection, see Chapter 7 for Guderley reflection.

1.4 Paradoxes

There are several paradoxes arising from the work of von Neumann, known as von Neumann paradoxes. One of them is: On the one-hand, there is a triple point structure from physical and numerical simulations in the case of reflection of a weak incident shock on a ramp of small angle; While on the other hand, theoretical arguments indicate that it cannot exist. See [39].

1.5 The 8th millennium priceless problem

There are seven open problems proposed by the Clay Mathematics Institute. Here we propose one more problem and call it the 8th millennium priceless problem: i.e., the two-dimensional Riemann problem (see Figure 1.5):

Find an entropy (or physical) solution for the adiabatic Euler system with a gamma greater than one and initial data being constant along any ray (or four constants in the four quadrants).

For more details, see [25, 45].

2 Characteristic decomposition of the pseudo-steady case

We present a brief overview of the method of characteristics and generalize it to handle the self-similar two-dimensional Euler system. We

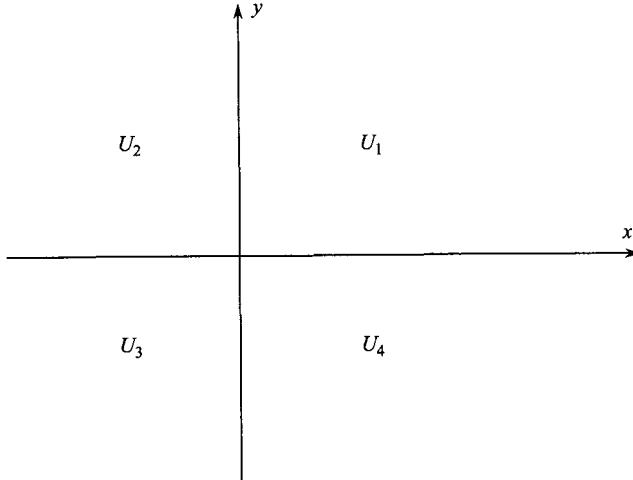


Figure 1.5 A common Riemann datum.

characterize its simple waves.

2.1 Riemann problems

Consider the full Euler system for an ideal fluid from Section 1. Let (r, θ) be the polar coordinate in the plane. Let the domain $\Omega \subset \mathbb{R}^2$ be either the whole plane or sectorial: $\Omega = \{(r, \theta) | r > 0, \theta \in (\theta_1, \theta_2)\}$ for a pair $\theta_1, \theta_2 \in (0, 2\pi)$, with $\theta_1 < \theta_2$.

Consider the initial data:

$$(p, \rho, u, v)|_{t=0} = (p_0, \rho_0, u_0, v_0)(\theta), \quad \text{all } \theta \text{ or } \theta_1 < \theta < \theta_2.$$

That is, the data are independent of the radius r . *Riemann problems* are Cauchy problems for the Euler system with these initial data, which is the most general form that we shall consider. When the domain has boundary, we will provide boundary conditions such as the no-flow boundary condition.

We seek self-similar solutions to the above problem in the variables $(\xi, \eta) = (x/t, y/t)$. The system then becomes

$$\begin{cases} -\xi\rho_\xi - \eta\rho_\eta + (\rho u)_\xi + (\rho v)_\eta = 0, \\ -\xi(\rho u)_\xi - \eta(\rho u)_\eta + (\rho u^2 + p)_\xi + (\rho uv)_\eta = 0, \\ -\xi(\rho v)_\xi - \eta(\rho v)_\eta + (\rho vu)_\xi + (\rho v^2 + p)_\eta = 0, \\ -\xi(\rho E)_\xi - \eta(\rho E)_\eta + (\rho Eu + pu)_\xi + (\rho Ev + pv)_\eta = 0, \end{cases} \quad (2.1)$$

while the initial condition becomes boundary condition at infinity

$$\lim(p, \rho, u, v)(\xi, \eta) = (p_0, \rho_0, u_0, v_0)(\theta), \quad \text{all } \theta \text{ or } \theta_1 < \theta < \theta_2$$

as we send (ξ, η) to infinity while holding $\eta = \xi \tan \theta$ fixed. (Examples of explicit solutions are given later.)

In the self-similar plane and for smooth solutions, the system takes the form:

$$\begin{cases} \frac{1}{\rho} \partial_s \rho + u_\xi + v_\eta = 0, \\ \partial_s u + \frac{1}{\rho} p_\xi = 0, \\ \partial_s v + \frac{1}{\rho} p_\eta = 0, \\ \frac{1}{\gamma p} \partial_s p + u_\xi + v_\eta = 0, \end{cases} \quad (2.2)$$

where

$$\partial_s := (u - \xi) \partial_\xi + (v - \eta) \partial_\eta.$$

We call $(u - \xi, v - \eta)$ pseudo-flow directions, as opposed to the other two characteristic directions, called (pseudo-)wave characteristics. We easily derive

$$\partial_s(p\rho^{-\gamma}) = 0. \quad (2.3)$$

So the entropy $p\rho^{-\gamma} = A$ is constant along pseudo-flow lines. For a region Ω whose pseudo-flow lines come from a constant state, we obtain that the entropy is constant in the region.

Consider the self-similar vorticity $W = v_\xi - u_\eta$ in an isentropic region. From the two transport equations

$$\partial_s u + \gamma A \rho^{\gamma-2} \rho_\xi = 0, \quad \partial_s v + \gamma A \rho^{\gamma-2} \rho_\eta = 0,$$

we obtain by differentiation that

$$\partial_s W - W + W(u_\xi + v_\eta) = 0.$$

Using the equation $\partial_s \rho + \rho(u_\xi + v_\eta) = 0$, we obtain

$$\partial_s \frac{W}{\rho} = \frac{W}{\rho}. \quad (2.4)$$

Thus zero W at a boundary will result in zero value for W in the region the pseudo-flow characteristics flow in. Hence, for a region whose pseudo-flow lines come from a constant state, the vorticity must be zero everywhere. So the region is irrotational and isentropic.

In the physical space (t, x, y) , the physical vorticity $\omega = v_x - u_y$ satisfies

$$\omega_t + (u\omega)_x + (v\omega)_y + \left(\frac{p_y}{\rho}\right)_x - \left(\frac{p_x}{\rho}\right)_y = 0. \quad (2.5)$$

The physical vorticity ω has zero source of production in the isentropic case when $\left(\frac{p_y}{\rho}\right)_x - \left(\frac{p_x}{\rho}\right)_y = 0$. The two vorticity are related by $t\omega = v_\xi - u_\eta = W$.

2.2 Isentropic system

Consider the two-dimensional, isentropic compressible Euler equations

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2 + p)_x + (\rho u v)_y = 0, \\ (\rho v)_t + (\rho u v)_x + (\rho v^2 + p)_y = 0, \end{cases} \quad (2.6)$$

where ρ is the density, (u, v) is the velocity and p is the pressure taken as the function of density, $p(\rho) = A\rho^\gamma$, for some constant A and a gas constant $\gamma > 1$. We investigate the pseudo-steady case of (2.6); i.e., the solution depends on the self-similar variables $(\xi, \eta) = (x/t, y/t)$. Then (2.6) becomes

$$\begin{cases} -\xi\rho_\xi - \eta\rho_\eta + (\rho u)_\xi + (\rho v)_\eta = 0, \\ -\xi(\rho u)_\xi - \eta(\rho u)_\eta + (\rho u^2 + p)_\xi + (\rho u v)_\eta = 0, \\ -\xi(\rho v)_\xi - \eta(\rho u)_\eta + (\rho u v)_\xi + (\rho v^2 + p)_\eta = 0. \end{cases} \quad (2.7)$$

This flow is often used in physical applications and numerical schemes, and two-dimensional Riemann problems as well.

We use, instead of ρ , the enthalpy

$$i = \frac{c^2}{\gamma - 1} = \frac{A\gamma}{\gamma - 1} \rho^{\gamma-1}$$

as one of the dependent variables. Let us rewrite (2.7) for smooth solutions as

$$\begin{cases} (u - \xi)i_\xi + (v - \eta)i_\eta + 2\kappa i(u_\xi + v_\eta) = 0, \\ (u - \xi)u_\xi + (v - \eta)u_\eta + i_\xi = 0, \\ (u - \xi)v_\xi + (v - \eta)v_\eta + i_\eta = 0, \end{cases} \quad (2.8)$$

where $\kappa = (\gamma - 1)/2$.

One may assume further that the flow is irrotational:

$$u_\eta = v_\xi. \quad (2.9)$$

Then, we insert the second and third equations of (2.8) into the first one to deduce the system

$$\begin{cases} (2\kappa i - (u - \xi)^2)u_\xi - (u - \xi)(v - \eta)(u_\eta + v_\xi) + (2\kappa i - (v - \eta)^2)v_\eta = 0, \\ u_\eta - v_\xi = 0, \end{cases} \quad (2.10)$$

supplemented by the Bernoulli's law

$$i + \frac{1}{2}((u - \xi)^2 + (v - \eta)^2) = -\varphi, \quad \varphi_\xi = u - \xi, \quad \varphi_\eta = v - \eta. \quad (2.11)$$

2.3 Some explicit solutions

We present the explicit solutions of planar waves and the Suchkov interaction.

Planar elementary waves. We list waves of the 1-D Riemann problem in the 2-D setting.

- (i) Constant states: $(\rho, u, v) = \text{constant}$ for the isentropic Euler.
- (ii) Backward and forward rarefaction waves:

$$R_{\mp}(\xi) : \begin{cases} \xi = u \mp c, & (c = \sqrt{p'(\rho)}) \\ \frac{du}{d\rho} = \mp \frac{c}{\rho}, \\ v = \text{constant}. \end{cases}$$

Suchkov explicit solution (1958 [34]). For an incline angle θ_s satisfying

$$\tan^2 \theta_s = \frac{3 - \gamma}{\gamma + 1},$$

the solution inside the interaction zone has the explicit form

$$c = \left(1 + \frac{\gamma - 1}{2 \sin \theta_s}\xi\right) \tan^2 \theta_s; \quad u = \left(\xi - \frac{1}{\sin \theta_s}\right) \tan^2 \theta_s; \quad v = \eta.$$

The characteristics are straight lines. The vacuum boundary is straight, located at $\xi = -\frac{2 \sin \theta_s}{\gamma - 1}$. See Figure 2.1.

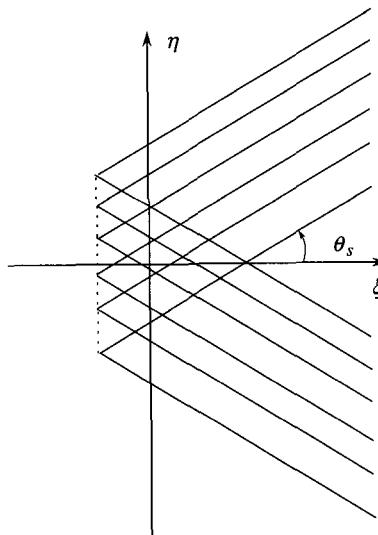


Figure 2.1 Suchkov explicit solution.

Axially symmetric solutions. From Zhang-Zheng (1997), the exact solutions of the two dimensional compressible Euler equations (2.6) for $p = A\rho^\gamma$, $A > 0$, $1 < \gamma < 3$, are valid only for $\gamma = 2$ and take the form

$$u = \frac{x+y}{2t}, \quad v = \frac{-x+y}{2t}, \quad \rho = \frac{r^2}{8At^2}$$

in $r \leq 2t\sqrt{p'_0}$, and

$$\begin{aligned} u &= \left(2tp'_0 \cos \theta + \sqrt{2p'_0} \sqrt{r^2 - 2t^2 p'_0} \sin \theta \right) / r, \\ v &= \left(2tp'_0 \sin \theta - \sqrt{2p'_0} \sqrt{r^2 - 2t^2 p'_0} \cos \theta \right) / r, \quad r > 2t\sqrt{p'_0} \\ \rho &= \rho_0, \end{aligned} \quad (2.12)$$

where $\rho_0 > 0$ is an arbitrary parameter, $p'_0 \equiv p'(\rho_0)$, and (r, θ) is the polar coordinates $x = r \cos \theta$, $y = r \sin \theta$. The solutions have the initial data:

$$\begin{aligned} u(x,y,0) &= \sqrt{2p'(\rho_0)} \sin \theta, \\ v(x,y,0) &= -\sqrt{2p'(\rho_0)} \cos \theta, \\ \rho(x,y,0) &= \rho_0. \end{aligned} \quad (2.13)$$

The particle trajectories in the cone $r < 2t\sqrt{p'(\rho_0)}$ given by

$$\begin{cases} \frac{dx}{dt} = u(x,y,t), & \frac{dy}{dt} = v(x,y,t), \\ r(t_0) = r_0, & \theta(t_0) = \theta_0, \end{cases}$$

where $r_0 \leq 2t_0\sqrt{p'(\rho_0)}$, take the form

$$\begin{aligned} r &= r_0 e^{\theta_0 - \theta}, \\ \theta &= \theta_0 - \frac{1}{2} \ln \frac{t}{t_0}, \quad \frac{r_0^2}{4t_0 p'(\rho_0)} \leq t < \infty \end{aligned}$$

in the polar coordinates and valid in the time interval indicated. The number of revolutions of these spirals approach infinity as $r_0 \rightarrow 0+$ in $0 < t \leq t_0$ for fixed $t_0 > 0$.

2.4 A characteristic decomposition

We present a characteristic decomposition of the potential flow equation in the self-similar plane. The decomposition allows for a proof that any wave adjacent to a constant state is a simple wave for the adiabatic Euler system. This result is a generalization of the well-known result on 2-d steady potential flow and a recent similar result on the pressure gradient system [10]. It has the potential for construction of more complex waves. According to Courant and Friedrichs, “simple waves are most important tools for the solutions of flow problems; simple waves and their generalizations apparently have not been sufficiently emphasized in mathematical studies of hyperbolic differential equations” ([7], p.40).

2.4.1 Introduction to the method of characteristics

Consider solving

$$\begin{cases} u_t + a(t, x)u_x = 0, & t > 0, x \in \mathbb{R} \\ u(0, x) = u_0(x). \end{cases} \quad (2.14)$$

Define the solution $x = x(t; x_0)$ to

$$\frac{dx}{dt} = a(t, x), \quad x(0) = x_0 \in \mathbb{R}$$

to be a *characteristic curve*. Examine a solution $u(t, x)$ to (2.14) along a characteristic $u(t, x(t; x_0))$ with differentiation:

$$\frac{du(t, x(t; x_0))}{dt} = u_t + u_x \frac{dx}{dt} = u_t + au_x = 0.$$

So we have

$$u(t, x(t; x_0)) = u(0, x(0; x_0)) = u(0, x_0) = u_0(x_0).$$

Assuming that $a(t, x)$ is a smooth and bounded function, so that the characteristics cover the upper plane $t > 0, x \in \mathbb{R}$ exactly once, we have a unique solution. This is the *method of characteristics*.

Consider the scalar conservation law

$$\begin{cases} u_t + f(u)_x = 0, \\ u(0, x) = 0. \end{cases} \quad (2.15)$$

The equation can be rewritten for smooth solutions to be $u_t + f'(u)u_x = 0$. So for any given smooth solution $u(t, x)$ we have the characteristics as

$$\frac{dx}{dt} = f'(u(t, x)), \quad x(0) = x_0,$$

denoted by $x = x(t; x_0)$ as before. Along the characteristics we find that

$$\frac{du(t, x(t; x_0))}{dt} = 0,$$

thus $u(t, x(t; x_0)) = u_0(x_0)$, and so the speed of the characteristics $f'(u) = f'(u_0(x_0))$ are constant along the characteristics, which then implies that the characteristics are straight lines. Assuming $f'' > 0$ and $u_0(x)$ is bounded, continuous and increasing, we have found a unique solution by tracing the characteristics.

The one-dimensional wave equation

$$u_{tt} - c^2 u_{xx} = 0 \quad (2.16)$$

with constant speed c has an interesting decomposition

$$(\partial_t + c\partial_x)(\partial_t - c\partial_x)u = 0, \quad \text{or} \quad (2.17)$$

$$(\partial_t - c\partial_x)(\partial_t + c\partial_x)u = 0 \quad (2.18)$$

known from elementary text books. One can rewrite them as

$$\partial_+\partial_- u = 0 \quad \text{or} \quad \partial_-\partial_+ u = 0, \quad (2.19)$$

where $\partial_{\pm} = \partial_t \pm c\partial_x$. Sometimes, the same fact is written in Riemann invariants

$$\partial_t R + c\partial_x R = 0, \quad \partial_t S - c\partial_x S = 0 \quad (2.20)$$

for the *Riemann invariants*

$$R := (\partial_t - c\partial_x)u, \quad S := (\partial_t + c\partial_x)u. \quad (2.21)$$

Now consider a linear system of equations with constant coefficients

$$\mathbf{u}_t + A\mathbf{u}_x = 0,$$

where A is an $n \times n$ matrix of real numbers with n distinct eigenvalues $\lambda_1 < \lambda_2 < \dots < \lambda_n$ and n linearly independent left eigen vectors ℓ_i ($i = 1, 2, \dots, n$). We can multiply the system of equations with ℓ_i and obtain

$$(\ell_i \mathbf{u})_t + \lambda_i (\ell_i \mathbf{u})_x = 0,$$

so the system is diagonalized by the Riemann invariants $\{\ell_i \mathbf{u}\}_{i=1}^n$.

For a pair of system of hyperbolic conservation laws

$$\begin{bmatrix} u \\ v \end{bmatrix}_t + \begin{bmatrix} f(u, v) \\ g(u, v) \end{bmatrix}_x = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad (2.22)$$

it is known that a pair of Riemann invariants exist so that the system can be rewritten as

$$\begin{cases} \partial_t R + \lambda_1(u, v)\partial_x R = 0, \\ \partial_t S + \lambda_2(u, v)\partial_x S = 0, \end{cases} \quad (2.23)$$

where (R, S) are the Riemann invariants and the λ 's are the two eigenvalues of the system.

These decompositions and Riemann invariants are useful in the construction of solutions, for example, the construction of D'Alembert formula, and proof of development of singularities ([19]). An example of

the system is the system of isentropic irrotational steady two-dimensional Euler equations for compressible ideal gases

$$\begin{cases} (c^2 - u^2)u_x - uv(u_y + v_x) + (c^2 - v^2)v_y = 0, \\ u_y - v_x = 0 \end{cases} \quad (2.24)$$

supplemented by Bernoulli's law

$$\frac{c^2}{\gamma - 1} + \frac{u^2 + v^2}{2} = k^2, \quad (2.25)$$

where $\gamma > 1$ is the gas constant while $k > 0$ is an integration constant. This system has two unknowns (u, v) , and by introducing the polar coordinates

$$u = q \cos \Theta, \quad v = q \sin \Theta, \quad (2.26)$$

we can find that a pair of Riemann invariants are

$$R_i = \Theta + (-1)^i \int \sqrt{q^2 - c^2}/(qc) dq, \quad (2.27)$$

where $c^2 = c^2(q)$ from Bernoulli's law. (The indefinite integral has explicit form.)

Following the existence of Riemann invariants, any solution adjacent to a constant state is a simple wave. A *simple wave* means a solution (u, v) that depends on a single parameter rather than the pair parameters (x, y) .

To find out why systems of three or more equations do not in general have Riemann invariants, we consider

$$\mathbf{u}_t + A(\mathbf{u})\mathbf{u}_x = 0, \quad (2.28)$$

where A has n distinct eigenvalues λ_i and linearly independent left eigenvectors $\ell_i (i = 1, \dots, n)$ with component $\ell_i = \{\ell_{ij}\}_{j=1}^n$. We obtain similarly

$$\ell_i \mathbf{u}_t + \lambda_i \ell_i \mathbf{u}_x = 0 \quad (i = 1, \dots, n). \quad (2.29)$$

When A is a constant matrix and so ℓ_i are constants, we take $R_i = \ell_i \mathbf{u} (\equiv \sum_j \ell_{ij} u_j)$ for the Riemann invariants. We now wish for some R_i so that

$$\sum_j \ell_{ij} (u_j)_t = (R_i)_t \quad (2.30)$$

or

$$\nabla_{\mathbf{u}} R_i = \ell_i. \quad (2.31)$$

This may not be always possible. But we can choose the ℓ_i with non-zero factors $\varphi_i(\mathbf{u})$, so (2.31) can become

$$\nabla_{\mathbf{u}} R_i = \varphi_i \ell_i. \quad (2.32)$$

For $n = 2$, the solvability condition for (2.32) is

$$\operatorname{curl} (\varphi_i \ell_i) = 0, \quad (2.33)$$

which in general has a nonzero solution for $\varphi_i(\mathbf{u})(i = 1, 2)$. But for $n > 2$, there are more compatibility conditions than the number of variables φ_i , so in general there are no solutions to (2.32).

Despite the general difficulty, the success for the pressure gradient system stirs the desire to consider the pseudo-steady isentropic irrotational Euler system which has three equations with source terms,

$$\begin{cases} (\rho U)_\xi + (\rho V)_\eta = -2\rho, \\ (\rho U^2 + p(\rho))_\xi + (\rho UV)_\eta = -3\rho U, \\ (\rho UV)_\xi + (\rho V^2 + p(\rho))_\eta = -3\rho V, \end{cases} \quad (2.34)$$

where $(\xi, \eta) = (x/t, y/t)$, $(U, V) = (u - \xi, v - \eta)$ is the pseudo-velocity, and the pressure $p = p(\rho)$ is the function of the density ρ . No explicit forms of the Riemann invariants are found, but decompositions similar to $\partial_+ \partial_- \lambda_- = m \partial_- \lambda_-$ hold for some m , where λ_- is an eigenvalue.

We use the characteristic decomposition to establish that a wave adjacent to a constant state must be a simple wave for the pseudo-steady irrotational isentropic Euler system. A *simple wave* for this case is such that one family of pseudo-wave characteristics are straight lines and the physical quantities velocity, speed of sound, pressure, and density are constant along the wave characteristics. Further, using the fact that entropy and vorticity are constant along the pseudo-flow characteristics (the pseudo-flow lines), our irrotational result extends to the adiabatic full Euler system.

2.4.2 Decomposition

Now let c be the speed of sound and $(U, V) = (u - \xi, v - \eta)$ be the pseudo-velocity. We can rewrite the equations of motion (2.10), (2.11) in a new form

$$\begin{bmatrix} u \\ v \end{bmatrix}_\xi + \begin{bmatrix} \frac{-2UV}{c^2-U^2} & \frac{c^2-V^2}{c^2-U^2} \\ -1 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}_\eta = 0 \quad (2.35)$$

to draw as much parallelism to the steady case as possible. We emphasize the mixed use of the variables (U, V) and (u, v) , i.e., (U, V) is used in the coefficients while (u, v) is used in differentiation. This way we obtain zero on the right-hand side for the system.

The eigenvalues are similar as before:

$$\frac{d\eta}{d\xi} = \Lambda_\pm = \frac{UV \pm \sqrt{c^2(U^2 + V^2 - c^2)}}{U^2 - c^2}. \quad (2.36)$$

The left eigenvectors are

$$\ell_{\pm} = [1, \Lambda_{\mp}]. \quad (2.37)$$

And we have similarly

$$\partial^{\pm} u + \Lambda_{\mp} \partial^{\pm} v = 0, \quad (2.38)$$

where $\partial^{\pm} = \partial_{\xi} + \Lambda_{\pm} \partial_{\eta}$. (The notation $\partial_{\pm} = \partial_u + \lambda_{\pm} \partial_v$ are reserved for the hodograph plane, see later.)

Our Λ_{\pm} now depend on more than (U, V) . But, let us regard Λ_{\pm} as a simple function of three variables $\Lambda_{\pm} = \Lambda_{\pm}(U, V, c^2)$ as given in (2.36). Thus we need to build differentiation laws for c^2 . We can directly obtain from (2.11) that

$$\left(\frac{c^2}{\gamma - 1} \right)_{\xi} + U u_{\xi} + V v_{\xi} = 0, \quad \left(\frac{c^2}{\gamma - 1} \right)_{\eta} + U u_{\eta} + V v_{\eta} = 0. \quad (2.39)$$

We have

$$\partial^{\pm} c^2 = -(\gamma - 1)(U \partial^{\pm} u + V \partial^{\pm} v). \quad (2.40)$$

So we move on to compute

$$\begin{aligned} \partial^{\pm} \Lambda_{\pm} &= \partial_U \Lambda_{\pm} \partial^{\pm} U + \partial_V \Lambda_{\pm} \partial^{\pm} V + \partial_{c^2} \Lambda_{\pm} \partial^{\pm} c^2 \\ &= \partial_U \Lambda_{\pm} (\partial^{\pm} u - 1) + \partial_V \Lambda_{\pm} (\partial^{\pm} v - \Lambda_{\pm}) + \partial_{c^2} \Lambda_{\pm} \partial^{\pm} c^2 \\ &= \partial_U \Lambda_{\pm} \partial^{\pm} u + \partial_V \Lambda_{\pm} \partial^{\pm} v + \partial_{c^2} \Lambda_{\pm} \partial^{\pm} c^2 - \partial_U \Lambda_{\pm} - \partial_V \Lambda_{\pm} \Lambda_{\pm}. \end{aligned} \quad (2.41)$$

We need to handle the term $\partial_U \Lambda_{\pm} + \partial_V \Lambda_{\pm} \Lambda_{\pm}$. We show it is zero. Recalling that

$$(c^2 - U^2)\Lambda^2 + 2UV\Lambda + c^2 - V^2 = 0, \quad (2.42)$$

and regarding that Λ depends on the three quantities (U, V, c^2) independently, we can easily find

$$\Lambda_U = \frac{\Lambda(U\Lambda - V)}{\Lambda(c^2 - U^2) + UV}, \quad \Lambda_V = -\frac{U\Lambda - V}{\Lambda(c^2 - U^2) + UV}. \quad (2.43)$$

Thus

$$\Lambda_U + \Lambda \Lambda_V = 0. \quad (2.44)$$

Therefore we end up with

$$\partial^{\pm} \Lambda_{\pm} = [\partial_U \Lambda_{\pm} - \Lambda_{\mp}^{-1} \partial_V \Lambda_{\pm} - (\gamma - 1) \partial_{c^2} \Lambda_{\pm} (U - \Lambda_{\mp}^{-1} V)] \partial^{\pm} u. \quad (2.45)$$

So the factor $\partial^{\pm} u$ is present in $\partial^{\pm} \Lambda_{\pm}$. Thus, if one of the quantities (u, v, c^2) is a constant along Λ_{\pm} , so are all the rest.

Proposition 2.1 (Commutator relation). We have

$$\partial^- \partial^+ I - \partial^+ \partial^- I = \frac{\partial^- \Lambda_+ - \partial^+ \Lambda_-}{\Lambda_- - \Lambda_+} (\partial^- I - \partial^+ I). \quad (2.46)$$

Proof. Compute

$$\begin{aligned} \partial^- \partial^+ I &= (\partial_\xi + \Lambda_- \partial_\eta)(\partial_\xi + \Lambda_+ \partial_\eta)I \\ &= I_{\xi\xi} + \Lambda_- I_{\xi\eta} + \Lambda_+ I_{\xi\eta} + \Lambda_- \Lambda_+ I_{\eta\eta} + \partial_\xi \Lambda_+ I_\eta + \Lambda_- \partial_\eta \Lambda_+ I_\eta. \end{aligned} \quad (2.47)$$

Similarly we have

$$\partial^+ \partial^- I = I_{\xi\xi} + \Lambda_+ I_{\xi\eta} + \Lambda_- I_{\xi\eta} + \Lambda_- \Lambda_+ I_{\eta\eta} + \partial_\xi \Lambda_- I_\eta + \Lambda_+ \partial_\eta \Lambda_- I_\eta. \quad (2.48)$$

The difference is

$$\partial^- \partial^+ I - \partial^+ \partial^- I = (\partial^- \Lambda_+ - \partial^+ \Lambda_-) I_\eta. \quad (2.49)$$

We can use the representation

$$I_\eta = \frac{\partial^- I - \partial^+ I}{\Lambda_- - \Lambda_+} \quad (2.50)$$

to complete the proof. \square

Theorem 2.1 (Characteristic decomposition). There holds

$$\partial^+ (\partial^- u) + \frac{\partial^+ \Lambda_- - \partial^- \Lambda_+}{\Lambda_+ - \Lambda_-} (\partial^- u) = \frac{\Lambda_+ \Lambda_-}{\Lambda_+ - \Lambda_-} \left[\frac{\partial^- \Lambda_-}{\Lambda_-^2} \partial^+ u - \frac{\partial^+ \Lambda_+}{\Lambda_+^2} \partial^- u \right]. \quad (2.51)$$

$$\partial^\pm \Lambda_\pm = [\partial_U \Lambda_\pm - \Lambda_\mp^{-1} \partial_V \Lambda_\pm - (\gamma - 1) \partial_{c^2} \Lambda_\pm (U - \Lambda_\mp^{-1} V)] \partial^\pm u. \quad (2.52)$$

$$\begin{aligned} \partial^\pm \Lambda_\mp &= [\partial_U \Lambda_\mp - \frac{1}{\Lambda_\mp} \partial_V \Lambda_\mp - (\gamma - 1) \partial_{c^2} \Lambda_\mp (U - \frac{1}{\Lambda_\mp} V)] \partial^\pm u \\ &\quad - \frac{2(U \Lambda_\mp - V)}{c^2 - U^2}. \end{aligned} \quad (2.53)$$

(Keep formulas at this level for structure recognition, e.g., we can utilize concavity properties of the characteristics.)

Proof. We use the commutator relation on u to find

$$\partial^+ \partial^- u = \partial^- \partial^+ u + \frac{\partial^+ \Lambda_- - \partial^- \Lambda_+}{\Lambda_+ - \Lambda_-} (\partial^+ u - \partial^- u). \quad (2.54)$$

We use the Euler equations $\partial^\pm u + \Lambda_\mp \partial^\pm v = 0$ to bring the term $\partial^- \partial^+ u$ back to $\partial^+ \partial^- u$ as follows. Use the equation to obtain

$$\begin{aligned}\partial^- \partial^+ u &= \partial^-(-\Lambda_- \partial^+ v) = -[\partial^- \Lambda_- \partial^+ v + \Lambda_- \partial^- \partial^+ v] \\ &= \frac{\partial^- \Lambda_-}{\Lambda_-} \partial^+ u - \Lambda_- \partial^- \partial^+ v.\end{aligned}\tag{2.55}$$

Use the commutator to obtain

$$\partial^- \partial^+ v = \partial^+ \partial^- v + \frac{\partial^- \Lambda_+ - \partial^+ \Lambda_-}{\Lambda_- - \Lambda_+} (\partial^- v - \partial^+ v).\tag{2.56}$$

Now use the equation to convert all the v back to u :

$$\partial^- \partial^+ v = -\partial^+ \left(\frac{\partial^- u}{\Lambda_+} \right) - \frac{\partial^- \Lambda_+ - \partial^+ \Lambda_-}{\Lambda_- - \Lambda_+} \left(\frac{1}{\Lambda_+} \partial^- u - \frac{1}{\Lambda_-} \partial^+ u \right).\tag{2.57}$$

Combining the last few steps we obtain

$$\begin{aligned}\partial^- \partial^+ u &= \frac{\partial^- \Lambda_-}{\Lambda_-} \partial^+ u + \frac{\Lambda_-}{\Lambda_+} \partial^+ \partial^- u - \frac{\Lambda_-}{\Lambda_+^2} \partial^+ \Lambda_+ \partial^- u + \\ &\quad \Lambda_- \frac{\partial^- \Lambda_+ - \partial^+ \Lambda_-}{\Lambda_- - \Lambda_+} \left(\frac{1}{\Lambda_+} \partial^- u - \frac{1}{\Lambda_-} \partial^+ u \right).\end{aligned}\tag{2.58}$$

Place that into the first expression we obtain an equation for $\partial^+ \partial^- u$ and so we solve for it to yield the decomposition of the theorem.

The second equality has been proved in the preceding paragraphs.

The proof of the third equality is like this. Similar to (2.41), we have

$$\begin{aligned}\partial^+ \Lambda_- &= \partial_U \Lambda_- \partial^+ U + \partial_V \Lambda_- \partial^+ V + \partial_{c^2} \Lambda_- \partial^+ c^2 \\ &= [\partial_U \Lambda_- - \frac{1}{\Lambda_-} \partial_V \Lambda_- - (\gamma - 1) \partial_{c^2} \Lambda_- (U - \frac{1}{\Lambda_-} V)] \partial^+ u \\ &\quad - (\partial_U \Lambda_- + \Lambda_+ \partial_V \Lambda_-),\end{aligned}\tag{2.59}$$

and similar to (2.43) we obtain

$$\partial_U \Lambda_- + \Lambda_+ \partial_V \Lambda_- = \partial_U \Lambda_- + \frac{1}{\Lambda_-} \frac{c^2 - V^2}{c^2 - U^2} \partial_V \Lambda_- = \frac{2(U \Lambda_- - V)}{c^2 - U^2}.\tag{2.60}$$

This completes the proof. \square

This way we see potential for a direct approach to the gas expansion into the vacuum (see [24, 34]), and a possible way for other patches.

Theorem 2.2 (Simple wave). (Li, Zhang, and Zheng [27], 2006) The solution adjacent to a constant state to the pseudo-steady Euler system is a simple wave, in which one family of characteristics are straight lines along each of which the variables (u, v, c) are constant.

3 The hodograph transformation and the interaction of rarefaction waves

We analyze the self-similar isentropic irrotational Euler via the hodograph transformation. We diagonalize the system of equations in the phase space. We use these equations to analyze the binary interaction of arbitrary planar rarefaction waves, which includes the classical problem of a wedge of gas expanding into vacuum.

3.1 Primary system

Consider the two-dimensional isentropic compressible Euler system

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2 + p)_x + (\rho u v)_y = 0, \\ (\rho v)_t + (\rho u v)_x + (\rho v^2 + p)_y = 0, \end{cases} \quad (3.1)$$

where $p(\rho) = A\rho^\gamma$ where $A > 0$ will be scaled to be one and $\gamma > 1$ is the gas constant.

Here is a list of our notations: ρ density, p pressure, (u, v) velocity, $c = \sqrt{\gamma p/\rho}$ speed of sound, $i = c^2/(\gamma - 1)$ enthalpy, γ gas constant, $(\xi, \eta) = (x/t, y/t)$ the self-similar (or pseudo-steady) variables, φ pseudo-velocity potential, θ wedge half-angle, and

$$U = u - \xi, V = v - \eta, \kappa = (\gamma - 1)/2, m = (3 - \gamma)/(\gamma + 1), \theta_s = \arctan \sqrt{m}.$$

Letters C , C_1 and C_2 denote generic constants.

Our primary system is system (3.1) in the self-similar variables (ξ, η) :

$$\begin{cases} (u - \xi)i_\xi + (v - \eta)i_\eta + 2\kappa i(u_\xi + v_\eta) = 0, \\ (u - \xi)u_\xi + (v - \eta)u_\eta + i_\xi = 0, \\ (u - \xi)v_\xi + (v - \eta)v_\eta + i_\eta = 0. \end{cases} \quad (3.2)$$

We assume further that the flow is irrotational:

$$u_\eta = v_\xi. \quad (3.3)$$

Then, we insert the second and third equations of (3.2) into the first one to deduce the system,

$$\begin{cases} (2\kappa i - U^2)u_\xi - UV(u_\eta + v_\xi) + (2\kappa i - V^2)v_\eta = 0, \\ u_\eta - v_\xi = 0, \end{cases} \quad (3.4)$$

supplemented by pseudo-Bernoulli's law

$$i + \frac{1}{2}((u - \xi)^2 + (v - \eta)^2) = -\varphi, \quad \varphi_\xi = u - \xi, \quad \varphi_\eta = v - \eta. \quad (3.5)$$

The system can then be written as a single second-order equation for φ :

$$(2\kappa i - \varphi_\xi^2)\varphi_{\xi\xi} - 2\varphi_\xi\varphi_\eta\varphi_{\xi\eta} + (2\kappa i - \varphi_\eta^2)\varphi_{\eta\eta} = \varphi_\xi^2 + \varphi_\eta^2 - 4\kappa i, \quad (3.6)$$

where

$$i + \frac{1}{2}|\nabla\varphi|^2 + \varphi = 0.$$

3.2 The concept of hodograph transformation

The original form of a hodograph transformation is for a homogeneous quasi-linear system of two first-order equations for two known variables (u, v) in two independent variables (x, y) . By regarding (x, y) as functions of (u, v) and assuming that the Jacobian does not vanish nor is infinity, one can re-write the system for the unknowns (x, y) in the variables (u, v) , which is a linear system if the coefficients of the original system do not depend on (x, y) . See the book of Courant and Friedrichs [7]. Specifically, consider the system of two equations of the form,

$$\begin{pmatrix} u \\ v \end{pmatrix}_x + A(u, v; x, y) \begin{pmatrix} u \\ v \end{pmatrix}_y = 0, \quad (3.7)$$

where the coefficient matrix $A(u, v; x, y)$ is

$$A(u, v; x, y) = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \quad (3.8)$$

We introduce the hodograph transformation,

$$T : (x, y) \rightarrow (u, v). \quad (3.9)$$

Differentiating the identities $u = u(x(u, v), y(u, v))$, $v = v(x(u, v), y(u, v))$ with respect to u, v , we solve for u_x, u_y, v_x, v_y to find

$$u_x = y_v/j, \quad u_y = -x_v/j, \quad v_x = -y_u/j, \quad v_y = x_u/j, \quad j := x_u y_v - x_v y_u.$$

Then (3.7) is reduced to the system

$$\begin{pmatrix} y_v \\ -y_u \end{pmatrix} + A(u, v; x, y) \begin{pmatrix} -x_v \\ x_u \end{pmatrix} = 0. \quad (3.10)$$

Obviously, if the coefficient matrix A does not depend on (x, y) , (3.10) becomes a linear system for the unknowns (x, y) .

3.2.1 The hodograph transformation for the pseudo-steady Euler

The idea of hodograph transformation does not obviously generalize to other systems such as system (3.4) of more than two simple equations or for inhomogeneous systems.

For (3.4), one realizes that the three variables (i, u, v) are functions of (ξ, η) , so one can still try to use (u, v) as the independent variables and regard (ξ, η) as functions of (u, v) and ultimately regard i as a function of (u, v) . This was done in 1958 in a paper [34] by Pogodin, Suchkov and Ianenko. The implementation is as follows. Let the hodograph transformation be

$$T : (\xi, \eta) \rightarrow (u, v) \quad (3.11)$$

for (3.2) and reverse the roles of (ξ, η) and (u, v) . Differentiating the identities

$$\xi = \xi(u(\xi, \eta), v(\xi, \eta)), \quad \eta = \eta(u(\xi, \eta), v(\xi, \eta))$$

with respect to ξ, η , we find

$$\xi_u = v_\eta/J, \quad \xi_v = -u_\eta/J, \quad \eta_u = -v_\xi/J, \quad \eta_v = u_\xi/J; \quad J = u_\xi v_\eta - v_\xi u_\eta.$$

Inserting these into system (3.4), we obtain

$$\begin{cases} (2\kappa i - U^2)\eta_v + UV(\xi_v + \eta_u) + (2\kappa i - V^2)\xi_u = 0, \\ \xi_v - \eta_u = 0. \end{cases} \quad (3.12)$$

The difficulty here is that i , as a function of u and v , cannot be determined explicitly and point-wise. But we can obtain something else. Differentiating pseudo-Bernoulli's law (3.5) with respect to u, v , we obtain

$$\begin{aligned} \xi - u &= i_u, \\ \eta - v &= i_v. \end{aligned}$$

(3.13)

These interesting identities provide an explicit correspondence between the physical plane and the hodograph plane provided that the transformation T is not degenerate.

We continue to differentiate (3.13) with respect to u and v :

$$\begin{aligned} \xi_u &= 1 + i_{uu}, & \xi_v &= i_{uv}, \\ \eta_u &= i_{uv}, & \eta_v &= 1 + i_{vv}, \end{aligned} \quad (3.14)$$

and inserting these into the first equations of (3.12) to obtain,

$$(2\kappa i - i_u^2)i_{vv} + 2i_u i_v i_{uv} + (2\kappa i - i_v^2)i_{uu} = i_u^2 + i_v^2 - 4\kappa i.$$

(3.15)

This is a very interesting second order partial differential equation for i alone. So the study of irrotational, pseudo-steady and isentropic fluid flow can proceed along (3.15).

We point out for the case $\gamma = 1$ that the dependent variable $i = \ln \rho$, instead of $i = c^2/(\gamma - 1)$, is used [23]. Then we can obtain a similar equation for i ,

$$(1 - i_u^2)i_{vv} + 2i_u i_v i_{uv} + (1 - i_v^2)i_{uu} = i_u^2 + i_v^2 - 2. \quad (3.16)$$

3.2.2 Steady Euler

The steady isentropic and irrotational Euler system of (3.1) has the form

$$\begin{cases} (2\kappa i - u^2)u_x - uv(u_y + v_x) + (2\kappa i - v^2)v_y = 0, \\ u_y - v_x = 0, \end{cases} \quad (3.17)$$

where i is given by Bernoulli's law

$$\frac{u^2 + v^2}{2} + i = \frac{k_0}{2}, \quad (3.18)$$

where k_0 is a constant. See [7]. Using the hodograph transformation from (x, y) to (u, v) , we obtain a linear system,

$$\begin{cases} (2\kappa i - u^2)y_v + uv(x_v + y_u) + (2\kappa i - v^2)x_u = 0, \\ x_v - y_u = 0. \end{cases} \quad (3.19)$$

The hodograph transformation is valid in the region of non-simple waves. Differentiating Bernoulli's law (3.18), we obtain

$$-u = i_u, \quad -v = i_v. \quad (3.20)$$

This can also be obtained formally from (3.13) by regarding the steady flow as the limit of unsteady flow (3.1) in $t \rightarrow \infty$. We see that (3.20) is a trivial consequence of (3.18). Comparing (3.20) with (3.13), we see that it is much more difficult to convert the hodograph plane of the steady case back into the physical plane than the pseudo-steady case. However, system (3.19) has more advantage over (3.12) of the pseudo-steady case because i is expressed in an explicit form by Bernoulli's law (3.18). In sum, the phase space structure of the steady case is trivial and its difficulty is finding the conversion (x, y) as a function of (u, v) ; while the conversion formula for the pseudo-steady case is explicit, its work is to find its phase space structure.

3.2.3 Similarity to one-dimensional problems

The current approach parallels the procedure that is used to find centered rarefaction waves to genuinely nonlinear strictly hyperbolic systems of conservation laws in one space dimension. Recall for a one-dimensional system $\mathbf{u}_t + f(\mathbf{u})_x = 0$ of n equations, a centered rarefaction wave takes the form $\xi = \lambda_k(\mathbf{u})$ for a $k \in [1, n]$ and the state variable \mathbf{u} satisfies the system of ordinary differential equations $(f'(\mathbf{u}) - \lambda_k(\mathbf{u})I)\mathbf{u}_\xi = 0$, whose solutions are rarefaction wave curves in the phase space. The development (or inversion) of the phase space solutions onto the ξ -axis requires the monotonicity of $\lambda_k(\mathbf{u})$ along the vector field of the k -th right eigenvector r_k ; i.e., the genuine nonlinearity. For the self-similar 2-D Euler system, we have a pair $\xi = u + i_u, \eta = v + i_v$ from (3.13) in place of $\xi = \lambda_k(\mathbf{u})$; and the second-order partial differential equation (3.15) in place of the ordinary differential system. For inversion to the physical space, we show that the Jacobian J_T^{-1} of (3.29) does not vanish.

3.3 Characteristics in both planes

We assert that the characteristics of (3.4) are mapped into the characteristics of (3.12) by the hodograph transformation (3.11). And there holds

$$\lambda_\pm = -\frac{1}{\Lambda_\mp}, \quad (3.21)$$

where, the eigenvalues of (3.4) are

$$\Lambda_\pm = \frac{(u - \xi)(v - \eta) \pm c\sqrt{(u - \xi)^2 + (v - \eta)^2 - c^2}}{(u - \xi)^2 - c^2}, \quad (3.22)$$

while the eigenvalues of (3.12) are

$$\lambda_\pm = \frac{i_u i_v \pm c\sqrt{(i_u^2 + i_v^2 - c^2)}}{c^2 - i_v^2}. \quad (3.23)$$

By using (3.13), it is easy to see (3.21). For the correspondence between Λ_\pm and λ_\pm , we let $\eta = \eta(\xi)$ be a characteristic curve in the (ξ, η) plane with $\frac{d\eta}{d\xi} = \Lambda_+$ and be mapped onto a curve $v = v(u)$. Then, using (3.14) and (3.21), we have

$$\Lambda = \frac{d\eta}{d\xi} = \frac{\eta_u + \eta_v \frac{dv}{du}}{\xi_u + \xi_v \frac{dv}{du}}, \quad (3.24)$$

i.e.,

$$\frac{dv}{du} = -\frac{\xi_u \Lambda_+ - \eta_u}{\xi_v \Lambda_+ - \eta_v} = -\frac{(1 + i_{uu})\Lambda_+ - i_{uv}}{i_{uv}\Lambda_+ - (1 + i_{vv})} = -\frac{i_{uu} + 1 + \lambda_- i_{uv}}{i_{uv} + \lambda_- (i_{vv} + 1)}. \quad (3.25)$$

We rewrite (3.15) as

$$i_{uu} + 1 + (\lambda_- + \lambda_+)i_{uv} + \lambda_-\lambda_+(i_{vv} + 1) = 0. \quad (3.26)$$

Then we conclude

$$\frac{dv}{du} = \lambda_+. \quad (3.27)$$

Similarly we obtain the correspondence between Λ_- and λ_- .

We remark that we will establish in the pseudo-steady case that the transform is not degenerate, i.e.,

$$J_T(u, v; \xi, \eta) = \frac{\partial(u, v)}{\partial(\xi, \eta)} = u_\xi v_\eta - u_\eta v_\xi \neq 0 \quad (3.28)$$

in regions of non-simple waves, to be detailed later. In the direction from (u, v) plane to the (ξ, η) plane, it is more direct to compute

$$J_T^{-1}(u, v; \xi, \eta) = \xi_u \eta_v - \xi_v \eta_u \neq 0. \quad (3.29)$$

3.4 Phase space system of equations

In this section we use the inclination angles of characteristics as useful variables to rewrite (3.15) in the hodograph plane. We proceed as follows. We first transform the second order equation (3.15) into a first-order system of equations. Introduce

$$X = i_u, \quad Y = i_v. \quad (3.30)$$

Then we deduce a 3×3 system of first order equations,

$$\begin{bmatrix} 2\kappa i - Y^2 & XY & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ i \end{bmatrix}_u + \begin{bmatrix} XY & 2\kappa i - X^2 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ i \end{bmatrix}_v = \begin{bmatrix} X^2 + Y^2 - 4\kappa i \\ 0 \\ X \end{bmatrix}. \quad (3.31)$$

This system is equivalent to (3.15) for C^1 solutions if the given datum for Y is compatible with the datum for i_v . The characteristic equation is

$$(2\kappa i - Y^2)\lambda^2 - 2XY\lambda + 2\kappa i - X^2 = 0 \quad (3.32)$$

besides the trivial factor λ . This system has three eigenvalues:

$$\begin{aligned} \lambda_0 &= 0, \\ \frac{dv}{du} = \lambda_\pm &= \frac{XY \pm \sqrt{2\kappa i(X^2 + Y^2 - 2\kappa i)}}{2\kappa i - Y^2} \\ &= \frac{2\kappa i - X^2}{XY \mp \sqrt{2\kappa i(X^2 + Y^2 - 2\kappa i)}}, \end{aligned} \quad (3.33)$$

from which we deduce that (3.31) is hyperbolic if $X^2 + Y^2 - 2\kappa i > 0$ provided that $i > 0$ and $2\kappa i - Y^2 \neq 0$ (or $2\kappa i - X^2 \neq 0$). If $2\kappa i - Y^2 = 0$ or $2\kappa i - X^2 = 0$, then the solutions are planar rarefaction waves. The three left eigenvectors associated with (3.33) are

$$l_0 = (0, 0, 1), \quad l_{\mp} = (1, \lambda_{\pm}, 0). \quad (3.34)$$

We multiply (3.31) by the left eigen-matrix $M = (l_+, l_-, l_0)^T$ (hereafter the superscript T means transpose) from the left-hand side to obtain the “characteristic form”

$$\begin{cases} X_u + \lambda_- Y_u + \lambda_+(X_v + \lambda_- Y_v) = \frac{X^2 + Y^2 - 4\kappa i}{2\kappa i - Y^2}, \\ X_u + \lambda_+ Y_u + \lambda_-(X_v + \lambda_+ Y_v) = \frac{X^2 + Y^2 - 4\kappa i}{2\kappa i - Y^2}, \\ i_u = X. \end{cases} \quad (3.35)$$

Introduce the inclination angles α, β ($-\pi/2 < \alpha, \beta < \pi/2$) of Λ_+ and Λ_- -characteristics by

$$\tan \alpha = \Lambda_+, \quad \tan \beta = \Lambda_-. \quad (3.36)$$

Note that, see (3.21),

$$\Lambda_+ = -\frac{1}{\lambda_-}, \quad \Lambda_- = -\frac{1}{\lambda_+}, \quad (3.37)$$

so that $\lambda_+ = -\cot \beta$ and $\lambda_- = -\cot \alpha$, from which we have the following proposition.

Proposition. We have

$$X = c \frac{\cos \frac{\alpha+\beta}{2}}{\sin \frac{\alpha-\beta}{2}}, \quad Y = c \frac{\sin \frac{\alpha+\beta}{2}}{\sin \frac{\alpha-\beta}{2}}. \quad (3.38)$$

Proof. Because of homogeneity, we can let $c = 1$. Thus the equations are

$$\frac{XY + \sqrt{X^2 + Y^2 - 1}}{1 - Y^2} = -\cot \beta, \quad \frac{XY - \sqrt{X^2 + Y^2 - 1}}{1 - Y^2} = -\cot \alpha. \quad (3.39)$$

The difference and the product of the two equations are

$$\frac{2\sqrt{X^2 + Y^2 - 1}}{1 - Y^2} = \cot \alpha - \cot \beta, \quad \frac{X^2 - 1}{Y^2 - 1} = \cot \alpha \cot \beta.$$

Thus we use $X^2 - 1 = (Y^2 - 1) \cot \alpha \cot \beta$ to remove X^2 to yield

$$2\sqrt{Y^2 + (Y^2 - 1) \cot \alpha \cot \beta} = (1 - Y^2)(\cot \alpha - \cot \beta).$$

Squaring the equation, multiplying it with $\sin^2 \alpha \sin^2 \beta$, we obtain

$$Y^4 \sin^2(\alpha - \beta) - 2Y^2(\sin^2 \alpha + \sin^2 \beta) + \sin^2(\alpha + \beta) = 0.$$

We then solve the quadratic equation to find

$$Y = \pm \frac{\sin \alpha \pm \sin \beta}{\sin(\alpha - \beta)}.$$

We choose the plus signs for our preference. We can then use the half-angle formula to yield

$$Y = c \sin\left(\frac{\alpha + \beta}{2}\right) / \sin\left(\frac{\alpha - \beta}{2}\right).$$

Then we use the sum of the equations of (3.39) to find

$$2XY = (Y^2 - 1)(\cot \alpha + \cot \beta)$$

to yield a unique X as seen. This completes the proof. \square

We observe that the variables α, β are “Riemann invariants” for (3.35). In fact, we can write (3.35) as

$$\begin{aligned} \partial_+ \alpha &= \frac{\gamma + 1}{4c} \cdot \frac{\sin(\alpha - \beta)}{\sin \beta} \cdot \left[m - \tan^2 \frac{\alpha - \beta}{2} \right], \\ \partial_- \beta &= \frac{\gamma + 1}{4c} \cdot \frac{\sin(\alpha - \beta)}{\sin \alpha} \cdot \left[m - \tan^2 \frac{\alpha - \beta}{2} \right], \\ \partial_0 c &= \frac{\gamma - 1}{2} \cdot \frac{\cos \frac{\alpha + \beta}{2}}{\sin \frac{\alpha - \beta}{2}}, \end{aligned} \quad (3.40)$$

where we use the notations of directional derivatives,

$$\partial_+ = \frac{\partial}{\partial u} + \lambda_+ \frac{\partial}{\partial v}, \quad \partial_- = \frac{\partial}{\partial u} + \lambda_- \frac{\partial}{\partial v}, \quad \partial_0 = \frac{\partial}{\partial u}, \quad (3.41)$$

and keep the letter m for

$$m = \frac{3 - \gamma}{1 + \gamma}. \quad (3.42)$$

We further introduce the normalized directional derivatives along characteristics,

$$\bar{\partial}_+ = (\sin \beta, -\cos \beta) \cdot (\partial_u, \partial_v), \quad \bar{\partial}_- = (\sin \alpha, -\cos \alpha) \cdot (\partial_u, \partial_v). \quad (3.43)$$

Using them, we write (3.40) as

$$\begin{aligned} \bar{\partial}_+ \alpha &= \frac{\gamma + 1}{4c} \cdot \sin(\alpha - \beta) \cdot \left[m - \tan^2 \frac{\alpha - \beta}{2} \right] =: G(\alpha, \beta, c), \\ \bar{\partial}_- \beta &= G(\alpha, \beta, c), \end{aligned} \quad (3.44)$$

$$\partial_0 c = \frac{\gamma - 1}{2} \cdot \frac{\cos \frac{\alpha + \beta}{2}}{\sin \frac{\alpha - \beta}{2}}.$$

A few remarks are in order. We note that we have an alternative expression for

$$G(\alpha, \beta, c) = \frac{1}{c} \tan \frac{\alpha - \beta}{2} (\cos(\alpha - \beta) - \kappa).$$

We note further that

$$\bar{\partial}_+ c = -\kappa, \quad \bar{\partial}_- c = \kappa. \quad (3.45)$$

This means that the first two equations of (3.44) are essentially decoupled from the third c -equation. Further, system (3.40) is linearly degenerate in the sense of Lax [18]. For the particular case that $\tan((\alpha - \beta)/2) = \sqrt{m}$ for $1 < \gamma < 3$, and α and β are constants, the first two equations are satisfied. In fact, the explicit solutions of Suchkov [37] in the expansion problem of a wedge of gas into vacuum is such a case, see Remark 3.1 in Section 3.6. The variables (α, β) might be called *Riemann variables*, with signature that system (3.44) is diagonalized.

The mapping $(X, Y) \rightarrow (\alpha, \beta)$ is bijective as long as system (3.35) is hyperbolic.

We summarize the above as follows.

Theorem 3.1. The two-dimensional pseudo-steady, irrotational, isentropic flow (3.15) can be transformed into a linearly degenerate system of first order partial differential equations (3.40) or (3.44) provided that the transform $(X, Y) \rightarrow (\alpha, \beta)$ is invertible, i.e., system (3.35) is hyperbolic.

Proof of (3.40). We find easily that $\partial_+ \alpha = \sin^2 \alpha \partial_+ \lambda_-$. Holding λ as a function of three independent variables (X, Y, c^2) in the characteristic equation (3.32) we find

$$\begin{aligned} \partial_X \lambda &= \frac{X + \lambda Y}{\lambda(2\kappa i - Y^2) - XY}, \\ \partial_Y \lambda &= \lambda \lambda_X, \quad \partial_{c^2} \lambda = -\frac{\lambda^2 + 1}{2\lambda(2\kappa i - Y^2) - 2XY}. \end{aligned} \quad (3.46)$$

We then use

$$\begin{aligned} \partial_+ \lambda_- &= \partial_X \lambda_- \partial_+ X + \partial_Y \lambda_- \partial_+ Y + \partial_{c^2} \lambda_- \partial_+ c^2 \\ &= \partial_X \lambda_- (\partial_+ X + \lambda_- \partial_+ Y) + \partial_{c^2} \lambda_- 2\kappa (X + \lambda_+ Y). \end{aligned}$$

We use (3.35) and (3.46) to simplify it to obtain (3.40). The proof of (3.40) is complete. \square

Regarding $\bar{\partial}_- \alpha$ and $\bar{\partial}_+ \beta$, we are unable to obtain explicit expressions for them like (3.44). But we have second-order equations. By direct computations, we first obtain

Lemma 3.1 (Commutator relation of ∂_{\pm}). For any quantity $I = I(u, v)$, we have

$$\partial_{-}\partial_{+}I - \partial_{+}\partial_{-}I = \frac{\partial_{-}\lambda_{+} - \partial_{+}\lambda_{-}}{\lambda_{-} - \lambda_{+}}(\partial_{-}I - \partial_{+}I). \quad (3.47)$$

Lemma 3.2 (Commutator relation of $\bar{\partial}_{\pm}$). For any quantity $I = I(u, v)$, we have,

$$\bar{\partial}_{-}\bar{\partial}_{+}I - \bar{\partial}_{+}\bar{\partial}_{-}I = \frac{1 - \cos(\alpha - \beta)}{\sin(\alpha - \beta)}(\bar{\partial}_{-}I + \bar{\partial}_{+}I)\bar{\partial}_{+}\alpha, \quad (3.48)$$

where $\bar{\partial}_{+}\alpha$ is given in (3.44). Noting $\bar{\partial}_{+}\alpha = \bar{\partial}_{-}\beta$ in (3.44), we can also use $\bar{\partial}_{-}\beta$ in (3.48).

Using these commutator relations, we easily derive the Theorem 3.2.

Theorem 3.2. Assume that the solution of (3.44) $(\alpha, \beta) \in C^2$. Then we have

$$\begin{aligned} \bar{\partial}_{+}\bar{\partial}_{-}\alpha + W\bar{\partial}_{-}\alpha &= Q(\alpha, \beta, c), \\ -\bar{\partial}_{-}\bar{\partial}_{+}\beta + W\bar{\partial}_{+}\beta &= Q(\alpha, \beta, c), \end{aligned} \quad (3.49)$$

where $W(\alpha, \beta, c)$ and $Q(\alpha, \beta, c)$ are

$$W(\alpha, \beta, c) = \frac{\gamma + 1}{4c} \left[(m - \tan^2 \omega) (3 \tan^2 \omega - 1) \cos^2 \omega + 2 \tan^2 \omega \right],$$

$$Q(\alpha, \beta, c) = \frac{(\gamma + 1)^2}{16c^2} \sin(2\omega) (m - \tan^2 \omega) (3 \tan^2 \omega - 1), \quad (3.50)$$

$$\omega = \frac{\alpha - \beta}{2}.$$

Proof. The proof is simple. Recall from (3.45) that

$$\bar{\partial}_{+}c = -\kappa, \quad \bar{\partial}_{-}c = \kappa. \quad (3.51)$$

Then we apply the commutator relation to obtain (setting $I = \alpha$ in (3.48))

$$\bar{\partial}_{+}\bar{\partial}_{-}\alpha = \bar{\partial}_{-}\bar{\partial}_{+}\alpha + \frac{1 - \cos(\alpha - \beta)}{\sin(\beta - \alpha)}(\bar{\partial}_{-}\alpha + \bar{\partial}_{+}\alpha)\bar{\partial}_{-}\beta. \quad (3.52)$$

Using the expressions of $\bar{\partial}_{+}\alpha$ and $\bar{\partial}_{-}\beta$ in (3.44), we compute directly to yield the result in (3.49) and the proof of Theorem 3.2 is complete. \square

3.5 Planar rarefaction waves

We present in concise form planar rarefaction waves of the Euler system. Let $R_{12}^+(\xi)(\eta > v_1)$ and $R_{12}^-(\xi)(\eta < v_1)$ connect the two states (p_1, ρ_1, u_1, v_1) and (p_2, ρ_2, u_2, v_1) , denoted by (1) and (2) respectively in Figure 3.1, via a planar wave in the ξ variable. The solution is isentropic so that the entropy $S := p\rho^{-\gamma} = p_1\rho_1^{-\gamma} = p_2\rho_2^{-\gamma}$ is constant in the solution. The solution from (3.4) and (3.5) has the formula

$$R_{12}^\pm(\xi) : \begin{cases} \xi = u + \sqrt{p'(\rho)}, \quad (\xi_2 = u_2 + c_2 \leq \xi \leq \xi_1 = u_1 + c_1) \\ u = u_1 + \int_{\rho_1}^\rho \rho^{-1} \sqrt{p'(\rho)} d\rho, \quad (0 \leq \rho_2 \leq \rho \leq \rho_1) \\ v = v_2 = v_1, \\ \eta > v_1 \text{ or } \eta < v_1. \end{cases} \quad (3.53)$$

The sonic curve of these two waves is a straight segment

$$\eta = v_1, \quad \xi \in (\xi_2, \xi_1), \quad (3.54)$$

which is often called a *sonic stem*.

The (pseudo-wave) characteristics are given by $d\eta/d\xi = \Lambda_\pm$ where $\Lambda_+ = \infty$ in R_{12}^+ and

$$\Lambda_- = \frac{(\eta - v_1)^2 - p'(\rho)}{2\sqrt{p'(\rho)}(\eta - v_1)}, \quad \eta > v_1.$$

Since

$$\frac{d\xi}{d\rho} = \frac{2p'(\rho) + \rho p''(\rho)}{2\rho\sqrt{p'(\rho)}}, \quad \text{i.e.,} \quad \xi = \xi_2 + \frac{\gamma+1}{\gamma-1}\sqrt{\gamma S}(\rho^{(\gamma-1)/2} - \rho_2^{(\gamma-1)/2}),$$

we obtain

$$\frac{d(\eta - v_1)^2}{d\rho} = \frac{\gamma+1}{2\rho}[(\eta - v_1)^2 - \gamma S \rho^{\gamma-1}],$$

which yields the negative family of characteristics

$$\eta - v_1 = \begin{cases} \sqrt{\left(C + \frac{\gamma(\gamma+1)}{3-\gamma}S\rho^{(\gamma-3)/2}\right)\rho^{(\gamma+1)/2}}, & \text{if } \gamma \neq 3, \\ \rho\sqrt{C - 6S\ln\rho}, & \text{if } \gamma = 3, \end{cases} \quad (3.55)$$

where C is an arbitrary constant and S is the entropy. Note that there is a straight characteristic curve among them (when $C = 0, \rho_2 = 0$) for $\gamma \in (1, 3)$; that is

$$\eta - v_1 = \frac{\gamma-1}{\sqrt{(3-\gamma)(\gamma+1)}}(\xi - u_2). \quad (3.56)$$

Above this straight curve, the characteristics are convex; and below it they are concave. Further, if $\rho_2 = 0$, then all the negative family of characteristics converge to the point $(\xi, \eta) = (u_2, v_2)$ with the same asymptotic slope as in (3.56) for $\gamma \in (1, 3)$. For $\gamma > 3$, the constant C is always positive and the asymptotic slope is plus infinity.

If $\rho_2 > 0$, then the sonic curve consists of the sonic circle of the state (p_2, ρ_2, u_2, v_2) and the sonic stem. The characteristics and sonic curves are drawn in Figure 3.1.

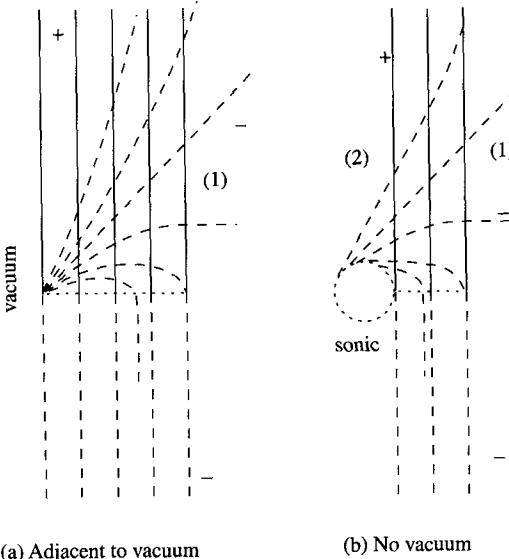


Figure 3.1 Characteristics in planar rarefaction waves for $1 < \gamma < 3$. Dashed lines are characteristics of the negative family, solid lines are those of the plus family; while dotted lines are sonic curves located at $\eta = v_1$.

3.6 The gas expansion problem

We now use the hodograph transformation to study the expansion of a wedge of gas into vacuum. The problem was studied in [37, 32, 22, 24, 23, 28]. Here are some notations in this section: We use $\theta_s \in [0, \pi/2]$, m_0 , defined by

$$\tan \theta_s = \sqrt{(3 - \gamma)/(\gamma + 1)}, \quad m_0 = 1/\sqrt{m}, \quad (3.57)$$

for $1 \leq \gamma \leq 3$; and $\theta_s \equiv 0$ for $\gamma > 3$.

3.6.1 Planar rarefaction waves in a given direction

First we prepare our planar rarefaction waves. Given a direction (n_1, n_2) with $n_1^2 + n_2^2 = 1$, and a positive constant ρ_1 . Consider the initial data for (3.1)

$$(\rho, u, v)(x, y, 0) = \begin{cases} (\rho_1, 0, 0), & \text{for } n_1 x + n_2 y > 0, \\ \text{vacuum}, & \text{for } n_1 x + n_2 y < 0. \end{cases} \quad (3.58)$$

The solution of (3.1) and (3.58) takes the form, see [25],

$$(\rho, u, v)(x, y, t) = \begin{cases} (\rho_1, 0, 0), & \zeta > 1, \\ (\rho, u, v)(\zeta), & -1/\kappa \leq \zeta \leq 1, \\ \text{vacuum}, & \zeta < -1/\kappa, \end{cases} \quad (3.59)$$

where $\zeta = n_1 \xi + n_2 \eta$, $(\xi, \eta) = (x/t, y/t)$, and the solution (c, u, v) has been normalized so that $c_1 = 1$. The rarefaction wave solution $(\rho, u, v)(\zeta)$ satisfies

$$\zeta = n_1 u + n_2 v + c, \quad \frac{n_1}{\kappa} c - u = \frac{n_1}{\kappa}, \quad \frac{n_2}{\kappa} c - v = \frac{n_2}{\kappa}. \quad (3.60)$$

Note that this rarefaction wave corresponds to a segment in the hodograph plane, $n_2 u - n_1 v = 0$, $-n_1/\kappa \leq u \leq 0$.

In particular, when we consider the rarefaction wave propagates in the x -direction, i.e., $(n_1, n_2) = (1, 0)$, this wave can be expressed as

$$x/t = u + c, \quad c = \kappa u + 1, \quad v \equiv 0, \quad -1/\kappa \leq u \leq 0. \quad (3.61)$$

That is, in the hodograph (u, v) plane, this rarefaction wave is mapped onto a segment $v \equiv 0$, $-1/\kappa \leq u \leq 0$, on which we have

$$i = \frac{1}{2\kappa}(\kappa u + 1)^2, \quad i_u = \kappa u + 1, \quad i_{uu} = \kappa. \quad (3.62)$$

3.6.2 A wedge of gas

We place the wedge symmetrically with respect to the x -axis and the sharp corner at the origin, as in Figure 3.2(a). This problem is then formulated mathematically as seeking the solution of (3.1) with the initial data

$$(i, u, v)(t, x, y)|_{t=0} = \begin{cases} (i_0, u_0, v_0), & -\theta < \delta < \theta, \\ (0, \bar{u}, \bar{v}), & \text{otherwise,} \end{cases} \quad (3.63)$$

where $i_0 > 0$, u_0 and v_0 are constant, (\bar{u}, \bar{v}) is the velocity of the wave front, not being specified in the state of vacuum, $\delta = \arctan y/x$ is the

polar angle, and θ is the half-angle of the wedge restricted between 0 and $\pi/2$. This can be considered as a two-dimensional Riemann problem for (3.1) with two pieces of initial data (3.63). As we will see below, this problem is actually the interaction of two whole planar rarefaction waves. See Figure 3.2(b). We note that the solution we construct is valid for any “portion” of (3.63) as the solution is hyperbolic.

The gas away from the sharp corner expands into the vacuum as planar rarefaction waves R_1 and R_2 of the form $(i, u, v)(t, x, y) = (i, u, v)(\zeta)$ ($\zeta = (n_1 x + n_2 y)/t$) where (n_1, n_2) is the propagation direction of waves. We assume that initially the gas is at rest, i.e., $(u_0, v_0) = (0, 0)$. Otherwise, we replace (u, v) by $(u - u_0, v - v_0)$ and (ξ, η) by $(\xi - u_0, \eta - v_0)$ in the following computations (see also (3.2)). We further assume that the initial sound speed is unit since the transformation $(u, v, c, \xi, \eta) \rightarrow c_0(u, v, c, \xi, \eta)$ with $c_0 > 0$ can make all variables dimensionless. Then the rarefaction waves R_1 , R_2 emitting from the initial discontinuities l_1 ,

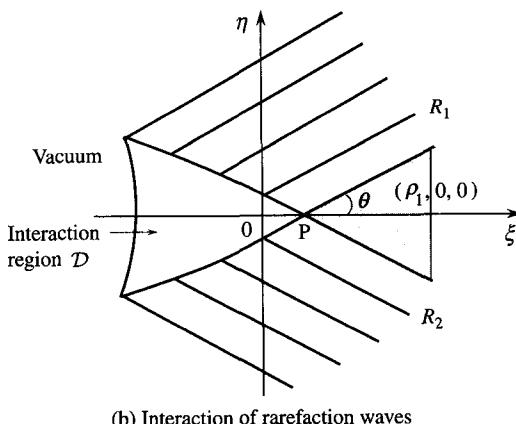
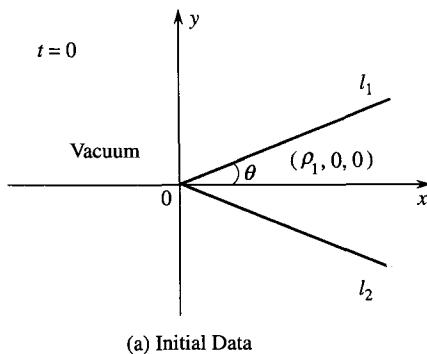


Figure 3.2 The expansion of a wedge of gas.

l_2 are expressed in (3.60) with $(n_1, n_2) = (\sin \theta, -\cos \theta)$ and $(n_1, n_2) = (\sin \theta, \cos \theta)$ respectively. These two waves begin to interact at $P = (1/\sin \theta, 0)$ in the (ξ, η) plane due to the presence of the sharp corner and a wave interaction region, called the *wave interaction region* \mathcal{D} , is formed to separate from the planar rarefaction waves by k_1, k_2 ,

$$\begin{aligned} k_1 : (1 - \kappa^2)\xi_1^2 - (\kappa\eta_1 + 1)^2 &= C_\gamma(\kappa\eta_1 + 1)^{(\kappa+1)/\kappa}, \\ (\xi_1 > 0, -1 \leq \eta_1 \leq 1/\kappa), \\ k_2 : (1 - \kappa^2)\xi_2^2 - (\kappa\eta_2 + 1)^2 &= C_\gamma(\kappa\eta_2 + 1)^{(\kappa+1)/\kappa}, \\ (\xi_2 > 0, -1/\kappa \leq \eta_2 \leq 1), \end{aligned} \quad (3.64)$$

where k_1 and k_2 are two characteristics from P , associated with the nonlinear eigenvalues of system (3.2), see [25, 41], and the constant C_γ is

$$C_\gamma = 2^{(3-\gamma)/(\gamma-1)}(\gamma+1)^{-2/(\gamma-1)}(3+\gamma^2)\gamma^{-(\gamma+1)/(\gamma-1)}, \quad (3.65)$$

and

$$\begin{cases} \xi_1 = \xi \cos \theta + \eta \sin \theta, \\ \eta_1 = -\xi \sin \theta + \eta \cos \theta, \end{cases} \quad \begin{cases} \xi_2 = \xi \cos \theta - \eta \sin \theta, \\ \eta_2 = \xi \sin \theta + \eta \cos \theta. \end{cases} \quad (3.66)$$

So, the wave interaction region \mathcal{D} is bounded by k_1, k_2 and the interface of gas with vacuum. The solution outside \mathcal{D} consists of the constant state (i_0, u_0, v_0) , the vacuum, and the planar rarefaction waves R_1 and R_2 .

Problem A. *Find a solution of (3.2) inside the wave interaction region \mathcal{D} , subject to the boundary values on k_1 and k_2 , which are determined continuously from the rarefaction waves R_1 and R_2 .*

This problem is a Goursat-type problem for (3.2) since k_1 and k_2 are characteristics. Note also that initial data (3.63) is irrotational, we conclude that the flow is always irrotational provided that it is continuous. So the irrotationality condition (3.3) holds.

3.6.3 A wedge of gas in the hodograph plane

Our strategy to solve this problem is to use the hodograph transformation, solve the associated problem in the hodograph plane, and show that the hodograph transformation is invertible.

For this purpose, we need to map the wave interaction region \mathcal{D} in the (ξ, η) plane into a region Ω in the (u, v) plane. Notice that the mapping

of the planar rarefaction waves R_1 and R_2 into (u, v) plane are exactly two segments

$$\begin{aligned} H_1 : u \cos \theta + v \sin \theta &= 0, \quad (-\sin \theta / \kappa \leq u \leq 0) \text{ and} \\ H_2 : u \cos \theta - v \sin \theta &= 0, \quad (-\sin \theta / \kappa \leq u \leq 0). \end{aligned}$$

The boundary values of c on H_1 , H_2 , are

$$c|_{H_1} = 1 + \kappa v' =: c_0^1, \quad c|_{H_2} = 1 + \kappa v'' =: c_0^2, \quad (3.67)$$

where $v' = u \sin \theta - v \cos \theta$ and $v'' = u \sin \theta + v \cos \theta$. Obviously,

$$0 \leq c_0^1, c_0^2 \leq 1. \quad (3.68)$$

Thus the wave interaction region Ω is bounded by H_1 , H_2 and the interface of vacuum connecting D and E in the hodograph (u, v) -plane, see Figure 3.3. We define Ω more precisely to contain the boundaries H_1 and H_2 , but not the vacuum boundary $c = 0$.

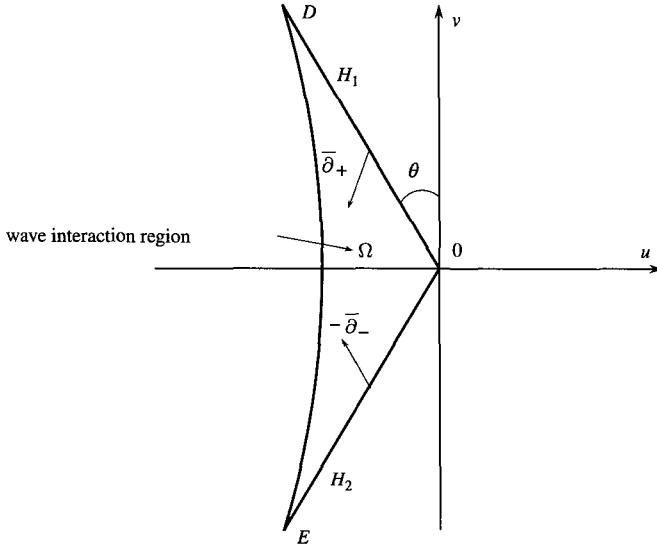


Figure 3.3 Wave interaction region in the hodograph plane.

Boundary conditions. We need to derive the necessary boundary conditions on H_1 and H_2 respectively. This can be done simply by using coordinate transformations

$$\tan \alpha = \Lambda_+, \quad \tan \beta = \Lambda_-, \quad \lambda_{\pm} \Lambda_{\mp} = -1$$

and the characteristics distribution of Section 3.5. Rotating Figure 3.1 clockwise by $\pi/2 - \theta$, we see easily that $\alpha|_{H_1} = \theta, \beta|_{H_2} = -\theta$. Thus the

boundary values for α, β on H_1 and H_2 are

$$\alpha|_{H_1} = \theta, \quad \beta|_{H_2} = -\theta. \quad (3.69)$$

The boundary values of c on H_1 and H_2 are given in (3.67). Now Problem A becomes Problem B.

Problem B. *Find a solution (α, β, c) of (3.40) with boundary values (3.69) and (3.67), in the wave interaction region Ω in the hodograph plane.*

The values of β on H_1 or α on H_2 can be integrated from the system. We estimate the boundary values (3.69) and (3.67).

Lemma 3.3 (Boundary data estimate). For the boundary data (3.69) on the boundaries H_i , $i = 1, 2$, we have the following estimates:

(i) If $\theta < \theta_s$, we have

$$2\theta \leq (\alpha - \beta)|_{H_i} \leq 2\theta_s. \quad (3.70)$$

(ii) If $\theta > \theta_s$, we have

$$2\theta_s \leq (\alpha - \beta)|_{H_i} \leq 2\theta. \quad (3.71)$$

Proof. It follows from the convexity of the characteristics. The extreme values of β are determined at the ends of the characteristics, i.e., either the starting value or the ending (at the vacuum) asymptotic value

$$\tan \beta_e := \frac{\gamma - 1}{\sqrt{(3 - \gamma)(\gamma + 1)}}.$$

This β_e is related to m by

$$\tan^2 \frac{\frac{\pi}{2} - \beta_e}{2} = m.$$

Then the proof is complete. \square

3.6.4 Local existence

The local existence of solutions at the origin $(u, v) = (0, 0)$ follows routinely from the idea [31, Chapter 2] or [40]. We need only to check the compatibility condition to this problem, i.e.,

$$\frac{1}{\lambda_+} \left[l^0 \cdot \partial_+ K - \kappa \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega} \right] = \frac{1}{\lambda_-} \left[l^0 \cdot \partial_- K - \kappa \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega} \right] \quad (3.72)$$

at $(u, v) = (0, 0)$, where $K = (\alpha, \beta, c)^\top$ and $l^0 = (0, 0, 1)$. That is, we need to check if this is true

$$\frac{1}{\lambda_+} \left[\partial_+ c - \kappa \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega} \right] = \frac{1}{\lambda_-} \left[\partial_- c - \kappa \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega} \right]. \quad (3.73)$$

This is obviously true by using (3.45). Hence we have the following Lemma.

Lemma 3.4 (Local existence). There is a $\delta > 0$ such that the C^1 -solution of (3.40), (3.67), and (3.69) exists uniquely in the region $\bar{\Omega} = \{(u, v) \in \Omega; -\delta < u < 0\}$, where δ depends only on the C^0 and C^1 norms of α, β on the boundaries H_1 and H_2 .

We do not give the proof. For details, see [31, Chapter 2], [40], or Section 4.

3.6.5 Statement of main existence

Next we will extend the local solution to the whole region Ω . Therefore some *a priori* estimates on the C^0 and C^1 norms of α, β and i , are needed. The norm of i comes from the norms of α and β , see the third equation of (3.40). Therefore we need only the estimate on α and β . Recall that the derivation of (3.40) is based on the strict hyperbolicity of the flow, $i > 0$. These will be achieved when we estimate the C^0 norms of α and β , see subsection 3.6.6. The main existence theorem is stated as follows. Let l be the interface of the gas with vacuum.

Theorem 3.3 (Global existence in the hodograph plane). There exists a solution $(\alpha, \beta, i) \in C^1$ to the boundary value problem (3.40) with boundary values (3.67) and (3.69)(Problem B) in Ω . The vacuum interface l exists and is Lipschitz continuous.

We prove this theorem by two steps. We estimate the solution itself in subsection 3.6.6 and then proceed with estimates on the gradients in subsection 3.6.7. The proof of Theorem 3.3 is also given in subsection 3.6.7.

After we solve Problem B, we establish the inversion of hodograph transformation in subsection 3.6.8, which establishes the existence of the gas expansion problem, Problem A.

Theorem 3.4 (Global existence in the physical plane). There exists a solution $(c, u, v) \in C^1$ of (3.2) for the gas expansion problem (Problem A) in the wave interaction region \mathcal{D} in the physical plane, the (ξ, η) -plane.

3.6.6 The maximum norm estimate on (α, β, c)

We estimate the solution (α, β, c) itself, i.e., the C^0 norm of α , β and c . We adopt the method of invariant regions, see [36] or section 5. We shall consider the case $0 < \theta < 2\theta_s$ in this lecture notes. See our paper [28] for more.

Lemma 3.5 (Invariant square). Suppose that there exists a C^1 solution $(\alpha(u, v), \beta(u, v), c(u, v))$ to problem (3.40), (3.67) and (3.69) in Ω . Suppose that $0 < \theta < 2\theta_s$. Then the C^0 -norms of α and β have uniform bounds:

- (i) If $\theta \leq \theta_s$, we have $\theta \leq \alpha \leq 2\theta_s - \theta, -2\theta_s + \theta \leq \beta \leq -\theta$;
- (ii) If $\theta_s < \theta < 2\theta_s$, then we have $2\theta_s - \theta \leq \alpha \leq \theta, -\theta \leq \beta \leq \theta - 2\theta_s$.

Proof. For convenience, let us use system (3.44). For the first case, we construct a square bounded by L_1 , L_2 and their symmetric parts with respect to the line $\alpha - \beta = 2\theta_s$, as shown in Figure 3.4(a). By Lemma 3.3, we note that L_1 corresponds to H_1 and L_2 to H_2 . On the boundary L_1 , L_2 of this region, we have

$$G(\alpha, \beta, c) > 0, \quad \text{on } L_1, L_2. \quad (3.74)$$

On the other hand, the sign of G reverses on the counter parts. Note that the vector $(\sin \beta, -\cos \beta)$ on H_1 points toward the interior of Ω , and the vector $(\sin \alpha, -\cos \alpha)$ on H_2 points towards outside of Ω , see Figure 3.3. We conclude that such a square is invariant, see Figure 3.4(a).

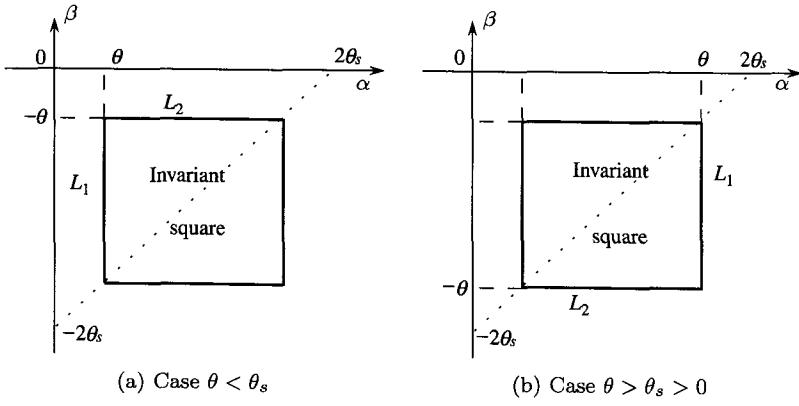


Figure 3.4 Invariant regions.

Similarly, we can treat the second case $\theta_s < \theta < 2\theta_s$. We construct a square as shown in Figure 3.4(b). The square resides in the fourth quadrant because $\theta < 2\theta_s$. Then we have

$$G(\alpha, \beta, c) < 0, \quad \text{on } L_1, L_2.$$

Therefore, the square is invariant.

However, the above proof has a problem since $G(\alpha, \beta, c)$ vanishes at the ends of L_1 and L_2 . To fix it, we consider a larger square and modify the system. Consider the first case for example. Let $\chi(\tau)$ be a nonnegative function in $C_c^1(-1, 1)$ with maximum value $\chi(0) = 1$. Let

$$f_\epsilon = \frac{4c}{\gamma + 1} G(\alpha, \beta, c) + \epsilon \chi\left(\frac{\alpha - (\theta - 2\epsilon)}{\epsilon}\right) - \epsilon \chi\left(\frac{\alpha - (2\theta_s - \theta + 2\epsilon)}{\epsilon}\right) \quad (3.75)$$

for any small $\epsilon > 0$. Consider replacing the equations of (3.44) with the modified ones, e.g. the first one with

$$\frac{4c}{\gamma + 1} \bar{\partial}_+ \alpha = f_\epsilon. \quad (3.76)$$

And consider the square with the left side on $\alpha = \theta - 2\epsilon$ and the right side on $\alpha = 2\theta_s - \theta + 2\epsilon$. We require ϵ small enough so that $\theta - 2\epsilon > 0$. On the left side we have $\bar{\partial}_+ \alpha > 0$ in the square so enlarged in all four sides. Thus we have $\alpha \geq \theta - 2\epsilon$. Sending ϵ to zero, we obtain $\alpha \geq \theta$. Similarly we obtain estimates on the other three sides. The proof is complete.

Corollary 3.1 (Invariant triangle). Suppose $\theta \in (\pi/6, 2\theta_s)$ ($\gamma < 2$). For solutions (α, β, c) of (3.44), (3.69) and (3.67), we have:

- (i) If $\theta \in (\pi/6, \theta_s)$, then $G(\alpha, \beta, c) > 0, \bar{\partial}_+ \alpha > 0, \bar{\partial}_- \beta > 0, \bar{\partial}_- \alpha > 0, \bar{\partial}_+ \beta > 0$ in Ω .
- (ii) If $\theta \in (\theta_s, 2\theta_s)$, then $G(\alpha, \beta, c) < 0, \bar{\partial}_+ \alpha < 0, \bar{\partial}_- \beta < 0, \bar{\partial}_- \alpha < 0, \bar{\partial}_+ \beta < 0$ in Ω .

Proof. For $\theta \in (\pi/6, \theta_s)$, and by the invariant square, we find that $3 \tan^2 \omega - 1 > 0$. Thus Q has the same sign with G in Theorem 3.2. Assuming $G > 0$, we find that $\bar{\partial}_- \beta > 0, \bar{\partial}_- \alpha > 0$. Thus the two variables α, β are both decreasing along a characteristic of the negative family. Since $\alpha - 2\theta_s < \beta$ on the boundary, there will never be $\alpha - 2\theta_s = \beta$ in finite length of time (along the characteristic), because at such an encounter, the derivative of β would be zero, but that of α is still strictly decreasing. For the other case, the signs of the derivatives are reversed, while β will be going up to get closer to $\alpha - 2\theta_s$, but never will get to it in finite step. \square

Remark 3.1. If the angle of the wedge θ and the adiabatic index γ are related by

$$\tan^2 \theta = \frac{3 - \gamma}{\gamma + 1}, \quad (3.77)$$

for $1 < \gamma < 3$, i.e., $\theta = \theta_s$, then boundary value (3.69) becomes constant $(\alpha, \beta)|_{H_j} = (\theta, -\theta)$, $j = 1, 2$. In this case the invariant region shrinks to

a point $(\theta, -\theta)$ on the line $\alpha - \beta = 2\theta_s$. Note that the source terms of (3.44) vanish on the boundaries H_1, H_2 . We obtain an explicit solution

$$c = 1 + \frac{\kappa}{\sin \theta} u, \quad (3.78)$$

where $-\sin \theta / \kappa \leq u \leq 0$. We further use (3.13) to get an explicit solution for the original gas expansion problem

$$\begin{aligned} c &= 1 + \frac{\kappa(\xi \sin \theta - 1)}{\kappa + \sin^2 \theta}, \\ u &= \frac{\sin \theta(\xi \sin \theta - 1)}{\kappa + \sin^2 \theta}, \\ v &= \eta. \end{aligned} \quad (3.79)$$

This solution was first observed in [37].

Remark 3.2. In the proof of Lemma 3.5, we observe that

$$\frac{\cos((\alpha + \beta)/2)}{\sin \omega} > \delta \quad (3.80)$$

for some constant $\delta > 0$. It follows from the third equation of (3.44) that

$$c < 1 + \delta u \quad (3.81)$$

for $u < 0$ and thus c vanishes at $u > -1/\delta$. Therefore there exists a curve $u = u(v)$ such that $c(u(v), v) = 0$ where $u = u(v)$ is well-defined in the (u, v) plane. This is the interface of gas and vacuum.

Remark 3.3. If $\theta < \pi/6, \gamma < 2$, then the characteristics in the (u, v) plane are neither convex nor concave: Their convexity types change along their paths. See Fig.3.5.

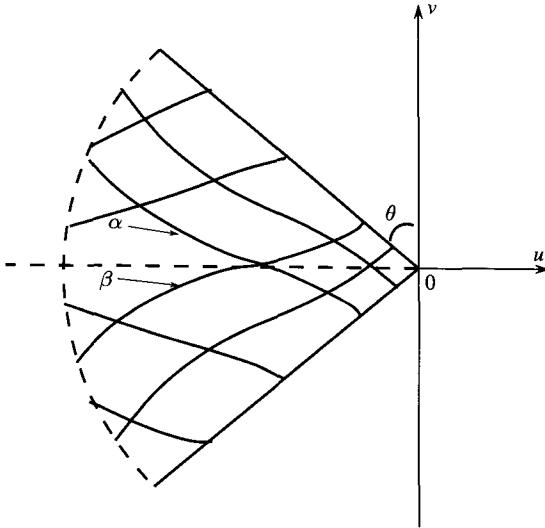
Corollary 3.2. For the gas expansion problem, the mapping $(X, Y) \rightarrow (\alpha, \beta)$ is bijective in the whole region Ω .

Proof. It suffices to check the non-degeneracy of the Jacobian from $(X, Y) \rightarrow (\alpha, \beta)$,

$$J(X, Y; \alpha, \beta) = -\frac{1}{\sin^2 \omega} \cdot \cot \omega. \quad (3.82)$$

In view of Lemma 3.5, we obtain the conclusion. \square

Corollary 3.2 show that we can convert system (3.31) into system (3.40) and therefore use system (3.40) or (3.44) to discuss Problem B in the hodograph plane.

Figure 3.5 Changes of convexity types of λ_{\pm} -characteristics.

3.6.7 Gradient estimates and the proof of Theorem 3.3

In order to establish the existence of smooth solutions in the whole wave interaction region Ω , we need to establish gradient estimates for system (3.40) or (3.44). Due to the degeneracy of interface l , we cut off a sufficient thin strip between the interface l and the level set of $c = \epsilon$, $\epsilon > 0$. The remaining sub-domain is denoted by Ω_ϵ , in which $c > \epsilon$. We first show that there is a unique solution on Ω_ϵ . Then we extend the solution to Ω by using the argument of the arbitrariness of $\epsilon > 0$.

Lemma 3.6 (Gradient estimate). Assume that there is a C^1 solution (α, β) in Ω_ϵ to system (3.40) or (3.44) with boundary values (3.69) and (3.67). Then the C^1 norm of α and β has a uniform bound $C = C(\theta, \gamma)$:

$$\|(\alpha, \beta)\|_{C^1(\Omega_\epsilon)} \leq C/\epsilon^2. \quad (3.83)$$

Proof. We use (3.49) to integrate $\bar{\partial}_-\alpha$ and $\bar{\partial}_+\beta$ along λ_+ and λ_- -characteristics respectively. Noting (3.45), we know that the integral path has a limited length. Also we note that Q has a uniform bound C/ϵ^2 in Ω_ϵ . Then we deduce that $\bar{\partial}_-\alpha$ and $\bar{\partial}_+\beta$ are uniformly bounded in Ω_ϵ ,

$$|\bar{\partial}_-\alpha| < C/\epsilon^2, \quad |\bar{\partial}_+\beta| < C/\epsilon^2. \quad (3.84)$$

On the other hand, since G has a bound C/ϵ in Ω_ϵ (see (3.44)), so are $\bar{\partial}_+\alpha$ and $\bar{\partial}_-\beta$,

$$|\bar{\partial}_+\alpha| < C/\epsilon, \quad |\bar{\partial}_-\beta| < C/\epsilon. \quad (3.85)$$

Hence using the identities

$$\begin{aligned}\partial_u &= -\sin^{-1}(\alpha - \beta)(\cos \alpha \bar{\partial}_+ - \cos \beta \bar{\partial}_-), \\ \partial_v &= -\sin^{-1}(\alpha - \beta)(\sin \alpha \bar{\partial}_+ - \sin \beta \bar{\partial}_-),\end{aligned}\quad (3.86)$$

and using the hyperbolicity $\alpha \neq \beta$ in Ω_ϵ , we conclude that $\partial_u \alpha$, $\partial_v \alpha$, $\partial_u \beta$ and $\partial_v \beta$ are uniformly bounded in Ω_ϵ , as expressed in (3.83). \square

Lemma 3.7 (Modulus estimate). Assume that the solution $(\alpha, \beta) \in C^1(\Omega_\epsilon)$. Then we have the following modulus estimate,

$$\|(\alpha, \beta)\|_{C^{1,1}(\Omega_\epsilon)} < C/\epsilon^2, \quad (3.87)$$

where $\|\cdot\|_{C^{1,1}(\Omega_\epsilon)}$ represents the norm of the space of functions whose C^1 -derivatives are Lipschitz continuous.

Proof. Using (3.49), we follow [10] or [30] to obtain

$$\|\bar{\partial}_- \alpha\|_{C^{0,1}(\Omega_\epsilon)} < C/\epsilon^2, \quad \|\bar{\partial}_+ \beta\|_{C^{0,1}(\Omega_\epsilon)} < C/\epsilon^2. \quad (3.88)$$

Then we use the same approach to derive

$$\|\bar{\partial}_+ \alpha\|_{C^{0,1}(\Omega_\epsilon)} < C/\epsilon^2, \quad \|\bar{\partial}_- \beta\|_{C^{0,1}(\Omega_\epsilon)} < C/\epsilon^2. \quad (3.89)$$

Thus the identities (3.86) are used to yield (3.87). \square

Proof of Theorem 3.3 With the classical technique in [30] or [10], we obtain the “global” solution in Ω_ϵ by the extension from the local solution.

In view of Lemma 3.4, we obtain a local solution (α, β, c) in $\Omega_\delta = \{(u, v) \in \Omega_\epsilon; -\delta < u < 0\}$. We take a level set of c , denoted by Υ_c , in Ω_δ . On this curve, (α, β, c) is known from the local solution and $(\alpha, \beta) \in C^1(\Upsilon_c)$ in view of Lemma 3.7. Then our problem becomes to find a solution of (3.40) in the remaining region, subject to the data on H_1 , H_2 and Υ_c .

Denote the slope of Υ_c by s_0 ,

$$s_0 := \frac{dv}{du} = -\frac{c_u}{c_v} = -\cot\left(\frac{\alpha + \beta}{2}\right). \quad (3.90)$$

Then we have

$$\frac{1}{s_0} - \frac{1}{\lambda_-} = \frac{\sin \omega}{\cos \frac{\alpha+\beta}{2} \cos \alpha} > 0, \quad \frac{1}{s_0} - \frac{1}{\lambda_+} = -\frac{\sin \omega}{\cos \frac{\alpha+\beta}{2} \cos \beta} < 0. \quad (3.91)$$

This shows that the level set Υ_c is not a characteristic and λ_\pm -characteristics always points toward the right hand side of Υ_c . Thus, we follow the proof of Lemma 4.1 in [10, Page 294], using Lemmas 3.6 and 3.7, to finish the proof of the existence of solutions in Ω_ϵ .

Owing to the arbitrariness of width $\epsilon > 0$, we use the contradiction argument to show that the C^1 solution (α, β, c) can be extend to the whole region Ω .

The discussion of vacuum boundary is left in subsubsection 3.6.10.

□

3.6.8 Inversion

We now discuss the inversion of the hodograph transformation, i.e., the Jacobian $J_T^{-1}(u, v; \xi, \eta)$ in (3.29) does not vanish for the gas expansion problem.

We look at the hodograph transformation $T : (\xi, \eta) \rightarrow (u, v)$. The mapping (3.13) defines a domain via $\xi = u + i_u$, $\eta = v + i_v$. We need to show that no two points map to one,

$$J_T^{-1}(u, v; \xi, \eta) = \xi_u \eta_v - \xi_v \eta_u = (1 + i_{uu})(1 + i_{vv}) - i_{uv}^2 \neq 0. \quad (3.92)$$

We calculate, on the one hand, multiplying (3.15) with $(1 + i_{uu})$,

$$\begin{aligned} & (2\kappa i - i_u^2)i_{uv}^2 + 2i_u i_v i_{uv}(1 + i_{uu}) + (2\kappa i - i_v^2)(1 + i_{uu})^2 \\ &= (2\kappa i - i_u^2)[i_{uv}^2 - (1 + i_{uu})(1 + i_{vv})]. \end{aligned} \quad (3.93)$$

On the other hand, from (3.23) and (3.32) we have

$$\begin{aligned} & (2\kappa i - i_u^2)i_{uv}^2 + 2i_u i_v i_{uv}(1 + i_{uu}) + (2\kappa i - i_v^2)(1 + i_{uu})^2 \\ &= (2\kappa i - i_v^2)(\partial_+ i_u + 1)(\partial_- i_u + 1). \end{aligned} \quad (3.94)$$

Then we obtain

$$\begin{aligned} J_T^{-1}(u, v; \xi, \eta) &= -\frac{(\partial_+ X + 1)(\partial_- X + 1)}{\lambda_- \lambda_+} \\ &= -\frac{(\bar{\partial}_+ X + \sin \beta)(\bar{\partial}_- X + \sin \alpha)}{\cos \alpha \cos \beta}, \end{aligned} \quad (3.95)$$

by using the definition of $\bar{\partial}_{\pm}$, see (3.43). This is parallel to (3.139). Therefore, in order to show that $J_T^{-1}(u, v; \xi, \eta)$ does not vanish, it is equivalent to prove that:

Lemma 3.8. The non-degeneracy of the Jacobian $J_T^{-1}(u, v; \xi, \eta)$ is equivalent to

$$\bar{\partial}_+ X + \sin \beta \neq 0 \text{ and } \bar{\partial}_- X + \sin \alpha \neq 0. \quad (3.96)$$

Recall the expression of X in terms of α, β in (3.38). Then we compute

$$\begin{aligned}\bar{\partial}_+ X + \sin \beta &= -\kappa \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega} - \frac{1+\kappa}{2} \cot \omega \cos \beta [m - \tan^2 \omega] \\ &\quad + \sin \beta + \frac{c}{2} \frac{\cos \alpha}{\sin^2 \omega} \bar{\partial}_+ \beta, \\ \bar{\partial}_- X + \sin \alpha &= \kappa \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega} + \frac{1+\kappa}{2} \cot \omega \cos \alpha [m - \tan^2 \omega] \\ &\quad + \sin \alpha - \frac{c}{2} \frac{\cos \beta}{\sin^2 \omega} \bar{\partial}_- \alpha.\end{aligned}\tag{3.97}$$

They are easily simplified to be

$$\begin{aligned}\bar{\partial}_+ X + \sin \beta &= -\frac{1+\kappa}{\sin(2\omega)} \cos \alpha + \frac{c}{2} \frac{\cos \alpha}{\sin^2 \omega} \bar{\partial}_+ \beta = \frac{c}{2} \frac{\cos \alpha}{\sin^2 \omega} [\bar{\partial}_+ \beta - Z], \\ \bar{\partial}_- X + \sin \alpha &= \frac{1+\kappa}{\sin(2\omega)} \cos \beta - \frac{c}{2} \frac{\cos \beta}{\sin^2 \omega} \bar{\partial}_- \alpha = -\frac{c}{2} \frac{\cos \beta}{\sin^2 \omega} [\bar{\partial}_- \alpha - Z],\end{aligned}\tag{3.98}$$

where

$$Z := \frac{1+\kappa}{c} \tan \omega.\tag{3.99}$$

Note that on the boundary H_1, H_2 , the values $\bar{\partial}_+ \beta$ and $\bar{\partial}_- \alpha$ are, respectively,

$$\bar{\partial}_+ \beta|_{H_2} \equiv 0, \quad \bar{\partial}_- \alpha|_{H_1} \equiv 0.\tag{3.100}$$

Therefore (3.96) follows from the following Lemma.

Lemma 3.9. We have

$$\begin{aligned}(\bar{\partial}_+ + W)(Z - \bar{\partial}_- \alpha) &= \frac{1+\kappa}{2c \cos^2 \omega} (Z - \bar{\partial}_+ \beta), \\ (-\bar{\partial}_- + W)(Z - \bar{\partial}_+ \beta) &= \frac{1+\kappa}{2c \cos^2 \omega} (Z - \bar{\partial}_- \alpha),\end{aligned}\tag{3.101}$$

and

$$\bar{\partial}_+ \beta < Z, \quad \bar{\partial}_- \alpha < Z\tag{3.102}$$

in the region Ω , see Figure 3.3.

Proof. The identities are obtained by computation. By the boundary condition on Z and (3.100) we obtain the inequalities (3.102). Here are the computations. By $\bar{\partial}_\pm c = \mp \kappa$ we have

$$\bar{\partial}_+ \frac{1}{c} \tan \omega = \frac{\kappa}{c^2} \tan \omega + \frac{1}{c \cos^2 \omega} \cdot \frac{1}{2} (\bar{\partial}_+ \alpha - \bar{\partial}_+ \beta).$$

Inserting the expression of $\bar{\partial}_+\alpha$, we find

$$\bar{\partial}_+ \frac{1}{c} \tan \omega = \frac{1}{2c \cos^2 \omega} \left\{ \frac{\kappa}{c} \sin(2\omega) + \frac{1+\kappa}{2c} \sin(2\omega) [m - \tan^2 \omega] - \bar{\partial}_+ \beta \right\}.$$

Inserting the expression of W , we have

$$(\bar{\partial}_+ + W)Z = -\frac{1+\kappa}{2c \cos^2 \omega} \bar{\partial}_+ \beta + Q + \frac{(1+\kappa)^2}{4c^2} \frac{\sin(2\omega)}{\cos^2 \omega} (\tan^2 \omega + m + \frac{2\kappa}{1+\kappa}).$$

Noting that $m + \frac{2\kappa}{1+\kappa} = 1$, and by Theorem 3.2, we obtain the first identity. The proof of the second one is similar.

To establish the inequalities, we note that both inequalities hold at the origin, and thus they hold in a neighborhood of the origin. Let $c = \epsilon \in (0, 1)$ be the first level curve on which at least one of the two strict inequalities becomes equality. Note that the level curves of c are transversal to the characteristics. We integrate identities (3.101) in the domain $\epsilon < c < 1$ along characteristics to yield strict inequalities (3.102), resulting in a contradiction. Thus, both strict inequalities must hold up to $c = 0$. \square

Bounded and non-vanishing Jacobian guarantees local one-to-one, but not always global one-to-one, as seen in the mapping $Z \rightarrow \exp Z$ where $Z = x + iy$ and i is the imaginary unit, on the domain $\{(x, y) \mid x \in (1, 2), y \in (0, 100)\}$. It maps a long strip to an annulus bounded by radii e and e^2 with multiple coverage, while the Jacobian never vanishes in the rectangle. We establish the global one-to-one, which is guaranteed by the monotonicity of ξ and η along characteristics. In fact, $\partial_{\pm} \xi = 1 + \partial_{\pm} X \neq 0$ (or $\pm \bar{\partial}_{\pm} \xi < 0$) following the above lemma. From (3.12) we have $\partial_+ \xi = -\lambda_- \partial_+ \eta$, $\partial_- \xi = -\lambda_+ \partial_- \eta$, thus ξ and η have the same monotonicity along a plus characteristic curve since $-\lambda_- > 0$, but opposite monotonicity along a minus characteristics since $-\lambda_+ < 0$. For any two points in the interaction zone in the (u, v) plane, there exist two characteristic curves connecting the two points. Either ξ or η is monotone along the connecting path. Thus, no two points from the (u, v) domain maps to one point in the (ξ, η) plane.

3.6.9 Proof of Theorem 3.4

The above estimates are sufficient for the proof of Theorem 3.4. For completeness, we sum it as follows. First we use the hodograph transformation (3.11) to convert Problem A into Problem B. Since the interaction region \mathcal{D} in Figure 3.2(b) is a wave interaction region, the Jacobian $J_T(u, v; \xi, \eta)$ does not vanish in view of Theorem 3.7, so the hodograph transformation (3.11) is valid. Then we solve Problem B in Theorem 3.3. In Lemmas 3.9 and 3.8, we show that the hodograph

transformation $J_T(u, v; \xi, \eta)$ is invertible. Thus the proof of Theorem 3.4 is complete. \square

3.6.10 Properties of the solutions

a. Convexity of characteristics in the physical plane

Now we discuss the convexity of Λ_{\pm} -characteristics in the mixed wave region \mathcal{D} , in the (ξ, η) plane. It is a rather simple way to look at this from the correspondence between the (ξ, η) plane and the (u, v) plane.

Consider the hodograph transformation T of (3.11). We note, by using the chain rule, that

$$\frac{\partial}{\partial u} + \lambda_+ \frac{\partial}{\partial v} = \left(\frac{\partial \xi}{\partial u} + \lambda_+ \frac{\partial \xi}{\partial v} \right) \frac{\partial}{\partial \xi} + \left(\frac{\partial \eta}{\partial u} + \lambda_+ \frac{\partial \eta}{\partial v} \right) \frac{\partial}{\partial \eta}. \quad (3.103)$$

We rewrite (3.12) as

$$\frac{\partial \xi}{\partial u} + \lambda_+ \frac{\partial \xi}{\partial v} = -\lambda_- \left(\frac{\partial \eta}{\partial u} + \lambda_+ \frac{\partial \eta}{\partial v} \right). \quad (3.104)$$

Using (3.13), we have

$$\frac{\partial \xi}{\partial u} + \lambda_+ \frac{\partial \xi}{\partial v} = \partial_+ X + 1. \quad (3.105)$$

Thus we derive a differential relation from (3.103), by noting $\Lambda_+ = -1/\lambda_-$,

$$\bar{\partial}_+ = (\bar{\partial}_+ X + \sin \beta) \left(\frac{\partial}{\partial \xi} + \Lambda_+ \frac{\partial}{\partial \eta} \right). \quad (3.106)$$

Similarly, we have

$$\bar{\partial}_- = (\bar{\partial}_- X + \sin \alpha) \left(\frac{\partial}{\partial \xi} + \Lambda_- \frac{\partial}{\partial \eta} \right). \quad (3.107)$$

Acting (3.106) on Λ_+ and (3.107) on Λ_- as well as using the definition of α, β (i.e., $\Lambda_+ = \tan \alpha, \Lambda_- = \tan \beta$), we obtain

$$\begin{aligned} \left(\frac{\partial}{\partial \xi} + \Lambda_+ \frac{\partial}{\partial \eta} \right) \Lambda_+ &= \frac{1}{\cos^2 \alpha} \cdot (\bar{\partial}_+ X + \sin \beta)^{-1} \cdot \bar{\partial}_+ \alpha, \\ \left(\frac{\partial}{\partial \xi} + \Lambda_- \frac{\partial}{\partial \eta} \right) \Lambda_- &= \frac{1}{\cos^2 \beta} \cdot (\bar{\partial}_- X + \sin \alpha)^{-1} \cdot \bar{\partial}_- \beta. \end{aligned} \quad (3.108)$$

Therefore, the convexity of Λ_{\pm} -characteristics is determined by two factors respectively. By Corollary 3.1, the signs of $\bar{\partial}_+ \alpha$ and $\bar{\partial}_- \beta$ just depend on the relation between the wedge angle θ and the index θ_s , i.e.,

$$\bar{\partial}_+ \alpha < 0, \quad \bar{\partial}_- \beta < 0, \quad (3.109)$$

if $\theta \in (\theta_s, 2\theta_s)$; and

$$\bar{\partial}_+ \alpha > 0, \quad \bar{\partial}_- \beta > 0, \quad (3.110)$$

if $\theta \in (\pi/6, \theta_s)$. In view of Lemma 3.9, we have

$$\bar{\partial}_+ X + \sin \beta < 0, \quad \bar{\partial}_- X + \sin \alpha > 0. \quad (3.111)$$

Hence we conclude

$$\left(\frac{\partial}{\partial \xi} + \Lambda_+ \frac{\partial}{\partial \eta} \right) \Lambda_+ > 0, \quad \left(\frac{\partial}{\partial \xi} + \Lambda_- \frac{\partial}{\partial \eta} \right) \Lambda_- < 0, \text{ for } \theta \in (\theta_s, 2\theta_s) \quad (3.112)$$

and

$$\left(\frac{\partial}{\partial \xi} + \Lambda_+ \frac{\partial}{\partial \eta} \right) \Lambda_+ < 0, \quad \left(\frac{\partial}{\partial \xi} + \Lambda_- \frac{\partial}{\partial \eta} \right) \Lambda_- > 0, \text{ for } \theta \in (\pi/6, \theta_s). \quad (3.113)$$

Theorem 3.5. The Λ_{\pm} -characteristics in the wave interaction region \mathcal{D} of (ξ, η) plane have fixed convexity types:

- (i) If $\theta \in (\theta_s, 2\theta_s)$, the Λ_+ -characteristics are convex and the Λ_- -characteristics are concave.
- (ii) If $\theta \in (\pi/6, \theta_s)$, the Λ_+ -characteristics are concave and the Λ_- -characteristics are convex.
- (iii) If $\theta = \theta_s$, the solution has the explicit form (3.78) and all characteristics are straight.

b. Regularity of the vacuum boundary

Recall that formula (3.13) transforms the solution (α, β, c) from the (u, v) plane back into the (ξ, η) -plane. Note that (α, β) , and thus c_u and c_v , are uniformly bounded for $1 < \gamma < 3$, and that c tends to zero with a rate much faster than c_u, c_v for $\gamma \geq 3$. We conclude that on the vacuum boundary, the (u, v) coordinates coincide with the (ξ, η) coordinates. In fact, by using (3.13), we have

$$\xi = u + i_u = u + \frac{c}{\kappa} c_u = u, \quad \eta = v + i_v = v + \frac{c}{\kappa} c_v = v. \quad (3.114)$$

We prove that the vacuum boundary is Lipschitz continuous. Let us consider the curve $\{(u, v) \mid i(u, v) = \epsilon > 0\}$ for all small positive ϵ . Differentiating the equation $i(u(v), v) = \epsilon$ with respect to v , we find

$$\frac{du}{dv} = -\frac{Y}{X} = -\tan\left(\frac{\alpha + \beta}{2}\right). \quad (3.115)$$

Since $|\alpha + \beta| < \pi/2$ uniformly with respect to $\epsilon > 0$, the level curve $i(u, v) = \epsilon$ has a bounded derivative and in the limit as $\epsilon \rightarrow 0+$ converges to a Lipschitz continuous vacuum boundary.

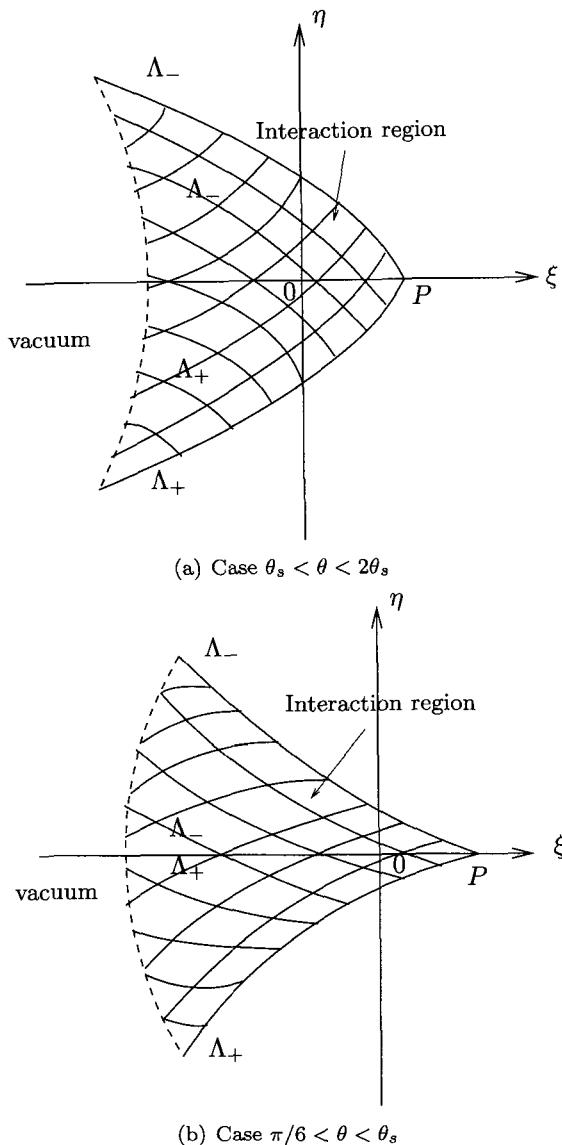


Figure 3.6 Convexity types of the characteristics and the vacuum boundaries in the (ξ, η) plane are opposite to each other in the two cases.

c. Relative location

For the explicit solution of the case $\theta = \theta_s$, the vacuum boundary is a vertical segment. Now we hold θ fixed and consider varying γ so that $\theta < \theta_s$. Then we find α and β lie on the left-hand side of the line

$\alpha - \beta = 2\theta_s$ in the $\alpha - \beta$ phase plane. By the formula $i_v = Y$ and the location of the boundary data, we have $Y < 0$ on the upper half of the wedge, thus i is monotone decreasing in v on the upper half, hence the vacuum boundary is on the left of the Suchkov boundary and of a concave type. Similarly, the other case $\theta > \theta_s$ has the opposite result.

Theorem 3.6. Let the vacuum boundary be represented as $\xi = \xi(\eta)$. Then it is Lipschitz continuous. It is less than the Suchkov solution boundary and is convex if $\theta < \theta_s$, but it is concave and greater than the Suchkov solution boundary for $\theta > \theta_s$.

d. Characteristics on the vacuum boundary

We already know that the sound speed c attains zero in a finite range of u . Conversely, from (3.45) we deduce that the lengths of λ_{\pm} -characteristics are limited. Then in view of (3.44) it can be seen that on the vacuum boundary there must hold

$$\alpha - \beta = 2\theta_s. \quad (3.116)$$

In fact, if (3.116) were not true, then $\bar{\partial}_+ \alpha$ and $\bar{\partial}_- \beta$ would become infinity as (u, v) approaches the vacuum boundary, which forces (α, β) to reach the line $\alpha - \beta = 2\theta_s$ in the (α, β) -plane. See Figure 3.6. In particular, for $1 < \gamma < 3$, a Λ_+ -characteristic line has a non-zero intersection angle with a Λ_- -characteristic line. However, if $\gamma \geq 3$, we have $\alpha = \beta$ on the vacuum boundary such that they are all tangent to the vacuum boundary.

3.7 Summary remarks

We have considered the phase space equation (3.14)

$$(2\kappa i - i_u^2)i_{vv} + 2i_u i_v i_{uv} + (2\kappa i - i_v^2)i_{uu} = i_u^2 + i_v^2 - 4\kappa i \quad (3.117)$$

known from 1958 for the enthalpy i with the inverse of the hodograph transformation (3.13)

$$\xi = u + i_u, \quad \eta = v + i_v \quad (3.118)$$

for the two-dimensional self-similar isentropic irrotational Euler system. Upon introducing the variables of inclination angles of characteristics and normalized characteristic derivatives, we have changed the second-order phase space equation to a first order system (3.44),

$$\begin{aligned} \bar{\partial}_+ \alpha &= G(\alpha, \beta, c), & \bar{\partial}_- \beta &= G(\alpha, \beta, c), \\ \partial_0 c &= \kappa \cos\left(\frac{\alpha+\beta}{2}\right) / \sin\left(\frac{\alpha-\beta}{2}\right), \end{aligned} \quad (3.119)$$

where

$$G(\alpha, \beta, c) = \frac{1 + \kappa}{2c} \cdot \sin(\alpha - \beta) \cdot \left[m - \tan^2 \frac{\alpha - \beta}{2} \right].$$

Derivatives of the variables α and β along directions not represented in (3.119) are provided by the higher-order system (3.49). We use these infrastructure to construct solutions to binary interactions of planar waves in the phase space and show that the Jacobian of the inverse of the hodograph transformation does not vanish, so we obtain in particular a global solution to the gas expansion problem with detailed shapes and positions of the vacuum boundaries and characteristics.

The invariant regions in the phase space revealed in the process have more potential than what has been used here. For example, we will use them to handle binary interactions of simple waves, which will lead to the eventual construction of global solutions to some four-wave Riemann problems that will not have vacuum in their data, see a forthcoming paper [29].

We have made a comparison of the pair (3.117) and (3.118) to the pair of eigenvalue $\xi = \lambda(u)$ and wave curve system $(\lambda - f'(u))u' = 0$ for the one-dimensional system $u_t + f(u)_x = 0$ from Lax [18]. The wave curves of the one-dimensional case correspond to surfaces in the phase space (i, u, v) . It will be a very interesting next step to find out the phase space structure that involves subsonic domains and shock waves as well as the hyperbolic surfaces.

It is proven in [28] that the solution in the hodograph plane can be transformed back to the physical self-similar plane for all $\gamma > 1$ and the vacuum boundary is a Lipschitz continuous curve which is monotone in the upper and lower parts of the wedge respectively. They also determine explicitly the relative location of the vacuum boundary with respect to the vertical position of the explicit solution of Suchkov [37]. Moreover, they have drawn a clear picture of the distribution of characteristics.

This section is based heavily on the joint work [28] with Jiequan Li. Some of the subsections are written by Jiequan. The author thanks Jiequan very much for his help.

3.8 Appendix A: simple waves

3.8.1 Concept of simple waves

Now we recall some facts about simple waves. Simple waves were systematically studied, e.g. in [18], for hyperbolic systems in two independent variables,

$$u_t + A(u)u_x = 0, \quad (3.120)$$

where $u = (u_1, \dots, u_n)^\top$, the $n \times n$ matrix $A(u)$ has real and distinct eigenvalues $\lambda_1 < \dots < \lambda_n$ for all u under consideration. They are defined as a special family of solutions of the form

$$u = U(\phi(x, t)). \quad (3.121)$$

The function $\phi = \phi(x, t)$ is scalar. Substituting (3.121) into (3.120) yields

$$U'(\phi)\phi_t + A(U(\phi))U'(\phi)\phi_x = 0, \quad (3.122)$$

which implies that $-\phi_t/\phi_x$ is an eigenvalue of $A(U(\phi))$ and $U'(\phi)$ is the associated eigenvector. This concludes that in the (x, t) plane a simple wave is associated with a kind of characteristic field, say, λ_k , and spans a domain in which characteristics of the k -kind are straight along which the solution is constant.

The property of simple waves can be analyzed by using *Riemann invariants*. A Riemann invariant is a scalar function $w = w(x, t)$ satisfying the following condition,

$$r_k \cdot \text{grad } w = 0 \quad (3.123)$$

for all values of u , where r_k is the k -th right eigenvector of A . Using the Riemann invariants, it can be shown that *a state in a domain adjacent to a domain of constant state is always a simple wave*.

In general, system (3.120) is not endowed with a *full coordinate system of Riemann invariants* such that it is transformed into a diagonal form [9]. Note that in (3.120) the coefficient matrix A depends on u only. Once A depends on x and t as well as u , the treatment in [18] and [9] breaks down. For example, we are unable to use the same techniques to show that it is a simple wave to be adjacent to a constant state.

3.8.2 Simple waves for pseudo-steady Euler equations

We introduce in a traditional manner a *simple wave* for (3.4) as a solution $(u, v) = (u, v)(\xi, \eta)$ that is constant along the level set $l : l(\xi, \eta) = C$ for some function $l(\xi, \eta)$, where C is constant. That is, this solution has the form:

$$(u, v)(\xi, \eta) = (F, G)(l(\xi, \eta)). \quad (3.124)$$

Inserting this into (3.4) gives

$$\begin{pmatrix} (2\kappa i - U^2)l_\xi - UVl_\eta, -UVl_\xi + (2\kappa i - V^2)l_\eta \\ l_\eta \\ l_\xi \end{pmatrix} \begin{pmatrix} F' \\ G' \end{pmatrix} = 0. \quad (3.125)$$

Here we use $U := u - \xi$, $V := v - \eta$ for short. It turns out that $(F', G') = (0, 0)$ or there exists a singular solution for which the coefficient matrix

becomes singular. The former just gives a trivial constant solution. But for the latter, $l(\xi, \eta)$ satisfies

$$(2\kappa i - U^2)l_\xi^2 - 2UVl_\xi l_\eta + (2\kappa i - V^2)l_\eta^2 = 0, \quad (3.126)$$

i.e.,

$$\frac{l_\xi}{l_\eta} = \frac{UV \pm \{2\kappa i(U^2 + V^2 - 2\kappa i)\}^{1/2}}{U^2 - 2\kappa i} =: \Lambda_\pm, \quad (3.127)$$

which implies that the level curves $l(\xi, \eta) = C$ are characteristic lines, and

$$F' + \Lambda_\pm G' = 0 \quad (3.128)$$

holds along each characteristic line $l(\xi, \eta) = C$ locally at least.

In a recent paper by Li, Zhang, Zheng [27], the pseudo-steady full Euler is shown to have a characteristic decomposition. Let us quote several identities from that paper. First, the flow will be isentropic and irrotational adjacent to a constant state. Then the pseudo-characteristics are defined as

$$\frac{d\eta}{d\xi} = \frac{UV \pm c\sqrt{U^2 + V^2 - c^2}}{U^2 - c^2} \equiv \Lambda_\pm. \quad (3.129)$$

Here c is the speed of sound $c^2 = \gamma p/\rho$. Regarding Λ_\pm as simple straight functions of the three independent variables (U, V, c^2) , we have

$$\partial_U \Lambda = \Lambda(U\Lambda - V)/\Theta, \quad \partial_V \Lambda = (V - U\Lambda)/\Theta, \quad \partial_{c^2} \Lambda = -(1 + \Lambda^2)/(2\Theta), \quad (3.130)$$

where $\Theta := \Lambda(c^2 - U^2) + UV$. Then we further obtain

$$\partial^\pm u + \Lambda_\mp \partial^\pm v = 0, \quad (3.131)$$

$$\partial^\pm c^2 = -2\kappa (U\partial^\pm u + V\partial^\pm v), \quad (3.132)$$

$$\partial^\pm \Lambda_\pm = [\partial_U \Lambda_\pm - \Lambda_\mp^{-1} \partial_V \Lambda_\pm - 2\kappa(U - V/\Lambda_\mp) \partial_{c^2} \Lambda_\pm] \partial^\pm u, \quad (3.133)$$

where $\partial^\pm = \partial_\xi + \Lambda_\pm \partial_\eta$. We keep ∂_\pm for later use in the hodograph plane. Thus, if one of the quantities (u, v, c^2) is a constant along Λ_- , so are the remaining two and Λ_- . The same is true for the plus family Λ_+ . Hence we have the following Proposition.

Proposition 3.1 (Section 4, [27]). For the irrotational and isentropic pseudo-steady flow (3.2) or (3.4), we have the following characteristic decomposition

$$\partial^+ \partial^- u = h \partial^- u, \quad \partial^- \partial^+ u = g \partial^+ u, \quad (3.134)$$

where $h = h(\xi, \eta)$ and $g = g(\xi, \eta)$ are some functions. Similar decompositions hold for v , c^2 and Λ_\pm . We further conclude that simple waves are waves such that one family of characteristic curves are straight along which the physical quantities (u, v, c^2) are constant.

Appendix B: convertibility

We are now ready to discuss the non-degeneracy of hodograph transformation (3.11).

Theorem 3.7 (Sufficient and Necessary Condition). Let the irrotational, isentropic and pseudo-steady fluid flow (3.2) be smooth at a point $(\xi, \eta) = (\xi_0, \eta_0)$. Then the Jacobian $J_T(u, v; \xi, \eta)$ of the hodograph transformation (3.11) vanishes in a neighborhood of the point if the flow is a simple wave in the neighborhood. Conversely, if the Jacobian $J_T(u, v; \xi, \eta)$ vanishes in a neighborhood of the point, then the flow is a simple wave in the neighborhood.

Proof. Assume first that $c^2 - V^2 \neq 0$ at $(\xi, \eta) = (\xi_0, \eta_0)$. We compute

$$\begin{aligned} J_T(u, v; \xi, \eta) &= u_\xi v_\eta - u_\eta v_\xi \\ &= -\frac{1}{c^2 - V^2} \cdot [((c^2 - U^2)u_\xi - 2UVu_\eta)u_\xi] - u_\eta^2 \\ &= -\frac{1}{c^2 - V^2} \cdot [(c^2 - U^2)u_\xi^2 - 2UVu_\xi u_\eta + (c^2 - V^2)u_\eta^2]. \end{aligned} \quad (3.135)$$

Therefore, the degeneracy of the transformation implies

$$(c^2 - U^2)u_\xi^2 - 2UVu_\xi u_\eta + (c^2 - V^2)u_\eta^2 = 0. \quad (3.136)$$

It follows that

$$-\frac{u_\xi}{u_\eta} = \Lambda_\pm. \quad (3.137)$$

That is

$$u_\xi + \Lambda_+ u_\eta = 0, \quad \text{or} \quad u_\xi + \Lambda_- u_\eta = 0 \quad (3.138)$$

at (ξ_0, η_0) . For the former, we deduce that $\partial^+ u = 0$ along the whole Λ_+ -characteristic line through (ξ_0, η_0) in view of (3.134) in Proposition 3.1, and so do $\partial^+ v$ and $\partial^+ c$. Therefore, we conclude that the wave is a simple wave associated with Λ_+ .

Conversely, if a point $(\xi, \eta) = (\xi_0, \eta_0)$ is in the region of a simple wave, then equation (3.138) holds either for the plus or minus families. From there we go up the derivation to find that the Jacobian vanishes in the same neighborhood.

The case that $c^2 - (v - \eta)^2 = 0$ is a special planar simple wave. Therefore the conclusion follows naturally. \square

We comment that the Jacobian $J_T(u, v; \xi, \eta)$ can be factorized as

$$J_T(u, v; \xi, \eta) = -\frac{1}{\Lambda_- \Lambda_+} \partial^+ u \cdot \partial^- u = -\partial^+ v \cdot \partial^- v. \quad (3.139)$$

4 Local solutions for quasilinear systems

We present the classical results of the existence of C^1 smooth solutions to Cauchy, Goursat, and mixed type problems for quasilinear hyperbolic systems of equations in two independent variables.

4.1 Introduction

We consider the system

$$\frac{\partial u}{\partial x} + A(x, y, u) \frac{\partial u}{\partial y} = c(x, y, u), \quad (4.1)$$

where

$$A = (a_{ij})_{n \times n}, \quad u = (u_i)_{i=1}^n, \quad c = (c_i)_{i=1}^n$$

for $(x, y, u) \in \mathcal{D}$ where \mathcal{D} is a bounded closed domain in \mathbb{R}^{2+n} .

Definition 4.1. A system (4.1) is called **hyperbolic** in \mathcal{D} if for all $(x, y, p) \in \mathcal{D}$, there exists n real eigenvalues

$$\lambda_1(x, y, p) \leq \lambda_2(x, y, p) \leq \cdots \leq \lambda_n(x, y, p) \quad (4.2)$$

and n linearly independent left eigenvectors $\ell^{(i)}(x, y, p)$ ($i = 1, 2, \dots, n$) such that

$$\ell^{(i)}(\lambda_i I - A) = 0 \quad (i = 1, 2, \dots, n). \quad (4.3)$$

A system (4.1) is called **strictly hyperbolic** in \mathcal{D} if the eigenvalues are distinct for all $(x, y, p) \in \mathcal{D}$.

Definition 4.2. A system (4.1) is called **continuously differentiable** if all

$$A(x, y, p), \quad c(x, y, p), \quad \lambda_i(x, y, p), \quad \ell^{(i)}(x, y, p) \quad (4.4)$$

are continuously differentiable in \mathcal{D} (i.e., first-order derivatives are continuous).

Proposition 4.1. If the matrix $A(x, y, p)$ is continuously differentiable on \mathcal{D} , and system (4.1) is strictly hyperbolic, then $\{\lambda_i\}_{i=1}^n$ are all continuously differentiable, and $\{\ell^{(i)}\}_{i=1}^n$ can be chosen to be continuously differentiable.

We omit its proof.

We multiply system (4.1) with the left eigenvectors to derive the *characteristic form* (or *characteristic relation*)

$$\ell^{(i)} \left(\frac{\partial u}{\partial x} + \lambda_i \frac{\partial u}{\partial y} \right) = b_i, \quad (i = 1, 2, \dots, n) \quad (4.5)$$

or

$$\ell^{(i)} \frac{du}{di} = b_i, \quad (4.6)$$

where

$$\frac{d}{di} := \frac{\partial}{\partial x} + \lambda_i \frac{\partial}{\partial y}, \quad b_i := \ell^{(i)} c.$$

If $\{\ell^{(i)}\}_{i=1}^n$ are continuous on \mathcal{D} , then

$$|\det(\ell^{(i)})| \geq \alpha > 0$$

for some constant positive $\alpha > 0$.

For a given solution $u(x, y)$, we define the *characteristics* of system (4.1) by

$$\frac{dy}{dx} = \lambda_i(x, y, u). \quad (4.7)$$

Cauchy problem. Let Γ be a smooth curve in the (x, y) plane. Let u be given on Γ such that none of the λ_i 's is tangent to Γ . Find u . A special case is when Γ is a segment of the axis $x = 0$ and the initial data is

$$u(0, y) = \phi(y),$$

where $\phi(y)$ is a given function.

Goursat problem (characteristic boundary value problem). Let Γ_1, Γ_2 be two smooth curves on the half-plane $x \geq 0$, which pass through the origin $(0, 0)$. Assume $\Gamma_1(x) \leq \Gamma_2(x)$ ($x \geq 0$), u is given on both Γ_1 and Γ_2 such that Γ_1 is a characteristic curve of the first family λ_1 and Γ_2 is a characteristic curve of the n -th family λ_n , and the characteristic relations (4.6) hold on both Γ_i ($i = 1, 2$). Find u in the sectorial domain formed by Γ_1 and Γ_2 , see Figure 4.1.

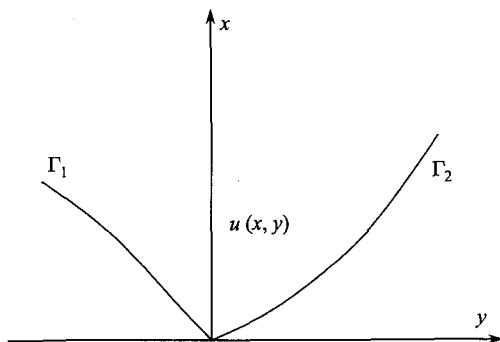


Figure 4.1 Goursat problem: Find u in the sector.

Mixed characteristic boundary value and initial value problem. On Γ_1 and Γ_2 are given the values of u , such that Γ_1 is a characteristic curve of the first family λ_1 and the characteristic relation hold on it, while all characteristic directions $\lambda_2, \dots, \lambda_n$ on Γ_2 point into the sectorial domain (in the positive x axis). Find u in the sectorial domain.

4.2 Existence of solutions to the Cauchy problem

We study the Cauchy problem for system (4.1) for which the initial value is

$$u(0, y) = \phi(y), \quad y_1 \leq y \leq y_2, \quad (4.8)$$

where $y_1 < y_2$ and

$$|\phi(y) - u^{(0)}| \leq \frac{1}{2}M \quad \text{in } y_1 \leq y \leq y_2 \quad (4.9)$$

for a constant vector $u^{(0)}$ and some real number $M > 0$. The norm of a (column) vector is defined to be the largest of the norms of its components.

Assume we have two continuously differentiable functions $y_1(x), y_2(x)$ on $[0, \delta_0]$ for some $\delta_0 > 0$, satisfying $y_1(0) = y_1, y_2(0) = y_2, y_1(x) < y_2(x)$ for $x \in [0, \delta_0]$. Let

$$R(\delta) : 0 \leq x \leq \delta, \quad y_1(x) \leq y \leq y_2(x),$$

where $\delta \in (0, \delta_0]$.

Assume system (4.1) is continuously differentiable and hyperbolic on

$$R_M(\delta_0) := \{(x, y, u) \mid (x, y) \in R(\delta_0), |u - u^{(0)}| \leq M\}.$$

Assume $\phi(y)$ is continuously differentiable in $[y_1, y_2]$.

Definition 4.3. We call $R(\delta)$ a *strong determinate domain* of system (4.1), if each characteristic curve passing through any point $(x_0, y_0) \in R(\delta)$ ($x_0 > 0$) of system (4.1) for any continuously differentiable function $u(x, y)$ on $R(\delta)$ satisfying

$$|u(x, y) - u^{(0)}| \leq M, \quad (4.10)$$

remain in $R(\delta)$ for $x \leq x_0$ until its intersection with the initial segment $y_1 \leq y \leq y_2$.

We see obviously that $R(\delta)$ is a strong determinate domain if $R(\delta_0)$ is one and $\delta \in (0, \delta_0)$. If (4.1) is semi-linear, then the domain formed by the lines $x = 0, x = \delta$, and the characteristics of the first and last families (through y_2 and y_1 respectively) is a strong determinate domain.

We use the iteration proposed by Courant and Lax [8]

$$d_i[\ell^{(i)}(v)u] = b_i(v) + v d_i \ell^{(i)}(v), \quad (i = 1, 2, \dots, n) \quad (4.11)$$

with data $u(0, y) = \phi(y)$, where $d_i := \partial_x + \lambda_i(v)\partial_y$, which yields a mapping

$$u = Tv.$$

A solution to Cauchy problem (4.1), (4.8) is the same as a fixed point of T .

Theorem 4.1. Assume $\phi(y)$ is continuously differentiable on $[y_1, y_2]$ with a bound M in (4.9). Assume system (4.1) is hyperbolic and continuously differentiable on $R_M(\delta_0)$ for some $\delta_0 > 0$. Assume $R(\delta_0)$ is a strong determinate domain of (4.1). Then there exists a $\delta > 0$, such that problem (4.1), (4.8) has a continuously differentiable solution on $R(\delta)$.

4.2.1 Primary representations

Let $v(x, y)$ be an *admissible function* on $R(\delta)$ ($\delta \leq \delta_0$); i.e., $v(x, y)$ is continuously differentiable on $R(\delta)$, satisfying the initial condition (4.8), and the inequality (4.10).

Let $y = y_i(x; \xi, \eta)$ be the characteristic $C_i(v)$:

$$\frac{dy}{dx} = \lambda_i(x, y, v(x, y)), \quad y(\xi) = \eta \quad (4.12)$$

for $(\xi, \eta) \in R(\delta)$. The characteristic y_i exists for all $0 \leq x \leq \xi$ and $(x, y_i(x; \xi, \eta)) \in R(\delta)$ because $R(\delta_0)$ is a strong determinate domain.

Along $C_i(v)$ from 0 to ξ , we integrate (4.11) and use initial data (4.8) to obtain

$$\ell^{(i)}(v)u = \ell^{(i)}(\phi(y_i(0; \xi, \eta)))\phi(y_i(0; \xi, \eta)) + \int_0^\xi [b_i(v) + v \frac{d}{dx} \ell^{(i)}(v)] dx \quad (4.13)$$

for $i = 1, 2, \dots, n$. Here the variables in the integrand are $(x, y_i(x; \xi, \eta))$ while the details of $\ell^{(i)}(\phi(y_i(0; \xi, \eta)))$ are $\ell^{(i)}(0, y_i(0; \xi, \eta), \phi(y_i(0; \xi, \eta)))$.

Once we show that u from (4.13) is C^1 , then it will be a C^1 solution to (4.11) with (4.8), and the unique one, thus the mapping T will be established. We show that u has continuous derivatives u_ξ, u_η on $R(\delta)$. We consider u_η first. The continuity of u_η is not obvious from the representation (4.13) since we assume only that system (4.1) and v are C^1 and already first-order derivatives are used in the expression. But formally we can differentiate

$$\int_0^\xi v \frac{d}{dx} \ell^{(i)}(v) dx \quad (4.14)$$

with respect to η to obtain

$$\begin{aligned} & v(x, y_i(x; \xi, \eta)) \frac{\partial}{\partial \eta} \ell^{(i)}(x, y_i(x; \xi, \eta), v(x, y_i(x; \xi, \eta)))|_0^\xi \\ & + \int_0^\xi \left[\frac{\partial v}{\partial \eta} \frac{d}{dx} \ell^{(i)}(v) - \frac{dv}{dx} \frac{\partial \ell^{(i)}(v)}{\partial \eta} \right] dx, \end{aligned} \quad (4.15)$$

from which and (4.13) we infer the continuity of u_η . We used integration by parts in the process in obtaining (4.15). Rigorous derivation of (4.15) is to use difference quotient, integration by parts for the difference quotient, and then take limit. For more details, see Chapter 6, Sect. 3 of L. C. Evans [11] or Hartman and Winter [15].

The continuity of u_ξ is similar. So we have obtained the C^1 regularity of u .

Note that we have

$$\frac{\partial y_i(x; \xi, \eta)}{\partial \eta} = \exp \left(\int_\xi^x \frac{\partial \lambda_i(v)}{\partial y} dx \right). \quad (4.16)$$

From (4.13), (4.14), (4.15) we obtain

$$\begin{aligned} \ell^{(i)}(v) \frac{\partial u}{\partial \eta} &= \phi'(y_i(0; \xi, \eta)) \ell^{(i)}(\phi(y_i(0; \xi, \eta))) \exp \left(\int_\xi^0 \frac{\partial \lambda_i(v)}{\partial y} dx \right) + (v - u) \\ &\cdot \frac{\partial \ell^{(i)}(v)}{\partial \eta} + \int_0^\xi \left[\frac{\partial b_i(v)}{\partial y} + \frac{\partial v}{\partial y} \frac{d \ell^{(i)}(v)}{dx} - \frac{dv}{dx} \frac{\partial \ell^{(i)}(v)}{\partial y} \right] \exp \left(\int_\xi^x \frac{\partial \lambda_i(v)}{\partial y} dx \right) dx \end{aligned} \quad (4.17)$$

for $i = 1, 2, \dots, n$. From (4.11) we obtain

$$\ell^{(i)}(v) \frac{\partial u}{\partial \xi} = b_i(v) + (v - u) \left[\frac{\partial \ell^{(i)}(v)}{\partial \xi} + \lambda_i(v) \frac{\partial \ell^{(i)}(v)}{\partial \eta} \right] - \lambda_i(v) \ell^{(i)}(v) \frac{\partial u}{\partial \eta} \quad (4.18)$$

for $i = 1, 2, \dots, n$. Expressions (4.13), (4.17), (4.18) are the representations that we want to establish. We point out that notations like $\frac{\partial \ell^{(i)}(v)}{\partial y}$ represent the partial derivatives with respect to y of the composition $\ell^{(i)}(x, y, v(x, y))$.

4.2.2 Primary estimates

Let $B(t_1, t_2, \dots, t_m)$ be a function (scalar, vector, or matrix) in a domain of \mathbb{R}^m . Denote the 0th-order and 1st-order norms respectively by

$$\|B\|_0 = \sup |B|, \quad \|B\|_1 = \|B\|_0 + \|B_{t_1}\|_0 + \dots + \|B_{t_m}\|_0.$$

The norm of a constant matrix (a_{ij}) is the maximum of the row sums $\sum_j |a_{ij}|$.

Let $\Sigma_R(\delta)$ be

$$\Sigma_R(\delta) := \{v \in C^1(R(\delta)) \mid v(0, y) = \phi(y), \|v - u^{(0)}\|_0 \leq M, \|v - u^{(0)}\|_1 \leq R\}.$$

Lemma 4.1. Under the conditions of Theorem 4.1, there exist constants $\delta, R, \beta \in (0, 1)$ such that the map T maps $\Sigma_R(\delta)$ into $\Sigma_R(\delta)$ and

$$\|Tv - Tv'\|_0 \leq \beta \|v - v'\|_0 \quad \text{for any } v, v' \in \Sigma_R(\delta).$$

The constants δ, R, β depend only on the 1st-order norms of $\lambda_i, \ell^{(i)}, c$ on the domain $R_M(\delta_0)$, ϕ on the interval $[y_1, y_2]$, and the (hyperbolicity) number α from

$$|\det(\ell^{(i)})| \geq \alpha > 0. \quad (4.19)$$

Proof. In order to reduce the number of constants, we will use k to denote various constants that depend only on the 1st-order norms of $\lambda_i, \ell^{(i)}, c$ on the domain $R_M(\delta_0)$, ϕ on the interval $[y_1, y_2]$, and the number α .

From (4.11) we find

$$\begin{aligned} \ell^{(i)}(v)[u(\xi, \eta) - \phi(y_i(0, \xi, \eta))] = \\ [\ell^{(i)}(\phi(y_i(0; \xi, \eta))) - \ell^{(i)}(v(\xi, \eta))] \phi(y_i(0; \xi, \eta)) + \int_0^\xi [b_i(v) + v \frac{d}{dx} \ell^{(i)}(v)] dx \end{aligned} \quad (4.20)$$

for $i = 1, 2, \dots, n$. It is easy to see

$$|\ell^{(i)}(\phi(y_i(0, \xi, \eta))) - \ell^{(i)}(v(\xi, \eta))| \leq kR\delta \quad (i = 1, 2, \dots, n).$$

So we obtain from (4.20) that

$$|\ell^{(i)}(v)[u(\xi, \eta) - \phi(y_i(0, \xi, \eta))]| \leq kR\delta + k(1 + R)\delta.$$

From (4.19) we further obtain

$$|u(\xi, \eta) - \phi(y_i(0, \xi, \eta))| \leq k(1 + R)\delta. \quad (4.21)$$

From (4.9) we obtain

$$\|u - u^{(0)}\|_0 \leq M/2 + k(1 + R)\delta. \quad (4.22)$$

For the derivatives we note

$$|\phi(y_i(0, \xi, \eta)) - v(\xi, \eta)| \leq kR\delta. \quad (4.23)$$

From (4.21) we obtain further

$$\|u - v\|_0 \leq k(1 + R)\delta. \quad (4.24)$$

From here and (4.17) we obtain

$$\begin{aligned} |\ell^{(i)}(v) \frac{\partial u}{\partial \eta}| &\leq k e^{k(1+R)\delta} + kR(1+R)\delta + kR(1+R)e^{k(1+R)\delta}\delta \\ &= k e^{k(1+R)\delta} + kR(1+R)(1 + e^{k(1+R)\delta})\delta, \end{aligned}$$

thus

$$\left\| \frac{\partial u}{\partial \eta} \right\|_0 \leq k e^{k(1+R)\delta} + kR(1+R)(1+e^{k(1+R)\delta})\delta. \quad (4.25)$$

From this, (4.18) and (4.24), we obtain

$$\left\| \frac{\partial u}{\partial \xi} \right\|_0 \leq k + k e^{k(1+R)\delta} + kR(1+R)(1+e^{k(1+R)\delta})\delta. \quad (4.26)$$

Combining (4.22), (4.25) and (4.26) we obtain

$$\|u - u^{(0)}\|_1 \leq M/2 + k + k(1+R)\delta + k e^{k(1+R)\delta} + kR(1+R)(1+e^{k(1+R)\delta})\delta. \quad (4.27)$$

Let R be large, e.g.

$$R > M/2 + 2k,$$

then select small δ , so as to make the right hand side of (4.22) be less or equal to M , and the right-hand side of (4.27) be less or equal to R . These choices of R, δ depend only on k , thus they depend on the 1st-order norms of $\lambda_i, \ell^{(i)}, c, \phi$ and α .

For the contraction property, we let $u = Tv, u' = Tv'$. Then

$$\frac{d}{di} [\ell^{(i)}(v)u] = b_i(v) + v \frac{d}{di} \ell^{(i)}(v) \quad (i = 1, 2, \dots, n),$$

$$\frac{d}{di'} [\ell^{(i)}(v')u'] = b_i(v') + v' \frac{d}{di'} \ell^{(i)}(v') \quad (i = 1, 2, \dots, n),$$

where $\frac{d}{di} = \partial_x + \lambda_i(v)\partial_y, \frac{d}{di'} = \partial_x + \lambda_i(v')\partial_y$. Thus we obtain

$$\begin{aligned} \frac{d}{di} [\ell^{(i)}(v)(u - u')] &= b_i(v) + v \frac{d}{di} \ell^{(i)}(v) - \frac{d}{di} \ell^{(i)}(v) \cdot u' - \ell^{(i)}(v) \frac{du'}{di} \\ &= b_i(v) - b_i(v') + (v - v') \frac{d}{di} \ell^{(i)}(v) - \frac{d}{di} [\ell^{(i)}(v) - \ell^{(i)}(v')] \cdot u' \\ &\quad + b_i(v') + v' \frac{d}{di} \ell^{(i)}(v) - \frac{d}{di} \ell^{(i)}(v') \cdot u' - \ell^{(i)}(v) \frac{du'}{di} \\ &= [b_i(v) - b_i(v')] + (v - v') \frac{d}{di} \ell^{(i)}(v) - \frac{d}{di} [\ell^{(i)}(v) - \ell^{(i)}(v')] \cdot (u' - v') \\ &\quad + b_i(v') + v' \frac{d}{di} \ell^{(i)}(v) - \frac{d}{di} \ell^{(i)}(v') \cdot u' - \ell^{(i)}(v) \frac{du'}{di} - v' \frac{d}{di} [\ell^{(i)}(v) - \ell^{(i)}(v')]. \end{aligned}$$

We then use the equation for (u', v') to obtain

$$\begin{aligned} b_i(v') &= \frac{d}{di'} [\ell^{(i)}(v')u'] - v' \frac{d}{di'} \ell^{(i)}(v') \\ &= \frac{d}{di} [\ell^{(i)}(v')u'] + (\lambda(v') - \lambda(v)) \frac{\partial}{\partial y} [\ell^{(i)}(v')u'] - v' \frac{d}{di'} [\ell^{(i)}(v) - \ell^{(i)}(v')]. \end{aligned}$$

Combining them we obtain

$$\begin{aligned} \frac{d}{di} [\ell^{(i)}(v)(u - u')] &= [b_i(v) - b_i(v')] + (v - v') \frac{d}{di} \ell^{(i)}(v) \\ &\quad - \frac{d}{di} [\ell^{(i)}(v) - \ell^{(i)}(v')] \cdot (u' - v') + (\ell^{(i)}(v') - \ell^{(i)}(v)) \frac{du'}{di} \\ &\quad + v'(\lambda(v) - \lambda(v')) \frac{\partial}{\partial y} \ell^{(i)}(v') + (\lambda(v') - \lambda(v)) \frac{\partial}{\partial y} [\ell^{(i)}(v')u'] \end{aligned}$$

($i = 1, 2, \dots, n$). Integrating it along $C_i(v)$, noting $u - u' = 0$ at $x = 0$, using integration by parts in the first term, we obtain

$$\ell^{(i)}(v)(u - u') = (v' - u')[\ell^{(i)}(v) - \ell^{(i)}(v')] + \int_0^{\xi} \Phi^{(i)} \cdot (v - v') dx \quad (4.28)$$

for $i = 1, 2, \dots, n$, where $\Phi^{(i)}$ is a vector function, which is bounded by a quadratic form of the 1st-order norms of $\lambda_i, \ell^{(i)}, c, u', v, v'$. Applying (4.24) to

$$\|v' - u'\|_0 \leq k(1 + R)\delta$$

we obtain from (4.28) that

$$|\ell^{(i)}(v)(u - u')| \leq k(1 + R)\delta\|v - v'\|_0 \quad i = 1, 2, \dots, n.$$

Hence we have

$$\|u - u'\|_0 \leq \beta\|v - v'\|_0,$$

where $\beta = k(1 + R)\delta < 1$ when δ is sufficient small. The proof of Lemma 4.1 is complete. \square

4.2.3 Estimates on modulus of continuity

We introduce the concept of modulus of continuity, see Hartman and Wintner [15]. For any function $B(t_1, t_2, \dots, t_m)$ on a domain $S \subset \mathbb{R}^m$, let

$$\omega(h; B) := \sup |B(t_1, t_2, \dots, t_m) - B(t'_1, t'_2, \dots, t'_m)|$$

$$(t_1, t_2, \dots, t_m), (t'_1, t'_2, \dots, t'_m) \in S, |t_k - t'_k| \leq h,$$

which is referred to as the *modulus of continuity* of B on S .

It is apparent that B is uniformly continuous if and only if $\omega(h; B) \rightarrow 0$ ($h \rightarrow 0$); a family of $\{B\}$ is equicontinuous if and only if there is an $\omega(h)$ with the property $\omega(h) \rightarrow 0$ ($h \rightarrow 0$) such that $\omega(h; B) \leq \omega(h)$ for every B in the family. Other properties are listed below.

1. If f is Lip-continuous, then $\omega(h; f) \leq Nh$, N is Lip constant; $\omega(h; f) \leq \|f\|_1 h$.
2. $\omega(h; f \pm g) \leq \omega(h; f) + \omega(h; g)$.
3. $\omega(h; fg) \leq \|f\|_0 \omega(h; g) + \|g\|_0 \omega(h; f)$.
4. $\omega(h; f/g) \leq \frac{\|f\|_0}{a^2} \omega(h; g) + \frac{1}{a} \omega(h; f)$, for a scalar function g with $|g| \geq a > 0$.
5. $\omega(h; f(g)) \leq \omega(\omega(h; g); f) \leq \|f\|_1 \omega(h; g)$.
6. $\omega(h; F) \leq \delta \omega(h; f) + \|f\|_0 [\omega(h; \varphi) + \omega(h; \psi)]$ for

$$F(x, y) = \int_{\psi(x, y)}^{\varphi(x, y)} f(t, x, y) dt \quad (0 \leq \varphi, \psi \leq \delta).$$

7. $\omega(ch; f) \leq ([c] + 1)\omega(h; f)$, $c > 0$ is a constant, $[c]$ represents the integer part of c .

Lemma 4.2. Under the hypothesis of Theorem 4.1, the map T has the property: For any $u^{(1)} \in \Sigma_R(\delta)$, let $u^{(j)} = Tu^{(j-1)}$, then the sequences $\{u_x^{(j)}\}, \{u_y^{(j)}\}$ are equicontinuous on $R(\delta)$. Here δ, R are as in Lemma 4.1, δ may be smaller, but processes the same properties of Lemma 4.1.

Proof. From (4.17) we obtain

$$\ell^{(i)}(v) \frac{\partial u}{\partial \eta} = \sum_{j=1}^5 \varphi_j(\xi, \eta) \quad (i = 1, 2, \dots, n), \quad (4.29)$$

where

$$\begin{aligned} \varphi_1 &= \phi'(y_i(0; \xi, \eta)) \ell^{(i)}(\phi(y_i(0; \xi, \eta))) \exp \left(\int_\xi^0 \frac{\partial \lambda_i(v)}{\partial y} dx \right), \\ \varphi_2 &= (v - u) \frac{\partial \ell^{(i)}(v)}{\partial \eta}, \\ \varphi_3 &= \int_0^\xi \frac{\partial b_i(v)}{\partial y} \exp \left(\int_\xi^x \frac{\partial \lambda_i(v)}{\partial y} dx \right) dx, \\ \varphi_4 &= \int_0^\xi \frac{\partial v}{\partial y} \frac{d\ell^{(i)}(v)}{dx} \exp \left(\int_\xi^x \frac{\partial \lambda_i(v)}{\partial y} dx \right) dx, \\ \varphi_5 &= - \int_0^\xi \frac{dv}{dx} \frac{\partial \ell^{(i)}(v)}{\partial y} \exp \left(\int_\xi^x \frac{\partial \lambda_i(v)}{\partial y} dx \right) dx. \end{aligned}$$

Using the properties 1–7 on the modulus of continuity, we can obtain

$$\begin{aligned} \omega(h; \varphi_1) &\leq k[\omega(h) + \delta \omega(h; v_y)], \\ \omega(h; \varphi_2) &\leq k[\omega(h) + \delta \omega(h; v_y)], \\ \omega(h; \varphi_3) &\leq k[\omega(h) + \delta \omega(h; v_y)], \\ \omega(h; \varphi_4) &\leq k[\omega(h) + \delta(\omega(h; v_x) + \omega(h; v_y))], \\ \omega(h; \varphi_5) &\leq k[\omega(h) + \delta(\omega(h; v_x) + \omega(h; v_y))], \end{aligned} \quad (4.30)$$

where k , as before, depends only on the 1st-order norms of $\lambda_i, \ell^{(i)}, c, \phi$, and R, δ , and α . The term $\omega(h)$, vanishing as $h \rightarrow 0$, is independent of $v \in \Sigma_R(\delta)$.

We point out that there is a factor δ for every term in (4.30) that involves either $\omega(h; v_x)$ or $\omega(h; v_y)$. This is because, for the term $\omega(h; \varphi_2)$, there holds

$$\|v - u\|_0 \leq k\delta$$

(see (4.24)). For the other four terms, it is because the integrals involving v_x, v_y have intervals of integration less than δ . Property 6 of the modulus of continuity comes in handy.

From (4.29), (4.30) we obtain

$$\omega(h; \ell^{(i)} u_y) \leq k\omega(h) + k\delta[\omega(h; v_x) + \omega(h; v_y)],$$

hence

$$\omega(h; u_y) \leq k\omega(h) + k\delta[\omega(h; v_x) + \omega(h; v_y)]. \quad (4.31)$$

From (4.18), (4.31) we obtain

$$\omega(h; u_x) \leq k\omega(h) + k\delta[\omega(h; v_x) + \omega(h; v_y)]. \quad (4.32)$$

Adding the two above, we obtain

$$\omega(h; u_x) + \omega(h; u_y) \leq k\omega(h) + q[\omega(h; v_x) + \omega(h; v_y)], \quad (4.33)$$

where $q = k\delta$. We choose a small δ to make $q < 1$.

Applying (4.33) on $u = u^{(j)}, v = u^{(j-1)}$, we obtain an iteration formula relating $\omega(h; u_x^{(j)}) + \omega(h; u_y^{(j)})$ and $\omega(h; u_x^{(j-1)}) + \omega(h; u_y^{(j-1)})$. Using the iteration formula repeatedly, noting $q < 1$, we obtain

$$\omega(h; u_x^{(j)}) + \omega(h; u_y^{(j)}) \leq \frac{k}{1-q}\omega(h) + \omega(h; u_x^{(1)}) + \omega(h; u_y^{(1)}).$$

The proof of Lemma 4.2 is complete. \square

Proof of Theorem 4.1. We notice that T is not established as a compact mapping. But we can manage as follows. The strict contraction property yields that the full sequence $\{u^{(j)}\}$ is convergent. Thus we focus on the sequences $\{u_x^{(j)}\}, \{u_y^{(j)}\}$. There is a subsequence of them that converges. Use this subsequence in (4.13) on the right-hand side and pass to the limit. The left-hand side involves no derivatives, thus converges, too. So we have a solution. This completes the proof of Theorem 4.1. \square

We remark that we can make T a continuous mapping from a compact convex set of a Banach space to itself and therefore Schauder's fixed point theorem can apply.

4.2.4 Lipschitz data

Theorem 4.2. Let system (4.1) be a Lipschitz continuous hyperbolic system on $R_M(\delta_0)$. Let $\phi(y)$ be Lipschitz continuous on $[y_1, y_2]$. Let $R(\delta_0)$ be a strong determinate domain for (4.1). Then there exists a $\delta > 0$, such that problem (4.1) and (4.8) have a Lipschitz continuous solution on $R(\delta)$, satisfying the equations almost everywhere.

For a proof, see [40].

4.3 Goursat problem

We assume that system (4.1) is continuously differentiable and strictly hyperbolic on the domain

$$R_M(\delta_0) := \{(x, y, u) \mid (x, y) \in R(\delta_0), |u| \leq M\};$$

and $R(\delta)$ is

$$R(\delta) := \{(x, y) \mid x \in [0, \delta], y \in [y_1(x), y_2(x)]\},$$

where $\Gamma_1 : y = y_1(x)$, $\Gamma_2 : y = y_2(x)$ ($x \in [0, \delta_0]$) are two $C^1[0, \delta_0]$ curves passing through the origin $(0, 0)$ and satisfying $y_1(x) < y_2(x)$ for $x \in (0, \delta_0]$.

Let $\phi^{(1)}(x), \phi^{(2)}(x)$ be from $C^1[0, \delta_0]$ satisfying $\phi^{(1)}(0) = \phi^{(2)}(0) = 0$ and

$$\|\phi^{(1)}\|_0 \leq M/2, \quad \|\phi^{(2)}\|_0 \leq M/2. \quad (4.34)$$

Assume that $\phi^{(1)}(x), \phi^{(2)}(x)$ are such that Γ_1, Γ_2 are the first (smallest) and last (largest) characteristics:

$$\frac{dy_1}{dx} = \lambda_1(x, y_1, \phi^{(1)}(x)), \quad \frac{dy_2}{dx} = \lambda_n(x, y_2, \phi^{(2)}(x)) \quad (4.35)$$

and the characteristic relations

$$\begin{cases} \ell^{(1)}(x, y_1, \phi^{(1)}(x)) \frac{d\phi^{(1)}}{dx} = b_1(x, y_1, \phi^{(1)}(x)) \\ \ell^{(n)}(x, y_2, \phi^{(2)}(x)) \frac{d\phi^{(2)}}{dx} = b_n(x, y_2, \phi^{(2)}(x)) \end{cases} \quad (4.36)$$

hold. The *Goursat problem* is to find a solution to system (4.1) in the domain $R(\delta)$ ($\delta \leq \delta_0$) satisfying

$$u(x, y_1(x)) = \phi^{(1)}(x), \quad u(x, y_2(x)) = \phi^{(2)}(x). \quad (4.37)$$

Compatibility conditions. For smooth solutions to exist for systems with more than two characteristic directions, we need to impose more conditions. Two directional derivatives along the directions λ_1, λ_n at the origin determine all other directional derivatives along λ_k ($k = 2, 3, \dots, n - 1$). Thus we can solve $\nabla u(0, 0)$ from

$$\begin{aligned} \frac{d\phi^{(1)}}{dx}(0) &= \frac{\partial u(0,0)}{\partial x} + \lambda_1(0) \frac{\partial u(0,0)}{\partial y}, \\ \frac{d\phi^{(2)}}{dx}(0) &= \frac{\partial u(0,0)}{\partial x} + \lambda_n(0) \frac{\partial u(0,0)}{\partial y} \end{aligned}$$

to find

$$\frac{\partial u(0,0)}{\partial x} = \frac{\lambda_n^0 \phi_x^{(1)} - \lambda_1^0 \phi_x^{(2)}}{\lambda_n^0 - \lambda_1^0}, \quad \frac{\partial u(0,0)}{\partial y} = \frac{\phi_x^{(2)} - \phi_x^{(1)}}{\lambda_n^0 - \lambda_1^0},$$

and use the characteristic relation

$$\ell^{(k)}(0) \left[\frac{\partial u(0,0)}{\partial x} + \lambda_k(0) \frac{\partial u(0,0)}{\partial y} \right] = b_k(0) \quad (k = 2, 3, \dots, n - 1)$$

to obtain

$$\ell^{(k)}(0) \left[\frac{\lambda_n^0 \phi_x^{(1)} - \lambda_1^0 \phi_x^{(2)}}{\lambda_n^0 - \lambda_1^0} + \lambda_k^0 \frac{\phi_x^{(2)} - \phi_x^{(1)}}{\lambda_n^0 - \lambda_1^0} \right] = b_k(0) \quad (4.38)$$

($k = 2, 3, \dots, n-1$) or its alternative forms

$$\begin{aligned} \ell^{(k)}(0) \left[\frac{\lambda_n^0 - \lambda_k^0}{\lambda_n^0 - \lambda_1^0} \phi_x^{(1)} + \frac{\lambda_k^0 - \lambda_1^0}{\lambda_n^0 - \lambda_1^0} \phi_x^{(2)} \right] &= b_k(0), \\ \frac{1}{\lambda_n^0 - \lambda_k^0} \left[\ell^{(k)}(0) \frac{d\phi^{(2)}(0)}{dx} - b_k(0) \right] &= \frac{1}{\lambda_1^0 - \lambda_k^0} \left[\ell^{(k)}(0) \frac{d\phi^{(1)}(0)}{dx} - b_k(0) \right] \end{aligned} \quad (4.39)$$

($k = 2, 3, \dots, n-1$), which are called compatibility conditions, where $\lambda_1(0) = \lambda_1^0 = \lambda_1(0, 0, 0)$, etc., and $u_x(0, 0)$, $u_y(0, 0)$ stand for the limiting values at $(0, 0)$.

Sometimes, the boundary Γ_2 may not be the last characteristic, in which case more compatibility conditions are needed to insure the existence of a smooth solution. Consider the system

$$\frac{\partial u_1}{\partial x} + \frac{\partial u_1}{\partial y} = 0, \frac{\partial u_2}{\partial x} + 2 \frac{\partial u_2}{\partial y} = 0, \frac{\partial u_3}{\partial x} + 3 \frac{\partial u_3}{\partial y} = x. \quad (4.40)$$

Let the value $(0, 0, 0)$ be given on $\Gamma_1 : y = x$ (the first characteristic) and $\Gamma_2 : y = 2x$ (the second characteristic). Obviously, the characteristic relations hold on Γ_1 and Γ_2 (To satisfy the characteristic relations, it is sufficient if u_1 is constant on Γ_1 and u_2 is constant on Γ_2 , the others are arbitrary). The compatibility condition (4.39) are satisfied for this set-up. But, if it were to have a smooth solution, then we integrate the third equation along the third characteristic from a point (x_1, y_1) on Γ_1 to a point (x_2, y_2) on Γ_2 , $x_2 > x_1 > 0$, to result in a contradiction

$$0 = u_3(x_2, y_2) - u_3(x_1, y_1) = \int_{x_1}^{x_2} x \, dx = \frac{1}{2}(x_2^2 - x_1^2).$$

Therefore, we shall consider only the case that Γ_2 is the last characteristic.

Theorem 4.3. Goursat problem has a solution in $C^1(R(\delta))$ for some $\delta > 0$ provided that the compatibility condition (4.38) holds.

For a pair of equations ($n = 2$), there is no need for any compatibility conditions, and the Goursat problem always has a smooth solution.

For a proof of the theorem, we start with the set of admissible functions. A function $v \in C^1(R(\delta))$ is *admissible* if it satisfies the boundary condition (4.37) and

$$\|v\|_0 \leq M. \quad (4.41)$$

Lemma 4.3. The set of admissible functions is non-empty, provided that $\delta > 0$ is sufficiently small.

Proof. We introduce the coordinate transform

$$\alpha = y - y_1(x), \quad \beta = y - y_2(x),$$

where we assume that $y_1(x), y_2(x)$ have been smoothly extended to all $|x| \leq \delta_0$. According to the strict hyperbolicity and relations (4.35), we deduce that the Jacobian of this transform is not zero at the origin. Therefore, for small $\delta > 0$, the transform maps $R(\delta)$ one-to-one into a right triangle $\alpha = 0, \beta = 0, \beta - \alpha = y_1(\delta) - y_2(\delta)$ in the α, β plane. After the transform, the existence problem has become: Given C^1 functions $\phi^{(2)}(\alpha), \phi^{(1)}(\beta)$, with $\phi^{(1)}(0) = \phi^{(2)}(0)$, on the intervals $0 \leq \alpha \leq y_2(\delta) - y_1(\delta)$ and $y_1(\delta) - y_2(\delta) \leq \beta \leq 0$ respectively, satisfying (4.34), prove that there is a C^1 function $v(\alpha, \beta)$ on the triangle satisfying $v(\alpha, 0) = \phi^{(2)}(\alpha), v(0, \beta) = \phi^{(1)}(\beta)$ and (4.41). But obviously $v(\alpha, \beta) = \phi^{(1)}(\beta) + \phi^{(2)}(\alpha)$ is a choice. The proof is complete. \square

Now let v be an admissible function on $R(\delta)$. Consider the transform $u = Tv$ by

$$\frac{d}{di} [\ell^{(i)}(v)u] = b_i(v) + v \frac{d}{di} \ell^{(i)}(v) \quad (i = 1, 2, \dots, n) \quad (4.42)$$

with the boundary condition (4.37).

Let $y = y_i(x; \xi, \eta)$ denote the i -th characteristic curve $C_i(v)$ passing through $(\xi, \eta) \in R(\delta)$. $R(\delta)$ is divided into $(n - 1)$ parts by the $(n - 2)$ interior characteristics passing through the origin. Let $R^{(k)}(\delta)$ denote the closed part between $C_k(v)$ and $C_{k+1}(v)$ ($1 \leq k \leq n - 1$).

From $(\xi, \eta) \in R^{(k)}(\delta)$ we draw downward characteristics $C_i(v)$ ($1 \leq i \leq n$). When $i \leq k$, $C_i(v)$ intersects Γ_2 at some point $(x_{i,o}, y_2(x_{i,o}))$:

$$y_2(x_{i,o}) = y_i(x_{i,o}; \xi, \eta) \quad (i \leq k). \quad (4.43)$$

When $i \geq k + 1$, $C_i(v)$ intersects Γ_1 at some point $(x_{i,o}, y_1(x_{i,o}))$:

$$y_1(x_{i,o}) = y_i(x_{i,o}; \xi, \eta) \quad (i \geq k + 1). \quad (4.44)$$

Because

$$\frac{dy_2(x_{i,o})}{dx} - \frac{\partial}{\partial x} y_i(x_{i,o}; \xi, \eta) \neq 0 \quad (i \leq k),$$

$$\frac{dy_1(x_{i,o})}{dx} - \frac{\partial}{\partial x} y_i(x_{i,o}; \xi, \eta) \neq 0 \quad (i \geq k + 1),$$

we can solve $x_{i,o}$ from (4.43), (4.44) to obtain

$$x_{i,o} = f_i(\xi, \eta) \quad (i \leq k), \quad (4.45)$$

$$x_{i,o} = g_i(\xi, \eta) \quad (i \geq k+1) \quad (4.46)$$

and f_i, g_i are C^1 on $R^{(k)}(\delta)$:

$$\frac{\partial f_i}{\partial \eta} = \frac{1}{\lambda_n(v) - \lambda_i(v)} \frac{\partial y_i}{\partial \eta} \quad (i \leq k), \quad (4.47)$$

$$\frac{\partial g_i}{\partial \eta} = \frac{1}{\lambda_1(v) - \lambda_i(v)} \frac{\partial y_i}{\partial \eta} \quad (i \geq k+1). \quad (4.48)$$

Integrating (4.42) along $C_i(v)$, we obtain

$$\ell^{(i)}(v)u = \begin{cases} \ell^{(i)}(f_i, y_2(f_i), \phi^{(2)}(f_i))\phi^{(2)}(f_i) + \int_{f_i}^{\xi} [b_i(v) + v \frac{d}{dx} \ell^{(i)}(v)] dx \\ \quad (i \leq k), \\ \ell^{(i)}(g_i, y_1(g_i), \phi^{(1)}(g_i))\phi^{(1)}(g_i) + \int_{g_i}^{\xi} [b_i(v) + v \frac{d}{dx} \ell^{(i)}(v)] dx \\ \quad (i \geq k+1). \end{cases} \quad (4.49)$$

It is easy to find for $i \leq k$ that

$$\begin{aligned} & \frac{\partial}{\partial \eta} [\ell^{(i)}(f_i, y_2(f_i), \phi^{(2)}(f_i))\phi^{(2)}(f_i)] \\ &= \ell^{(i)} \frac{d\phi^{(2)}}{dx} \frac{\partial f_i}{\partial \eta} + \phi^{(2)} [\ell_x^{(i)} + \ell_y^{(i)} y'_2 + \ell_u^{(i)} \frac{d\phi^{(2)}}{dx}] \frac{\partial f_i}{\partial \eta}, \end{aligned} \quad (4.50)$$

where $\ell_x^{(i)}, \ell_y^{(i)}, \ell_u^{(i)}$ represent the partial derivatives when x, y, u are independent.

Using the technique from Section 4.2, we find that the partial derivative of

$$F_i(\xi, \eta) = \int_{f_i}^{\xi} \left[b_i(v) + v \frac{d}{dx} \ell^{(i)}(v) \right] dx \quad (i \leq k) \quad (4.51)$$

has formula (for $i \leq k$)

$$\begin{aligned} \frac{\partial F_i}{\partial \eta} &= - [b_i(v) + v \frac{d}{dx} \ell^{(i)}(v)]|_{x=f_i} \frac{\partial f_i}{\partial \eta} + v \frac{\partial}{\partial \eta} \ell^{(i)}(v)|_{x=f_i}^{\xi} \\ &+ \int_{f_i}^{\xi} \left[\frac{\partial b_i(v)}{\partial y} + \frac{\partial v}{\partial y} \frac{d\ell^{(i)}(v)}{dx} - \frac{dv}{dx} \frac{\partial \ell^{(i)}(v)}{\partial y} \right] \frac{\partial y_i}{\partial \eta} dx \\ &= v \frac{\partial \ell^{(i)}(v)}{\partial \eta} - b_i(f_i, y_2(f_i), \phi^{(2)}(f_i)) \frac{\partial f_i}{\partial \eta} - \phi^{(2)}(f_i) \{ \ell_y^{(i)} + \ell_u^{(i)} v_y \} \frac{\partial y_i}{\partial \eta} \\ &- \phi^{(2)}(f_i) \{ \ell_x^{(i)}(f_i, y_2(f_i), \phi^{(2)}(f_i)) + \ell_u^{(i)} v_x + \lambda_i(v) [\ell_y^{(i)} + \ell_u^{(i)} v_y] \} \frac{\partial f_i}{\partial \eta} \\ &+ \int_{f_i}^{\xi} \left[\frac{\partial b_i(v)}{\partial y} + \frac{\partial v}{\partial y} \frac{d\ell^{(i)}(v)}{dx} - \frac{dv}{dx} \frac{\partial \ell^{(i)}(v)}{\partial y} \right] \frac{\partial y_i}{\partial \eta} dx. \end{aligned} \quad (4.52)$$

Since $v(x, y_2(x)) = \phi^{(2)}(x)$, thus we have

$$v_x + v_y y'_2 = \frac{\partial \phi^{(2)}}{\partial x}. \quad (4.53)$$

Combining (4.49), (4.50), (4.52), (4.53) with (4.47), (4.48) we obtain for $(\xi, \eta) \in R^{(k)}(\delta)$ that

$$\begin{aligned} \ell^{(i)}(v) \frac{\partial u}{\partial \eta} &= \left[\ell^{(i)}(f_i, y_2(f_i), \phi^{(2)}(f_i)) \frac{d\phi^{(2)}(f_i)}{dx} - b_i \right] \frac{\partial f_i}{\partial \eta} + (v - u) \cdot \\ &\quad \frac{\partial \ell^{(i)}(v(\xi, \eta))}{\partial \eta} + \int_{f_i}^{\xi} \left[\frac{\partial b_i(v)}{\partial y} + \frac{\partial v}{\partial y} \frac{d\ell^{(i)}(v)}{dx} - \frac{dv}{dx} \frac{\partial \ell^{(i)}(v)}{\partial y} \right] \frac{\partial y_i}{\partial \eta} dx \quad (i \leq k). \end{aligned} \quad (4.54)$$

By the same token, we have

$$\begin{aligned} \ell^{(i)}(v) \frac{\partial u}{\partial \eta} &= \left[\ell^{(i)}(g_i, y_1(g_i), \phi^{(1)}(g_i)) \frac{d\phi^{(1)}(g_i)}{dx} - b_i \right] \frac{\partial g_i}{\partial \eta} + (v - u) \cdot \\ &\quad \frac{\partial \ell^{(i)}(v(\xi, \eta))}{\partial \eta} + \int_{g_i}^{\xi} \left[\frac{\partial b_i(v)}{\partial y} + \frac{\partial v}{\partial y} \frac{d\ell^{(i)}(v)}{dx} - \frac{dv}{dx} \frac{\partial \ell^{(i)}(v)}{\partial y} \right] \frac{\partial y_i}{\partial \eta} dx \quad (i \geq k+1). \end{aligned} \quad (4.55)$$

Thus the function in (4.49) has continuous partial derivative u_η on each $R^{(k)}(\delta)$. On the boundaries of $R^{(k)}(\delta)$ in the interior of $R(\delta)$, the continuity of u_η follows from the compatibility condition (4.39) and formulas (4.47), (4.48). By the same token we obtain a similar formula for $\ell^{(i)}(v)u_\xi$ and we deduce that u_ξ is continuous on $R^{(k)}(\delta)$. Thus u is the unique $C^1(R(\delta))$ solution of (4.42), (4.37).

From (4.42) we obtain for $1 \leq i \leq n$

$$\ell^{(i)}(v) \frac{\partial u}{\partial \xi} = b_i(v) + (v - u) \left[\frac{\partial \ell^{(i)}(v)}{\partial x} + \lambda_i(v) \frac{\partial \ell^{(i)}(v)}{\partial y} \right] - \lambda_i(v) \ell^{(i)}(v) \frac{\partial u}{\partial \eta}. \quad (4.56)$$

Let

$$\Sigma_R(\delta) := \{v \in C^1(R(\delta)) \mid \|v\|_0 \leq M, \|v\|_1 \leq R, v = \phi^{(j)} \text{ on } \Gamma_j, j = 1, 2\}$$

and define

$$u^{(k)} = Tu^{(k-1)}, \quad k = 2, 3, \dots, \quad u^{(1)} \in \Sigma_R(\delta).$$

Lemma 4.4. There exists positive constants $\beta \in (0, 1), \delta, R$ such that the map T maps $\Sigma_R(\delta)$ to itself,

$$\|Tv - Tv'\|_0 \leq \beta \|v - v'\|_0, \quad \text{for any } v, v' \in \Sigma_R(\delta),$$

and the sequences $\{u_x^{(k)}, u_y^{(k)}\}_{k=1}^\infty$ are equicontinuous on $R(\delta)$ for any fixed $u^{(1)} \in \Sigma_R(\delta)$. The constants δ, R, β depend only on the $C^1(R_M(\delta_0))$ norms of $\lambda_i, \ell^{(i)}, c$, the $C^1[0, \delta_0]$ norms of $\phi^{(1)}, \phi^{(2)}$, and the hyperbolicity $\alpha = \min |\det(\ell_j^{(i)})|$ on $R_M(\delta_0)$.

Proof. It is similar to the proof of Lemmas 4.1 and 4.2. We use (4.49), (4.54), (4.55) on $R^{(k)}(\delta)$ for the estimates. For the estimate of

modulus of continuity of u_η , we notice especially that there are formulas

$$\frac{\partial f_i}{\partial \eta} = \frac{1}{\lambda_n(v) - \lambda_i(v)} \exp \left(\int_{\xi}^{f_i} \frac{\partial \lambda_i(v)}{\partial y} dt \right) (x = f_i) \quad (i \leq k),$$

$$\frac{\partial g_i}{\partial \eta} = \frac{1}{\lambda_1(v) - \lambda_i(v)} \exp \left(\int_{\xi}^{g_i} \frac{\partial \lambda_i(v)}{\partial y} dt \right) (x = g_i) \quad (i \geq k+1).$$

Therefore derivatives of v appear only in the integrands whose length of integration is less than δ (see the point of emphasis in the proof of Lemma 4.2). The rest of the proof is similar to that of Lemma 4.2. The proof of the lemma is complete. \square

The proof of the theorem is straightforward based on Lemmas 4.3 and 4.4.

4.4 Mixed initial-boundary value problem

We seek a solution to the problem: Given two curves $\Gamma_j : y = y_j(x) \in C^1([0, \delta_0])$ and two vector functions $\phi^{(j)}(x) \in C^1([0, \delta_0])(j = 1, 2)$ satisfying

$$\frac{dy_1}{dx} = \lambda_1(x, y_1, \phi^{(1)}), \quad (4.57)$$

$$\ell^{(1)}(x, y_1(x), \phi^{(1)}) \frac{d\phi^{(1)}}{dx} = b_1(x, y_1, \phi^{(1)}), \quad (4.58)$$

and all characteristic directions on Γ_2 corresponding to $\phi^{(2)}$ of system (4.1) point into the sectorial domain $R(\delta_0)$. Find a solution in $R(\delta)(\delta \leq \delta_0)$ to system (4.1) with data (4.37).

Compatibility condition. We need the condition

$$\frac{1}{y'_2(0) - \lambda_k^0} \left[\ell_0^{(k)} \frac{d\phi^{(2)}(0)}{dx} - b_k^0 \right] = \frac{1}{\lambda_1^0 - \lambda_k^0} \left[\ell_0^{(k)} \frac{d\phi^{(1)}(0)}{dx} - b_k^0 \right] \quad (4.59)$$

($2 \leq k \leq n$) for the existence of a smooth solution.

Theorem 4.4. The mixed initial-boundary value problem has a solution in $C^1(R(\delta))$ provided that compatibility condition (4.59) hold.

Note that there are $n - 1$ compatibility conditions, which is one more than for the Goursat problem. Thus compatibility condition is needed even for a system of two equations for the mixed initial-boundary value problem.

In addition, we can let $\Gamma_2 : y = y_2(x)$ be a segment $0 \leq y \leq y_0$ on the y axis and specify $u = \phi^{(2)}(y) \in C^1[0, y_0]$ with $\phi^{(2)}(0) = 0$ for a similar existence result as Theorem 4.4; The compatibility condition is now

$$\ell_0^{(k)} \frac{d\phi^{(2)}(0)}{dx} = \frac{1}{\lambda_1^0 - \lambda_k^0} \left[\ell_0^{(k)} \frac{d\phi^{(1)}(0)}{dx} - b_k^0 \right] \quad (2 \leq k \leq n). \quad (4.60)$$

The presentation of subsections 1–4 is based on the work of Wang and Wu [40].

4.5 Application to 2-D Euler

Consider the unknowns (α, c, β) in the system

$$\begin{aligned} \frac{\partial \alpha}{\partial u} - \cot \beta \frac{\partial \alpha}{\partial v} &= \frac{\gamma + 1}{4c} \cdot \frac{\sin(\alpha - \beta)}{\sin \beta} \cdot (m - \tan^2 \omega), \\ \frac{\partial c}{\partial u} &= \frac{\gamma - 1}{2} \frac{\cos \frac{\alpha + \beta}{2}}{\sin \omega}, \\ \frac{\partial \beta}{\partial u} - \cot \alpha \frac{\partial \beta}{\partial v} &= \frac{\gamma + 1}{4c} \cdot \frac{\sin(\alpha - \beta)}{\sin \alpha} \cdot (m - \tan^2 \omega), \end{aligned} \quad (4.61)$$

where $\gamma > 1$, $m = \frac{3-\gamma}{1+\gamma}$, $\omega = (\alpha - \beta)/2$. Fix $\theta \in (0, \pi/2)$. Consider two segments

$$\begin{aligned} H_1 : u \cos \theta + v \sin \theta &= 0 \quad (-\frac{2}{\gamma-1} \sin \theta \leq u \leq 0) \quad \text{and} \\ H_2 : u \cos \theta - v \sin \theta &= 0 \quad (-\frac{2}{\gamma-1} \sin \theta \leq u \leq 0). \end{aligned}$$

The boundary values of c on H_1 , H_2 , are

$$c|_{H_1} = c|_{H_2} = 1 + \frac{\gamma - 1}{2 \sin \theta} u. \quad (4.62)$$

Obviously, we note that $0 \leq c \leq 1$ on the boundaries. We want to seek a solution in the wave interaction region Ω bounded by H_1 , H_2 and the interface of vacuum connecting D and E in the (u, v) -plane, see Figure 3.2.

The boundary values for α , β on H_1 and H_2 are

$$\alpha|_{H_1} = \theta, \quad \beta|_{H_2} = -\theta. \quad (4.63)$$

The values of β on H_1 or α on H_2 can be integrated from the system. For example, we have on H_1 the equation

$$\frac{d\beta}{du} = \frac{\gamma + 1}{4c} \cdot \frac{\sin(\theta - \beta)}{\sin \theta} \cdot (m - \tan^2 \omega)$$

with $\beta(0, 0) = -\theta$ and c from (4.62). It is obvious that a local smooth solution exists for β on H_1 .

Problem B. Find a solution (α, c, β) of (4.61) with boundary values (4.62) and (4.63), in the wave interaction region Ω .

Theorem 4.5. A local C^1 solution (α, c, β) exists at the origin in Ω .

Proof. We need only to verify the compatibility condition (4.39): We let $x = -u$, $y = v$, and rewrite the system into

$$\begin{aligned} \frac{\partial \alpha}{\partial x} + \cot \beta \frac{\partial \alpha}{\partial y} &= -\frac{\gamma+1}{4c} \cdot \frac{\sin(\alpha-\beta)}{\sin \beta} \cdot (m - \tan^2 \omega), \\ \frac{\partial c}{\partial x} &= -\frac{\gamma-1}{2} \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega}, \\ \frac{\partial \beta}{\partial x} + \cot \alpha \frac{\partial \beta}{\partial y} &= -\frac{\gamma+1}{4c} \cdot \frac{\sin(\alpha-\beta)}{\sin \alpha} \cdot (m - \tan^2 \omega), \end{aligned} \quad (4.64)$$

then $\lambda_1 = \cot \beta$, $\lambda_2 = 0$, $\lambda_3 = \cot \alpha$; $\ell_j^{(i)} = \delta_{ij}$, $\Gamma_1 = H_2$, $\Gamma_2 = H_1$. The compatibility condition materializes as

$$\frac{1}{\cot \theta} \left[\frac{d}{dx}|_{\Gamma_2} c - \left(-\frac{\gamma-1}{2} \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega} \right) \right] = \frac{1}{\cot \beta} \left[\frac{d}{dx}|_{\Gamma_1} c - \left(-\frac{\gamma-1}{2} \frac{\cos \frac{\alpha+\beta}{2}}{\sin \omega} \right) \right] \quad (4.65)$$

at the point $(x, y) = (0, 0)$ and $(\alpha, \beta) = (\theta, -\theta)$. But we have

$$\frac{d}{dx}|_{\Gamma_2} c = -\frac{d}{du}|_{H_1} c = -\frac{d}{du}(c_{H_1}) = -\frac{\gamma-1}{2} \frac{1}{\sin \theta},$$

and a similar equation holds on H_2 , so each side of (4.65) is zero at the origin, thus the compatibility condition holds. By the local existence theorem for the Goursat problem (Theorem 4.3), we have a smooth solution at the origin. This completes the proof. \square

We can use Theorem 4.4 for the mixed-initial-boundary value problem to extend the local solution beyond the origin. But we need *a priori* maximum bounds on (α, c, β) to obtain global solution (until vacuum forms).

5 Invariant regions for systems

We present a basic tool for establishing invariant regions for systems of partial differential equations including reaction-diffusion-convection equations and quasilinear hyperbolic systems of equations.

5.1 Basic theorems

Consider the system

$$v_t = Dv_{xx} + Mv_x + f(v, t), \quad (x, t) \in \Omega \times \mathbb{R}_+ \quad (5.1)$$

with the initial condition

$$v(x, 0) = v_0(x), \quad x \in \Omega. \quad (5.2)$$

Here Ω is an open interval in \mathbb{R} , $D = D(v, x)$, and $M = M(v, x)$ are matrix-valued functions defined on open subset $U \times V \subset \mathbb{R}^n \times \Omega$, $D \geq 0$, $v = (v_1, v_2, \dots, v_n)$ and f is a smooth function from $U \times \mathbb{R}_+$ into \mathbb{R}^n . If Ω is not all of \mathbb{R} , we will assume that v satisfies specific boundary conditions; e.g., Dirichlet, Neumann, or even Goursat boundary conditions (a.k.a. characteristic boundary values). We assume that this problem has a local (in time) solution on some subset X of $C^1(\Omega)$; i.e., given a function $v_0 \in X$, there is a $\delta > 0$ and a smooth solution $v(x, t)$ of (5.1), (5.2) defined for $x \in \Omega$ and $t \in [0, \delta)$, such that $v(\cdot, t) \in X, 0 \leq t < \delta$.

Definition 5.1. A closed subset $\Sigma \subset \mathbb{R}^n$ is called a **(positively) invariant region** for the local solution defined by (5.1), (5.2) if any solution $v(x, t)$ having all of its boundary and initial values in Σ , satisfies $v(x, t) \in \Sigma$ for all $x \in \Omega$ and for all $t \in [0, \delta)$.

We shall assume that **condition K** is met for the couple (X, Σ) : For any $u \in X$, there is a compact set $K \subset \Omega$ such that $u(x) \in \Sigma$ for all $x \notin K$.

For example, let us consider the simple heat equation $u_t = u_{xx}$, where X is the space C_0^2 of C^2 functions which tend to zero as $|x| \rightarrow \infty$, and

$$\Sigma = \{u : -1 \leq u \leq 1\}.$$

Then the maximum principle shows that Σ is invariant, and condition K is met for this (X, Σ) . Observe that condition K is always valid if we consider (5.1) on a bounded domain, with the standard boundary conditions; the condition is needed only to recover some measure of compactness when we are on unbounded domains.

The invariant regions Σ will be made up of intersections of “half spaces;” i.e., we consider regions Σ of the form

$$\Sigma = \cap_{i=1}^m \{v \in U : G_i(v) \leq 0\}, \quad (5.3)$$

where G_i are smooth real-valued functions defined on open subsets of U , and for each i , the gradient ∇G_i never vanishes.

Let $\Omega = \mathbb{R}$. Now if there is a solution v of (5.1), (5.2) with (boundary data and) initial data in Σ for all $x \in \Omega$, which is *not* in Σ for all $t > 0$, then there is a function G_i , a time $t_0 > 0$, and a point $x_0 \in \mathbb{R}$ (using condition K) such that for $t \leq t_0$ and $x \in \mathbb{R}$, $G_i(v(x, t)) \leq 0$, and for any $\epsilon > 0$, there is a $t' \in (t_0, t_0 + \epsilon)$, such that $G_i(v(x_0, t')) > 0$.

Thus, if the assumptions

$$G_i(v(x_0, t)) < 0 \quad \text{for } 0 \leq t < t_0 \quad \text{and } G_i(v(x_0, t_0)) = 0, \quad (5.4)$$

together imply that

$$\frac{\partial G_i(v)}{\partial t} < 0 \quad \text{at } (x_0, t_0), \quad (5.5)$$

then Σ must be invariant.

Let us introduce another definition.

Definition 5.2. The smooth function $G : \mathbb{R}^n \rightarrow \mathbb{R}$ is called **quasi-convex** at v if $\eta \cdot \nabla^2 G(v)\eta \geq 0$ whenever η is such that $\eta \cdot \nabla G(v) = 0$.

If $G = G(v_i)$ depends only on one component of v and ∇G does not vanish at a point, then it is quasi-convex at that point.

Here is our main theorem

Theorem 5.1. Let Σ be defined by (5.3), and suppose that for all $t \in \mathbb{R}_+$ and for every $v_0 \in \partial\Sigma$ (so $G_i(v_0) = 0$ for some i), the following conditions hold:

- (1) ∇G_i at v_0 is a common left eigenvector of $D(v_0, x)$ and $M(v_0, x)$ for all $x \in \mathbb{R}$.
- (2) If $\nabla G_i D(v_0, x) = \mu \nabla G_i$ with $\mu \neq 0$, then G_i is quasi-convex at v_0 .
- (3) $\nabla G_i \cdot f < 0$ at v_0 for all $t \in \mathbb{R}_+$.

Then Σ is invariant for (5.1).

Proof. Let $G = G_i$ for simplicity in notation. To show that Σ is invariant, we assume that (5.4) holds and we shall show (5.5). Thus, at (x_0, t_0) ,

$$\frac{\partial G(v)}{\partial t} = \nabla G \cdot v_t = \nabla G(Dv_{xx} + Mv_x + f).$$

Now since ∇G is a left eigenvector of D and M , we have at $v_0 = v(x_0, t_0)$,

$$\nabla G \cdot D = \mu \nabla G \quad \text{and} \quad \nabla G \cdot M = \lambda \nabla G.$$

This implies that

$$\frac{\partial G(v)}{\partial t} = \mu \nabla G \cdot v_{xx} + \lambda \nabla G \cdot v_x + \nabla G \cdot f. \quad (5.6)$$

Now we make the key claim that at (x_0, t_0)

$$\nabla G \cdot v_x = 0. \quad (5.7)$$

To see that, define $h(x) = G(v(x, t_0))$; then $h(x_0) = 0$, and $h'(x) = \nabla G \cdot v_x(x, t_0)$. If $h'(x_0) > 0$, then $h(x) > 0$ for $x > x_0$ provided $|x - x_0|$

is small. Thus $G(v(x, t_0)) > 0$ for x close to x_0 , and so $G(v(x, t)) > 0$ for $|t - t_0| < \epsilon$ for some $\epsilon > 0$; in particular $G(v(x, t)) > 0$ for some x and some $t < t_0$. This violates (5.4). Similarly $h'(x_0) < 0$ is impossible. Thus $\nabla G \cdot v_x(x_0, t_0) = h'(x_0) = 0$ and this proves the claim.

Observe too that with $h(x)$ as defined above, $h''(x_0) \leq 0$; otherwise we would arrive at a contradiction similar to the one above. It follows that

$$0 \geq h''(x_0) = \nabla^2 G(v_x, v_x) + \nabla G \cdot v_{xx}. \quad (5.8)$$

Now suppose that $\mu \neq 0$; then $\mu > 0$, so from the second hypothesis, together with our claim, we find that $\nabla^2 G(v_x, v_x) \geq 0$ at (x_0, t_0) . Therefore from (5.8), $\nabla G \cdot v_{xx} \leq 0$ at (x_0, t_0) . Thus (5.6) gives

$$\frac{\partial G(v)}{\partial t} \leq \nabla G \cdot f < 0$$

in view of the third hypothesis. This completes the proof. \square

Remarks (i) We could have replaced hypotheses (2) and (3) by (2') and (3') where

(2') If $\nabla G_i D(v_0, x) = 0$ then (3) is still assumed; If $\nabla G_i D(v_0, x) = \mu \nabla G_i$ with $\mu \neq 0$, then G_i is *strongly quasi-convex* at v_0 ; i.e., $\nabla G_i(v_0)(\eta, \eta) > 0$ whenever $\eta \neq 0$ and $\nabla G_i(v_0)(\eta) = 0$; and

(3') $\nabla G_i \cdot f \leq 0$ at v_0 .

It seems we need condition (3) for the case $D \equiv 0$; cf. [36, p.202].

(ii) If D and M are diagonal matrices, and $G_i = u_i - c_i$ for some constant c_i , then G_i is everywhere quasi-convex, and ∇G_i is a left eigenvector of both D and M . Therefore the half space

$$\{u : u_i - c_i \leq 0\}$$

is invariant for (5.1), provided that

$$f_i(u_1, u_2, \dots, u_{i-1}, c_i, u_{i+1}, \dots, u_n) < 0,$$

where f_i is the i -th component of f . This gives the following useful corollary.

Corollary 5.1. (a) Suppose that D and M are diagonal matrices. Then any region of the form

$$\Sigma = \cap_{i=1}^m \{u : a_i \leq u_i \leq b_i\} \quad (5.9)$$

is invariant for (5.1), provided that f points strictly into Σ on $\partial\Sigma$; i.e., provided that hypothesis (3) of the theorem is valid.

(b) If D is the identity matrix and $M = 0$, then any convex region Σ , in which f points into Σ on $\partial\Sigma$, is invariant for (5.1).

We shall refer to such an invariant region (5.9) as an *invariant rectangle*.

One can relax condition (3) to (3') by the so-called f -stability of the system, see [36]. The idea of the f -stability is to consider the system

$$v_t = Dv_{xx} + Mv_x + f + \epsilon h,$$

where $\epsilon > 0$ and the vector field $h(v)$ is such that $dG_i(h) < 0$ on $\partial\Sigma$ for each $i \in \{1, 2, \dots, m\}$. And then let $\epsilon \rightarrow 0+$. One can also derive necessary conditions which must hold if Σ is an invariant set. See also [36]. The necessary conditions are almost the same as the sufficient conditions. Similar conclusions for more general systems in several space variables are possible, see [3]. The conditions are very restrictive. In our applications, we recommend the bootstrapping method as an alternative. This subsection is adapted with correction from [36].

5.2 Examples

We suggest to find invariant regions for the ordinary differential equations $y' = y(1-y)$; $y' = 1 - \frac{1}{x+1} - y$; $y' = 1 - e^{-x} - y$; $y' = 1 - e^{-2x} - y$.

Consider also the Goursat problem for the isothermal Euler system

$$\begin{cases} \alpha_t + \cot \beta \alpha_x = -\frac{\sin(\alpha - \beta)}{2 \sin \beta} \cdot \left[1 - \tan^2 \frac{\alpha - \beta}{2} \right], \\ \beta_t + \cot \alpha \beta_x = -\frac{\sin(\alpha - \beta)}{2 \sin \alpha} \cdot \left[1 - \tan^2 \frac{\alpha - \beta}{2} \right], \end{cases} \quad (5.10)$$

where $\alpha = \theta_0 \in (0, \frac{\pi}{2})$ on the line $t = \cot \theta_0 x, t > 0$, and $\beta = -\theta_0$ on $t = -\cot \theta_0 x, t > 0$.

6 The pressure gradient system

We derive the pressure gradient system and discuss its subsonic solutions.

6.1 Introduction

There has been remarkable progress in the study of the full (adiabatic) compressible Euler equations in multi-dimensions, see the work of Shuxing Chen, Jiequan Li, Zhouping Xin and Huicheng Yin, Yongqian Zhang, and Guiqiang Chen and Feldman. There are also natural simplifications of the Euler system such as the isentropic case or the irrotational (potential) flow equations, steady flows, or the unsteady transonic small disturbance equation (UTSD). We wish to derive the pressure gradient system of equations. They offer different perspectives for the full Euler system.

6.1.1 Derivation

Recall that the full Euler system for an ideal and polytropic gas takes the form

$$\begin{aligned}\rho_t + \nabla \cdot (\rho u) &= 0, \\ (\rho u)_t + \nabla \cdot (\rho u \otimes u + pI) &= 0, \\ (\rho E)_t + \nabla \cdot (\rho Eu + pu) &= 0,\end{aligned}\tag{6.1}$$

where

$$E := \frac{1}{2}|u|^2 + \frac{1}{\gamma - 1} \frac{p}{\rho},$$

where $\gamma > 1$ is the gas constant.

We let

$$\varepsilon := \frac{1}{\gamma - 1}$$

and look for an asymptotic solution of the form

$$\begin{aligned}\rho &= \rho_0 + \varepsilon \rho_1 + O(\varepsilon^2), \\ u &= \quad \varepsilon u_1 + O(\varepsilon^2), \\ p &= \quad \varepsilon p_1 + O(\varepsilon^2).\end{aligned}\tag{6.2}$$

This scaling corresponds to sound speeds of the order $O(1)$:

$$c = \sqrt{\gamma p / \rho} = O(\varepsilon^0).$$

Thus, we are scaling space and time variables by the same factor (order $O(1)$) to study acoustic phenomena.

The leading-order perturbation equation from conservation of mass is

$$(\rho_0)_t = 0,$$

so

$$\rho_0 = \rho_0(x).$$

The leading-order equation from conservation of momentum at $O(\varepsilon)$ and conservation of energy at $O(\varepsilon^2)$ form the *variable-density pressure gradient system*:

$$\begin{cases} (\rho_0 u_1)_t + \nabla p_1 = 0, \\ \left(\frac{1}{2} \rho_0 |u_1|^2 + p_1 \right)_t + \nabla \cdot (p_1 u_1) = 0. \end{cases}\tag{6.3}$$

For smooth solutions, the energy equation can be reduced to

$$(p_1)_t + p_1 \nabla \cdot u_1 = 0.\tag{6.4}$$

One can eliminate u_1 to form a single equation for p_1 :

$$\left(\frac{(p_1)_t}{p_1} \right)_t - \nabla \cdot \left(\frac{1}{\rho_0} \nabla p_1 \right) = 0.$$

If we let $\rho_0 = 1$, then we obtain a nice equation — the *pressure gradient equation*:

$$\left(\frac{(p_1)_t}{p_1} \right)_t - \Delta p_1 = 0. \quad (6.5)$$

This asymptotic regime (6.2) lacks strong physical sense because there is no such physical material for very large γ . There is no other nonphysical concerns though. For example, it is physical to have $p = \varepsilon p_1 + O(\varepsilon^2) \rightarrow 0$ since p is an independent variable and one can adjust the temperature to achieve it.

One guiding principle in asymptotics is this philosophy: One can look at a physical process *at any scale* one wishes — The point is whether you find anything interesting there.

6.1.2 Progress of research

Both Cauchy and Riemann problems for systems (6.3) or (6.5)) are open.

- Peng Zhang, Jiequan Li, and Tong Zhang (journal DCDS) have given a set of conjectures (with numerics) for solutions to the four-wave Riemann problem for these systems, see the book by Li et. al. [25] or Section 9.3 of Zheng's book [45].

The self-similar coordinates $\xi = x/t$, $\eta = y/t$ can reduce the Riemann problem by one dimension. However, both equations (6.3), (6.5) and even their linearized versions are of mixed type in the self-similar coordinates. One major difficulty in proving the conjectures is this change of type combined with the nonlinearity of the systems.

- Zheng ([42], 1997) has established the existence of solutions in the elliptic region. Equation (6.5) in the self-similar coordinates (ξ, η) takes the form

$$(P - \xi^2)P_{\xi\xi} - 2\xi\eta P_{\xi\eta} + (P - \eta^2)P_{\eta\eta} + \frac{1}{P}(\xi P_\xi + \eta P_\eta)^2 - 2(\xi P_\xi + \eta P_\eta) = 0. \quad (6.6)$$

The eigenvalues of the coefficient matrix of the second order terms of (6.6) can be found to be P and $P - \xi^2 - \eta^2$. Zheng proved in [42] the existence of a weak solution for equation (6.6) in any open, bounded and convex region $\Omega \subset \mathbb{R}^2$ with smooth boundary and the degenerate boundary datum

$$P|_{\partial\Omega} = \xi^2 + \eta^2 \quad (6.7)$$

provided that the boundary of Ω does not contain the origin $(0, 0)$.

- Kyungwoo Song ([26], 2003) has removed the restriction on the origin and the complete smoothness of the boundary. Kim and Song ([17], 2004) have obtained regularity of the solution in the interior of the domain and continuity up to and including the boundary.

• Dai and Zhang ([10], 2000) have obtained the interaction of two rarefaction waves adjacent to the vacuum.

• Zhen Lei and Yuxi Zheng ([21], 2006) have clarified the location of the vacuum boundary for the interaction of two rarefaction waves adjacent to the vacuum, improving the result of Dai and Zhang ([10].

• Seunghoon Bang (Ph.D. thesis, 2007 [1]) has obtained three wave interactions including interaction of a planar wave with a simple wave.

• Zheng has established the existence of a global solution involving a shock as a free boundary ([43], 2003). The regular reflection of a shock hitting a wedge of large angle has also been established, see ([44], 2006). This motivates the regular reflection results of Chen and Feldman, Elling and Liu, and Kim for the potential flow.

Jiequan Li has been able to use the result as motivation to solve the gas expansion to vacuum problem for the full Euler system in the hodograph plane. As we have seen in Chapter 3, we have converted the solution from the hodograph plane back to the self-similar plane.

6.1.3 One-dimensional planar waves

For any $p_1 > p_2 \geq 0$ and real u_1, v_1 , we can associate a one-dimensional rarefaction wave (see Figure 6.1):

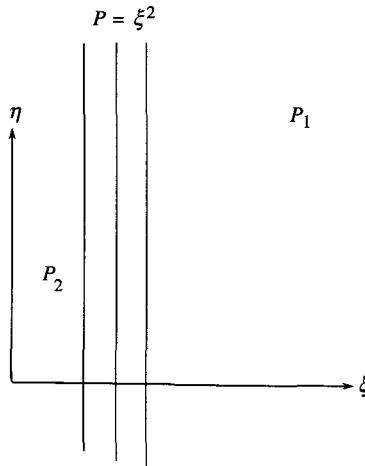


Figure 6.1 A planar wave.

$$\begin{cases} p = \xi^2, \\ u = 2(\xi - \sqrt{p_1}) + u_1, \\ v = v_1. \end{cases} \quad (\sqrt{p_2} < \xi < \sqrt{p_1}) \quad (6.8)$$

6.2 Two-dimensional Riemann problems

We consider the interaction of four rarefaction waves, see Figure 6.2. Note the subsonic region in the center. For binary interactions of planar rarefaction waves, see Dai and Zhang [10]. See Lei and Zheng [21] for more regularity estimate and extension of the solution of Dai and Zhang to a global solution. See Bang [1] for the interaction of a planar rarefaction wave with a simple wave.

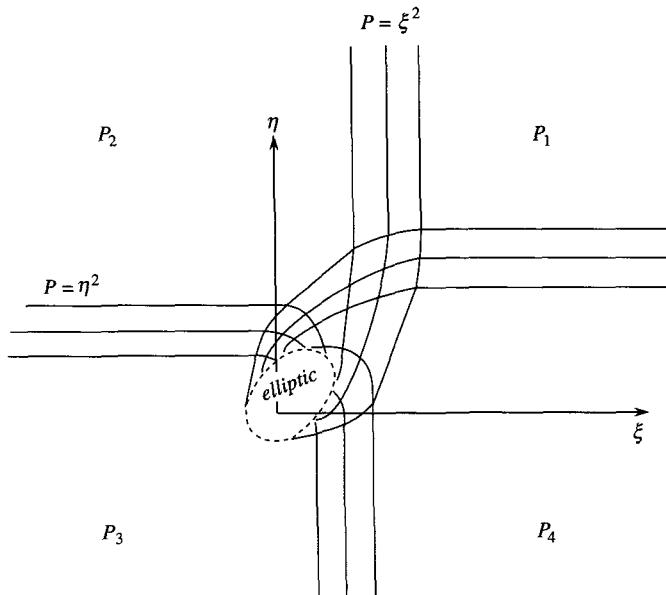


Figure 6.2 An elliptic region in the solution of a Riemann problem.

6.3 Subsonic region

We consider the problem: Find a weak or smooth solution $u(x, y)$ to the problem

$$\begin{cases} (u - x^2)u_{xx} - 2xyu_{xy} + (u - y^2)u_{yy} \\ \quad + \frac{1}{u}(xu_x + yu_y)^2 - 2(xu_x + yu_y) = 0 \text{ in } \Omega, \\ u|_{\partial\Omega} = x^2 + y^2, \end{cases} \quad (6.9)$$

where the region $\Omega \subset \mathbb{R}^2$ is open, bounded, and convex with boundary $\partial\Omega \in C^{2,\alpha}$ for some $\alpha \in (0, 1)$.

We assume that the origin $(0, 0)$ does not lie on the boundary of Ω . Our result is below.

Theorem 6.1 (Existence of subsonic solutions). There exists a positive weak solution $u \in H_{loc}^1(\Omega)$ to problem (6.9) with $u \in C_{loc}^{0,\alpha}(\Omega)$. It takes on the boundary value in the sense $[u - (x^2 + y^2)]^{3/2} \in H_0^1(\Omega)$. Furthermore, it has

- (i) maximum principle: $\min_{\partial\Omega}(x^2 + y^2) \leq u(x, y) \leq \max_{\partial\Omega}(x^2 + y^2)$,
- (ii) interior ellipticity: $u(x, y) - (x^2 + y^2) > 0$ in Ω .

Remark. $H_{loc}^1(\Omega)$ denotes the space of all functions $u \in H^1(\Omega')$ for any $\Omega' \subset\subset \Omega$, i.e. $\Omega' \subset \Omega$ and the closure $\overline{\Omega'} \subset \Omega$. $C_{loc}^{0,\alpha}(\Omega)$ means the same.

Proof. 1. We introduce the function

$$K(x, y, z) := \begin{cases} z, & \text{if } z \geq x^2 + y^2, \\ x^2 + y^2, & \text{if } z < x^2 + y^2. \end{cases}$$

This function is Lipschitz continuous in \mathbb{R}^3 . Now consider the screened problem

$$\begin{cases} (K(x, y, u) - x^2 + \varepsilon)u_{xx} - 2xyu_{xy} + (K(x, y, u) - y^2 + \varepsilon)u_{yy} \\ \quad + \frac{1}{N(u)}(xu_x + yu_y)^2 - 2(xu_x + yu_y) = 0 \quad \text{in } \Omega, \\ u|_{\partial\Omega} = x^2 + y^2, \end{cases} \quad (6.10)$$

where $\varepsilon > 0$ is a parameter, $N(v) \in C^3(\mathbb{R})$ is bounded from both above and below with a positive lower bound, and $N(v) = v$ in the interval $[\min_{\partial\Omega}(x^2 + y^2), \max_{\partial\Omega}(x^2 + y^2)]$.

We find easily that the equation in (6.10) is uniformly elliptic in Ω , since the two eigenvalues of the matrix

$$\mathbf{A} = \begin{pmatrix} K - x^2 + \varepsilon & -xy \\ -xy & K - y^2 + \varepsilon \end{pmatrix}$$

are $\Lambda = K(x, y, u) + \varepsilon$, and $\lambda = K(x, y, u) + \varepsilon - (x^2 + y^2) \geq \varepsilon$. Then the existence of a $C^{2,\alpha}(\overline{\Omega})$ solution $u^{K,\varepsilon,N}(x, y)$ of (6.10) follows from Theorem 15.12 of Gilbarg and Trudinger [GiTr] in the special case of two space dimensions in which the Hölder continuity of K is sufficient.

2. The solutions $u^{K,\varepsilon,N}$ satisfy the maximum principle

$$\min_{\partial\Omega}(x^2 + y^2) \leq u^{K,\varepsilon,N}(x, y) \leq \max_{\partial\Omega}(x^2 + y^2) \quad \text{in } \Omega,$$

see Gilbarg-Trudinger [GiTr]. Therefore the $N(u)$ regularization degenerates to u itself. So the solutions $u^{K,\varepsilon,N}$ are independent of N . We therefore drop the superscript N . We thus have $C^{2,\alpha}(\overline{\Omega})$ solutions $u^{K,\varepsilon}$

to the problem

$$\begin{cases} (K(x, y, u) - x^2 + \varepsilon)u_{xx} - 2xyu_{xy} + (K(x, y, u) - y^2 + \varepsilon)u_{yy} \\ \quad + \frac{1}{u}(xu_x + yu_y)^2 - 2(xu_x + yu_y) = 0, \\ u|_{\partial\Omega} = x^2 + y^2. \end{cases} \quad (6.11)$$

We next show that the $K(x, y, u)$ screening in fact degenerates to u also for the solutions $u^{K,\varepsilon}(x, y)$. Let

$$F(x, y) := u^{K,\varepsilon}(x, y) - (x^2 + y^2).$$

We have $F|_{\partial\Omega} = 0$. We claim that $F > 0$ in Ω . Otherwise there exists a point $(x_0, y_0) \in \Omega$, which may depend on (K, ε) , such that $F(x_0, y_0) \leq 0$, $F_x(x_0, y_0) = F_y(x_0, y_0) = 0$, and

$$(K(x, y, u^{K,\varepsilon}(x, y)) - x^2 + \varepsilon)F_{xx} - 2xyF_{xy} + (K(x, y, u^{K,\varepsilon}(x, y)) - y^2 + \varepsilon)F_{yy} \geq 0$$

at (x_0, y_0) . Rewriting those conditions in terms of u and using equation (6.11), we find

$$u^{K,\varepsilon}(x_0, y_0) \leq x_0^2 + y_0^2,$$

$$\frac{\partial u^{K,\varepsilon}}{\partial x}(x_0, y_0) = 2x_0, \quad \frac{\partial u^{K,\varepsilon}}{\partial y}(x_0, y_0) = 2y_0,$$

and

$$2(K(x, y, u^{K,\varepsilon}(x, y)) - x^2 + \varepsilon) + 2(K(x, y, u^{K,\varepsilon}(x, y)) - y^2 + \varepsilon) + \frac{(2x_0^2 + 2y_0^2)^2}{u^{K,\varepsilon}(x_0, y_0)} - 2(2x_0^2 + 2y_0^2) \leq 0 \quad \text{at } (x_0, y_0).$$

But the last inequality is false because

$$\begin{aligned} K(x, y, u^{K,\varepsilon}(x, y)) - x^2 + \varepsilon &\geq \varepsilon, \\ K(x, y, u^{K,\varepsilon}(x, y)) - y^2 + \varepsilon &\geq \varepsilon, \\ \frac{(2x_0^2 + 2y_0^2)^2}{u^{K,\varepsilon}(x_0, y_0)} - 2(2x_0^2 + 2y_0^2) &\geq \frac{(2x_0^2 + 2y_0^2)^2}{x_0^2 + y_0^2} - 2(2x_0^2 + 2y_0^2) \geq 0. \end{aligned}$$

Hence we have proved that

$$u^{K,\varepsilon}(x, y) > x^2 + y^2 \quad \text{in } \Omega.$$

Therefore our solutions $u^{K,\varepsilon}$ do not even depend on K . So we drop the K dependence and we have existence of $C^{2,\alpha}(\bar{\Omega})$ solutions $u^\varepsilon(x, y)$ to the problem

$$\begin{cases} (u - x^2 + \varepsilon)u_{xx} - 2xyu_{xy} + (u - y^2 + \varepsilon)u_{yy} \\ \quad + \frac{1}{u}(xu_x + yu_y)^2 - 2(xu_x + yu_y) = 0, \\ u|_{\partial\Omega} = x^2 + y^2. \end{cases} \quad (6.12)$$

3. We now establish one of the two major estimates on u^ε independently of $\varepsilon > 0$. We introduce the function

$$\varphi^\varepsilon(x, y) = u^\varepsilon - (x^2 + y^2).$$

We find that φ^ε satisfies the equation

$$\begin{cases} (\varepsilon + \varphi + y^2)\varphi_{xx} - 2xy\varphi_{xy} + (\varepsilon + \varphi + x^2)\varphi_{yy} + \\ \frac{(x\varphi_x + y\varphi_y - 2\varphi)^2}{\varphi + x^2 + y^2} + 2(x\varphi_x + y\varphi_y) + 2x^2 + 2y^2 + 4\varepsilon = 0, \\ \varphi|_{\partial\Omega} = 0, \end{cases} \quad (6.13)$$

where we have dropped the superscript ε for simplicity. We know that $0 < \varphi^\varepsilon \leq \max_{\partial\Omega}(x^2 + y^2)$ in Ω . We show that

$$\int \int_{\Omega} \varphi^\varepsilon |\nabla \varphi^\varepsilon|^2 dx dy \leq C \quad \text{independent of } 1 \geq \varepsilon > 0. \quad (6.14)$$

In fact, we multiply equation (6.13) with φ^ε to find

$$\begin{aligned} & [(\varphi + y^2 + \varepsilon)\varphi\varphi_x - xy\varphi\varphi_y]_x + [-xy\varphi\varphi_x + (\varphi + x^2 + \varepsilon)\varphi\varphi_y]_y \\ & - (2\varphi + y^2 + \varepsilon)\varphi_x^2 + 2xy\varphi_x\varphi_y - (2\varphi + x^2 + \varepsilon)\varphi_y^2 + \frac{\varphi(x\varphi_x + y\varphi_y)^2}{\varphi + x^2 + y^2} \\ & + 3\varphi(x\varphi_x + y\varphi_y) - \frac{4\varphi^2}{\varphi + x^2 + y^2}(x\varphi_x + y\varphi_y) + \frac{4\varphi^3}{\varphi + x^2 + y^2} \\ & + 2\varphi(x^2 + y^2) + 4\varphi\varepsilon = 0. \end{aligned} \quad (6.15)$$

Integrating over Ω and using the zero boundary condition of φ , we obtain

$$\begin{aligned} & \int \int_{\Omega} [(2\varphi + y^2 + \varepsilon)\varphi_x^2 - 2xy\varphi_x\varphi_y + \\ & \quad (2\varphi + x^2 + \varepsilon)\varphi_y^2 - \frac{\varphi(x\varphi_x + y\varphi_y)^2}{\varphi + x^2 + y^2}] dx dy \\ & = \int \int_{\Omega} \left[\left(3\varphi - \frac{4\varphi^2}{\varphi + x^2 + y^2} \right) (x\varphi_x + y\varphi_y) \right. \\ & \quad \left. + \frac{4\varphi^3}{\varphi + x^2 + y^2} + 2\varphi(x^2 + y^2) + 4\varphi\varepsilon \right] dx dy. \end{aligned}$$

We further simplify the integral on the left-hand side in the above equation to obtain

$$\begin{aligned} & \int \int_{\Omega} \frac{2\varphi + x^2 + y^2}{\varphi + x^2 + y^2} [(\varphi + y^2)\varphi_x^2 - 2xy\varphi_x\varphi_y + (\varphi + x^2)\varphi_y^2] dx dy \\ & \quad + \varepsilon \int \int_{\Omega} (\varphi_x^2 + \varphi_y^2) dx dy \\ & = \int \int_{\Omega} \left[\left(3\varphi - \frac{4\varphi^2}{\varphi + x^2 + y^2} \right) (x\varphi_x + y\varphi_y) \right. \\ & \quad \left. + \frac{4\varphi^3}{\varphi + x^2 + y^2} + 2\varphi(x^2 + y^2) + 4\varphi\varepsilon \right] dx dy. \end{aligned}$$

Using the fact $\varphi > 0$ in Ω and is bounded from above by $\max_{\partial\Omega}(x^2 + y^2)$, we find

$$\int \int_{\Omega} \varphi(\varphi_x^2 + \varphi_y^2) \leq C \int \int_{\Omega} \varphi(|\varphi_x| + |\varphi_y|) dx dy + C,$$

where C is a constant depending on Ω , but independent of $\varepsilon \in (0, 1]$.

A weighted Cauchy-Schwarz inequality on the right-hand side yields

$$\int \int_{\Omega} \varphi (\varphi_x^2 + \varphi_y^2) dx dy \leq \frac{1}{2} \int \int_{\Omega} \varphi (\varphi_x^2 + \varphi_y^2) dx dy + C,$$

which further yields

$$\int \int_{\Omega} \varphi^\varepsilon (\varphi_x^{\varepsilon 2} + \varphi_y^{\varepsilon 2}) dx dy \leq C(\Omega).$$

So (6.14) is proved.

4. We need our next major estimate to be able to draw convergent subsequences of φ^ε in $H_{loc}^1(\Omega)$ from the above estimate. We establish ellipticity of u^ε uniformly for $\varepsilon \in (0, 1]$ in the interior of Ω . Let $\xi(x, y)$ be any nonnegative function in $C^3(\Omega)$ with zero boundary data $\xi|_{\partial\Omega} = 0$. We claim that there exists a small positive number $\beta > 0$, independent of $\varepsilon \in (0, 1]$ such that

$$\eta^\varepsilon(x, y) := u^\varepsilon(x, y) - (x^2 + y^2) - \beta \xi(x, y) > 0 \quad \text{in } \Omega. \quad (6.16)$$

In fact, we can take $\beta > 0$ so small that

$$\beta \left(2 \max_{\partial\Omega} (x^2 + y^2) + 1 \right) \max_{\Omega} |D^2 \xi| < \min_{\partial\Omega} (x^2 + y^2), \quad (6.17)$$

where $D^2 \xi$ represents all second order derivatives, and

$$\beta \max_{\Omega} \left| \frac{2x^2 + 2y^2 + \beta x \xi_x + \beta y \xi_y}{x^2 + y^2 + \beta \xi} (x \xi_x + y \xi_y - 2\xi) \right| < \min_{\partial\Omega} (x^2 + y^2). \quad (6.18)$$

We note that $\min_{\partial\Omega} (x^2 + y^2) > 0$ since the origin $(0, 0)$ does not belong to $\partial\Omega$. We use contradiction method to prove (6.16). Suppose $\eta^\varepsilon(x, y)$ is not positive in Ω . There must be a minimum point in the interior of Ω . At the minimum point,

$$\begin{aligned} u^\varepsilon &\leq x^2 + y^2 + \beta \xi(x, y), \\ u_x^\varepsilon &= 2x + \beta \xi_x, \\ u_y^\varepsilon &= 2y + \beta \xi_y, \end{aligned}$$

and

$$(u^\varepsilon - x^2 + \varepsilon) \eta_{xx}^\varepsilon - 2xy \eta_{xy}^\varepsilon + (u^\varepsilon - y^2 + \varepsilon) \eta_{yy}^\varepsilon \geq 0.$$

Using equation (6.12) for u^ε , we find

$$\begin{aligned} &(u^\varepsilon - x^2 + \varepsilon) \eta_{xx}^\varepsilon - 2xy \eta_{xy}^\varepsilon + (u^\varepsilon - y^2 + \varepsilon) \eta_{yy}^\varepsilon \\ &+ 2(u^\varepsilon - x^2 + \varepsilon) + 2(u^\varepsilon - y^2 + \varepsilon) \\ &+ \beta [(u^\varepsilon - x^2 + \varepsilon) \xi_{xx} - 2xy \xi_{xy} + (u^\varepsilon - y^2 + \varepsilon) \xi_{yy}] \\ &+ \frac{(2x^2 + 2y^2 + \beta x \xi_x + \beta y \xi_y)^2}{\eta^\varepsilon + (x^2 + y^2) + \beta \xi} \\ &- 2(2x^2 + 2y^2 + \beta x \xi_x + \beta y \xi_y) = 0. \end{aligned} \quad (6.19)$$

We observe that in the above equation, the first three terms together is nonnegative; the next two terms together gives

$$\begin{aligned} 2(u^\varepsilon - x^2 + \varepsilon) + 2(u^\varepsilon - y^2 + \varepsilon) &\geq 2(u^\varepsilon - x^2 - y^2) + 2u^\varepsilon \\ &\geq 2u^\varepsilon \geq 2\min_{\partial\Omega}(x^2 + y^2) > 0. \end{aligned}$$

The terms in the bracket can be bounded by

$$\begin{aligned} \beta |[(u^\varepsilon - x^2 + \varepsilon)\xi_{xx} - 2xy\xi_{xy} + (u^\varepsilon - y^2 + \varepsilon)\xi_{yy}]| \\ \leq \beta (2\max_{\partial\Omega}(x^2 + y^2) + 1) \max_{\Omega} |D^2\xi| \\ \leq \min_{\partial\Omega}(x^2 + y^2), \end{aligned}$$

where we used (6.17). All the remaining terms in the equation have the estimate

$$\begin{aligned} &\frac{(2x^2 + 2y^2 + \beta x\xi_x + \beta y\xi_y)^2}{\eta^\varepsilon + \beta\varepsilon + x^2 + y^2} - 2(2x^2 + 2y^2 + \beta x\xi_x + \beta y\xi_y) \\ &\geq \frac{(2x^2 + 2y^2 + \beta x\xi_x + \beta y\xi_y)^2}{x^2 + y^2 + \beta\varepsilon} - 2(2x^2 + 2y^2 + \beta x\xi_x + \beta y\xi_y) \\ &= \beta \frac{2x^2 + 2y^2 + \beta x\xi_x + \beta y\xi_y}{x^2 + y^2 + \beta\varepsilon} (x\xi_x + y\xi_y - 2\xi) \end{aligned}$$

whose absolute value is less than $\min_{\partial\Omega}(x^2 + y^2)$ by (6.18). Thus the equation for η^ε (6.19) is violated. Hence $\eta^\varepsilon > 0$ in Ω . Therefore

$$u^\varepsilon(x, y) - (x^2 + y^2) \geq \beta\xi(x, y) > 0 \quad \text{in } \Omega$$

for a small $\beta > 0$ independent of $\varepsilon > 0$.

5. Hence in any $\Omega' \subset\subset \Omega$, we have

$$\int \int_{\Omega'} |\nabla \varphi^\varepsilon|^2 dx dy \leq C.$$

We therefore have a subsequence of $\{\varphi^\varepsilon\}_{\varepsilon>0}$, still denoted by $\{\varphi^\varepsilon\}_{\varepsilon>0}$ which converges weakly to a function $\varphi \in H_{loc}^1(\Omega)$,

$$\begin{aligned} \varphi^\varepsilon &\rightharpoonup \varphi \quad \text{in } L_{loc}^2(\Omega), \\ \nabla \varphi^\varepsilon &\rightharpoonup \nabla \varphi \quad \text{in } L_{loc}^2(\Omega). \end{aligned}$$

Using a theorem found in Evans [Ev1] (see the appendix for more details), we can improve the weak convergence to strong convergence:

$$\nabla \varphi^\varepsilon \rightarrow \nabla \varphi \text{ in } L_{loc}^2(\Omega).$$

Thus φ satisfies the equation in problem (6.9) in the sense of distributions.

Or we can change (6.12) to a form without needing the strong convergence:

$$[(1 - \frac{x^2 - \varepsilon}{u})u_x - \frac{xy}{u}u_y]_x + [-\frac{xy}{u}u_x + (1 - \frac{y^2 - \varepsilon}{u})u_y]_y + 2(xu_x + yu_y) = 0. \quad (6.20)$$

We have trivially

$$\int \int_{\Omega} \left| \nabla \left((\varphi^\varepsilon)^{3/2} \right) \right|^2 dx dy \leq C, \quad (\varphi^\varepsilon)^{3/2} \Big|_{\partial\Omega} = 0.$$

It follows that

$$\int \int_{\Omega} \left| \nabla \left(\varphi^{3/2} \right) \right|^2 dx dy \leq C, \text{ and } \varphi^{3/2} \Big|_{\partial\Omega} = 0 \text{ in trace sense.}$$

Therefore

$$\varphi^{3/2} \in H_0^1(\Omega).$$

By Lemma 15.4 of Gilbarg-Trudinger [GiTr] (i.e., the Krylov estimate), we obtain the interior estimate $u \in C_{loc}^{0,\alpha}(\Omega)$. \square

Remarks. **1.** We do not obtain higher regularity than $u \in C_{loc}^{0,\alpha}(\Omega)$ despite that the equation in (6.9) is elliptic in the interior, because the quadratic nonlinear term $(xu_x + yu_y)^2$ is only in L^1 . **2.** The regularity assumption on the smoothness of the boundary can be greatly reduced. We use the regularity for quoting the simplest elliptic theorems.

6.4 Four-wave interaction

Taking four rarefaction waves to interaction, Seunghoon Bang [1] has obtained the solution of the interaction up to the boundary of the domain of hyperbolic determinacy. The *domain of hyperbolic determinacy* is a region where both families of characteristics can be traced back to infinity in the self-similar plane. See Figure 6.3. In Figure 6.3, the interactions ABMC, EOKG, and DFJN are all settled. The waves MCEO and MBDN are continuous simple waves which have one straight family of characteristics. The simple wave KGIA, adjacent to the sonic curve, involves characteristics that start from the sonic curve and end at the sonic curve, therefore KGIA is not quite determined. A similar situation occurs in the zone KOcb, and two additional places above the symmetric axis $\eta = \xi$. Wave reflection occurs in these zones, which are compressive that may result in shock formation. These zones might be named Erebus zones or styx.

7 Open problems

We propose some concrete open problems.

We have proposed open problems in Section 1, but they may be too rough to start with for a student. Here we describe in more concrete terms a few interesting problems.

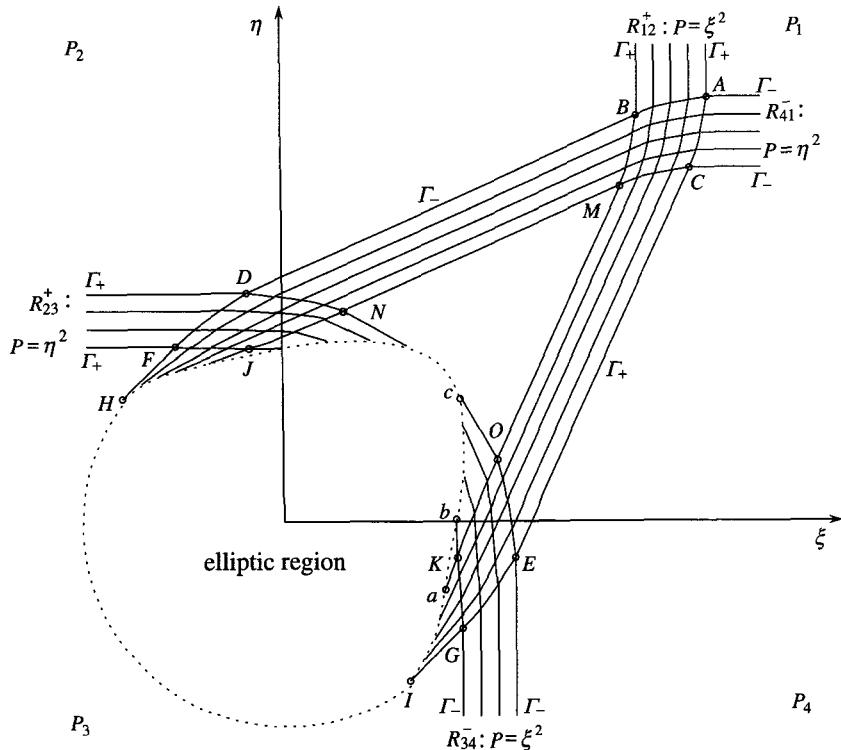


Figure 6.3 Hyperbolic solution of Bang (Courtesy of Bang).

1. Consider the steady isentropic and irrotational 2-D Euler system in an infinite channel with a smooth bump inward, see Figure 7.1. The flow from infinity is subsonic and remains subsonic, except near the bump, where it becomes supersonic since the bump makes the channel narrower. The supersonic patch contains a compressive family of characteristics toward the rear end (right half of the figure), which may form a shock. The characteristics method may play a major role in the construction of the supersonic patch.

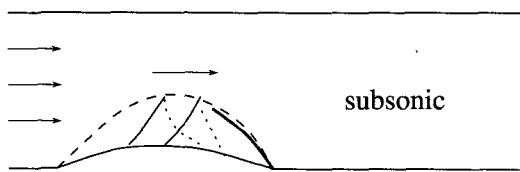


Figure 7.1 Steady flow in a channel with a bump.

2. Consider the interaction of four rarefaction waves of the self-similar Euler system. This lecture notes has done the interaction of two rarefaction waves. A note of caution is that Jiequan Li and I are working on it, slowly. Furthermore, shock waves may form near the sonic boundary, see the recent article of Glimm et. al. [14]. Yet, shock-free interaction with non-vacuum subsonic domain and non-constant hyperbolic region does exist, see Zheng [46].

3. Consider using the self-similar coordinates directly, rather than using the hodograph plane, in the construction of solutions to the self-similar Euler system. The reason we do not use it is that we find that the complexity of the formulas in the self-similar plane is overwhelming. We believe that it can be simplified. Xiao Chen and I have made some progress, see [6].

4. von Neumann paradoxes. John Hunter and Allen M. Tesdall [16] propose to use a rarefaction wave fan to resolve the von Neumann paradox associated with the Mach triple point configuration at small wedge angle and weak incident shock, mentioned in Chapter one. Once the rarefaction wave fan is used, the paradox is settled, but the fan will cause reflection off the sonic curve, which forms more shock waves, which will hit the Mach stem to cause more rarefaction wave fans, and then their reflections off the sonic curve, and therefore more shocks and more rarefaction wave fans, so on and so forth, to result in a chain of rarefaction wave fans, reflections, and shocks. This type of reflection is called Guderley reflection. See Figure 7.2. The method of characteristics for the self-similar Euler system should play a major role in the construction of the hyperbolic solutions near the wave fans at the triple point.

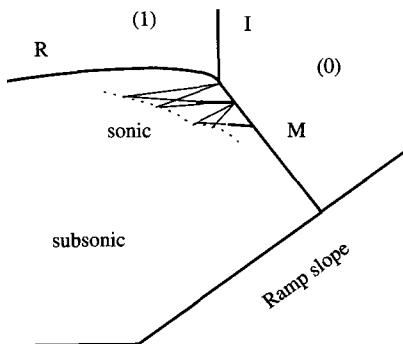


Figure 7.2 Guderley reflection.

5. Interaction of four slip lines in the same rotation direction. See Li [25] or Zheng [45].
6. Similar decomposition for the three dimensional Euler system.

Epilogue: Stories

The lecture notes are filled with technical computations. The classroom presentation, however, followed the idea of characteristics and Riemann variables. Indeed, we followed the same idea in our original research through the model of pressure-gradient system. The success in the pressure gradient system greatly motivates us to look for the diagonal form and Riemann variables in the Euler system. The model of pressure-gradient system plays a key step. Without it, we would not be encouraged enough to go far enough to find the diagonal form of the self-similar Euler system in the hodograph plane. The success of the pressure gradient system is the key motivation. The model building is the spirit of all the success presented in this lecture notes. To work on hard problems, building a ladder problem is a good idea, as said by S. T. Yau. The model building of the pressure gradient system starts with Tong Zhang and some of his students and post-doctoral associates, and John Hunter offered the asymptotic derivation.

Associated with the technical computations and the model building, it seems it is appropriate to tell the story of “Fishing of Zhuang-Zi” from ancient Chinese philosophy. It says that Zhuang-Zi was fishing while a boy was watching nearby. Zhuang-Zi offered the boy to take some fish home, but the boy wanted to have Zhuang-Zi’s fishing gears. Zhuang-Zi was amazed at the boy’s long-term thinking, but said: “Fishing gears is a better choice in the long run than taking fish, but do you know the fishing method? Learning the method will benefit you for a life time, but the gears will not.” I think the ladder building idea might benefit this summer school’s diverse 121 students more than the technical steps.

The method of characteristics for two dependent variables typically involves the bootstrapping method. Bootstrapping originally means a person trying to pull himself off the ground by pulling his bootstraps (the slips on the rear ends of the boots, designed to make it easier to put the boots on), which is apparently impossible. In mathematics, it seems it means that two variables get both good estimates from seemingly nothing. Typical way to get the good estimate is to start with little on one variable, using it to get something good for the other variable, and then use that to get better estimate on the first variable. Using the process again and again, one can eventually find very good estimates for both. So it is more like shoe-lacing. A story describes this process best. It is about two Jewish owners of a liquor store in a small village. One day these two people went to town to purchase one-week’s liquor for selling for the following week. The profit of the liquor would feed both families of the men. But on their way home from the town, it started to get cold and miserable. One guy started to think about drinking a bit of the liquor, but he did not have any money and he knew that

both of them needed the liquor to make money to feed their families. However, the guy searched his pockets and miraculously found 50 cents. He said to the other guy: “I gave you this much money, could I buy a half cup of liquor?” The other guy looked at the money, thought for a moment, and said: “It is a good deal. Being a Jew, I certainly take the business.” So the first guy got a small amount of liquor. Liquor made him warm, vigorous, and happy. The other guy looked at his happy friend and started to get jealous and felt bad about himself because he did not have the 50 cents as his friend had. But, then, it occurred to him, that he had just got 50 cents in the previous business deal. So he said to his friend: “I gave you this much money, could I buy a half cup of liquor?” The first guy looked at the money, thought for a moment, and said: “It is a good deal. Being a Jew, I certainly take the business.” So the second guy got a small amount of liquor. Thereafter the two kept using the 50 cents to buy more liquor, and just as they got to their village, the liquor were all consumed, and the two were very happy.

References

- [1] S. Bang, Rarefaction wave interaction of the pressure-gradient system, Ph.D thesis, Penn State University, 2007.
- [2] G. Ben-dor and I.I. Glass, Domains and boundaries of non-stationary oblique shock wave reflection, *J. Fluid Mech.*, (1), 92, (1979), 459–496; (2), 96, (1980), 735–756.
- [3] K. Chuey, C. Conley, and J. Smoller, Positively invariant regions for systems of nonlinear diffusion equations, *Ind. U. Math. J.*, **26** (1977), 373–392.
- [4] T. Chang, G.Q. Chen and S.L. Yang, On the 2-D Riemann problem for the compressible Euler equations, I. Interaction of shock waves and rarefaction waves, *Disc. Cont. Dyna. Syst.*, 1, no. 4, (1995), 555–584.
- [5] S.X. Chen, Z.P. Xin and H.C. Yin, Global shock waves for the supersonic flow past a perturbed cone, *Comm. Math. Phys.*, 228, No. 1, (2002), 47–84.
- [6] X. Chen and Y. Zheng, The direct approach to the interaction of rarefaction waves of the two-dimensional Euler equations, 2007.
- [7] R. Courant and K.O. Friedrichs, Supersonic flow and shock waves, Interscience Publishers, Inc., New York, 1948.
- [8] R. Courant and P. Lax, Cauchy’s problem for non-linear hyperbolic differential equations in two independent variables, *Annali di matematica*, **40** (1955).

- [9] C. Dafermos, Hyperbolic Conservation Laws in Continuum Physics (Grundlehren der mathematischen Wissenschaften), 443, Springer, 2000.
- [10] Z. Dai and T. Zhang, Existence of a global smooth solution for a degenerate Goursat problem of gas dynamics, *Arch. Ration. Mech. Anal.*, 155 (2000), 277–298.
- [11] L.C. Evans, Partial Differential Equations, AMS, 2002, 310–313.
- [12] I.I. Glass and J.P. Sislian, Nonstationary flows and shock waves (Oxford Engineering Science Series), 1994.
- [13] H.M. Glaz, P. Colella, I.I. Glass and R.L. Deschambault, A numerical study of oblique shock-wave reflections with experimental comparisons, *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, 398 (1985), 117–140.
- [14] J. Glimm, X. Ji, J. Li, X. Li, P. Zhang, T. Zhang and Y. Zheng, Transonic shock formation in a rarefaction Riemann problem for the 2-D compressible Euler equations, preprint, submitted for publication (2007).
- [15] P. Hartman and A. Wintner, On hyperbolic partial differential equations, *Amer. Jour. Math.*, 74 (1952).
- [16] J.K. Hunter and A.M. Tesdall, The Mach reflection of weak shocks, XXI ICTAM, 15–21 August 2004, Warsaw, Poland.
- [17] E.H. Kim and K. Song, Classical solutions for the pressure-gradient equation in the non-smooth and nonconvex domain, *J. Math. Anal. Appl.*, 293 (2004), 541–550.
- [18] P. Lax, Hyperbolic systems of conservation laws II, *Communications on Pure and Applied Mathematics*, Vol. X (1957), 537–566.
- [19] P. Lax, Development of singularities of solutions of nonlinear hyperbolic partial differential equations, *J. Mathematical Phys.*, 5 (1964), 611–613.
- [20] P. Lax and X. Liu, Solutions of two-dimensional Riemann problem of gas dynamics by positive schemes, *SIAM J. Sci. Compt.*, 19, no. 2, (1998), 319–340.
- [21] Z. Lei and Y. Zheng, A complete global solution to the pressure gradient equation *J. Differential Equations*, 236 (2007), 280–292.
- [22] L.E. Levine, The expansion of a wedge of gas into a vacuum, *Proc. Camb. Phil. Soc.*, 64, (1968), 1151–1163.
- [23] J. Li, Global Solution of an Initial-value Problem for Two-dimensional Compressible Euler Equations, *Journal of Differential Equations*, Vol. 179, No. 1, 178–194, 2002.

- [24] J. Li, On the two-dimensional gas expansion for compressible Euler equations, SIAM J. Appl. Math., **62** (2001), 831–852.
- [25] J. Li, T. Zhang and S. Yang, The two-dimensional Riemann problem in gas dynamics, Pitman monographs and surveys in pure and applied mathematics 98, Addison Wesley Longman limited, 1998.
- [26] K. Song, The pressure-gradient system on non-smooth domains Comm. Partial Differential Equations. 28 (2003), 199–221.
- [27] J. Li, T. Zhang and Y. Zheng, Simple waves and a characteristic decomposition of the two dimensional compressible Euler equations, Commu Math Phys, 267 (2006), 1–12.
- [28] J. Li and Y. Zheng, Interaction of rarefaction waves of the two-dimensional self-similar Euler equations, Arch. Rat. Mech. Anal. (to appear).
- [29] J. Li and Y. Zheng, Global solutions to some two-dimensional Riemann problems for the compressible isentropic Euler equation, work in preparation.
- [30] T. Li, Global classical solutions for quasilinear hyperbolic systems, John Wiley and Sons, 1994.
- [31] T. Li and W. Yu, Boundary value problem for quasilinear hyperbolic systems, Duke University, (1985).
- [32] A.G. Mackie, Two-dimensional quasi-stationary flows in gas dynamics, Proc. Camb. Phil. Soc., 64, (1968), 1099–1108.
- [33] A. Majda and E. Thomann, Multi-Dimensional shock fronts for second order wave equations, Comm. PDE., 12 (7) (1987), 777–828.
- [34] I.A. Pogodin, V.A. Suchkov and N.N. Ianenko, On the traveling waves of gas dynamic equations, J. App.. Math. Mech., 22, (1958), 256–267.
- [35] C.W. Schulz-Rinne, J.P. Collins and H.M. Glaz, Numerical solution of the Riemann problem for two-dimensional gas dynamics, SIAM J. Sci. Compt., 4, no. 6, (1993), 1394–1414.
- [36] J. Smoller, Shock waves and reaction-diffusion equations, 2nd ed., Springer-Verlag, (1994).
- [37] V.A. Suchkov, Flow into a vacuum along an oblique wall, J. Appl. Math. Mech., 27 (1963), 1132–1134.
- [38] M. Van Dyke, An album of fluid motion, Parabolic Press, Inc., 10th ed. (1982).
- [39] J. von Neumann, Collected works, Pergamon press.
- [40] R. Wang and Z. Wu, On mixed initial boundary value problem for quasilinear hyperbolic system of partial differential equations in two

independent variables (in Chinese), *Acta Scientiarum, Naturalium* of Jinlin University, (1963), 459–502.

- [41] T. Zhang and Y. Zheng, Conjecture on the structure of solution of the Riemann problem for two-dimensional gas dynamics systems, *SIAM J. Math. Anal.*, **21** (1990), 593–630.
- [42] Y. Zheng, Existence of solutions to the transonic pressure-gradient equations of the compressible Euler equations in elliptic regions, *Comm. Partial Differential Equations*. **22**(1997), 1849–1868.
- [43] Y. Zheng, A global solution to a two-dimensional Riemann problem involving shocks as free boundaries, *Acta Mathematicae Applicatae Sinica* (English series), **19** (2003), 559–572.
- [44] Y. Zheng, Two-dimensional regular shock reflection for the pressure gradient system of conservation laws, *Acta Mathematicae Applicatae Sinica* Vol. **22**(2006), no. 2, 177–210 (English series).
- [45] Y. Zheng, Systems of Conservation Laws: Two-Dimensional Riemann Problems, **38** PNLDE, Birkhäuser, Boston, 2001.
- [46] Y. Zheng, Absorption of characteristics by sonic curve of the two-dimensional Euler equations, dedicated to Professor Ta-Tsien Li (Daqian Li) on his 70th birthday, *Disc. Cont. Dyna. Syst.*, 2007.

This page intentionally left blank

Nonlinear Conservation Laws, Fluid Systems and Related Topics

Series in Contemporary Applied Mathematics
CAM 13

This book is a collection of lecture notes on Nonlinear Conservation Laws, Fluid Systems and Related Topics delivered at the 2007 Shanghai Mathematics Summer School held at Fudan University, China, by world's leading experts in the field.

The volume comprises five chapters that cover a range of topics from mathematical theory and numerical approximation of both incompressible and compressible fluid flows, kinetic theory and conservation laws, to statistical theories for fluid systems. Researchers and graduate students who want to work in this field will benefit from this essential reference as each chapter leads readers from the basics to the frontiers of the current research in these areas.

Higher Education Press

www.hep.com.cn

World Scientific

www.worldscientific.com

7292 hc

ISBN-13 978-981-4273-27-5

ISBN-10 981-4273-27-9



9 789814 273275